Ethical Automaticity

Michael Brownstein and Alex Madva

Philosophy of the Social Sciences 2012 42: 68 originally published online 15 December 2011
DOI: 10.1177/0048393111426402

Philosophy of the Social Sciences
42(1) 68–98
© The Author(s) 2012
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0048393111426402
http://pos.sagepub.com
SAGE

Ethical Automaticity

Michael Brownstein[1] and Alex Madva[2]

Abstract

Social psychologists tell us that much of human behavior is automatic. It is natural to think that automatic behavioral dispositions are ethically desirable if and only if they are suitably governed by an agent's reflective judgments. However, we identify a class of automatic dispositions that make normatively self-standing contributions to praiseworthy action and a well-lived life, independently of, or even in spite of, an agent's reflective judgments about what to do. We argue that the fundamental questions for the "ethics of automaticity" are what automatic dispositions are (and are not) good for and when they can (and cannot) be trusted.

Keywords

automaticity, implicit bias, alief, virtue

1. Introduction

Much of what we do, we do automatically. We often act effortlessly, efficiently, uncontrollably, and unconsciously, whether in setting and pursuing goals, making moral evaluations, or "reading" the gestures and body lan-

Received 21 September 2011

[1]New Jersey Institute of Technology, Newark, NJ
[2]Columbia University, New York, NY

Corresponding Author:
Michael Brownstein, Department of Humanities, NJIT, University Heights, Newark, NJ, 07102-1982
Email: msb@njit.edu

guage of others.[1] With the mounting evidence that automaticity is pervasive comes the recognition that our automatic dispositions are often discordant with our reflective judgments. For example, you might be an "aversive racist" who embraces the goal of being egalitarian but also demonstrates all kinds of prejudiced automatic dispositions. Research on automaticity is predominantly concerned with cases of discordance like this, in which an agent's automatic dispositions seem to consist in merely causal, nonrational mechanisms. Philosophical and psychological interest in the *ethical* upshots of automaticity has centered on how we can regulate these putatively nonrational mechanisms in accordance with our considered ends. It is often thought, or taken for granted, that an automatic disposition is ethically desirable if and only if it is suitably governed by an agent's reflective judgments.

Adjusting the automatic to suit the reflective is the right thing to do in cases like aversive racism. But many automatic dispositions do not fit this mold. We identify a class of automatic dispositions that make normatively self-standing contributions to praiseworthy action and a well-lived life. In some cases, these dispositions promote praiseworthy action *in spite of* being discordant with an agent's reflective judgments; in other cases, they promote praiseworthy action when there simply are not any relevant reflective judgments to be concordant *with*. These ethical automatic dispositions are flexibly adaptive to changes in the environment and capable of certain kinds of error. Consequently, the fundamental questions for the "ethics of automaticity" are not simply how to regulate, control, or change automatic dispositions but rather *what* automatic dispositions are (and are not) good for and *when* they can (and cannot) be trusted.

2. The Ethics of Automaticity in Social Psychology and Philosophy

When social psychologists turn their attention to the ethics of automaticity, their focus tends to be on the study of the regulation of unwanted inclinations and impulses. This is a natural consequence of the prevailing theoretical constructs in the discipline. Roland Deutsch and Fritz Strack (2010, 63) claim, for example, 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." This psychological discordance

[1] On automatic goal setting, see Bargh, Chen, and Burrows (1996) and Dijkersthuis, Chartrand, and Aarts (2007). On automatic moral evaluation, see, for example, Greene and Haidt (2002) and Levy and Bayne (2004). On body language and gesture, see Goldin-Meadow and Beilock (2010) and Aviezer et al. (in press).

arises because implicit cognition is essentially automatic, uncontrollable, and outside of awareness, whereas full-fledged beliefs and intentions can form and revise in the light of an agent's reflective deliberation. Cases of discordance result in "irrational behavior," where Deutsch and Strack define "irrational" as "the case in which behavior occurs against the actor's explicit beliefs" (70).[2]

The research paradigm of belief-behavior discordance has been fruitful. It helps to illuminate, for example, how significant racial disparities can persist in the United States even though most Americans sincerely disavow racism. Social psychologists have made a compelling case that one facet of this complex problem stems from the fact that many Americans occupy the conflicted state of aversive racism. Despite their avowed egalitarianism, aversive racists are more likely to hire a white job candidate over an equally qualified black candidate, more likely to find a black defendant guilty than a white defendant with equally incriminating evidence, and more likely to exhibit a range of discriminatory "microbehaviors" (e.g., they tend to make less eye contact, make more speech errors, and sit further away from black interlocutors).[3] Aversive racists tend to be unaware that they are prone to these prejudiced judgments and behaviors.

The causes of aversive racism are complex, but a principal source of the rogue automatic dispositions, in this and similar cases of belief-behavior discordance, is thought to be the agent's repeated exposure to biased representations of social groups.[4] Deutsch and Strack suggest (2010, 64-65) that an

[2] Deutsch and Strack's account of "reflective and impulsive processes" underlying social behavior has been called "the most influential model" in their field (Payne and Gawronski 2010). See the other chapters in the *Handbook of Implicit Social Cognition*—for example, Payne and Cameron's chapter (2010) explores the implications of this framework for issues of social justice.

[3] We discuss aversive racism, which was coined as such by Kovel (1970), in more depth in a companion article (Brownstein and Madva, "The Normativity of Automaticity," under review). For a survey of relevant empirical literature, see Pearson, Dovidio, and Gaertner (2009). For hiring bias, see Bertrand and Mullainathan (2003); for juror bias, see Levinson and Young (2010); 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, 446). Also see Huebner (2009) for a philosophically oriented review of relevant literature.

[4] Another principal source is thought to be evolved mechanisms for making in-group and out-group discriminations based on coalitional alliances (Kurzban, Tooby, and Cosmides 2001). The sources of unwanted automatic associations will differ from case to case, and such automatic dispositions will interact in complex ways.

individual can form automatic associations of "Arab" with "terror" in response to relentless media coverage, regardless whether the individual would reflectively judge that most Arabs are terrorists. Upon hearing "It is wrong to identify Arabs with terrorism," an individual may consciously conclude that the utterance is true on the basis of the evidence, even as the mere conjunction of the terms reinforces an automatic disposition to associate Arabs with terror. These unendorsed associations can subsequently shape her thought and microbehaviors in myriad ways.

In cases like this, it is natural to regard the automatic nature of implicit cognition as a source of irrationality, which is relevant to ethics insofar as automatic dispositions inhibit an agent from acting on the basis of her reflectively endorsed beliefs and intentions. Social psychologists have focused on what to do about cases like these. As Keith Payne and C. Daryl Cameron put it (2010, 456), "knowing how implicit cognition can cause our ethicality to corrode can also help us engage better moral self-regulation in pursuit of our ideals." Work on the social and political implications of automaticity has focused on how to counteract biasing effects in, for example, legal trials and hiring decisions (Jolls and Sunstein 2006). Work on the implications for health psychology, still in its nascent stages, has focused on how to counteract the automatic forces driving substance abuse and overeating (Wiers et al. 2010).

The idea that automaticity is primarily a source of irrationality has also largely carried over into attempts to integrate the empirical literature with philosophy of mind and ethics. In an innovative approach to belief-behavior discordance, Tamar Szabó Gendler (2008a, 2008b) hones in on a class of automatic dispositions that she calls "aliefs." More primitive than *be*lief, an *a*lief is a relatively inflexible disposition to react automatically to an apparent stimulus with certain fixed affective responses and behavioral inclinations (2008b, 557-60). In Gendler's parlance, a firm *be*liever that superstitions are bogus may yet be an abiding *a*liever who cowers before black cats and sidewalk cracks. An agent may be sincerely committed to antiracist *be*liefs but simultaneously harbor racist *a*liefs. What fundamentally distinguishes aliefs from beliefs is that, while each plays an important role in guiding behavior, beliefs are capable of being revised in light of the all-things-considered evidence and aliefs are not (Gendler 2008b, 566). Beliefs reflect what an agent takes to be true, while aliefs are yoked to how things merely seem. Aliefs are evidence insensitive in this way because they are automatic, associative, and arational (Gendler 2008a, 641-66). Their causal origins may be instinctual or habitual (Gendler 2008b, 568-70). In some cases, aliefs can change with changes in habit (566).

Gendler captures the overall structure of alief by suggesting that it has a distinctive kind of intentional content, with three components: a representation of some apparent state of affairs, an affective response, and a behavioral reaction, such as "Sidewalk crack! Scary! Avoid!" According to Gendler, states with this representational-affective-behavioral (*R*-*A*-*B*) content explain a wide array of otherwise-puzzling cases of belief-behavior discordance, including not only aversive racism but also phobias, fictional emotions, and bad habits (2008b, 554). In fact, Gendler suggests (2008a, 663) that aliefs are causally responsible for much of the "moment-by-moment management" of human behavior—whether that behavior is belief concordant or not.[5] The affective component of an activated alief may not be experienced as a fully articulated emotion, and the behavioral component may not be expressed as a fully intentional action, but even the most subtle aliefs may play a pivotal role in mind and behavior.

Despite their pervasive influence on behavior, Gendler argues (2008b, 572) that aliefs are fundamentally insensitive to norms and their ethical standing depends entirely on the extent to which they are brought "into line with our considered commitments." Aliefs are, on Gendler's view, neither good nor bad in and of themselves but only good or bad to the extent that they are in "harmony" with an agent's considered beliefs and intentions. Furthermore, not just any harmony will do. For Gendler, an agent's considered beliefs ought to be *in charge* of her aliefs: "the well-functioning aliever is one whose aliefs and beliefs largely coincide (or one whose ability to suppress contrary impulse is strong)" (2008a, 651).

We dub this the "top-down harmony" (TDH) view. At its core, TDH is the view that an automatic disposition such as an alief is in good ethical standing if and only if it is governed by an agent's considered beliefs and

[5] 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, Chartrand, & Aarts 2007). In this regard, research on automaticity is informed by Libet and colleagues' (1983) research suggesting that the initiation of action is unconscious and by Milner and Goodale's (1995) hypothesis that behavior is largely controlled by a nonconscious neural system. These neural dissociations are also thought to be true of nonhuman animals, and Gendler claims (2008a, 641) that "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." We are sympathetic with Gendler's claims about conceptual and developmental antecedence, but we lack the space to address them here.

ends.[6] It is ethically undesirable for aliefs to drive behavior in cases of discordance. Aliefs are only trustworthy guides to action in highly "stable, typical, and desirable" contexts in which an agent can safely assume that they will be largely belief concordant (Gendler 2008b, 554, 570-72).

Hence the principal ethical question Gendler raises (2008b, 554), in keeping with the predominant focus in social psychology, is "how we might regulate and respond to discordant alief." Her efforts to answer it involve interpreting relevant empirical findings in light of the storied ethical tradition reaching back to Plato, according to which harmony is achieved by mastering one's unreflective impulses. Gendler identifies (554) two principal strategies: an Aristotelian approach that "involves the cultivation of alternative habits through deliberate rehearsal" and an early-modern Cartesian approach that "involves the refocusing of attention through directed imagination." Gendler takes as paradigmatic the question of how a committed egalitarian might undo her automatic prejudices (i.e., discordant aliefs). The Aristotelian approach recommends that an agent identify particular prejudiced aliefs and implement novel habits to counteract them. In this vein, Gendler cites evidence that implicit measures of racial bias are significantly decreased after participants repeatedly practice negating stereotypic associations (Kawakami, Dovidio, Moll, Hermsen, and Russin 2000). The early-modern approach has less to do with overhauling aliefs in the long term and more with circumventing their pernicious influence in specific cases. A committed agent can activate belief-concordant aliefs by actively imagining more just or preferable states of affairs. In this vein, Gendler cites evidence that implicit measures of gender bias are significantly lower after subjects spend 5 minutes imagining a strong (hence counterstereotypical) woman (Blair, Ma, and Lenton 2001).

[6] For Gendler, the ethical import of alief and automaticity is effectively exhausted by the ethics of top-down harmony (TDH). Regarding possible exceptions to TDH, Gendler points out (2008b, 554) that some belief-discordant aliefs may be innocuous or therapeutic: "Sometimes this discord is deliberate and welcome: daydreaming, rollercoasters and therapy all exploit our capacity for belief-discordant alief." These are all cases, however, in which agents' beliefs are attuned to reality while their aliefs are not. It is an interesting question how such reality-*in*sensitive aliefs contribute to a well-lived life. Gendler does not consider the cases of interest to us, in which our aliefs get it right, either *independently of* or *in spite of* our beliefs.

The cultivation of alternative habits and the refocusing of attention are promising responses to the problems raised by aversive racism and phobic disorders.[7] As vital as these recommendations are for specific purposes, however, they substantially underestimate the role that automaticity plays in a well-lived life. Even though we disagree with some of Gendler's conclusions, we take alief to be a promising tool for illuminating the role that automaticity plays in ethical behavior and thus proceed in our discussion by giving an account of "ethical aliefs."[8]

3. The President-Elect's Grin

Often, an agent's habit-based automatic-affective responses—or, in a sense we will explain, her "feel" for what to do in a given situation—are self-standing guides to praiseworthy action. We take the following anecdote to exemplify a type of case that remains empirically underexplored. Where possible, we refer to relevant empirical research.

Those readers who watched President Obama's swearing in on 20 January 2009 might remember the series of slight flubs that took place between the

[7] It should be noted that endorsing the viability of alief as a psychological concept is not strictly necessary for taking these practical recommendations to heart. For example, Huebner (2009) reaches similar ethical conclusions despite drawing on different empirical research (e.g., Gilbert 1991) and construing the rogue automatic dispositions in different terms, as "stereotype-based judgments." Stereotype-based judgments issue from "Type-1 processes," which function like Gendler's aliefs in important respects; they are nonrational processes that unfold automatically and independently of an agent's "Type-2 processes"—that is, independently of the considered commitments that an agent would form upon reflection. In keeping with TDH, Huebner writes (2009, 75), "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." Both Gendler and Huebner, like Deutsch and Strack, construe the operative automatic dispositions as merely causal mechanisms, which are of ethical concern first and foremost because they have to be regulated by our higher powers of ratiocination.

[8] We provide a more thoroughgoing defense of alief and how best to revise it in a companion article. For criticism of alief, see Egan (forthcoming), Mandelbaum ("Against Alief," unpublished manuscript), Muller and Bashour (2011), and Schwitzgebel (2010a, 2010b).

(for just a few more moments) president-elect and the chief justice of the Supreme Court, John Roberts. Both men seemed overwhelmed by the moment. Obama started to respond with "I, Barack . . . " before Roberts had completed the first phrase of the oath, "I, Barack Hussein Obama, do solemnly swear . . . " The chief justice then botched the next phrase of the oath, saying, "that I will execute the office of president *to* the United States *faithfully*" rather than "that I will *faithfully* execute the office of president *of* the United States." The media chatter understandably focused on whether the oath "counted" or not, and it was privately readministered a day later.[9] But in the news coverage, something important was overlooked. With a slightly puzzled brow, Obama hesitated before repeating Roberts's botched phrasing, then smiled widely and nodded slightly to Roberts, as if to say, "It's okay, go on." These gestures received little explicit attention, but they defused what could have been a disastrously awkward situation. As much as spectators may have cringed in the moment, the tenor of the unfolding ceremony and the subsequent analyses of it could have been drastically different. What if Obama had reacted not with nonverbal cues of affirmation but with, say, the pressed lips and narrowed eyes of contained anger? What if he had rolled his eyes and interjected, "Ok, Chief, let's take it from the top?" In the midst of what was to be a fully scripted and carefully monitored situation, Obama's impromptu grin upheld the positive momentum of the moment, maintained a conversational, if not ideological, rapport between Obama and Roberts, and deftly kept the inauguration moving forward. It would be clear to anyone watching that Obama's grin neutralized the awkwardness and navigated the moment as well as possible. Despite his nervousness, his social adeptness was on display.

We take this to be a high-profile example of a humdrum way in which automatic, alief-like dispositions can be praiseworthy. Obama simply reacted to the demands of the situation. His skilled behavior resembles the way that one might react tactfully (rather than awkwardly) to a "close talker" who leans in a little too close, by subtly stepping backward; the way that one might show deference to an opponent (rather than disdain) after suffering a hard defeat in a tennis match, by offering a firm handshake; or the way that one might step into a busy store because an inchoate sense of danger is

[9] For a clip of the event and an "analysis" of the miscues by CNN's Jeanne Moos, see http://www.youtube.com/watch?v=EyYdZrGLRDs.

making one's hair stand on end, even though nothing is identifiably wrong.[10] Gendler does not explicitly discuss effortless social behaviors like these, but in important respects, they resemble the "moment to moment" behaviors aliefs are said to manage. Like aliefs, these unreflective responses are affect laden, in that they arise in response to a "felt sense" that things have gone awry, and, like aliefs, they are essentially automatic and recalcitrant to conscious control. Obama's perception of conversational awkwardness automatically activated impulses to smile and reassure.[11] Except for the contrast in ethical significance, Obama's gesture is similar to the aversive racist's microbehaviors. While an aversive racist might feel a subtle impulse to lean back and look away from an interracial interlocutor, Obama might (in other conversations) feel a subtle impulse to lean forward, "open" his posture, and make eye contact. In the aversive racist's case, these are undesirable aliefs, and they can contribute to real harm.[12] But how should we understand the cases in which aliefs are expressive, not of prejudice, but of social skill and even virtuosity?

4. Expressing Normative Attitudes

Although Obama's reaction received little explicit acknowledgment in the news, such gestures commonly *do* receive a form of acknowledgment, albeit not overtly. Roberts's ability to move forward with the oath of office

[10] De Becker (1998) argues that one important thing that agents 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. He emphasizes the accuracy of intuitions and feelings of anxiety in ambiguous contexts.

[11] Perhaps the intentional content of the operative alief in this case was something like "Uncomfortable interlocutor! Tension mounting! Smile!" It is not our concern to argue that the extension of aliefs per se should be widened but that the ethical relevance of automatic behavior should be widened to include cases of praiseworthy automatic action.

[12] See Valian (1998, 2005) for a compelling account of how the "accumulation of disadvantage" operates to maintain asymmetric power relations, although we disagree with her claim that the psychological processes involved are "purely cognitive rather than emotional or motivational" (Valian 2005, 198); in our view, automatic-affective dispositions play a key explanatory role.

was a form of tacit acknowledgment of Obama's skilled reaction.[13] Automatic appropriate behavior in this sense often calls for, and is met with, automatic appropriate responses. Agents can tacitly and automatically express and detect subtle forms of approval and disapproval. This interlocking chain of behavioral reactions enables engaged actors to continue to converse successfully without having to reflect on the situation, articulate and sort out what they are up to, or engage in any potentially self-defeating exercises of self-control.[14]

One of our aims is to figure out what this tacit form of acknowledgment in practice amounts to in theory. We propose to conceive the normative relations implicit in these interactions via analogy with the normative role commonly attributed to reactive attitudes. Reactive attitudes, such as indignation, anger, and gratitude, are attitudes that express the fact that we hold others responsible for what they do (Strawson 1974; Eshleman [2001] 2009). We can see this expression at work in the difference between feeling annoyed at the family dog for stepping on your toe and feeling angry at your friend for stepping on your toe. Your anger in the latter case is expressive of disapprobation. It carries

[13] See McIntosh et al. (2006, 295) for a discussion of the role that automatic facial reactions play in sustaining rapport, as well as for studies on individuals with autistic spectrum disorders who exhibit "an impairment in this basic automatic social-emotion process." See Lakens and Stel (2011) on the relationship between movement synchrony and attributions of social rapport and cohesion.

[14] Sarkissian (2010b) makes a similar point in response to "situationists," who stress the influence of subtle situational cues on behavior. Sarkissian explains that *our* behavior can often be the cue that influences others. Our mere gestures and tones of voice "not only affect how others react to us, but also thereby affect the kinds of reactions we face in turn," making possible a kind of reciprocal ethical bootstrapping (12). Sarkissian (2010a, 2010b) and Slingerland (2011) draw insights from Confucian literature regarding how these subtly encouraging dispositions can be cultivated with practice. We might take issue with the Confucian conclusion that we ought to be especially "attentive" to these subtle gestures. Intentional efforts to control one's behavior in these ways often backfire (Follenfant and Ric 2010; Huebner 2009, 82-83). Sarkissian (personal communication) suggests that backfiring itself can be avoided with sufficient practice. In our view, the ethical importance of these affect-laden behaviors is often best respected by not "attending to" them but *just feeling them*. We have in mind a kind of middle ground between the "pure flow," in which agents are just grass passively bending in the wind, and the active attention, in which agents are overtly trying to resist the influence of salient contextual features and be the wind blowing all the other blades of grass (i.e., agents) around.

within it a sense of normative violation; your friend *should not* have stepped on your toe. In a similar vein, an agent's automatic expressions of encouragement or discouragement can be thought of as expressing *implicit* reactive attitudes. An agent's unintentional affective responses activate behaviors that express tacit approval or disapproval of how the interaction is going. Implicit reactive attitudes—like the feeling leading Obama to grin—express a felt sense of rightness or wrongness about the situation. Automatic affective responses reflect long processes of enculturation and experience through which tacit feelings of right and wrong move agents to action.[15]

Furthermore, agents are often warranted, pro tanto, in acting on the basis of their implicit reactive attitudes. Barring further considerations, you are warranted for expressing anger at your friend for stepping on your toe. Just so, implicit reactive attitudes give agents pro tanto warrant for responding to the situation with tacit approval or disapproval. These feelings and gestures are just the same *automatic*, and an agent will typically be unable to control or even notice them, let alone *report* in any accurate or informative way about the source of the feeling, the grounds for the action, or even the occurrence of the bodily movement. Furthermore, engaged agents do not typically judge that their feelings are "pro tanto warranted"; they respond immediately, regardless of whether they would judge that their responses were appropriate. However, *we* third-party observers can see that Obama should have grinned as he did. Of course, the pro tanto warrant for automatic reactions of this sort can be defeated, as when an agent's feel for the situation is guided by prejudice or phobia. But the very fact that they can fail is evidence that they have a typically valuable, if defeasible, role to play in guiding appropriate interactional behavior.[16]

[15] See Bourdieu (1977) on the processes of embodied enculturation.

[16] See Arpaly (2004) and Wallace (1994) on the important, if defeasible, role that reactive attitudes play in guiding moral reflection. The warrant of felt tensions also bears analogy with the warrant of perception. Despite the systematic susceptibility of perception to error in specific contexts, an agent is pro tanto warranted to believe the testimony of her senses. See Appiah (2008) and Sarkissian (2010a) for similar analogies between the warrant of automaticity and perception. In the next section, we explain how automatic dispositions can be norm insensitive, but we intend to say more about the theoretical basis for their ethical standing in further work. Our sense is that praiseworthy automatic actions express virtuous moral character, while clueless or blameworthy automatic actions express deficient moral character. But we also think that in a large number of cases, including those we discuss in this article, praiseworthy automatic actions bring about good consequences and treat others as ends rather than means.

5. Self-Modification

Obama's alief can be genuinely praiseworthy, rather than just accidentally appropriate, because of the ways in which aliefs unfold and self-modify over time. Aliefs harbor their own proprietary modes of norm sensitivity, and this makes them proper subjects for ethical reflection.[17] A capacity for self-modification is a hallmark of norm sensitivity. Upstanding beliefs self-modify by revising in response to incoming evidence. Aliefs paradigmatically self-modify in two ways: changing automatically in response to subtle variations in the immediate environment and improving gradually in response to repeated experience. In both cases, the key to understanding the normativity of alief is the feedback and interplay of affective and behavioral components through time.

When one's interlocutor reacts with subtle confusion or disapproval, one can "feel" that things have gone awry. The beginning of this process is much like Gendler describes. Perception of a salient environmental stimulus elicits a felt sense of "tension" and activates behavioral responses (an alief with *R-A-B* content). However, the affective and behavioral components of such habit-based aliefs are not arbitrarily associated. Rather, they are integrally related. In paradigmatic cases, the activated behaviors are directed toward *alleviating* the agent's sense of tension. After behavioral responses are begun, an agent's felt sense of tension will change in turn, decreasing or increasing as the unfolding behaviors establish better or worse ways of carrying on. In a conversation, an agent might move forward and back until he finds the right spot. Although the agent may have no more than a peripheral awareness of this process, the interplay between felt tensions and behavioral adjustments makes aliefs capable of a distinctive kind of self-modification, which we refer to as "self-alleviation." Aliefs self-modify by, in effect, eliminating themselves. Felt tensions elicit behaviors aimed at reducing that tension. How an agent feels in a particular context is not a "one and done" reaction to one salient stimulus but rather an ongoing readjustment to the complexities of the unfolding situation. If an agent reacts to a felt tension only to find that the tension is *not* alleviated, the lingering discomfort constitutes a pro tanto signal that the reaction was not

[17] This is, of course, not to claim that every tokened alief and automatic behavior is intrinsically ethically relevant but that, as a class, they are norm sensitive and hence appropriate subjects for ethical reflection.

appropriate. Ceteris paribus, the sense of having reacted inappropriately sets other automatic reactions in motion, which are directed toward better ways of responding to the situation.

We use the term "felt tensions" to signify a specific class of automatic affective responses that are in a deep sense "geared" toward immediate behavioral reactions.[18] Felt tensions are marked by either positive or negative *valence*, which acts like a physiological reinforcer of possible behaviors.[19] The agent literally feels a (positive) attraction or (negative) repulsion to available courses of action. Thus, felt tensions are rarely, if ever, altogether unconscious.[20] They are typically felt but not noticed; we suspect that they are most likely to occupy focal awareness in cases of jarring belief discordance. Even in the most subtle cases, valent tensions make 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.[21] These felt bodily responses and inclinations toward behavioral readjustments are part of a coordinated response pattern that is automatically set in motion just as the agent begins to feel that things have gone awry.

The feedback provided by felt tensions and behavioral readjustments also makes possible a slow "fine-tuning" of an agent's ability to respond to future senses of tension. For example, the failure of a particular behavioral reaction to reduce a felt tension makes an agent likely to respond differently to a similar tension in the future. Token experiences of (un)alleviation are part of a gradual evolution. This gradual evolution helps to make sense of why some agents

[18] We borrow the concept of "felt tensions" from Dreyfus and Kelly (2007).

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

[20] Whether aliefs are ever completely unconscious is an empirical question, but we think the available evidence suggests at best that they are often not consciously accessed, rather than being inaccessible. We predict that they are at least phenomenally conscious.

[21] We draw this list of autonomic-physiological changes from Klaassen, Rietveld, and Topal (2010, 65). See also Barrett (2006).

(in some contexts) will be more responsive and flexible than others. Here we have in mind an example like Obama's social virtuosity, which he cultivated over time while developing a political career.[22]

The process of self-alleviation is rife with potential for error, which is another hallmark of norm sensitivity. Aliefs can fail in their own right and not only when they come out of concord with an agent's beliefs. For example, Obama could have rightly perceived the awkwardness of the moment but reacted with too much or too little affect, perhaps by smiling coldly or by feeling giddy. Alternatively, Obama might have felt an appropriate degree of affective tension but under- or overcompensated behaviorally, by nodding too subtly for Roberts to see or nodding so dramatically that he appeared insincere. Yet another possibility for error arises even if Obama felt the right tension and responded with an appropriate behavioral adjustment but failed to feel the right alleviation. He might have continued to sense that things were awry and continued to grin inflexibly.

Our account shows how aliefs can make a self-standing ethical contribution in these cases, rather than by virtue of being belief concordant. In contrast to TDH, our account makes no reference to the standing of aliefs relative to an agent's considered beliefs or reflectively endorsed ends. The praise that a skilled distance stander (or the president) is due when she effortlessly and automatically makes one of the countless gestures integral to the flow of a conversation does not derive from an achieved harmony with any reflective states. It appears that there is simply no such reflective state (about whether to

[22] If differences among agents' "patterns of feel" reflect differences in moral character (broadly construed), are agents morally responsible for them? What is the connection between moral responsibility and automatic action more generally? One might think that we are not responsible for our alief-driven behaviors insofar as they are unconscious. However, they are *not* actually unconscious (nor, for that matter, ever truly incorrigible); they are driven by a *feel* of tension. If there is an intimate connection between awareness and responsibility, then we are not completely "off the hook" for our alief-driven reactions. Of course, this does not mean we should throw people in jail for close talking; the morally appropriate reactions to these subtle infractions will often be *other* implicit behavioral responses. Also, there is conceptual room to judge that social virtuosos are praiseworthy for navigating these contexts and that the rest of us are ethically deficient in some sense, without our being *blameworthy* per se. See Madva (2012) for further discussion.

smile and nod, about how far to lean back as the close talker leans forward) present in experience. There is nothing for the alief-driven reaction to be harmonious *with*. And even if a relevant reflective state *were* present, it would lack the causal power to guide behavior in the right way. The opportunity for appropriate response is too brief and the execution of right action is too quick to be successfully guided by an occurrent reflective state.[23]

Aliefs are part of a normatively structured automatic readjustment to the immediate environment. Such automatic dispositions can be adequately attuned to the demands of the situation, at the right time in the right way, even when an agent has no particular commitments about how to act. They do not merely *cause* behaviors like ballistic reflexes; they provide pro tanto warrant to act in particular ways. The further question for empirical research and conceptual reflection is when the pro tanto warrant succeeds and fails. We have not compiled an exhaustive list of necessary and sufficient conditions, but we will make some specific suggestions in the context of responding to challenges for the sketch we have just given.

6. Objection: Obama's Grin *Does* Satisfy TDH

A critic of our view might concede that Obama's grin was not harmonious with an online reflective state but insist that we are looking for harmony in the wrong place. This critic might suggest that what makes Obama's automatic dispositions to smile and nod praiseworthy is not harmony with an occurrent reflective state but, rather, harmony with long-term intentions, like to be a sociable person. But emphasizing long-term intentions rather than occurrent reflective states is, in this context, problematic for at least three reasons.

First, it is not obvious how to pick out the relevant intention with which one's gestures are to be consistent without making it either impracticably specific or vacuously general. With which reflectively endorsed intentions are Obama's reactions supposed to be harmonious? "To be sociable" is too general to tell an agent much about how to handle a flubbed oath of office,

[23] Nor should this be thought to be an unhappy consequence, because beliefs may also lack not just the right sort of causal power but also access to the relevant normative information. It is typical of our automatic affective dispositions to be well suited to perceiving and responding appropriately to nonverbal cues, as these dispositions gradually develop and continually modulate in response to particular sociocultural contexts. Our capacity for reflective judgment is not well suited to this task.

and "reassure chief justices with a smile when they flub oaths of office" is uselessly specific.

Second, the over- or underspecificity of such intentions suggests their causal inefficacy. For TDH to get a grip, some appropriate causal relationship must hold between the long-term intentions and automatic dispositions. One's reflective attitudes must be *in charge*. Perhaps a long-term intention could make itself felt "in the moment" if an agent had cultivated habits concordant with it in advance. With enough practice, a good Aristotelian can make it the case that belief-concordant aliefs kick in automatically in response to salient contextual cues.[24] But making it the case that the intention *to be sociable* guided one in the moment would radically underdetermine the range of possible ways an agent might respond during a flubbed oath of office; the intention is simply too general to help an agent respond appropriately to the contextual particulars. And an agent could not feasibly avoid this problem by identifying and practicing for every possible contingency. Of course, both Obama and Roberts *did* practice beforehand and presumably *did* consider ways things might go wrong during the inauguration. Indeed, to the extent that Obama holds a relevant long-term intention to be sociable, don't most who aspire to public office, including Roberts? (Doesn't everyone?) And yet many people, similarly situated, would have handled the situation much less gracefully. A similarly situated agent who did not respond as Obama did would not thereby have *failed* to accomplish the intention to be sociable. Nor should we attribute Obama's success to an ability to harmonize such an intention with his automatic reactions. What makes this implausible is the difficulty of imagining how such intentions could be discernibly operative *in the moment*.

Third, to the extent that you find something admirable in how Obama handled the situation, we invite you to reflect on what the actual source of that admiration is. Is Obama praiseworthy *because* he is really good at harmonizing his winks and nods with his long-term intentions? Is this the source of our admiration in such cases? We doubt it. Even if some relevant causal connection does obtain between Obama's long-term intentions and his automatic dispositions, it would be quite forced to think that, in this case, the fact of that connection is the reason we find the automatic action admirable. It is more likely that we notice Obama's social adeptness because we sense that, if placed in a similar situation, we would be far less skillful.

[24] See Snow (2006) for an account of the relationship between automatic goal activation and "habitual virtuous action."

7. Objection: Obama's Grin Is Not Genuinely Ethical

Perhaps these cases of automatic social behavior pose no problem for TDH because Obama's grin and other interactional microbehaviors are not genuinely ethical at all. One might argue that aliefs are not truly norm sensitive because they can only respond appropriately in, as Gendler says, stable, typical, and desirable contexts. Alternatively, one might argue that these behaviors are just "social graces" of merely prudential significance—capable of expressing genuine skill but not *ethical* in any legitimate sense. We address these concerns in turn.

7.1 Flexibility

Gendler might respond that Obama's aliefs are not praiseworthy in the way that we suggested because they do not exhibit genuine flexibility in the face of changing circumstances. Perhaps Obama "got it right" simply because he is used to being in similar conversations and it is typical of his aliefs to be norm concordant in familiar environments (Gendler 2008b, 554). It is just a mere habit, which would have kicked in automatically once the familiar conditions obtained, in utter independence of whether there was any warrant, pro tanto or otherwise, for so acting. There would thus be nothing *normatively* remarkable going on.

But an objection along these lines would tread on an ambiguous sense of what counts as familiar. Was Obama in the highly familiar position of needing to diffuse an awkward moment or the radically novel position of being inaugurated president? Of course, even people in the most familiar of settings still flub routine social gestures and jumble their words. Familiarity cannot be sufficient to elicit well-executed automatic action. If we grant that certain key features of the environment were familiar, it does not follow that Obama's norm-sensitive reactions are fully *explained* by that familiarity. To the contrary, the fact that people can still make mistakes even when the conditions for action are ideal and the context is maximally familiar explains why these behaviors are essentially normative. As we argued above, capacity for error is an essential aspect of norm sensitivity; mere reflexes do not make mistakes.

Nevertheless, there is an obvious sense in which the environment was radically novel for both Roberts and Obama in that it was the first presidential swearing in for both of them. The enormity of the moment visibly affected both of them. So in this sense, familiarity is not even necessary for praiseworthy automatic action, because there was no relevantly similar situation with which either of them could have been familiar. Ultimately, whatever the familiar and unfamiliar elements were, they faced Roberts and Obama alike. Familiarity simply cannot explain Obama's gesture, since Roberts was in the same environment and did not display the same social adeptness. Part of what makes Obama's reaction impressive is precisely that he remained sensitive to the *relevant* familiar features, perhaps by staying attuned to Roberts's gestures, rather than being overwhelmed by all the contextually salient, but normatively disruptive, novelties.

7.2 Ethical vs. Prudent

Another objection might begin by conceding that Obama's grin was praiseworthy in some sense but not genuinely ethical. Perhaps it was a merely prudential act, a manifestation of "social graces," that reveals little about the *ethics* of automaticity. People do all kinds of useful things automatically, like smiling and nodding, but what are these behaviors to ethics?

Consider then the case of automatic *heroic* action. As Charles Goodstein (2007) put it,

> if you look at the history of most people who are designated heroes in the military and in other places, most of the time they say the reaction they had was without any mental preparation. It was spontaneous, it was without much consideration for the practicalities, the realities of the moment. I think they're honest when they say they don't think of themselves as heroes, they just reacted to something they saw as an emergency.[25]

One among many examples of automatic heroic actors is Wesley Autrey, the "subway hero," who saved the life of an epileptic man who, in the midst of a seizure, had fallen onto the tracks of the New York City subway. Autrey jumped onto the tracks himself and held the man's body down while the train passed just inches overhead. Like many other heroes, Autrey reported afterward that he "just reacted" immediately to the situation. Autrey's behavior was typical of courageous actors, who often "just react" without forethought,

[25] This view of heroes is common enough that it forms the background of Andrew Carnegie's stated mission for his Carnegie Hero Fund: "I do not expect to stimulate or create heroism by this fund, knowing well that heroic action is impulsive; but I do believe that, if the hero is injured in his bold attempt to serve or save his fellows, he and those dependent upon him should not suffer pecuniarily" (http://www.carnegiehero.org/fund_history.php).

without willing themselves to do something difficult, and without awareness of their decision-making processes. But surely heroes are paragons of ethical action. So if one were to grant that Obama's grin is praiseworthy but dismiss it as a mere expression of social skill in a ceremonial event, heroes present a relevantly similar case of automatic action with ethical gravity that is difficult to question.

Of course, Autrey's heroic feat involved many full-blown intentional actions, beyond the sorts of microbehaviors of distance-standing and aversive racists. But the *initiation* of Autrey's feat was fundamentally automatic. And like in the case of Obama's gesture, Autrey's alief could have failed in its own right. Autrey's alief itself could have failed affectively (by leading him to feel coldly indifferent or overly solicitous) or behaviorally (he might have tried to stop the train with his body like a superhero or by meekly yelling at it to stop) or by failing to be alleviated (he might have continued to hold the epileptic man to the ground after the train stopped).

We have claimed that cultivating a felt sense of what to do through repeated experience is vital to praiseworthy automatic action, but one might reasonably object that heroic feats like Autrey's are precisely the sorts of actions for which an agent cannot cultivate a feel. We agree, of course, that an aspiring hero can intentionally prepare for only a small subset of the possible scenarios calling for heroic action. It is an empirical question, albeit an inherently elusive one, what drives an agent to selfless heroic action, but we doubt, in many cases at least, that the prime mover is a standing intention to be heroic. Is the only alternative to concede that Autrey is just the kind of person who does that kind of thing? In fact, the capacity to perform these one-and-done feats may often be grounded in habitual felt tensions. Autrey, for example, suggests that his background in construction and extensive "work in confined spaces" may have enabled him to snap-judge accurately that he could fit under the train.[26] This is clearly speculative—Autrey's conjecture is made from a third-person perspective on his own actions (for more on this, see section 9)—but one would be hard pressed to come up with an occupation *better suited* to preparing an agent to respond accurately in that situation.

[26] See http://en.wikipedia.org/wiki/Wesley_Autrey. Of course, the applicability of his training in this way is not an ex post facto vindication of the ethical value of working in confined spaces. But in the absence of any alternatives, it provides a plausible explanation of how Autrey may have perceived the situation in a different light from onlookers.

Our admiration of Autrey's feat, like Obama's, is not grounded on a sense that Autrey harmonized his automatic dispositions with his considered beliefs. Presumably neither Obama nor Autrey *had* occurrent action-guiding beliefs about what they ought to do. Alternatively, one might reasonably claim that Autrey's aliefs were *discordant* with his considered beliefs. Consider that Autrey was with his two young children at the time. Suppose, prior to the event, he had been asked whether he would put his own life at enormous risk to save a stranger's, even if it meant leaving his daughters in the hands of anonymous subway riders. Might he have hesitated or hedged at least a little? He would have probably had to at least think it over, but there was no such hesitation when the time for action came. Would such belief discordance undermine the heroism of his action? No. Imagine a soldier who violates her battlefield code of conduct by helping an injured enemy.[27] Her putative belief discordance does not undermine the ethical standing of her action. Whether a given automatic behavior is praiseworthy is separate from whether it is the action an agent would reflectively judge worth doing.

8. Belief-Behavior Discordance

Do aliefs continue to be pro tanto warranted even in cases of full-blown conflict between alief and belief, as in aversive racism? Acting on such discordant aliefs is at best unintentionally harmful and at worst culpably immoral. One might think that the inevitable superiority of belief in cases of discordance shows that beliefs at least *ought* to be in charge, even if they often are not. But cases of discordance may be far less univocal than what the preponderance of the philosophical and psychological research suggests.

Social psychologists have much to tell us about well-meaning, clear-headed agents who bear regrettably biased dispositions but hardly anything to say about intellectually muddled agents who harbor morally upright dispositions. Conspicuously absent from the voluminous literature on belief-behavior discordance is research on "aversive egalitarians": agents who self-ascribe fully fledged prejudiced beliefs but unwittingly demonstrate automatic *egalitarian* dispositions. Nomy Arpaly's (2004) interpretation of the literary character Huckleberry Finn illustrates the kind of person we have in mind.

On Arpaly's reading (2004, 75), Huck's is a case of "inverse *akrasia*," in which an agent does the right thing in spite of his all-things-considered best

[27] Thanks to Katie Gasdaglis for this example.

judgment. Huck's dilemma is whether to turn in his friend Jim, an escaped slave. On the one hand, Huck believes that an escaped slave amounts to a stolen piece of property and that stealing is wrong. On the other, Huck is loyal to his friend. The result of his (perhaps less than ideal) deliberation is that he ought to turn Jim in, but Huck finds himself unable to do it. Against Bennett (1974), Arpaly argues persuasively that Huck's action is admirable rather than accidental. Could it be that Huck's behavior is guided by an ethical alief—for example, an affect-laden, automatic impulse to protect his friend, which (happily) trumped his racist beliefs? The roles that Huck's pro-Jim attitudes play in his deliberation, as unwelcome "gut reactions" that cannot be "reasoned away," have all the paradigmatic trappings of the automatic-affective processes that Gendler means to capture.[28] In this case, the ethical desirability of Huck's alief-like impulses *could not* derive from their concordance with his reflective judgments, because his judgments were wrong. Huck would have done better if he could have achieved something like bottom-up harmony, adjusting his reflective judgments in light of his automatic dispositions.

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 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.[29] To our knowledge,

[28] But perhaps the fact that Huck deliberates about what to do makes his example less than ideal for the purposes of illustrating praiseworthy automatic behavior. 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, Huck feels a lump in his throat, tension in his body, and sweat on his palms. Merely by looking at Jim, Huck feels repelled by what he believes he ought to do.

[29] Even such putatively objectionable automatic dispositions as disgust may be ethically desirable in certain contexts. The undesirable effects of disgust on social perception are well documented in, for example, Schnall et al. (2008) and Rozin, Haidt, and McCauley (2008). Sullivan (2006) argues that there is a deep link between experiences of "disgust" and what she calls the "unconscious habits of racism." She offers

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 underinvestigation of such empirical possibilities furnishes no evidence that they do not regularly occur, however. Such phenomena *should* be studied. There have, for example, been a handful of related studies on biased social attitudes that 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? Further research on cases like this will help us to better understand when aliefs can be trusted and when beliefs cannot.

9. Are Aliefs Reason Responsive?

If we take seriously the possibility that aliefs can get it right when beliefs get it wrong, one might be tempted to think this would show that aliefs are "reason responsive" after all. Arpaly, for one, argues that Huck's decision reflects responsiveness to reasons. Railton (2009) makes similar claims with respect to what he calls "practical competence" and "fluent agency." These philosophers might agree that TDH is false, on the grounds that putative aliefs can be reason responsive, but argue on precisely those grounds that alief is a bogus concept that cannot be coherently distinguished from belief. Schwitzgebel (2010a), for example, cites the intelligence of auto-

an intricate and convincing account of the phenomenology of perceiving the bodies of those who are unlike you as disgusting. What she does not discuss, however, is that a similar disgust mechanism might operate in the experience of the antiracist, the person who does not find the bodies of those unlike herself to be disgusting but instead finds *racism* disgusting. It is not hard to imagine such a person (or that we ourselves harbor many such affective associations). For her, the image of Alabama governor George Wallace barring black children from entering newly desegregated schools is literally *disgusting*. This experience of disgust might be just the thing that springs our imagined antiracist into automatic action.

matic responses in an argument against alief.[30] Construing automatic dispositions in this way, however, is misleading. In our view, aliefs are sensitive to ethical features of the world but not to those features qua reasons, because aliefs predictably fail to play the right sorts of cognitive roles in practical reflection.

There is a sense in which all of the agents we have discussed undeniably act "for reasons." Obama et al. surely act for reasons in the very general sense that their behaviors serve functions and ends that we (third-party observers) recognize to be valuable. For example, Obama's wink served the end of maintaining the momentum of his inauguration. This is an "external" or "objective" conception of reasons. It does not require that Obama be *guided* by the relevant reason, either consciously or unconsciously. There are likewise reasons for newborn babies to mimic their parents' gestures, even if newborns *have* no reasons whatsoever. Even the most canonically automatic behaviors may serve valuable teleofunctional roles. The question is not whether reasons can be found to justify those behaviors but whether those reasons *guide* them. To show that the operative automatic dispositions are reason responsive in a more substantive sense, one needs to make the case that Obama et al. acted *because* of some occurrent reason-responsive state.

Of course there is nothing approaching consensus in action theory or moral psychology on how to understand "subjective" or "motivating" reasons. On a view that we find intuitive, practical reasoning is the capacity for resolving, through reflections, questions about what to do (see, e.g., Wallace [2003] 2008). A state is reason responsive just insofar as it is capable, ceteris paribus, of figuring in practical reasoning in the right ways, perhaps by revising in proportion to the evidence, 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 the automatic dispositions we discuss here meet them. The states that drive automatic action seem, ceteris paribus, to be unavailable for report, unresponsive to undercutting evidence, and incapable of integrating inferentially with other states.

As we have argued, the affective states that drive ethical automatic action are not typically joined with the ability to *report* in any accurate or informative way about the source of one's feelings, the grounds an action, or even the occurrence of bodily movement. Although Autrey could speculate about his reasons for

[30] Schwitzgebel (2010b) has also argued that we—Brownstein and Madva—should "reframe [our] view as a *criticism* of the concept of alief, rather than an adaptation of it."

jumping onto the tracks, he offered them from a detached third-person perspective, reflecting on his own behavior and drawing inferences about it which cast him in a favorable (rather than reckless) light. The literature demonstrating the scope of confabulation in everyday life suggests that an agent is paradigmatically not a reliable reporter on the effects of automatic cognition on his choice and behavior.[31]

But perhaps a critic might suggest that praiseworthy automatic actions are sensitive to implicit or unacknowledged reasons. Arpaly argues that although Huck did not act on the reasons that he consciously considered, he unknowingly acted on the sum of the total reasons he holds. Let us take for granted that it is possible to act on unacknowledged reasons. Doing so would presumably involve acting on the basis of unconscious reason*ing*. Is it possible that Huck, Obama, and Autrey did so? Of course. But we think this notional possibility is not plausibly realized in cases of praiseworthy automatic action, because the motivating states in question are impressively *insensitive* to the full array of evidence, properly so called, and impressively *incapable* of integrating with other psychological states.[32]

In Huck's case, he took himself to have good reason to turn Jim in, but his attachment to Jim was precisely incapable of being moved by these considerations. His attachment could not be reversed; it was insensitive to what Huck took to be decisive evidence undercutting it. Nor could the state integrate properly with his other mental states. It was not that he *did* think of Jim's personhood or friendship in the context of deliberation but decide, all things considered, that he ought to turn Jim in just the same. Rather, Huck concluded that since he could not bring himself to do what he judged he ought, he was going to hell. This is again taking a third-person perspective on his actions rather than, from within practical reasoning, coming to a deliberative conclusion. The irreversibility and encapsulation of the state make it misleading to construe as reason responsive.

Obama and Autrey's actions also seemed evidence insensitive but in a somewhat different sense. It would seem otiose to describe Obama's reaction

[31] The locus classicus is Nisbett and Wilson (1977).

[32] While Gendler's examples fail to show that alief is insensitive to ethically relevant features in an agent's environment, they show persuasively that aliefs are not sensitive to those features qua *evidence*. In Gendler's cases, aliefs persist in the face of overwhelming evidence that they are not reflective of reality. This shows either that aliefs themselves are not sensitive to the evidence as such or that they are cognitively encapsulated from the agent's other psychological states (or both). Either way, aliefs would be largely incapable of responding to reasons.

as tracking the evidence. Obama did not grin, once he felt that he had accumulated sufficient considerations in favor of grinning (nor was his grin the result of a *failure* to track the evidence). In a context demanding immediate response, an agent typically only turns to considerations like these when the ordinary flow of events come undone. Had Obama grinned and Roberts subsequently recoiled in disgust, then we can imagine Obama considering the evidence before deciding what to do next. Construing automatic dispositions as reason responsive is misleading because it suggests that they are available for reflection and play certain cognitive roles that they seem not to play.[33] It overstates the similarities and conceals the important differences between the causal and normative contributions made, respectively, by an agent's automatic and reflective dispositions.

10. Conclusion

Further reflection and research should consider the specific types of cases in which aliefs are more likely to err or excel and in what senses they can or should be integrated with our standing reflective attitudes.

We have argued that automatic dispositions are ethically self-standing in that their ethical desirability does not depend on belief concordance. But ethical aliefs do not develop in *complete independence* of beliefs. Although aliefs are not straightforwardly accessible or responsive to beliefs, aliefs and beliefs indirectly influence each other in a variety of ways, and, often, they should. Gendler envisions the influence of alief on belief as a kind of

[33] We say more about the encapsulation and evidence insensitivity of alief in our companion piece. We can further see what is problematic with the general enthusiasm for attributing reason responsiveness to automatic acts by considering how Arpaly applies it even to cases of praiseworthy *athletic* action. On her view (2004, 53), for example, "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." This may be true of some "competent" players, but *experts*, by contrast, hit many shots in the absence of reasons for them. Many points in tennis matches are unique. For every situation, there is no one shot that is required. Better and worse shots are judged by their creativity, effectiveness, degree of difficulty, and even beauty. It would be bizarre to praise a shot as responsive to a reason. What would the reason be? A player does not receive praise because she has hit a stunningly rational shot. What would make a stunning shot *stunningly* rational? Such a perspective is at odds with what many players aim to do and what most spectators hope to see.

motivated "rationalization" on a par with cognitive dissonance (2008b, 578). But there is a sense in which Huck would have done better if he were somehow able to take his automatic-affective reactions into consideration. He would have been better served if he had been able somehow to revise his beliefs in light of his aliefs, toward a kind of bottom-up harmony. Perhaps Huck will simply go to his grave believing himself wicked for turning in an escaped slave. But perhaps, over time, Huck will begin to consider the "wisdom" of those affect-laden reactions that he could not reason away. By taking a quasi-theoretical stance toward himself, Huck might notice, just as we are trying to point out, the sensitivity, flexibility, and intelligence of his automatic dispositions. *How* to accomplish such an ethically appropriate integration of alief with belief, given their relative insularity, is an important question for future thought and research.

Although Huck's action was not in harmony with his beliefs, it was, in a different sense, "in harmony" with the demands of the situation. His automatic-affective responses, like Obama's and Autrey's, were "attuned" to their immediate environments and to the states of other agents. Musical analogies may be especially apt in such cases. Just as an expert musician can sense when a band member is losing the tempo and adjust appropriately, perhaps Obama sensed his interlocutor's departure from the tempo of the interaction and adjusted to bring him back into rhythm. Such sensitivity to the states of others may consist in as little as attunement to conversational discomfort or as much as attunement to profound suffering. In these cases, praiseworthy automatic actors establish an *ambient* harmony with the other agents in a shared situation.

Attaining ambient harmony between agents may often be at odds with aiming for harmony within. A deliberating agent like Huck might have been well served trying to harmonize his aliefs and beliefs, but an engaged agent like Obama would have been ill served trying to harmonize his impulses to nod and grin with his considered beliefs about the chief justice. Doing so would not help him become *more* of a social virtuoso, and it might make him less so, in much the same way that expert performance is typically degraded by overthinking. Aiming for internal harmony would inhibit his ability to respond—automatically, reliably, and ethically—to the callings of the context.

Acknowledgment

For all their invaluable feedback, we wish to thank James Bohman, John Christman, Katie Gasdaglis, Tamar Gendler, Christia Mercer, Mark Risjord, Raffaella de Rosa, Paul Roth, Eric Schwitzgebel, Shannon Sullivan, Stephen Turner, and Alison Wylie.

Versions of this article were presented at the 2011 Philosophy of Social Science Roundtable, the 2011 meeting of the Society for Philosophy and Psychology, the Rutgers University–Newark Philosophy Department Colloquium, and the New Jersey Institute of Technology Albert Dorman Colloquium Series. We are deeply grateful for the helpful comments and questions we received during these presentations.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Appiah, K. A. 2008. *Experiments in ethics*. Cambridge, MA: Harvard University Press.

Arpaly, N. 2004. *Unprincipled virtue: An inquiry into moral agency*. Oxford, UK: Oxford University Press.

Aviezer, H., S. Bentin, V. Dudareva, and R. R. Hassin. In press. The automaticity of emotional face-context integration. *Emotion*.

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

Barrett, L. 2006. Are emotions natural kinds? *Perspectives on Psychological Science* 1:28-58.

Bennett, J. 1974. The conscience of Huckleberry Finn. *Philosophy* 49:123-34.

Bertrand, M., and S. Mullainathan. 2003. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market. Working Paper 9873, National Bureau of Economic Research, New York.

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

Bourdieu, P. 1977. *Outline of a theory of practice*. Cambridge, UK: Cambridge University Press.

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

Dasgupta, N., and S. Asgari. 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.

Deutsch, R., and F. Strack. 2010. Building blocks of social behavior: Reflective and impulsive processes. In *Handbook of implicit social cognition: Measurement, theory, and applications*, edited by B. Gawronski and B. K. Payne, 62-79. New York: Guilford Press.

Dijksterhuis, A., T. L. Chartrand, and H. Aarts. 2007. Effects of priming and perception on social behavior and goal pursuit. In *Social psychology and the unconscious: The automaticity of higher mental processes,* edited by J. A. Bargh, 51-132. Philadelphia: Psychology Press.

Dovidio, J. F., K. Kawakami, and S. L. Gaertner. 2002. Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology* 82:62-68.

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

Egan, A. Forthcoming. Comments on Gendler's "The epistemic costs of implicit bias." *Philosophical Studies*.

Eshleman, A. (2001) 2009. Moral responsibility. In *Stanford encyclopedia of philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/entries/moral-responsibility.

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

Gendler, T. S. 2008a. Alief and belief. *Journal of Philosophy* 105:634-63.

———. 2008b. Alief in action (and reaction). *Mind and Language* 23:552-85.

Gilbert, D. T. 1991. How mental systems believe. *American Psychologist* 46:107-19.

Goldin-Meadow, S., and S. Beilock. 2010. Action's influence on thought: The case of gesture. *Perspectives on Psychological Science* 5:664-74.

Goodstein, C. 2007. Heroes rush in, but what would average Joe do? *New York Times*. January 7.

Greene, J., and J. Haidt. 2002. How (and where) does moral judgment work? *TRENDS in Cognitive Sciences* 6:517-23.

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

Jolls, C., and C. R. Sunstein. 2006. The law of implicit bias. Paper 1824, Faculty Scholarship Series. http://digitalcommons.law.yale.edu/fss_papers/1824.

Kawakami, K., J.F. Dovidio, J. Moll, S. Hermsen, A. Russin. 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.

Klaassen, P., E. Rietveld, and J. Topal. 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

Kurzban, R., J. Tooby, and L. Cosmides. 2001. Can race be erased? *Coalitional Computation and Social Categorization PNAS* 98:15387-92.

Lakens, D., and M. Stel. 2011. If they move in sync, they must feel in sync: Movement synchrony leads to attributed feelings of rapport. *Social Cognition* 29:1-14.

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

Levinson, J. D., and D. Young. 2010. Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence. *West Virginia Law Review* 112:307-50.

Levy, N., and T. Bayne. 2004. A will of one's own: Consciousness, control and character. *International Journal of Law and Psychiatry* 27:459-70.

Libet, B., C. A. Gleason, E. W. Wright, and D. K. Pearl. 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. Interpersonal fluency and implicit bias. Phd diss., Columbia University, New York.

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

McIntosh, D. N., A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger. 2006. When the social mirror breaks: Deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. *Developmental Science* 9:295-302.

Milner, D., and M. Goodale. 1995. *The visual brain in action*. Oxford, UK: Oxford University Press.

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

Nisbett, R. E., and T. D. Wilson. 1977. Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84:231-59.

Payne, B. K., and C. D. Cameron. 2010. Divided minds, divided morals: How implicit social cognition underpins and undermines our sense of social justice. In *Handbook of implicit social cognition: Measurement, theory, and applications*, edited by B. Gawronski and B. K. Payne, 445-60. New York: Guilford Press.

Payne, B. K., and B. Gawronski. 2010. A history of implicit social cognition: Where is it coming from? Where is it now? Where is it going? In *Handbook of implicit social cognition: Measurement, theory, and applications*, edited by B. Gawronski and B. K. Payne, 1-15. New York: Guilford Press.

Pearson, A. R., J. F. Dovidio, and S. L. Gaertner. 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 *Reasons for action*, edited by D. Sobel and S. Wall, 81-115. Cambridge, UK: Cambridge University Press.

Rozin, P., J. Haidt, and C. R. McCauley. 2008. Disgust. In *Handbook of emotions*, 3rd ed., edited by M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, 757-76. New York: Guilford Press.

Sarkissian, H. 2010a. Confucius and the effortless life of virtue. *History of Philosophy Quarterly* 27:1-16.

———. 2010b. Minor tweaks, major payoffs: The problems and promise of situationalism in moral philosophy. *Philosopher's Imprint* 10 (9): 1-15.

Schnall, S., J. Haidt, G. L. Clore, and A. H. Jordan. 2008. Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin* 34:1096-109.

Schwitzgebel, E. 2010a. Acting contrary to our professed beliefs, or the gulf between occurrent judgment and dispositional belief. *Pacific Philosophical Quarterly* 91:531-53.

———. 2010b. Response to Brownstein and Madva, "Alief and affordance: The normativity of automaticity." Paper presented at the Pacific Division Meeting of the American Philosophical Association, San Diego, California.

Slingerland, E. 2011. The situationist critique and early Confucian virtue ethics. *Ethics* 121:390-419.

Snow, N. 2006. Habitual virtuous actions and automaticity. *Ethical Theory and Moral Practice* 9:545-61.

Strawson, P. F. 1974. *Freedom and resentment and other essays*. London: Methuen.

Sullivan, S. 2006. *Revealing whiteness*. Bloomington: Indiana University Press.

Valian, V. 1998. *Why so slow? The advancement of women.* Cambridge, MA: MIT Press.

———. 2005. Beyond gender schemas: Improving the advancement of women in academia. *Hypatia* 20:198-213.

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

Wallace, J. 1994. *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

———. (2003) 2008. Practical reason. In *Stanford encyclopedia of philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/entries/practical-reason/.

Wiers, R. W., Houben, K., Roefs, A., de Jong, P., Hofman, W., and Stacy, A. W. 2010. Implicit cognition in health psychology: Why common sense goes out the window. In *Handbook of implicit social cognition: Measurement, theory, and applications*, edited by B. Gawronski and B. K. Payne, 463-88. New York: Guilford Press.

Bios

Michael Brownstein is assistant professor of philosophy at the New Jersey Institute of Technology. He is most interested in understanding how and when unconscious and unintended action can be skillful and even praiseworthy. His approach to this question is methodologically pluralist, drawing on the philosophy of action, mind and social science, empirical psychology, ethics, and phenomenology.

Alex Madva is a doctoral candidate in the Department of Philosophy at Columbia University. His dissertation examines the problems that automatic social biases raise for philosophy of mind, action, and ethics.