The Normativity of Automaticity

MICHAEL BROWNSTEIN AND ALEX MADVA

Abstract: While the *causal* contributions of so-called 'automatic' processes to behavior are now widely acknowledged, less attention has been given to their normative role in the guidance of action. We develop an account of the normativity of automaticity that responds to and builds upon Tamar Szabó Gendler's account of 'alief', an associative and arational mental state more primitive than belief. Alief represents a promising tool for integrating psychological research on automaticity with philosophical work on mind and action, but Gendler errs in overstating the degree to which aliefs are norm-insensitive.

1. Introduction

Interest in so-called 'automatic' behavior—which is effortless, efficient, uncontrolled, and relatively unconscious—has focused chiefly on cases of mismatch between agents' automatic actions and reflective states. In typical cases, agents behave contrary to their sincere intentions and avowals, seemingly driven by phobias, habits, or social prejudices.[1] Tamar Szabó Gendler (2008a,b) has offered some of the most vivid examples, including subjects who hesitate to eat a piece of fudge molded to resemble feces, despite acknowledging that the ugly fudge has the same ingredients as a piece they had just eaten (2008a, p. 636); sports fans who scream at their televisions while watching a taped game, despite acknowledging that their shouts can't transcend space and time in order to affect the game's outcome (2008b, pp. 553, 559); and 'aversive racists', who demonstrate prejudiced automatic dispositions despite sincerely avowing anti-racist commitments. Cases of mismatch like these seem to be pervasive and, often, pernicious.

On Gendler's account, mismatches between reflective states and automatic behaviors result from discordance between agents' beliefs and what she calls 'aliefs', which

For their invaluable feedback, the authors wish to thank Michael Brent, Taylor Carman, John Christman, Katie Gasdaglis, Tamar Gendler, Eric Katz, Sean Kelly, Christia Mercer, John Morrison, Eric Schwitzgebel, Shannon Sullivan and two anonymous reviewers for *Mind & Language*. Versions of this paper were presented at the 2011 Pacific Division meeting of the American Philosophical Association and the Philosophical Psychology Lab at Harvard University. We are deeply grateful for the helpful comments and questions we received during these presentations.

Address for correspondence: Michael Brownstein, Department of Humanities, NJIT, University Heights, Newark, NJ, 07102–1982, USA.
Email: msb@adm.njit.edu

[1] Such mismatches have been the subject of much research in social psychology (Bargh *et al.*, 1996; Deutsch and Strack, 2010; Pearson *et al.*, 2009), ecological psychology (Witt *et al.*, 2004), phenomenology (Dreyfus and Kelly, 2007; Rietveld, 2008a,b), ethics (Huebner, 2009), and philosophy of mind and action (Elga, manuscript; Gertler, 2011; Hunter, 2011; Peacocke, 1999, 2004; Rowbottom, 2007; Schwitzgebel, 2010).

Mind & Language, Vol. 27, No. 4 September 2012, pp. 410–434.

are automatic, motor-affective mental states more primitive than belief. While the concept of alief is a promising tool for understanding belief-behavior mismatches, Gendler overstates the extent to which aliefs are inflexible and norm-insensitive. We argue that these motor-affective states can adapt flexibly to changes in an agent's immediate environment and improve gradually over time. An alief is in good standing, we will argue, just insofar as its motor and affective components work in concert to reduce 'felt tensions', or experiences of 'disequilibrium' between an agent and her environment. These affective states self-correct over time by activating behaviors directed toward retrieving a felt sense of equilibrium. Everyday examples include moving around until one has found the right spot for viewing a painting or stepping backward in order to alleviate the subtle discomfort induced by a 'close-talker'. In these cases, aliefs play an integral normative role in the guidance of action, notwithstanding the fact that they are generally unnoticed and outside of immediate control.

After reviewing a few of Gendler's most suggestive examples of belief-behavior mismatch (2.1), we describe alief (2.2) and offer some brief arguments for why it explains these mismatches better than rival accounts (2.3). We then introduce new examples that mirror the affective and motor aspects of alief as Gendler describes them but that appear to exhibit norm-sensitivity (3.1), and we offer an alternative account of the intentional content of these (putatively) norm-sensitive aliefs (3.2). We go on to explain why alief, so construed, manifests a distinctive and legitimate sensitivity to norms (4.1 and 4.2), and how this norm-sensitivity differs from the evidence-sensitivity characteristic of belief (4.3). Aliefs are sensitive to subtle variations in the environment, but not to those variations *qua* evidence, because they are, *ceteris paribus*, incapable of responding properly to defeating considerations and of integrating properly with other psychological states. The evidence-insensitivity of alief is most visible in cases of mismatch between alief and belief, and we discuss how aliefs that successfully reduce felt tensions may nevertheless drive undesirable or even unethical behavior (5). But an agent's considered beliefs can lead her astray as well, and we conclude by considering cases, akin to Nomy Arpaly's (2004) examples of 'inverse *akrasia*', in which aliefs are *more* attuned to the demands of the situation than beliefs.[2]

2. Automaticity and Philosophy

2.1 Skywalkers and Aversive Racists

One of Gendler's most compelling examples of alief focuses on the 'Skywalk', a glass walkway extending 70 feet out from the rim of the Grand Canyon.[3] On the

[2] See also our companion paper, Brownstein and Madva, 2012, in which we specifically address the *ethical* significance of alief and automatic behavior. We identify a class of alief-like states that make normatively self-standing contributions to praiseworthy action and a well-lived life.

[3] Gendler's example derives from the early modern 'problem of the precipice' discussed by the likes of Hume, Pascal, and Montaigne. See Gendler, 2008a,b for discussion of the relevant literature.

one hand, tourists on the Skywalk presumably *believe* that they are safe. There is very little chance that they would venture out onto the platform otherwise. On the other hand, if their knees are knocking and they can't quite shake the impulse to get off, some facets of their behavior are somehow mismatched with their considered beliefs. Gendler argues that trembling Skywalkers are driven by 'belief-discordant' automatic mental states. So too are dessert lovers who refuse to eat feces-shaped fudge, sports fans who verbally abuse their televisions, and the many white Americans who sincerely avow anti-racist commitments but are nevertheless 'aversive racists'.[4] The latter are, for example, more likely to hire a white job candidate over an equally qualified black candidate and more likely to exhibit a range of discriminatory 'microbehaviors' (e.g. they tend to make less eye contact, commit more speech errors, and sit further away from black interlocutors).[5] According to Pearson *et al.*'s summary of the findings, people's sincerely avowed attitudes 'typically shape deliberative, well-considered responses for which people have the motivation and opportunity to weigh the costs and benefits of various courses of action', while their subtle prejudices 'typically influence responses that are more difficult to monitor or control' (2009, p. 9). What do trembling Skywalkers, fudge-avoiding dessert lovers, television-abusing sports fans, and aversive racists have in common?

2.2 Alief

On Gendler's view, all of these agents are driven by *aliefs*. Aliefs are states that dispose agents to respond automatically to apparent stimuli with certain fixed affective responses and behavioral inclinations (2.2.1).[6] Aliefs are, unlike beliefs, insensitive to evidence (2.2.2), but they can be changed over time with changes in habit (2.2.3). Aliefs are causally responsible for the brunt of moment-to-moment behavior (2.2.4).

[4] The term 'aversive racism' was coined by Kovel (1970). For a recent review, see Pearson *et al.*, 2009.

[5] For hiring bias, see Bertrand and Mullainathan, 2003 and for discriminatory unreflective behavior, see McConnell and Leibold, 2001 and Dovidio, Kawakami and Gaertner, 2002. The term 'microbehaviors' and a summary of the research can be found in Payne and Cameron, 2010, p. 446.

[6] According to Gendler, aliefs tend to share an array of common features. She writes, '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 . . . *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' (2008b, p. 557, original emphasis; see also 2008a, p. 641). We are sympathetic with Gendler's claims about conceptual and developmental antecedence but lack the space to address them here. Some critics worry that Gendler does not offer necessary and sufficient conditions for states of alief. See Egan, 2011; Mandelbaum, 2012; Muller and Bashour, 2011; and Schwitzgebel, 2010 for this and related concerns. A virtue of describing alief initially in terms of a cluster of characteristics is that doing so invites revisions to the core concept as future research unfolds. We take ourselves to be contributing to this effort.

2.2.1 Fixed Responses. Gendler characterizes alief as a relation between an agent and a distinctive kind of intentional content, with representational, affective, and behavioral (or *R-A-B*) components. Aliefs involve 'a cluster of dispositions to entertain simultaneously *R*-ish thoughts, experience *A*, and engage in *B*' (2008a, p. 645).[7] These components are associatively linked and automatically co-activating.[8] On the Skywalk, the mere perception of the steep drop 'activates a set of affective response patterns (feelings of anxiety) and motor routines (muscle contractions associated with hesitation and retreat)' (2008a, p. 640). The *R-A-B* content of this alief is something like, 'really high up, long long way down. Not a safe place to be! Get off!' (2008a, p. 635). Likewise, the sight of feces-shaped fudge 'renders occurrent a belief-discordant alief with the content: "dog-feces, disgusting, refuse-to-eat"' (2008a, p. 641).

2.2.2 Insensitivity to Evidence. The local activation of states with this *R-A-B* content is independent of what agents themselves recognize to be decisive evidence. Aliefs are *a*rational. The Skywalker's fear and trembling are unmoved by the overwhelming evidence that the walkway is safe. The fudge-avoider's disgust response 'runs counter to the subject's explicit belief that the object before her is composed of a substance that she considers delicious and appealing' (2008a, p. 641). Similarly, an agent who endorses anti-racist views on robust empirical grounds may nevertheless find that the mere perception of dark-skinned faces activates subtly unfavorable affective responses and motor routines (Amodio *et al.*, 2003).

The automatic-associative nature of alief provides a natural explanation for its evidence-insensitivity. In paradigmatic cases, certain fixed affective responses and behavioral inclinations are unavoidably activated by perceptions or thoughts of some salient cue. As we elaborate in (2.3), this evidence-insensitivity in turn explains why aliefs and beliefs differ (2008b, p. 566) and why aliefs and beliefs frequently come into conflict (2008b, p. 570).

2.2.3 Changes Over Time. The sources of aliefs may be innate or habitual (2008b, pp. 568–70). While activations of alief are unavoidable in specific contexts, they are malleable over time. In particular, aliefs may be influenced by changes in habit. For example, Gendler cites research that a committed anti-racist can reduce automatic stereotype activation by repeatedly negating stereotypic associations (Kawakami *et al.*, 2000). We elaborate on how aliefs change over time in (3) and (4). Gendler (2008a,b) also addresses the question of how one can go about changing

[7] Gendler uses the notion of content in an admittedly 'idiosyncratic way', leaving open whether the content is propositional or conceptual, but insisting that the content in some sense includes 'affective states and behavioral dispositions' (2008a, p. 635, n.4).

[8] The sense in which these states are *associative* is controversial (Mandelbaum, manuscript; Schwitzgebel, 2010). We are sympathetic with Gawronski and Bodenhausen's (2006) dual-process theory, according to which these processes are associative because they lack a 'subjective truth value'.

one's automatic dispositions—the 'ethics of alief'—as do we in Brownstein and Madva, 2012. Although our concern here is not with ethical normativity per se, understanding the norm-sensitivity of automatic dispositions will clearly shape ethical inquiry into the regulation of unwanted aliefs and, as we emphasize elsewhere, the promotion of desirable aliefs.

2.2.4 Causal Responsibility. Gendler proposes that alief, rather than belief, is primarily responsible for the 'moment-by-moment management' of behavior (2008a, p. 663). She grants that, 'belief plays an important role in the ultimate regulation of behavior', presumably, for example, in setting ends and determining means to reach them (2008a, p. 663). But she reasons that, 'if alief drives behavior in belief-discordant cases, it is likely that it drives behavior in belief-concordant cases as well' (2008a, p. 663.). Gendler's case here draws on the building consensus that automaticity guides much moment-to-moment behavior. So pervasive do psychologists think automaticity is in everyday life that some have openly wondered why we ever become focally aware of our behavior at all (e.g. Dijksterhuis *et al.*, 2007). In this regard, research on automaticity is informed by the extensive research suggesting that the initiation and online control of action is largely unconscious (Libet *et al.*, 1983; Milner and Goodale, 1995, 2008; Wegner, 2002). It is not within the scope of this paper to determine exactly *how much* moment-to-moment behavior is automatic. But why think that alief explains automatic behavior better than belief?

2.3 Judgment, Action, and Belief Attribution

Gendler's account of alief represents, *inter alia*, a distinctive proposal for making sense of long-standing problems for belief attribution. On the one hand, beliefs are something like what one takes to be true of the world.[9] On the other, beliefs are also thought to guide actions, together with one's desires and ends. What happens when these two roles, of truth-taking and of action-guiding, come apart? What do trembling Skywalkers and aversive racists *really* believe? In cases of mismatch, one might adopt: (a) a truth-taking view, which attributes beliefs on the basis of agents' reflective judgments and avowals (Gendler; Zimmerman, 2007); (b) an action-guiding view, which attributes beliefs on the basis of agents' spontaneous actions and emotions (Hunter, 2011); (c) a context-relative view, which takes both judgment and action to be relevant to belief attribution, and attributes to agents beliefs that vary wildly across contexts (Rowbottom, 2007); (d) a contradictory view, which takes both judgment and action to be independently sufficient for belief attribution, and attributes to agents contradictory beliefs (Egan, 2011; Gertler, 2011; Huddleston, 2011; Huebner, 2009; Muller and Bashour, 2011); or (e) an

[9] See Gilbert, 1991 for a psychological discussion of belief. See Schwitzgebel, 2006, 2010 for a review of contemporary philosophical approaches to belief.

indeterminacy view, which takes neither judgment nor action to be independently sufficient, and attributes to agents no determinate belief at all, but just some 'in-between' state (Elga, manuscript; Schwitzgebel, 2010).[10]

Each of these views fits more naturally with some cases of belief-behavior mismatch than others, but (a), the truth-taking view, which attributes belief on the basis of agents' reflective judgments and avowals, outperforms the alternatives in many prominent cases. Consider the Skywalker. That the Skywalk is safe is *both* what the agent judges to be true all things considered *as well as* what guides the agent's intentional decision to walk across it. At the same time, her emotional responses and behavioral inclinations 'go rogue'. In such cases, we can plausibly divide up the agent's dispositions between those that *are* and *are not* sensitive to the all-things-considered evidence (Gender, 2008b, p. 566). The Skywalker's reflective judgments regarding the Skywalk's safety can revise immediately with the incoming evidence. If credible reports emerged that someone actually fell through it, many fewer visitors would come! By contrast, the affective-behavioral aversion to the Skywalk is irredeemably yoked to how things perceptually *seem* to the agent, independently of whether she has ample reason to *judge* that the seeming is misleading.[11] Our 'reactive attitudes' toward such dispositionally muddled agents are also reasonable guides in determining what they believe (Zimmerman, 2007). Although we might judge the Skywalker to be phobic or lacking in self-control, we would not impugn her with ignorance or irrationality.

By contrast, proponents of (b), the action-guiding view, which attributes belief on the basis of agents' spontaneous actions and emotions, would have to argue that the Skywalker merely professed, wished, or imagined the platform to be safe. But in this case, if the Skywalker failed to believe that the platform was safe, or harbored any legitimate doubt, she would have to be clinically ill to decide to walk on it.[12]

[10] This sketch of the possible responses is drawn from Schwitzgebel (2010), who points out that a similar array of interpretive options arises in the literature on self-deception (see Deweese-Boyd, 2006/2008). Also see Gendler (2008a,b) for some discussion of historical predecessors to these contemporary views. It should be noted that, in what follows, we discuss the relevant views in terms of Gendler's example of the Skywalk, even though the authors listed in this paragraph do not focus on this example. These authors might argue that the Skywalk case is not exemplary of their views.

[11] See 2.2.2 above. Peacocke (2004, pp. 254–7) similarly endorses what he calls the 'belief-independence' of such emotional responses, citing Evans' (1982, pp. 123–4) discussion of the belief-independence of perception (evident in perceptual illusions like the Müller-Lyer). While Peacocke (2004) seems to defend (a), the truth-taking view of belief, he is sometimes cited as a defender of (b), an action-guiding view, because of his (1999, pp. 242–3) discussion of a case akin to aversive racism. It is natural to interpret Peacocke's considered position as *privileging* the role of judgment in belief attribution, while acknowledging that in some cases, so many of an agent's *other* decisions and actions may fail to cohere with her reflective judgments that it would be wrong to attribute the relevant belief to her.

[12] Or trying to impress a date, or fleeing from a greater threat, like the protagonist of an action film crossing a rickety bridge. Many different desires and beliefs might conjoin with the belief that the Skywalk was only probably safe to cause people to walk on it. But it is ad hoc to

Are the Skywalker's beliefs just (c) unstable across contexts? While it is surely the case that agents' beliefs often 'flip-flop' over time, the Skywalker seems to treat the Skywalk as both safe and unsafe in the same context. Perhaps, then, she (d) both believes and fails to believe that the Skywalk is safe. But attributing contradictory beliefs to her in this case seems to run up against Moore's paradox, the upshot of which is that an agent cannot *occurrently endorse* a proposition and its negation.[13] Is the Skywalker then (e) in an irreducibly vague state of 'in-between belief' (Schwitzgebel, 2010)? While problems associated with vagueness are ubiquitous and apply as much to ascriptions of belief and desire as they do to ascriptions of tallness and baldness, positing 'in-between' states seems to defer those problems rather than solve them.[14] Why not draw the distinctions as precisely as we can, and just acknowledge that there are outliers? The most pressing problem facing the contradictory and indeterminacy views is imagining how one could reasonably distinguish between them, a task comparable to distinguishing the view that a person is both tall and not tall from the view that the person is neither tall nor not tall. What could, in this context, be evidence for one over another?

The further question for the truth-taking view is how to understand those states that *aren't* expressed in an agent's reflective judgments and avowals. We think Gendler has rightly identified a range of motor-affective states that are automatic and relatively evidence-insensitive. But while there is much to recommend this novel approach to automaticity, Gendler errs in inferring from the fact that aliefs lack the evidence-sensitivity characteristic of considered beliefs that aliefs lack *any* norm-sensitivity whatsoever.

3. Aliefs in Good Standing

According to Gendler, aliefs are neither good nor bad in and of themselves, but only good or bad to the extent that we succeed in 'bringing them into line with our considered commitments' (2008b, p. 572; 2008a, p. 651). Call this the Dependency Thesis.

> (DT): An agent's aliefs are in good standing if and only if they are suitably regulated in accordance with her considered beliefs and ends.

DT suggests that, at best, a well-functioning alief is on par with a well-functioning thermostat. Both are non-normative systems the good standing of which derives

stipulate that Skywalkers actually possess such outlandish desires and beliefs. See also Gendler (2008a, pp. 654–6). Of course, some would-be Skywalkers might become so gripped by fear that they cannot actually step on the walkway, but the mere presence of *any* fear is problematic for belief attribution.

[13] But see Huddleston, 2011 and Muller and Bashour, 2011.

[14] See also Zimmerman, 2007, pp. 73–5.

from extrinsically given ends. Aliefs are only trustworthy in 'stable, typical, and desirable' contexts in which an agent can safely assume they are belief-concordant (2008b, pp. 554, 570–2). DT implies that it is undesirable or inappropriate for aliefs to drive behavior when they are mismatched with an agent's considered beliefs and reflectively endorsed ends.

Gendler's endorsement of DT for automatic states is part of a broader tendency.[15] Take, for example, what social psychologists Roland Deutsch and Fritz Strack say about implicit social cognition. They explain that, 'the paradigm of implicit social cognition rests on the notion that attitudes, prejudice, stereotypes, and the self may have an impact on behavior that sometimes opposes beliefs and intentions' (2010, p. 63). Such cases result in 'irrational behavior', where Deutsch and Strack then *define* 'irrational' as 'the case in which behavior occurs against the actor's explicit beliefs' (2010, p. 70).[16]

We aim to show, by contrast, that aliefs can be in good standing independently of whether they concord with beliefs. Aliefs are in good standing just insofar as they drive an agent to act in ways that alleviate a felt sense of 'disequilibrium' between herself and her environment. In the simplest cases (3.1), aliefs do so in the *absence* of belief altogether. These cases involve *neither* concordance *nor* discordance with considered beliefs or other reflective states, because no such states are implicated.

3.1 The Museumgoer

Imagine walking through a museum and encountering an enormous painting.[17] As soon as you see it, you feel an impulse to step back in order to get the painting into

[15] It should be noted that endorsing the viability of alief as a psychological concept is not strictly necessary for adopting a thesis relevantly similar to DT. For example, Huebner (2009) reaches similar normative conclusions despite drawing on different empirical research (e.g. Gilbert, 1991) and defending a contradictory view of belief attribution. Huebner describes the rogue automatic dispositions as 'stereotype-based judgments', which issue from 'Type-1 processes', and function like Gendler's aliefs in important respects. They are non-rational processes which unfold automatically and independently of an agent's 'Type-2 processes', i.e., independently of the considered commitments an agent would form upon reflection. In keeping with DT, Huebner writes: 'for those who acknowledge that many stereotype-based judgments are both misguided and unjustifiable, the important question to ask is whether egalitarian Type-2 processes can be recruited to override a stereotype-yielding Type-1 processes' (2009, p. 75). Both Gendler and Huebner construe the operative automatic dispositions as merely causal mechanisms, whose good standing depends on the extent to which they are regulated by our higher powers of ratiocination. For an endorsement of DT sympathetic to alief, see McKay and Dennett, who ask, 'Are such aliefs adaptive? Probably not. They seem to join other instances of "tolerated" side effects of imperfect systems' (2009, p. 500).

[16] Deutsch and Strack's account of the 'reflective and impulsive processes' underlying social behavior has been called 'the most influential model' in their field (Payne and Gawronski, 2010).

[17] We borrow this example from Dreyfus and Kelly (2007) who are in turn influenced by Merleau-Ponty (1962/2002). For more discussion on the physiological manifestations of automatic affect, see, for example, Barrett *et al.*, 2007.

view. There is a felt sense of 'tension' or 'disequilibrium' between yourself and the environment. Some artists seem to be aware of this automatic impulse. Consider the following caption, which was displayed by one of Barnett Newman's paintings:

> *Vir Heroicus Sublimus*, Newman's largest painting at the time of its completion, is meant to overwhelm the senses. Viewers may be inclined to step back from it to see it all at once, but Newman instructed precisely the opposite. When the painting was first exhibited, in 1951 ... Newman tacked to the wall a notice that read, 'There is a tendency to look at large pictures from a distance. The large pictures in this exhibition are intended to be seen from a short distance'.[18]

Newman seems to be anticipating the felt tension that inclines you to step back in order to improve your bodily orientation to the work. This felt tension is an amalgam of automatically unfolding physiological changes, such as muscle tension and autonomic arousal. As you move backward, you might squint your eyes or crane your neck, showing signs that your behavior is directed toward finding the spot that feels 'right' for taking in the painting. While feeling and responding to this sense of disequilibrium, your mind is largely occupied with the painting itself. You do not *notice* (in focal awareness) the impulse to step back, but you *feel* it nevertheless. In fact, it is unlikely that your causally operative mental states are available for report or deliberation. Ask the average museumgoer what the right distance to stand from an 8′x18′ painting is, and your question will likely be met with puzzled stares or confabulation.[19] As you approach the 'right' distance, though, the tension driving your movements subsides.

Felt tensions like these similarly guide action, in our view, when one assumes a sympathetic posture to listen to a friend in need, shifts to and fro to see whether the car will fit into a tight spot, or walks at a faster pace because of an inchoate sense of danger, even though nothing is identifiably wrong.[20] Compare the behavior of the museumgoer to the skills needed for 'distance-standing', or knowing how far to stand from an interlocutor. When you are too far or too close, a felt tension compels you to readjust. In all of these cases, the agent's behavior appears to be guided by a felt sense of rightness or wrongness. There is something *not right* in her bodily orientation to the environment, and this feeling of wrongness drives her to respond in particular ways. In this way, behavior-inducing felt tensions involve both 'descriptive' and 'directive' aspects (Millikan, 1995; Clark, 1997).[21] One and

[18] 'Abstract Expressionist New York', *Museum of Modern Art*, 2010–2011.

[19] Jeannerod (2006) makes a similar point about subjects who successfully catch objects falling at an accelerating rate but nevertheless report, due to a reliance on naïve physics, that those objects are falling at a constant speed.

[20] De Becker (1998) argues that one important thing agents in ambiguous contexts can do to protect themselves from robbery or assault is to heed their 'sixth sense' that things are amiss, rather than persuading themselves that their feelings are unjustified.

[21] Gendler also emphasizes that the different components of the content are operative 'all at once, in a single alief', but does not pursue the normative import of this point (2008b, p. 559).

the same state can both take the world to be a certain way and prescribe a certain way of responding. The museumgoer experiences the painting both as too big and as to-back-away-from.

The initial activation of these banal cases of directed felt tension and motor response mirror the affective and behavioral aspects of alief as Gendler describes them. Gendler might say that the perception of the painting activates an alief with the *R-A-B* content, 'Really big painting! Disorienting! Move away!' The close-talker activates an alief like, 'Bad breath! Awkward! Lean back!' A perceived stimulus automatically induces an affective-motor response. However, although this *R-A-B* content captures the *activation* of these states, it fails to account for the way the affective and behavioral components *unfold* and *interact* over time. The perception of the painting does not just activate a one-off motor routine or affective response. The sense of being overwhelmed by the painting activates behaviors aimed at reducing this disequilibrium, and will subside or intensify over time depending on how the agent moves. These subsequent changes in an agent's sense of rightness or wrongness will automatically activate further behaviors. Distance-standing is likewise a matter of continually making subtle adjustments in the face of increasing and decreasing senses of disequilibrium.[22]

That the alievers in our examples act appropriately does not seem to be accidental. An agent's sense of rightness or wrongness in a given situation will be sensitive to the context and to socio-cultural norms. The museumgoer will react differently given variations in the ambient lighting (to avoid glare and shadows), the location of the painting relative to other works, the presence and movements of fellow museumgoers, and the overall layout of the space. The distance-stander's feel will be informed by the mood, the subject matter, the personal history between interlocutors, etc. One can be especially adept or klutzy at adjusting appropriately to felt tensions. Some 'close-talkers', like those parodied on *Seinfeld*, chronically fail to exhibit the ordinary flexible self-modification that typically characterizes distance-standing.

In simple cases like the museumgoer, the appropriateness of the automatic reaction does not appear to derive from *concordance* with her considered beliefs and ends. The automatic affective-behavioral response thereby fails to satisfy DT. This is simply because no reflective state (regarding perceived distance from the painting) is present in her experience; there is nothing with which her automatic impulse can be in conflict or concord. The museumgoer moves away from the large painting to get the best view, without ever having to consider that her initial view was flawed. Of course, she may reflect on where to stand and judge that her automatic impulse

[22] We don't purport to explain how Gendler's account of alief applies to all of her examples, nor to explain whether our and Gendler's accounts will be perfectly coextensive. In particular, our argument that, as a class, aliefs are capable of being norm-sensitive, focuses on aliefs that are activated perceptually. Gendler plausibly claims that similar automatic states with *R-A-B* content can be activated 'internally' by thought and imagination. It is an interesting question whether such states are capable of the norm-sensitivity we describe here.

is misguided. But absent any defeating reflective considerations, her automatic impulse drives her to act at the right time in the right way, just by virtue of being appropriately responsive to the array of subtle environmental features. Moreover, her reflective states might themselves lack sufficient access to the ambient features relevant to determining the best position for viewing the painting.[23] Such reflective states can even interfere with her capacity to find the right spot, in much the same way that overthinking can impair expert athletic performance (DeCaro *et al.*, 2011).

3.2 *F-T-B-A* Content

While the *R-A-B* formula suffices for the *activation* of aliefs, it fails to capture the intrinsic and normative connections among the relata as they unfold and interact in response to changes in the immediate environment. In what follows, we emphasize the temporal element of alief because the unfolding in time of states of alief makes possible the adjustments and readjustments which are requisite for genuine norm-sensitivity, as we explain in (4). We will propose that, in paradigmatic cases, alief is a relation between an agent and '*F-T-B-A*' content: feature-tension-behavior-alleviation.

3.2.1 Feature. Aliefs are activated by salient environmental properties that make certain possibilities for action attractive. We refer generically to these alief-triggering environmental properties as 'features'. On this particular point, our account does not depart significantly from Gendler's. We prefer the term 'feature' to 'representation', but this is not because we intend to argue that aliefs are nonrepresentational. We omit discussion of representation because we do not think that the notion of representation illuminates anything *distinct* about alief.[24]

Which sorts of environmental features can activate aliefs and what makes a given feature salient are open empirical questions.[25] Features often become salient by virtue of their relation to practical goals and concerns. The features that rise to salience in guiding the museumgoer's moment-to-moment behavior as she rushes to find a bathroom may differ greatly from those that rise to salience when she

[23] See Sections 4 and 5 and references in 2.2.4 and note 19.

[24] How to attribute belief in cases of mismatch (2.3) is largely independent of whether belief is representational. One's stance on representationalism does not differentiate, for example, the view that beliefs frequently shift with context from the view that beliefs are frequently contradictory. And the truth-taking and action-guiding views could be seen *either* as privileging a certain set of dispositions in belief attribution *or* as privileging the set of representational states that explains those dispositions. Gendler herself describes her notion of representation as 'a thin one' (2008b, p. 559, n. 11; 2008a, p. 644). See Chemero, 2009 for discussion of how putatively representation-hungry behaviors can be explained by nonrepresentational mechanisms.

[25] For an overview of empirical issues regarding the role of salience in implicit cognition, see Moors *et al.*, 2010, pp. 22–30.

is ambling aimlessly. But features can become salient independently of goals, as when the mere sight of a sign for the restroom makes you feel like you have to go.[26]

3.2.2 Tension. On our view, 'tension' is not a blanket term for any affect or emotion, but refers to a specific class of automatic affective responses that are in a deep sense 'geared' towards immediate behavioral reactions. The Skywalk does not elicit some *arbitrary* affective response, but an embodied reaction of fear that repels the agent from the walkway. Much the same is true of the fudge-avoider and, regrettably, of the aversive racist. The affective components of Gendler's norm-discordant alievers are oriented toward actions that will reduce their discomfort.[27] These action-generating felt tensions are marked by either positive or negative *valence*, which acts like a physiological reinforcer of anticipated behaviors.[28] The agent literally feels a (positive) attraction or (negative) repulsion to various available courses of action. The museumgoer feels that 'things are not quite right' and moves in such a way as to retrieve equilibrium between herself and her environment. Even in this more subtle case, the valent tension makes an active contribution to phenomenal experience, together with an array of visceral 'low-level' bodily changes in an agent's autonomic nervous system, including changes in cardiopulmonary parameters, skin conductance, muscle tone, and endocrine and immune system activities (Klaasen *et al.*, 2010, p. 65; Barrett *et al.*, 2007). Felt tensions may be precipitating events of full-blown emotions, but they need not. (We suspect that the affective elements of alief are most likely to become focally conscious in jarring cases of belief-behavior

[26] We also think the range of practical 'concerns' that influence which environmental features rise to salience is broader or at least not coextensive with 'practical goals' per se (Rietveld, 2008b). For example, the features salient for distance-standing can be influenced by the shared mood of the conversation, independently of the interlocutors' goals. If the mood is hostile or awkward, interlocutors may stand further apart, but they need not be regulating their position in accordance with some goal, say, of staying away from boring acquaintances. We say more about the relation between reflective goals and automatic behavior in Brownstein and Madva, 2012.

[27] In a view he credits to Wittgenstein (1966), Erik Rietveld (2008a) discusses an experience of action-guiding tension, similar to our account of 'felt tensions', in terms of what he calls 'directed discontent'. Episodes of directed discontent offer agents the ability to make subtle action-guiding perceptual discriminations characterized by affective experiences of attraction or repulsion that unfold over time and are typically not reportable in propositional form (Klaassen *et al.*, 2010, p. 64). Rietveld helpfully contrasts directed discontent with another concept he traces to Wittgenstein, 'directed discomfort'. In short, directed discomfort characterizes a 'raw, undifferentiated rejection' of one's situation which does not give the agent an immediate sense of adequate alternatives (2008a, p. 980). Directed discontent, by contrast, is a feeling of tension accompanied immediately with opportunities for acting.

[28] For more on this 'affective force', see Varela and Depraz, 2005, p. 65. For discussion of the physiological explanation of action-initiating affective responses, see Prinz, 2004. Felt tensions are typically non-propositional, and so differ from the concept-laden emotional evaluations which Lazarus (1991) calls 'appraisals'. See Colombetti, 2007.

mismatch, as in those induced by confrontations with feces-shaped fudge and glass precipices.)

3.2.3 Behavior. Felt tensions set a range of anatomical and bodily reactions in motion, including limb movements, changes in posture, and vocalizations. The motor routines set in motion by felt tensions *can* rise to the level of fully-fledged actions.[29] The ordinary bodily changes and movements associated with the aliefs of Skywalkers, fudge-avoiders, and museumgoers are not, however, 'mere behaviors', but integral parts of coordinated response patterns that are oriented toward the *reduction* of tension, or a felt sense of 'alleviation', which we discuss in the next subsection. In an important sense, these are behavioral reactions *to* felt tensions, although both the affective and behavioral components of alief are temporally extended processes that overlap and influence each other reciprocally. The duration and vivacity of felt tensions influence the strength of the impulse to act and are in turn influenced by how the agent *does* act. As we explain, this reciprocal influence of tension and behavior stands in marked contrast to the relative insularity of alief from belief.

3.2.4 Alleviation. Behaviors responsive to felt tensions will or will not *alleviate* the sense of tension. As the behavior unfolds, one's felt sense of rightness or wrongness will change in turn, perhaps suggesting an improvement in one's relation to an environmental feature, or a failure to improve. The temporal unfolding and interplay between behavior and senses of tension is absent from Gendler's account of alief. This felt sense of (un)alleviated tension can feed back into further behaviors, as one continues to, say, crane one's neck or shift posture in order to get the best view of the painting. Once a tension is alleviated, the salient features of one's ambient environment may change as well, freeing one to pursue new practical concerns and opportunities for action. Alleviating the tension induced by a close-talker allows the distance-stander to focus on the subject matter of the conversation instead of on her own comportment. If one restrains one's behavior in some way, the sense of tension will, *ceteris paribus*, persist. The persistent force of unalleviated tension is evident in how the Skywalker continues to shiver and tremble while restraining the impulse to flee and how a fudge-avoider would feel as she brought the ugly snack to her lips.[30]

The interplay between tension and alleviation is key to understanding how aliefs can self-modify in a flexible and normative way, as we explain in the next section. The internal components of aliefs are not merely fortuitously associated, but form

[29] Exactly when alief-driven behaviors constitute fully-fledged actions, for which agents are directly responsible, is an important question. See Section 4 for brief discussion. See Brownstein and Madva, 2012; Madva, 2012; and Brownstein, in preparation (a).

[30] And if the museumgoer comes across a large painting along a narrow hallway, there may be no way to improve her relation to it and reduce her sense of tension, unless it is best seen obliquely, like the skull in the foreground of Holbein's (1533) painting, *The Ambassadors*. http://en.wikipedia.org/wiki/File:Holbein-ambassadors.jpg#filelinks

an integrated response to the local environment that unfolds over time. Experiences of felt tension initiate coordinated responses directed toward their own alleviation. This *self-alleviation* enables museumgoers and the like to adjust their behavior to the demands of the situation without the intervention of judgment or deliberation. It also explains how aliefs can 'learn' over time; token experiences of (un)alleviation contribute to a gradual 'fine tuning' of affective-behavioral associations to specific types of stimuli.

4. Normativity

Our account of *F-T-B-A* content is well-poised to bring out exactly what distinguishes alief-driven acts from brutely causal reflexes and 'mere behaviors', as well as from fully intentional, reason-based actions. This is in keeping with one of Gendler's stated aims, to show how alief '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)' (2008a, p. 555). We agree that this alternative is needed, but Gendler's emphasis on alief's norm-insensitivity renders it mysterious how alief can occupy this middle level of explanation.

Genuine norm-sensitivity requires that aliefs do more than simply 'get things right'. Thermostats and plants both respond to changes in their ambient environments and produce appropriate behaviors, but not by virtue of any genuine sensitivity to norms.[31] Aliefs are, unlike thermostats and plants, norm-sensitive insofar as they exhibit their own proprietary modes of flexible self-modification (4.1) and capacity for error (4.2). At the same time, norm-sensitivity requires *less* than practical rationality demands (4.3).

4.1 Self-Modification

It's often thought that the capacity for self-modification is a hallmark of norm-sensitivity. Paradigmatically, beliefs exhibit an ongoing self-modification by updating in response to incoming evidence and reasons. We agree that beliefs typically are, and aliefs typically are not, sensitive to evidence in this way. But it is a non sequitur to conclude that aliefs are therefore insensitive to changes in the world or without their own form of self-modification. There are two senses in which aliefs self-modify, both of which are evident in some of the cases that purport to show the inflexibility of alief-like states, such as cases of lagging habits. Consider Zimmerman's (2007) example of Hope, who continues to look for the trashcan under the sink even after she replaced it with a larger bin by the stove.

[31] There may be *something* 'normatively appropriate' in these phenomena, perhaps because they serve certain teleological roles. See Dretske, 1988 and Kacelnik, 2006.

The first, short-term type of self-modification is visible just after Hope's aliefs drive her to the wrong place. A salient feature of Hope's environment (*F*: 'coffee grinds!') induces feelings of tension (*T*: 'yuck!') and behavioral reactions (*B*: 'dispose under sink!'), but the response misfires and her tension persists unalleviated (*A*: 'still yuck!'). This unresolved tension in turn automatically activates further responses ('Not a trash can! Argh! Move to the stove! Ah … garbage dispensed'). Her automatic affective and motor responses are not just one-and-done reactions to a salient cue, but integrally related components that work in concert to guide her toward alleviation. Within immediate contexts, activated aliefs modify themselves by *eliminating* themselves. Rather than by updating to reflect the cumulative evidence, aliefs self-modify by compelling the agent to change her bodily orientation to the world so that the source of tension vanishes.[32]

Aliefs also self-modify in a more gradual way. An agent's sense of (un)alleviation in one context will contribute to the degree of attraction or repulsion she feels toward related actions in the future. This will strengthen or weaken the associative connections that obtain between perceptions of salient features, experiences of tension, initiation of motor routines, and new experiences of (un)alleviation. In this way, better or worse responses to felt tensions in particular situations guide agents toward *better and better* responses to felt tensions over time. This is true even when no belief revision is necessary. While Hope's beliefs about the trashcan's location update immediately, she need not have, and may be unable to form, accurate beliefs about the different amounts of force required for tossing out coffee grinds, banana peels and peach pits. The improvement in her ability to navigate through the kitchen efficiently, rather than awkwardly, is due to her gradually self-modifying aliefs. Her visceral sense of frustration when she errantly tosses some of the coffee grinds on the floor instead of in the bin will, *ceteris paribus*, make her less likely to repeat the mistake in the future.

4.2 Error

Another hallmark of normativity is the potential for error. DT dictates that aliefs can 'fail' only insofar as they fall out of concord with beliefs (as in cases of aversive racism). However, aliefs can also fail in their own way. There can be normatively deviant aliefs, just as there are normatively deviant beliefs.

In the case of the museumgoer, perhaps as soon as she steps back, she comes to feel too far from the painting. It could be that the appropriate behavior was to squint her eyes and tilt her head, rather than to move away. What determines whether the

[32] Consider also the fudge-avoider. Cringing and turning away from something that appears to be feces may seem inappropriate in a sense, but food safety isn't the only consideration driving the fudge-avoider's behavior. Even after the fudge-avoider embraces the belief that the food in front of her is safe to eat, it still *looks* gross. The alief is getting something right; visual presentation is a genuine part of gustatory experience. Cringing and turning away is the appropriate way to reduce the tension created when feces-shaped paraphernalia is thrust before one, even if it isn't feces.

reaction is adequate is whether it *alleviates* her sense of tension, regardless whether she ever judges that she is perceiving it from the right spot. A failure of this sort is manifest as *part* of her experience insofar as her sense of tension persists.

Automatic states with *F-T-B-A* content can succumb to error in a number of different ways. Suppose an interlocutor leans in to whisper something important. The addressee might rightly perceive the inward lean but under- or overreact *affectively*, by being coldly indifferent or disconcertingly solicitous. That would be a failure of feeling excessive or deficient tension, or feeling tension of the wrong sort (i.e. positive versus negative). Alternatively, the addressee might feel an appropriate degree of tension but under- or overcompensate *behaviorally*, by leaning in too little to hear the secret or leaning in so close that they bump heads. Yet another possibility for error arises even if the addressee feels the right tension and responds with an appropriate behavioral adjustment, but fails to feel the right *alleviation*. She might continue to fidget awkwardly after finding the optimal posture for listening to the whisperer. She can fail in any of these respects even when all of the conditions were right for her alief to reduce her tension appropriately (the conditions were, as Gendler would say, stable, typical and desirable). Somewhat ironically, the possibility that such aliefs can fail in *perfectly familiar* contexts shows that they are not ballistic causal reflexes but legitimately norm-sensitive responses, which are, no matter how well honed, always capable of getting things wrong.

One might object that these alleged failures of alief (failures to live up to the norms of reducing bodily tension) are better construed as failures of higher-order states to bring aliefs into line. In keeping with DT, Zimmerman (2007) suggests that phobias and lagging habits reflect failures to control or attend properly to one's actions. However, it's far from obvious that, when in perfectly normal conditions, an agent *should* be closely attending to or trying to control her (putatively *un*controllable!) automatic behaviors. Such exercises in self-control are as likely to be self-defeating as they are to be beneficial (Follenfant and Ric, 2010). In normal conditions, we can respond to context-determined opportunities for behavior automatically and appropriately, by, say, leaning in just the right amount as our interlocutor prepares to pass on the juicy gossip. But even in these conditions, when no attention or control is warranted, failure might still occur in any of the ways listed above. This would be a failure of alief proper, and not some other failure traceable back to attention, self-control, or other reflective states.

4.3 Belief, Evidence, and Reason

Given the weight that Gendler places on norm-insensitivity in her account of alief, one might take our case to count against the viability of alief as a psychological kind at all.[33] For example, Schwitzgebel (2010) cites the potential intelligence of putative aliefs to argue that aliefs cannot be coherently distinguished from beliefs:

[33] Schwitzgebel (2011) has voiced this concern about our argument, writing that we ought to 'reframe [our] view as a *criticism* of the concept of alief, rather than an adaptation of it'.

> Our habits, associations, and automatic responses *are*, to a substantial extent, responsive to evidence; and our verbal avowals or dispositions to judge are often *un*-responsive to evidence ... When I'm finally told that 'LOL' abbreviates 'laughing out loud' and not 'lots of love', my spontaneous responses do adjust, either swiftly or slowly. Evidence, whether presented verbally or encountered directly in the world, shapes my habits and associations, typically though not always in ways that we would rationally endorse if we considered it explicitly ... People judge in part automatically, associatively, and arationally, and they often show high intelligence in their habits and their unreflective, spontaneous responses (2010, pp. 539–540).

We agree that the distinction between belief and alief does not align neatly around 'smart' and 'dumb' ways of responding to the world. But we disagree when Schwitzgebel concludes that alief is an illegitimate concept because it separates out 'what is really an inseparable mix' (2010, p. 540). For one thing, it is not clear why the fact that a distinction has borderline cases should entail that there is no distinction at all. More substantively, we maintain that alief and belief are norm-sensitive in different senses, and are (or fail to be) in good standing in different ways. (And they are capable of coming into conflicts whose proper resolution is not always clear (p. 5).) To be in good standing, beliefs ought to 'fit' the available evidence. Aliefs in good standing move agents to reduce felt tensions in response to variations in the environment.[34]

Insofar as alief-like dispositions are responsive to changes in the environment, Schwitzgebel seems to infer that the relevant dispositions are changing in response to the evidence. But more needs to be said. That a state changes *when* the evidence changes does not indicate that that state is *responding* to the evidence. The state of

[34] One might worry that there are two 'directional' senses of normativity that need to be distinguished: an 'upstream' sense regarding the features to which aliefs and beliefs ought to *respond*, and a 'downstream' sense regarding the effects which aliefs and beliefs ought to *bring about*. It might then seem that our discussion here is incomplete, insofar as we are focusing on the 'causes' of belief (changes in evidence) but the 'effects' of alief (movements toward tension-reduction). This distinction is important, and much more could be said about it, but it is unclear how one could usefully separate these normative dimensions for the purpose of contrasting aliefs with beliefs. For one, the 'causes' and 'effects' of well-functioning aliefs are integrated; the content of alief includes mutually descriptive and directive aspects, simultaneously taking the world to be a certain way and prescribing a certain way of responding. Second, because beliefs often lack prescriptive content, it is not clear what 'effects' they ought to have. Perhaps well-formed beliefs ought to play some sort of role in the generation of behavior. But in this case, it is not clear how to separate aliefs and beliefs. Both aliefs and beliefs are implicated in intentional actions, and both can reliably give rise to situationally appropriate behavior as well. (See footnote 29 regarding whether aliefs, like beliefs, can give rise to fully-fledged actions.) Finally, it might be that the very idea that beliefs ought to have certain effects is strange. To say that beliefs ought to fit the evidence is to say, inter alia, that *changes* in evidence ought to cause *changes* in belief. The changes in belief are, in this case, *among* the normatively required effects. Thanks to an anonymous reviewer for raising this concern.

an agent's umbrella often changes when she acquires evidence that it is raining, but that doesn't show that the umbrella is changing in light of the evidence as such, i.e. changing *because* the available considerations justify doing so. The evidence-insensitivity of alief is made clear again and again in Gendler's examples, such as the Skywalker. The evidence makes it immediately, overwhelmingly clear that the Skywalk is safe, and the agent acknowledges as much. (The evidence here is not just linguistic or testimonial; she perceives the solidity of the walkway as she and others stand safely upon it.) But her aliefs are insensitive to the unambiguous evidence. It would not be enough to show that the aversion to the Skywalk would fade gradually if one stood on it constantly for days on end. This would show that the agent's aliefs were responding to changes in the world, but not to those changes *qua* evidence.[35] The same is true of our cases. The museumgoer's impulse to step back will likely persist even if she reads Newman's instructions that the artwork is meant to be seen from a short distance; the distance-stander will still feel compelled to lean back even if she believes that doing so might appear rude to her interlocutor.

It would also be misleading to suggest that the intelligence of some automatic-affective states makes them *reason-responsive* in some substantive sense. Peter Railton argues, for example, that the 'fluent agency' of artistic and athletic virtuosos—whose non-deliberative behaviors he likens to automatic-but-flexible reflexes—are 'clearly done *for reasons*, and, moreover, for reasons *as such*' (2009, pp. 97–98, emphasis in original).[36] He distinguishes acting for reasons 'as such' from acting in response to some deviant causal chain or 'robotically enacting a habit or routine' (2009, p. 98). But again, more needs to be said (see Brownstein, in preparation(b)). We agree with Railton that fluent agents (and alievers in good standing) are neither acting robotically nor deliberatively reasoning. And Railton is also surely right that there *are* reasons for these agents to act as they do. There are good reasons for the museumgoer to back up from the large painting, viz. to get the best view of it. The question, though, is whether this reason explains the museumgoer's action. In our view, for a reason to explain an action-guiding state in non-deliberative

[35] Even beliefs that fail to be moved by the evidence can play other recognizably cognitive roles in an agent's psychological economy. When one belief-like state fails to change in the face of unambiguously defeating considerations, some *other* belief will change. The agent may respond by discounting the evidence or considering a way in which the apparent inconsistencies might be resolved (Gawronski and Bodenhausen, 2006). Even the most stubborn beliefs must be capable of these sorts of inferential roles in order to be intelligible as beliefs. And aliefs don't do this. The Skywalker does not, on the basis of her alief, call the evidence of her safety into question, as if it they were two weights on a scale, and either one could be given up. She may recognize that her fear is out of step with the evidence, but this is just an observation of her own state based on the fact that her heart is pounding like crazy.

[36] Arpaly (2004) and Arpaly and Schroeder (2012) draw upon similar examples, to similar ends. On Arpaly's view, 'a major part of what it is to be a competent tennis player is to . . . act for good reasons rather than bad reasons in all your game-related actions' (2004, p. 53). This may be true of 'competent' players, but *experts*, in most cases, do not hit shots *for* reasons. When they do, they often choke. See Section 3.1 above and DeCaro *et al*., 2011.

contexts, the reason must be at least available for practical reasoning. On a view we find intuitive, practical reasoning is the capacity for resolving, through reflection, questions about what to do (see, e.g., Wallace, 2003/2008). An action-guiding state is reason-responsive just insofar as it is capable, *ceteris paribus*, of figuring in practical reasoning in the right ways, perhaps by mediating inferences about the appropriate means for achieving one's ends. A psychological state has to meet a number of conditions to play these roles, and we doubt that aliefs can meet them.

As we have described them, aliefs offer an agent a sense of what she 'ought' to do. This 'ought' is *part* of her experience. Aliefs elicit behaviors that manifest in phenomenal experience (perhaps peripherally, perhaps focally) as *better* or *worse* responses to feelings of tension. In this way, occurrent aliefs include mutually descriptive and directive aspects, simultaneously taking the world to be a certain way and prescribing a certain way of responding (see Section 3.1). Because the descriptive and directive are initially inseparable aspects of their intentional content, it is a difficult cognitive achievement for a deliberative agent to 'break aliefs down' into distinct and articulated beliefs and desires. This makes them impressively incapable of integrating inferentially with other psychological states, and provides a straightforward explanation for why they predictably fail to respond appropriately to defeating evidence or incorporate effectively into practical reflection.[37]

5. Harder Cases

We have suggested that aliefs can be norm-sensitive in virtue of their responsiveness to affective states of disequilibrium. Responsiveness to such affective states is flexible, self-modifying, and capable of error. It is a genuinely normative phenomenon.

But one might think this is pretty small potatoes. What about aversive racism and other cases of conflict between alief and belief in which aliefs seem to be getting things deeply wrong? Are these cases of aliefs functioning appropriately—sensitive to their proprietary norms—but insensitive to ethical norms? Does this undermine the normative status of alief altogether? One might be inclined to concede that aliefs have some *pro tanto* normative justification in cases like the museumgoer's, when aliefs operate in the absence of relevant beliefs. But surely, one might continue, because aliefs inevitably lag behind the evidence to which beliefs typically update

[37] We say more about why and how aliefs are incapable of inferential integration in this way in Brownstein and Madva, 2012; Madva, 2012; and Brownstein, in preparation(b). Empirical evidence suggests that aliefs are insensitive to the logical form of other states. For example, psychologists Gawronski and Bodenhausen argue that, 'the basic notion of these studies is that the mere co-occurrence between two objects can create a mental association between these objects, even though the validity of the implied relation is rejected at the propositional level' (2009, p. 207). An alternative to this associative view would be that aliefs are cognitively encapsulated belief-like states (Egan, 2011; Mandelbaum, 2012). Either way, they would be incapable of playing the necessary roles in practical reflection.

in a flash, the *pro tanto* justification of alief is wholly defeated when aliefs and beliefs collide head-on. Endorsing the normative authority of aliefs (such as those that dispose aversive racists to biased hiring practices and discriminatory microbehaviors) seems tantamount to what Gendler calls '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)' (2008b, p. 578).

There is no way of avoiding the fact that aliefs which successfully reduce felt tensions can lead to ethically repugnant behavior. An ethics of alief must address how best to regulate these unwanted aliefs. But the existence of ethically repugnant aliefs does not show that belief-driven behavior is always superior to alief-driven behavior or that only belief-driven behavior can be truly norm-sensitive. Alief is normatively 'subordinate' to belief if and only if, in cases of alief-belief conflict, an agent should act on the basis of her beliefs. But in some cases of belief-behavior mismatch, beliefs get things wrong while automatic motor-affective states get things right.

Consider an 'aversive egalitarian'.[38] This is an avowed racist who nevertheless behaves in egalitarian ways. Social psychologists have much to tell us about well-meaning, clearheaded agents who bear regrettably biased dispositions, but very little about intellectually muddled agents who harbor morally upright dispositions. Arpaly (2004) argues persuasively that the literary character Huckleberry Finn is just such a person. Huck's dilemma is whether to turn in his friend Jim, an escaped slave. Huck believes that an escaped slave amounts to a stolen piece of property and that stealing is wrong, but he is also loyal to his friend. The result of his (less-than-ideal) deliberation is that he ought to turn Jim in, but Huck finds himself unable to do it. Could it be that Huck's behavior is guided by an alief, one that (happily) has trumped his racist beliefs? The roles that Huck's pro-Jim attitudes play in his deliberation, as an unwelcome 'gut reaction' that can be neither internally justified nor 'reasoned away', have all the paradigmatic trappings of automatic-affective effects on reflective judgment.[39] Merely by looking at Jim, Huck feels repelled by what he believes he ought to do. Huck's last-second sensitivity to Jim's suffering or personhood isn't a lucky accident. It's genuinely admirable.

While Huck is fictional, there is good reason to think that he is not, in the relevant respects, unusual—and just how unusual he is remains an empirical question that ought to be explored. Huck resembles someone who reflectively

[38] The following discussion of aversive egalitarians overlaps with what we say in Section 8 of our companion paper (Brownstein and Madva, 2012).

[39] Perhaps the fact that Huck deliberates about what to do makes his example less than ideal for illustrating norm-sensitive automatic action. However, it would be quite easy to imagine a modified case in which Huck reflectively judges that it would be right to turn Jim in, and then at the very last moment, independently of his considered beliefs and outside of his control, automatically acts in a different way. As he is about to turn his friend in, imagine that Huck feels a lump in his throat, tension in his body, and the sweat on his palms.

judges that homosexuality is wrong on religious grounds but, perhaps because she has gay friends or family members, cannot help but show tacit sympathy (rather than disgust) toward images from a Gay Pride March. To our knowledge, no one has studied aversive egalitarians in this way. It is not surprising that there have not been many studies on people like this, because most people who come into psychology labs do not openly avow racism or homophobia at all. The under-investigation of such empirical possibilities furnishes no evidence that they don't regularly occur, however. There have, in fact, been a handful of related studies on biased social attitudes which many participants openly avow, such as associations of women with supportive qualities (e.g. nurturance) and men with leadership qualities (e.g. assertiveness). For example, Dasgupta and Asgari (2004) found that some female college students continued to explicitly endorse the view that women possess more supportive than leadership qualities, even after these associations were no longer apparent on implicit measures. Were the students' automatic dispositions flexibly tracking variations in the world, while their beliefs barely budged? One might object that this change in automatic dispositions seems to reflect the idea that students' 'gendered aliefs' were responding to a sort of evidence after all. Perhaps students tacitly judged the existence of their female professors to be evidence that women could also be assertive. These professors were certainly counterstereotypical exemplars, but we suspect that their influence on students' automatic associations may have been mediated more by repeated iterations of tension-reducing aliefs than by an inference (from counterstereotypical exemplars to falsity of stereotypes). The students might have felt more inclined to mimic the assertive postures of their female professors, for example. But these sorts of possibilities are not being empirically pursued. For who would pursue them if the automatic dispositions involved are considered norm-insensitive or even irrational by definition?[40]

6. Conclusion

The ethical dilemmas posed by cases of conflict between aliefs and beliefs are very real. But the implicated automatic-affective states are not in all cases so vulnerable to distortion or indifferent to reality. When agents' reflective beliefs fail to be properly sensitive to the evidence, their well-formed automatic affective responses may yet guide them in the right direction. Keeping in view the different respects in which

[40] In cases of alief-belief conflict, how can one tell which state, if any, is getting things right? We hazard an answer to this difficult question elsewhere (Brownstein and Madva, 2012), but it suffices for our purposes that *both* alief and belief can steer us wrong in such cases. In Gendler's cases, aliefs that successfully reduce tensions might be ethically undesirable. In other cases, beliefs may be misled by the evidence or fail to track it in the right way, while aliefs are attuned to something important. Presumably Huck's deliberation should have taken into account Jim's personhood, but similarly situated agents less sensitive to, say, the felt tensions of interpersonal behavior would have been unable to question or overcome such mistakes in judgment.

aliefs and beliefs can be and fail to be norm-sensitive will be vital for understanding the implications of automaticity in general. Further research and reflection into the 'ethics of automaticity', for example, must ask what our automatic dispositions are (and are not) good for and when they can (and cannot) be trusted.

Department of Humanities,
New Jersey Institute of Technology

Department of Philosophy,
Columbia University

References

Abstract Expressionist. New York, *Museum of Modern Art*, 2010–2011.

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.

Arpaly, N. 2004: *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Arpaly, N. and Schroeder, T. 2012: Deliberation and acting for reasons. *Philosophical Review*, 121(2), 209–239.

Bargh, J., Chen, M., and Burrows, L. 1996: The automaticity of social behavior. *Journal of Personality and Social Psychology*, 71, 230–44.

Barrett, L. F., Oschner, K. N. and Gross, J. J. 2007: On the automaticity of emotion. In J.A. Bargh (ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*. Philadelphia, PA: Psychology Press.

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.

Brownstein, M. In Preparation(a): Automatic agency.

Brownstein, M. In Preparation(b): Nondeliberative action.

Brownstein, M. and Madva, A. 2012: Ethical automaticity. *Philosophy of the Social Sciences*, 42, 68–98.

Chemero, A. 2009: *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Clark, A. 1997: *Being There*. Cambridge, MA: MIT Press.

Colombetti, G. 2007: Enactive Appraisal. *Phenomenology and the Cognitive Sciences*, 6, 527–46.

Dasgupta, N. and Asgari, S. 2004: Seeing is believing: exposure to counterstereotypic women leaders and its effect on automatic gender stereotyping. *Journal of Experimental Social Psychology*, 40, 642–58.

De Becker, G. 1998: *The Gift of Fear*. New York: Dell.

DeCaro, M., Albert, N., Thomas, R. and Beilok, S. 2011: Choking under pressure: multiple Routes to skill failure. *Journal of Experimental Psychology*, 140, 390–406.

Deutsch, R. and Strack, F. 2010: Building blocks of social behavior: reflective and impulsive processes. In B. Gawronski and B.K. Payne (eds), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press, 62–79.

Deweese-Boyd, I. 2006/2008: Self-Deception. Stanford Encyclopedia of Philosophy. In E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/self-deception/

Dijksterhuis, A., Chartrand, T. L. and Aarts, H. 2007: Effects of priming and perception on social behavior and goal pursuit. In J. A. Bargh (ed.), *Social Psychology and the Unconscious: The Automaticity of Higher Mental Processes*. Philadelphia, PA: Psychology Press, 51–132.

Dovidio, J. F., Kawakami, K. and Gaertner, S. L. 2002: Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology*, 82, 62–8.

Dretske, F. 1988: *Explaining Behavior*. Cambridge, MA: MIT Press.

Dreyfus, H. and Kelly, S. 2007: Heterophenomenology: heavy-handed sleight-of-hand. *Phenomenology and the Cognitive Sciences*, 6, 45–55.

Egan, A. 2011: Comments on Gendler's 'The epistemic costs of implicit bias'. *Philosophical Studies*, 156, 65–79.

Elga, A. Manuscript: Belief fragmentation.

Evans, G. 1982: *The Varieties of Reference*. Oxford: Oxford University Press.

Follenfant, A. and Ric, F. 2010: Behavioral rebound following stereotype suppression. *European Journal of Social Psychology*, 40, 774–82.

Gawronski, B. and Bodenhausen, G. V. 2006: Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5), 692–731.

Gawronski, B. and Bodenhausen, G. V. 2009: Operating principles versus operating conditions in the distinction between associative and propositional processes. *Behavioral and Brain Sciences*, 32, 207–08.

Gendler, T. S. 2008a: Alief and belief. *The Journal of Philosophy*, 105(10), 634–63.

Gendler, T. S. 2008b: Alief in action (and reaction). *Mind & Language*, 23, 552–85.

Gertler, B. 2011: Self-knowledge and the transparency of belief. In A. Hatzimoysis (ed.), *Self-Knowledge*. Oxford: Oxford University Press.

Gilbert, D. T. 1991: How mental systems believe. *American Psychologist*, 46, 107–119.

Huddleston, A. 2011: Naughty beliefs. *Philosophical Studies*, Online First, 24 February, doi: 10.1007/s11098-011-9714-5.

Huebner, B. 2009: Trouble with stereotypes for Spinozan minds. *Philosophy of the Social Sciences*, 39, 63–92.

Hunter, D. 2011: Alienated belief. *dialectica*, 65, 221–40.

Jeannerod, M. 2006: *Motor Cognition: What Actions Tell to the Self*. Oxford: Oxford University Press.

Kacelnik, A. 2006: Meanings of rationality. In S. Hurley and M. Nudds (eds), *Rational Animals?* Oxford: Oxford University Press.

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, 871–88.

Klaassen, P., Rietveld, E. and Topal, J. 2010: Inviting complementary perspectives on situated normativity in everyday life. *Phenomenology and the Cognitive Sciences*, 9, 53–73.

Kovel, J. 1970: *White Racism*. New York: Columbia University Press.

Lazarus, R. S. 1991: *Emotion and Adaptation*. New York: Oxford University Press.

Libet, B., Gleason, C. A., Wright, E. W. and Pearl, D. K. 1983: Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). The unconscious initiation of a freely voluntary act. Brain, 106, 623–42.

Madva, A. 2012: *The Hidden Mechanisms of Prejudice: Implicit Bias & Interpersonal Fluency*. PhD diss., Columbia University, New York.

Mandelbaum, E. 2012: Against alief Online First, 25 April, doi: 10.1007/s11098-012-9930-7.

McConnell, A. R. and Leibold, J. M. 2001: Relations among the implicit association test, discriminatory behavior, and explicit measure of racial attitudes. *Journal of Experimental Social Psychology*, 37, 435–42.

McKay, R. T. and Dennett, D. C. 2009: The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–561.

Merleau-Ponty, M. 1962/2002: *The Phenomenology of Perception,* trans. C. Smith. New York: Routledge.

Millikan, R. 1995: Pushmi-pullyu representations. *Philosophical Perspectives*, 9, 185–200.

Milner, A. D. and Goodale, M. A. 1995: *The Visual Brain in Action*. Oxford: Oxford University Press.

Milner, A. D. and Goodale, M. A. 2008: Two visual systems re-viewed. *Neuropsychologia*, 46, 774–85.

Muller, H. and Bashour, B. 2011: Why alief is not a legitimate psychological category. *Journal of Philosophical Research*, 36, 371–89.

Moors, A., Spruyt, A., and De Houwer, J. 2010: In search of a measure that qualifies as implicit: recommendations based on a decompositional view of automaticity. In B. Gawronski and B. K. Payne (eds), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press, 19–37.

Payne, B. K. and Cameron, C. D. 2010: Divided minds, divided morals: how implicit social cognition underpins and undermines our sense of social justice.

In B. Gawronski and B. K. Payne (eds), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press, 445–460.

Payne, B. K. and Gawronski, B. 2010: A history of implicit social cognition: where is it coming from? Where is it now? Where is it going? In B. Gawronski and B. K. Payne (eds), *Handbook of Implicit Social Cognition: Measurement, Theory, and Applications*. New York: Guilford Press, 1–15.

Peacocke, C. 1999: *Being Known*. Oxford: Oxford University Press.

Peacocke, C. 2004: *The Realm of Reason*. Oxford: Oxford University Press.

Pearson, A. R., Dovidio, J. F. and Gaertner, S. L. 2009: The nature of contemporary prejudice: insights from aversive racism. *Social and Personality Psychology Compass*, 3, 1–25.

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

Railton, P. 2009: Practical competence and fluent agency. In D. Sobel and S. Wall (eds), *Reasons for Action*. Cambridge: Cambridge University Press, 81–115.

Rietveld, E. 2008a: Situated normativity: the normative aspect of embodied cognition in unreflective action. *Mind*, 117, 973–1001.

Rietveld, E. 2008b: Unpublished dissertation. *Unreflective Action*.

Rowbottom, D. P. 2007: 'In-between believing' and degrees of belief. *Teorema*, 26, 131–7.

Schwitzgebel, E. 2006: Belief. In E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/belief/

Schwitzgebel, E. 2010: Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly*, 91, 531–553.

Schwitzgebel, E. 2011: Commentary on Brownstein and Madva, 'Alief and affordance: the normativity of automaticity'. Read at the Pacific Division meeting of the American Philosophical Association, 20 April 2011.

Varela, F. J. and Depraz, N. 2005: At the source of time: valence and the constitutional dynamics of affect. *Journal of Consciousness Studies,* 12, 61–81.

Wallace, J. 2003/2008: Practical reason. In E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. Available at: http://plato.stanford.edu/entries/practical-reason/

Witt, J. K., Proffitt, D. R. and Epstein, W. 2004: Perceiving distance: a role of effort and intent. *Perception*, 33, 570–90.

Wittgenstein, L. 1966: *Lectures and Conversations on Aesthetics, Psychology and Religious Belief*. Oxford: Blackwell.

Wegner, D. 2002: *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Zimmerman, A. 2007: The nature of belief. *Journal of Consciousness Studies*, 14(11), 61–82.