Jennifer Nagel
December 28, 2011

Gendler on alief

Forthcoming in: *Analysis Reviews*

It is not hard to explain what happens when the rubber hammer strikes a person's leg just below the kneecap. Carefully calculated action is quite different in character, but it is also naturally intelligible to us in its way. Somewhere between the clear extremes of mindless reflex and deliberate rational action, however, there lies an expanse which is harder to navigate, a murky zone where the mind lumbers forward on autopilot, and the line between genuine action and mere bodily motion seems strangely blurred. Tamar Gendler's *Imagination, Intuition and Philosophical Methodology* (2011) bravely explores this difficult territory, and does an excellent job of bringing some of its most peculiar features to light. The book as a whole is thickly studded with insights on a great variety of topics. In what follows, I will focus on the main conceptual innovation of its final chapters, the introduction of 'alief' as a distinctive mental state. According to Gendler, alief is the state that governs the mysterious penumbral range of human activity, while perhaps also guiding what we do in our 'brighter' moments to a greater extent than we might have supposed.

For the reader accustomed to the traditional separation between epistemic and motivational states, alief looks odd at first. Traditionally, beliefs are supposed to represent how things are, and to guide action in accordance with desire, where belief and desire can vary independently of one another. The concept of alief incorporates a mixture of epistemic and pragmatic elements: aliefs have some concatenation of what Gendler labels

as 'representational-affective-behavioral' (R-A-B) aspects composing their content. The mixture of these elements is illustrated by a series of vivid examples. Gendler describes the effect produced by venturing onto the glass-floored Grand Canyon Skywalk as an alief incorporating 'the visual appearance as of a cliff, the feeling of fear and the motor routine of retreat' (261); meanwhile, the task of throwing a dart at a picture of a loved one activates an alief with the content 'harmful action directed at beloved, dangerous and ill-advised, don't throw' (262). Given that the state of alief in each case incorporates three separately identifiable (R-A-B) aspects, one might wonder what is gained by lumping them together under a neologism. Here Gendler argues forcefully for the idea that human behavior is more intelligible if we see the theoretically separable visual appearances, emotional responses and motor impulses produced in these situations as coming to us in naturally pre-packaged bundles. It is a general feature of pre-packaged bundles that they may contain some elements we would not have chosen to purchase individually, but nature may well have its own reasons for installing clumps of responses in us as it does. If the substratum of human behavior is indeed naturally lumpy, the notion of alief may help us make better sense of ourselves.

One might wonder about the specific value of alief-shaped lumps in our explanations of action on seeing Gendler's most abstract description of their composition: 'A subject has an *occurrent alief* with representational-affective-behavioral content R-A-B when a cluster of dispositions to simultaneously entertain R-ish thoughts, experience A, and engage in B are activated—consciously or unconsciously—by some feature of the subject's internal or ambient environment.' (265) It is soon evident that this is just meant as the start of a characterization of alief. Thinking about Gendler's examples, it is clear that

there is a stronger relationship than mere simultaneity among the relevant representative, affective and motor responses: rather than just consisting in the co-occurrence of these activations, alief seems to have a sequential and causal internal structure. So it is not that the Skywalk situation just sparks a visual representation and at the same time an emotional response and a motor impulse, each of which are mutually independent. Rather, the sight of the vertical drop through the glass floor produces the anxiety which fuels a hesitation to step forward. This internal structure makes sense; indeed, it makes sense in a way which might seem to suggest that the best explanation of behavior in these circumstances will have to work by describing the interaction of these distinct factors, rather than simply appealing to the whole cluster in which they figure.

However, we are not necessarily forced to choose between a sheer cluster-based and a sequential component-based approach; indeed, Gendler in various places (for example at 291) explicitly endorses the legitimacy of pursuing both approaches in parallel. Even if the success of alief-based explanations ultimately rests on facts about the lower-level representational, affective and motor components that figure in them, appeal to alief is not necessarily otiose. The concept of alief can retain special explanatory value if there is something distinctive about the limits placed on these components within alief, or about the manner in which alief unites them, something which makes their interaction interestingly different from an outwardly similar action involving superficially similar components under deliberate rational control. In fact, Gendler's fuller characterization of alief isolates two central distinctive features of the state, one of which concerns a limitation on a component, and the other a way in which the components are related. The first defining feature of alief is a limit on its representational component: it is not subject to a

norm of accuracy in the way that belief is. The second defining feature of alief is the special manner in which it can link representation to behavior: without the mediation of desire.

Gendler in fact provides a longer list of further common characteristics—for example, that 'aliefs typically include an affective component' (288)—but these characteristics have a secondary status. They flesh out the picture of what alief is like, but they are not alief's essential and distinctive features, either because they are not necessary conditions on alief (so the 'typically' suggests that alief may fail to be affective) or because they do not set aliefs apart from other states (so presumably some emotional responses, such as moods of sadness or elation, include the affective without thereby counting as aliefs).

The two distinctive features of alief merit special attention, not least because they have some very interesting consequences. It is because the representational aspect of alief is not norm-governed that alief stands out as 'neither rational nor irrational'; it is because alief is directly action-generating that it cannot just be eliminated by factorization into simpler separate representational and conative elements which combine to produce action (288). These two aspects of alief will be examined in turn.

Alief differs from belief in its representational dimension by not involving acceptance, according to Gendler (268), and this characteristic is closely linked with failures of accuracy in alief. Asking the question 'does being in an alief state with the content R-A-B involve regarding it as true in some way that R is part of one's real or imagined environment?' Gendler concludes that 'Interestingly, the answer to this question turns out to be *no.'* (268) According to Gendler, 'alief just isn't reality-sensitive in the way

belief is. Its content doesn't track (one's considered impression of) the world.' (271) There is perhaps something a bit awkward about this formulation: saying that belief is supposed to *track* one's considered impression of the world makes it sound as though one's considered impression of the world has to be settled first in order to determine what to believe. Elsewhere a similar point is articulated in terms of evidence:

> All that is needed is to note that – whatever belief is—it is normatively governed by the following constraint: belief aims to 'track truth' in the sense that belief is subject to immediate revision in the face of changes in our all-things-considered evidence. When we gain new all-things-considered evidence—either as the result of a change in our evidential relation to the world, or as a result of a change in the (wider) world itself – the norms of belief require that our beliefs change accordingly. (296)

Here again, the notion of 'all-things-considered evidence' is something that we might not want to take for granted in the course of crafting a story about the norm of belief: it is not clear that we will be in a position to sort out genuine evidence from its misleading competitors without the guidance of a prior sense of what sorts of things it is right to believe. If I am settling what my all-things-considered evidence is, surely I am already doing a great deal of figuring out what to believe, as opposed to just building the basis for subsequent decisions about what to believe. However, it may be most charitable to read these passages as intended to characterize just one particular aspect of the norm of belief: whatever else should guide belief, new belief should be consistent with existing belief. We should aim to eliminate conflict between belief and our considered impressions or all-things-considered evidence (leaving open the question of how those considered impressions and all-things-considered bodies of evidence will themselves be governed, but presumably we will at least need to strive for internal consistency there already). It is no

small job to deliver a full account of the norms structuring belief, and Gendler cannot be faulted for having executed only part of this job in her work on alief.

What really matters for present purposes is the contrast between belief and alief, and here the message is easier to follow: 'Aliefs by their nature are insensitive to the possibility that appearances may misrepresent reality, and are unable to keep pace with variation in the world or with norm-world discrepancies. By contrast, beliefs, are, (modulo error) responsive to the way things are: not merely to the way things tend to be or the way things seem to be.' (302) This more objective characterization of the norm of belief is helpfully clear, as is the suggestion that aliefs are strictly driven by appearances. Indeed that latter suggestion helps to explain why the norm of consistency has no grip on them: appearances can be frankly inconsistent, for example in cases in which various senses present conflicting reports on an object. The glass skywalk presents a visual appearance of a vertical drop which is neither felt by the foot nor accepted on reflection, but it nonetheless maintains representational force within the sheltered and slavishly appearance-driven realm of alief.

In saying that belief demands acceptance regulated by a norm of consistency, Gendler is restricting the province of belief quite sharply. This restriction now helps to make sense of various aspects of her broader characterization of alief. For example, when she claims that alief is 'shared by human and non-human animals' (288), one initially wonders whether this is a merely secondary characteristic of alief, like its affective dimension, or something that is supposed to be distinctive of alief as opposed to belief. But if belief is possible only for creatures who are sensitive to the distinction between

appearance and reality, or capable of acceptance regulated by consistency with a considered impression of how things are, then it does not seem to be a state that nonhuman animals can attain.

It is somewhat radical to suggest that nonhuman animals lack beliefs. Intuitively, it might have seemed that a nonhuman animal could believe that, say, *there is food hidden in the basket* in just the same way as a human could, but presumably Gendler could hold that in applying our folk psychological classifications to nonhuman animals we naturally anthropomorphize them. What might be somewhat more difficult here is to explain the similarities in the apparent interaction between representational and conative states in humans and animals: if one characteristic of alief is that it can motivate independently of desire, then we may have some difficulties in explaining why animals with similar informational access to some hard-to-reach food cache will exhibit systematically different behavior depending on whether they are hungry or full. Gendler's theory could be elaborated in one of several different directions here: she could maintain that different aliefs are produced in the hungry and sated animal (after all, aliefs can be produced by some combination of internal and external stimuli); alternatively, she could note that alief *can* produce action independently of desire, but need not always do so, and may have effects that are heightened or dampened in the presence of various relevant conative states. There are potential costs associated with each of these options: because the first demands the ascription of a multiplicity of aliefs to explain different actions, it may end up making the alief-based theory less economical than its belief/desire-based rival. Because the second reintroduces an interaction between alief and desire, it may diminish some of the motivation for introducing a notion of alief as a special insulated state. If Gendler's

theory does have the consequence that nonhuman animals must get by on alief alone, it would be interesting to see which one of these possible avenues she would pursue, or whether she would resolve the difficulties here in some other way.

Meanwhile, the idea that belief requires acceptance does much to clarify a further element in the characterization of alief. The description of alief as associative and automatic in character raises an immediate question: on mainstream treatments of associative and systematic thinking, beliefs can also be formed associatively and automatically. To insist that belief is only ever produced by purely controlled and systematic reasoning would be to shrink the province of belief rather drastically: on such an account I could not, for example, believe that someone I see is a certain friend of mine on the basis of ordinary facial recognition. However, it is possible to defend a view in which personal-level acceptance is invariably controlled (cf. Frankish, 2009), so that more than merely associative and automatic activation would be required for me to reach a conscious judgment about the identity of my friend, although automatic processing would supply some of the input to such a judgment. It would be more unusual to maintain that belief always requires a moment of explicit or conscious judgment, but at least such a view would not insist that beliefs be entirely controlled in every aspect of their formation. A somewhat softer line here would not even insist not that states actually have passed through the bottleneck of conscious, explicit judgment or controlled acceptance in order to count as beliefs: perhaps beliefs are simply those states that are subject to the norms of controlled acceptance, whether or not they have actually been consciously judged. If there is some state within us that would yield (reasonable or unreasonable) conscious and explicit assent if we were directly prompted, then we will count this as a belief.

In articulating the distinction between alief and belief, Gendler does seem to identify belief with what we would explicitly answer if directly pressed on a question: "Did you *really* believe that there was really an ax-murderer approaching you? That throwing the dart at the photograph would harm your loved one? That the metal bars were not strong enough to hold you?' (296) Explicit answers are taken to reveal beliefs because they are produced on reflection, subject to some norms of consistency with all-things-considered evidence.

The interesting question now concerns the relationship between Gendler's apparently descriptive claims about the reality-sensitivity of belief and her suggestion that belief differs from alief in requiring (actual or potential) acceptance or endorsement, subject to a norm of consistency. It might seem trivially true that a state of mind will be more reality-sensitive (more objectively accurate, better able to keep pace with variation in the world) when it is produced by a way of thinking that is governed by a norm of consistency. But it is not obvious that aiming at consistency in fact enhances accuracy; whether or not it does so may depend on a variety of circumstances. Many of Gendler's examples focus on situations in which explicit acceptance is more accurate or sensitive to reality than implicit attitude: we accept that the thick glass floor will support us, and it will, while we feel that we are at risk, and are wrong about that. But the relationship between acceptance and accuracy is also worth studying in the light of cases in which implicit performance seems to outperform considered judgment.

In Bechara's Iowa Gambling Task, which Gendler discusses in connection with other issues in her chapter 14, participants are 'loaned' $2000 in facsimile money and instructed

to maximize profit as they play a game. The game is played by selecting cards, one at a time, at liberty from any of four decks; the cards specify how much money is won or lost. Unbeknownst to the naïve participant, two of the decks (A and B) are 'bad', yielding high initial gains but an overall pattern of loss, because of varied and sometimes large penalties. The other two decks (C and D) are 'good', yielding smaller initial gains but substantially smaller penalties, for an overall pattern of gain. Participants know the game will end at some point, but do not know how many cards they will be allowed to choose before the game is over (in fact, the game always ends after 100 cards). After the first twenty cards, participants are asked to tell the experimenter everything they know about the game, and how they feel about it; these questions are repeated at 10-card intervals thereafter.

Bechara and colleagues found that normal participants sampled all the decks initially, but soon began to exhibit different physiological responses to the good and bad decks, sweating slightly (more technically, showing higher anticipatory skin conductance responses, or SCRs) before selecting cards from the bad decks (Bechara, Damasio, Tranel, & Damasio, 1997). Early in the game, participants did not report any knowledge or feelings distinguishing the decks (the 'pre-hunch' period), despite showing different SCR responses to them. By roughly the 50th card, the average normal participant reported feeling better about the good decks, although still unsure of what was going on (the 'hunch' stage). By the 80th card, most normal participants were able to articulate their grounds for favoring the good decks (entering the 'conceptual' phase), and played in accordance with their explicit knowledge. The interesting finding for present purposes is that normal participants started to favor the better decks well before they could report any feelings or knowledge to justify their choices; at the pre-hunch stage they did not explicitly accept that decks A and B

were worse, but were already responding to this fact in the way they acted. (A somewhat different pattern of response was uncovered in patients with prefrontal lobe damage, who did not exhibit anticipatory SCRs, and persisted in choosing disadvantageously even after reaching the point of gaining explicit knowledge of the odds; Bechara and colleagues took these results to support the view that intelligent choice is guided by somatic signals correlated with SCR.)

These findings have been somewhat controversial; for example, other researchers have argued that the crude report measures used by Bechara and colleagues underestimated the declarative knowledge of their participants in the early stages (Maia & McClelland, 2004). But although these more recent studies have argued that successful performance only begins when some degree of declarative knowledge is accessible, they do not insist that this knowledge must be consciously accessed by the participant to guide intelligent choice. Maia and colleagues stress that their results are for example compatible with a model in which 'the same knowledge store that participants canvas to generate verbal reports can also directly feed a response-selection mechanism without the need for conscious intermediation' (2004, 16078-9). They furthermore agreed to some extent with Bechara about the poverty of the information available to verbal report, noting that participants started to chose advantageously even when they had only a feeling that certain decks were better, without being able to articulate the grounds of their preferences (for example, by alluding to the pattern of larger losses in the bad decks).

Bechara and colleagues' original findings suggested that there are conditions in which our automatic impulses could be more reality-sensitive than our conscious, explicit

judgments (see also Bechara, Damasio, Tranel, & Damasio, 2005, where they aim to respond to their critics). Maia and colleagues dispute those findings, but are careful to point out that explicit judgment is not a necessary means to intelligent behavior. Automatic and associative reasoning is swift, and can be extremely responsive to how things are; although more effortful and conscious thinking can handle certain kinds of problems that cannot be computed associatively, it is slow, and not invariably more accurate or 'reality-sensitive' (Evans, 2010; Gigerenzer, 2008). Depending on the circumstances, engaging in controlled cognition deliberately aiming at consistency may or may not make us more accurate.

If Bechara is right, then the reality-tracking gut responses of the participants were at odds with what they overtly accepted; if Maia and colleagues are right, then there was no such opposition, but overt acceptance was still unnecessary for the accuracy of the gut responses. Even assuming gut responses actually were swifter at tracking the patterns in the cards than considered judgments, it would be open to Gendler to retain her general view about the greater reality-sensitivity of belief over alief by classifying even the gut responses in these cases as beliefs: after all, they were formed in response to patterns of evidence, and evidence-sensitivity is in some places identified as the hallmark of belief (e.g. at 297). However, it is not entirely clear how this move would fit with the insistence that belief must always be guided by norms of consistency with one's considered impression of things.

Another set of questions arise on considering some experimental work done within a model of belief that has much in common with Gendler's own, the APE (associative-propositional evaluation) model put forward by Bertram Gawronski and colleagues

(Gawronski & Bodenhausen, 2006; Gawronski, Strack, & Bodenhausen, 2009). This model draws a line between implicit attitudes, which are 'automatic affective reactions resulting from the particular associations that are activated automatically when one encounters a relevant stimulus' (2006, 693) and explicit attitudes, which are generated by propositional reasoning, where 'cognitive consistency is exclusively a concern of propositional reasoning' (2006, 695). Gawronski and colleagues do not themselves identify beliefs with explicit attitudes, but the differences between their position and Gendler's on this point may be partly verbal; what is important is that they seem to share with Gendler a commitment to a fundamental split between two kinds of states, distinguished by the manner in which they are produced, an automatic way of thinking on the one hand, and a controlled and ideally consistency-driven way on the other.

Like Gendler's model, the APE model is motivated by the discovery of situations in which the two contrasted types of state come apart. But in addition to looking at situations in which our automatic responses can be led astray, Gawronski and colleagues have supported their model by examining the dark side of the consistency-seeking mechanisms which support our explicitly endorsed attitudes. One potentially problematic facet of consistency-seeking has been studied in cognitive dissonance theory: we experience discomfort when there is a certain kind of conflict between our expressed attitudes and overt behavior. A classic demonstration of this effect is the forced-compliance paradigm, in which experimental participants are paid to express an attitude they do not antecedently endorse. In the original experiment in this program (Festinger & Carlsmith, 1959), participants were obliged to spend an hour doing very tedious and repetitive tasks with spools and pegs. They were then asked to say some scripted remarks to the person they

thought was the next participant (actually a collaborator): "It was very enjoyable, I had a lot of fun, I enjoyed myself, it was very interesting, it was intriguing, it was exciting." Participants were paid either $1 or $20 to deliver this message, and were shortly afterwards asked by a supposedly unrelated interviewer in another room about the interest and intrigue of the original tasks. Participants who were paid *less* were more enthusiastic at this point, apparently because the adoption of a more positive attitude would be needed to subjectively justify their behavior. This effect has held up under many variations (for a review, see Harmon-Jones & Mills, 1999).

In support of their model, Gawronski and Strack examined the impact of forced compliance on explicit and implicit attitudes, and found that low (but not high) payment for the expression of an previously unendorsed attitude resulted in shifts in explicit attitudes, where implicit attitudes were unmoved throughout (Gawronski & Strack, 2004). Gawronski and Strack argue that implicit attitudes can remain unmoved under the forced compliance manipulation because the manipulation selectively targets the validation-oriented propositional reasoning that underpins explicit attitudes. The fact that this type of reasoning is governed by a norm of consistency brings costs as well as benefits for accuracy: under some manipulations, the drive for consistency may pull us away from the truth. It is something of an open question whether the artificial manipulations of the forced compliance paradigm have real-world analogues of a sort that should make us take this vulnerability of explicit attitude formation to be a real as opposed to merely hypothetical threat to accuracy. Arguably, social pressures may have some attitude-distorting function along these lines: if there are minor social rewards for the overt expression of certain attitudes, we may drift from saying things we do not mean to tailoring our explicit attitudes

to match what we say, in a manner that does not necessarily reflect the truth of those attitudes.

In the forced compliance paradigm, as in many of the cases discussed by Gendler, the implicit attitude is stable while the explicit one shows variation, but it is also worth examining situations in which the explicit attitude is stable while the implicit attitude shifts. Certain patterns of experience can change our automatic associations without shifting our explicit judgments. For example, in their work on the malleability of implicit racial attitudes, Olson and Fazio (2006) exposed their White American participants to patterns of stimuli which including many pairings of Black individuals with positive words and images and White individuals with negative words and images. These pairings were concealed amid a large stream of filler stimuli; believing that they were participating in research on attention, naïve participants were distracted during the conditioning by a demand to respond to the occasional appearance of an target image unrelated to the experimental hypothesis. The conditioning went 'under the radar': participants did not have explicit knowledge of the pattern of pairings they had witnessed, performing no better than chance when asked in an immediately subsequent phase of the experiment whether various race and valence pairings had been witnessed earlier. In the next experiment, a similar evaluative conditioning procedure was applied to the experimental participants, while control participants experienced the same stimuli without any systematic pairings between race and word valence. All participants then completed a priming measure of implicit attitude (Fazio, Jackson, Dunton, & Williams, 1995), which demands the rapid identification of words as negative or positive when presented after a prime consisting of a Black or White face. Control participants showed the typical White

American pattern of slower identification of positive terms after Black faces; in experimental participants, however, this bias was had been expunged. The evaluative conditioning procedure was successful in eliminating ordinary prejudicial associations. But although evaluative conditioning had an impact on implicit attitudes, it did not shift explicit racial attitudes: there was no significant different between the control and experimental groups in responses to a series of questionnaires about racial and other social topics. A further experiment established that the conditioning procedure had more than momentary impact, with implicit attitudes reflecting its effect even after a two-day delay, again despite the absence of any amelioration in explicit attitudes.

Although there is a tradition of seeing implicit attitudes as deeply ingrained by habit or heredity and resistant to new input from the world (e.g. Wilson, Lindsey, & Schooler, 2000), work like Olson and Fazio's suggests that the real story may be more complicated. Subtle patterns of experience can reshape implicit attitudes without attracting conscious attention or updating our overt propositional commitments. The way in which implicit attitudes respond to changes in the world is different from the way in which explicit attitudes do so, but there are interesting strengths and vulnerabilities in both kinds of thinking, and it is not clear that implicit attitudes should be singled out as 'unable to keep pace with variation in the world'. Arguably, explicit and implicit attitudes are both in their ways shaped by evidence, albeit somewhat different types of evidence; they are both also susceptible to the influence of mere appearance, although again in different ways. The APE model offers support for the basic idea that there are two kinds of attitude generated by two kinds of processing; its more fine-grained picture of the regulation of these attitudes is not exactly in keeping with Gendler's. It is not clear whether this poses a difficulty for

Gendler's model, or whether the main theoretical entities of the APE model are in the end somewhat distinct from Gendler's concepts of alief and belief, but a sharper picture of the commitments of Gendler's model could emerge from a direct comparison with rivals such as APE.

A contrast with other dual process models could also shed light on the second fundamental feature of alief. According to Gendler, alief is not factorizable into representational and conative elements because it can drive action independently of desire: the person who wants to step boldly onto the Skywalk may find himself shivering and immobile at the edge of it, and the participant in an experiment on disgust may find himself unable to lick the revoltingly-shaped chocolate. This feature of alief puts it in sharp contrast to belief, which is ordinarily thought to motivate only in conjunction with desire. Advocates of standard belief-desire models might worry about whether Gendler's examples really show desire-independent responses, or whether the cases she examines simply involve especially inflexible desires: perhaps the desire for self-preservation and the desire to avoid contamination motivate so strongly that they trigger action even on the basis of a merely apparent object, and cannot easily be outweighed by one's rational but weaker desires for a fun walk or a sweet snack. The usual orthogonal interaction between belief and desire might furthermore be harder to spot in these cases simply because the particular motivating desires they involve are rarely subject to significant variation.

Here it would be good to know more about the conative counterparts of alief, which Gendler mentions in passing as possible, but does not (yet) discuss in detail. Much empirical work on implicit attitudes (including Gawronski's, for example) focuses as heavily on preferences or desires as on representational states. It is unclear whether

dissociations between explicit attitudes and behavior are generally better explained by difficulties in figuring out what is actual, or difficulties in figuring out what is desirable. The development of alief does not of course bar the way to the development of a corresponding conative notion, as Gendler herself points out, but in the interest of finding deeper generality it might be useful to probe or develop these concepts in parallel.

There are some potential challenges – or testing grounds – for the notion of alief in recent work on implicit attitudes and behavior. Some of this work looks poised to take apart what alief puts together, or to introduce independent factors as decisive in the cases of interest. For example, it is argued that the impact of implicit attitudes on behavior is moderated by motivational factors. One recent and large-scale review concludes that "implicit measures will primarily predict behaviour under conditions of low opportunity or motivation to control, or when individuals rely on automatic processes to guide their behaviour for any other reason" (Friese, Hofmann, & Schmitt, 2009). There is nothing directly hostile to alief in the idea that motivation to control behavior might weaken the impact of alief on action: as Gendler characterizes it, alief is one of many simultaneously occurring states vying for influence on the behavior of an individual. But to the extent that alief is committed to bundling an implicit attitude to a motor response without motivational dependency, analyzing actions in terms of alief may make it harder to understand what is going on here.

A specific example may illustrate the sort of challenge we face. Spider phobia looks like just the sort of domain where the concept of alief would be useful, a domain in which people have feelings which are 'often extremely resistant to corrective verbal information' (Baeyens, Eelen, Crombez, & Van den Bergh, 1992, 134). For those with spider phobia,

harmless household spiders can provoke very strong feelings of fear and disgust, which are recognized even by the phobic individuals as irrational. Indeed, phobic individuals are as likely as nonphobic individuals to see spider phobia as irrational (Mayer, Merckelbach, & Muris, 2000). Phobic feelings have a direct impact on behavior: an otherwise desirable treat becomes aversive if a spider touches it, or even touches a sealed outer wrapper in which it is contained (de Jong & Muris, 2002). An alief-oriented approach to this phobia would be easiest to run if phobic individuals simply had stronger automatic negative associations with spiders; however, on various measures of implicit attitude, including the Affective Simon Paradigm and the IAT, there is no significant difference between the scores of phobic and nonphobic individuals (de Jong, van den Hout, Rietbroek, & Huijding, 2003). What determines whether someone has phobia is something at the level of control, rather than strength of automatic association: 'the nonfearful individual is the one who can override this automatic negative stereotype, whereas the phobic individual is the one who does not attempt or is not able to control it' (de Jong et al., 2003, 540). This understanding of phobia does not dispute the power of associative representations, but shows that variations in behavior are not explained by variations in associative representations on their own.

Of course it is open to Gendler to maintain that we all alieve in the same way as far as spiders are concerned, but some of us have the motivational resources to override these aliefs and act on the basis of beliefs instead. But the difficult question of when we shift from alief- to belief-driven behavior then becomes a pressing one. In particular, it is an interesting question whether cognitive control is a special case of behavioral control more generally, so that motivational factors would play roughly the same role in switching us

from alief to belief as they would in combining with belief to produce action. Existing psychological models of this domain, such as Sherman's Quadruple process model (Sherman et al., 2008), are surprisingly complex. Whether or not they are compatible with the retention of alief as a distinctive psychological state with explanatory value would depend in part on how an alief-based theory could explain transitions between alief- and belief-based behavior, and this is a hard question so far largely left open in Gendler's work.

It is not surprising that there are problems left open in the development of the theory of alief: this theory is novel, in some ways counterintuitive, and extremely ambitious. One aspect of this ambition deserves special mention in closing. The theory is motivated by a close study of cases in which there is some tension between our behavior and our reflective attitudes, but Gendler suggests that its application may in the end be much wider: in her view even the behavior that aligns with what we reflectively endorse could typically be driven by the more primitive state of alief (266, 281). This idea runs against our usual self-conception. Perhaps in part because folk psychology makes belief attribution feel so natural to us, it is hard to come to see ourselves as driven by some messier packages of representational, affective and motor signals, but Gendler may well be right that something like this is the truth of our condition. When we look closely enough, even the paradigm cases of rational attitude formation may be linked to affective and motor impulses; for example, there is evidence that the detection of syllogistic logical relations is partly enabled by emotional response (Morsanyi & Handley, in press). Meanwhile, according to one of the clearest models of the cognitive architecture of dual process theory, our conscious, serial way of thinking (System 2) does not exist above or independently of our unconscious, heuristic way of thinking (System 1): rather, System 2 is

itself realized in cycles of System 1 operations originally developed for the mental rehearsal of action (Carruthers, 2009). Our capacity for pure reasoning is on this view intimately connected with our capacity for motor control, a finding which cannot simply be read off the way in which this kind of reasoning presents itself to consciousness.

Whether or not the various rational and sub-rational elements in our thinking will end up being connected in a way that Gendler would recognize as alief, the method that she has pioneered in her recent essays stands out as an extremely promising way of making progress. Gendler is uncommonly alive to the possibility that we are not what we seem to ourselves to be. Her strategy of looking at the places where rationality is compromised, but not entirely absent—the dark outskirts of rationality—stands out as a singularly good way to set aside superficial appearances and gain deeper insight into the mind.

References:

Baeyens, F., Eelen, P., Crombez, G., & Van den Bergh, O. (1992). Human evaluative conditioning: Acquisition trials, presentation schedule, evaluative style and contingency awareness. *Behaviour Research and Therapy, 30*(2), 133-142.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (2005). The Iowa Gambling Task and the somatic marker hypothesis: some questions and answers. *Trends in cognitive sciences, 9*(4), 159-162.

Carruthers, P. (2009). An architecture for dual reasoning. In K. Frankish & J. Evans (Eds.), *In Two Minds: dual process and beyond* (pp. 109-127). Oxford: Oxford University Press.

de Jong, P., van den Hout, M., Rietbroek, H., & Huijding, J. (2003). Dissociations between implicit and explicit attitudes toward phobic stimuli. *Cognition & Emotion, 17*(4), 521-545.

de Jong, P. J., & Muris, P. (2002). Spider phobia:: Interaction of disgust and perceived likelihood of involuntary physical contact. *Journal of anxiety disorders, 16*(1), 51-65.

Evans, J. S. B. T. (2010). Intuition and Reasoning: A Dual-Process Perspective. *Psychological Inquiry, 21*(4), 313-326.

Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobstrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology, 69*(6), 1013.

Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203.

Frankish, K. (2009). Systems and levels: dual-system theories and the personalsubpersonal distinction. *In Two Minds: Dual Processes and Beyond*, 89-107.

Friese, M., Hofmann, W., & Schmitt, M. (2009). When and why do implicit measures predict behaviour? Empirical evidence for the moderating role of opportunity, motivation, and process reliance. *European Review of Social Psychology, 19*(1), 285-338.

Gawronski, B., & 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., & Strack, F. (2004). On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology, 40*(4), 535-542.

Gawronski, B., Strack, F., & Bodenhausen, G. V. (2009). Attitudes and Cognitive Consistency: The role of associative and propositional processes. In R. E. Petty & R. H. Fazio & P. Brinol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 85-117). Mahwah, NJ: Erlbaum.

Gendler, T. S. (2011). *Intuition, Imagination, and Philosophical Methodology*. New York: Oxford Univ Pr.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science, 3*(1), 20.

Harmon-Jones, E., & Mills, J. (1999). *An introduction to cognitive dissonance theory and an overview of current perspectives on the theory*. Washington: American Psychological Association.

Maia, T. V., & McClelland, J. L. (2004). A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the Iowa gambling task. *Proceedings of the National Academy of Sciences of the United States of America, 101*(45), 16075.

Mayer, B., Merckelbach, H., & Muris, P. (2000). Self-reported automaticity and irrationality in spider phobia. *Psychological reports, 87*(2), 395-405.

Morsanyi, K., & Handley, S. J. (in press). Logic feels so good -- I like it! Evidence for intuitive detection of logicality in syllogistic reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin, 32*(4), 421.

Sherman, J. W., Gawronski, B., Gonsalkorale, K., Hugenberg, K., Allen, T. J., & Groom, C. J. (2008). The self-regulation of automatic associations and behavioral impulses. *Psychological Review, 115*(2), 314.

Wilson, T., Lindsey, S., & Schooler, T. (2000). A model of dual attitudes. *Psychological Review, 107*(1), 101-126.