Why Implicit Attitudes Are (Probably) not Beliefs

Abstract

Should we understand implicit attitudes on the model of belief? I argue that implicit attitudes are (probably) members of a different psychological kind altogether, because they seem to be insensitive to the logical form of an agent's thoughts and perceptions. A state is sensitive to logical form only if it is sensitive to the logical constituents of the content of other states (e.g., operators like negation and conditional). I explain sensitivity to logical form and argue that it is a necessary condition for belief. Then I appeal to two areas of research that seem to show that implicit attitudes fail *spectacularly* to satisfy this condition—although there are persistent gaps in the empirical literature that leave matters inconclusive. I sketch an alternative account of how implicit attitudes respond to other mental states. Specifically, I propose that implicit attitudes are sensitive to relations of mere spatiotemporal contiguity in thought and perception, i.e., the spatial and temporal orders in which people think, see, or hear things.

1. Introduction: Madeleine meets Bob.

Imagine Madeleine seated at a computer in a psychology lab. She is learning about a fellow named Bob. She sees photos of Bob and reads about his pastimes and habits. Bob volunteers at an orphanage, assists the elderly, and fights against discriminatory laws that make it difficult for minorities to vote. When asked what she thinks of him, Madeleine says that Bob is agreeable. She is, apparently, pro-Bob. Unbeknownst to Madeleine, however, the computer has been flashing words such as "death," "hate," and "disgusting" before each photo. These words appear too quickly for Madeleine to recognize them consciously but long enough to register them subliminally. Given these subliminal perceptions, Madeleine has acquired a set of *anti*-Bob dispositions. Were she to interview him for a job, she would sit farther away and make less eye contact with him than she would another candidate. Were she to read Bob's résumé, she would dwell longer on his deficiencies than on his accomplishments. She would be less likely to consider him a good candidate for hire and more likely to think that he would end up in jail.[1]

[1] The case of Madeleine is based on Rydell and colleagues (2006), which measured the influence of subliminal conditioning on a timed association task (the Implicit Association Test) but not on more ecologically valid

The case of Madeleine and Bob foregrounds a tension in our understanding of belief. On the one hand, beliefs are thought to reflect what an agent takes to be true of the world. According to this view, Madeleine believes that Bob is agreeable. On the other hand, beliefs are thought to guide actions, together with an agent's desires and ends. According to this view, Madeleine does not believe that Bob is agreeable. In the example of Madeleine and Bob, the roles of truth-taking and of action-guiding come apart. Does Madeleine believe that Bob is agreeable, given what she judges to be true in light of the evidence? Or does she "really" believe that Bob is not agreeable, given how she unreflectively acts toward him? Does she believe both? Or perhaps neither?

Madeleine's ambivalence toward Bob shares a common structure with more troubling cases. Madeleine resembles someone who, like many members of liberal democracies, sincerely reports anti-racist beliefs while exhibiting predictable patterns of subtle racial bias in her thought and behavior. In these instances of "aversive racism" (Pearson and colleagues 2009), as in Madeleine's ambivalence toward Bob, agents' explicit reports seem to reflect their considered judgments, while their unreflective states pull them in undesirable directions. Psychologists refer to these unreflective states as "implicit attitudes," which they contrast with "explicit attitudes." Madeleine has pro-Bob explicit attitudes and anti-Bob implicit attitudes. Aversive racists have egalitarian explicit attitudes and prejudiced implicit attitudes.

It is clear that phenomena like aversive racism help sustain disparities between advantaged and disadvantaged social groups. For example, Dan-Olof Rooth and colleagues found that implicit *work-performance* biases in Sweden predicted real-world hiring

---

behaviors. For the influence of subliminal conditioning on seating distance and other nonverbal behaviors, see e.g., Kawakami et al. (2007b). For the influence of conditioning on résumé evaluation, see, e.g., Kawakami et al. (2007a).

discrimination against both Arab-Muslims (Rooth 2010) and obese individuals (Agerström and Rooth 2011). Employers who implicitly associated these social groups with laziness and incompetence were less likely to contact job applicants from these groups for an interview. In both cases, measures of implicit attitudes significantly predicted hiring discrimination over and above employers' self-reported attitudes. Even critics of the predictive power of measures of implicit racial and ethnic attitudes, such as Oswald et al. (2013), find that performances on these measures are small but consistent predictors of racially discriminatory behavior in medicine, law, education, and employment. As Valian (1998) and Greenwald, Banaji, and Nosek (forthcoming) explain, statistically small effects can add up to significant discriminatory impact, because they can affect considerable subsets of populations simultaneously and specific individuals repeatedly.

What, if anything, can we do to mitigate the effects of implicit biases? The correct answer to this question depends in part on the nature of the underlying psychological states. Combating implicit biases requires that we know what we are up against. Fully grasping their psychological nature requires understanding their modes of formation (what causes them to form, and how), modes of operation (how they work, e.g., what activates them in specific contexts, and how they interact with other mental states), modes of manifestation (how they are expressed in thought, feeling, and action, and how these expressions might be controlled), and modes of malleability (how they change over time) (citations removed). One pressing question is just how *belief-like* implicit attitudes are. For example, if implicit attitudes are belief-like, then perhaps we can combat them via rational argumentation. In this vein, Eric Mandelbaum (2013, forthcoming) cites evidence that implicit attitudes change in response to strong (persuasive) arguments and fail to change in response to weak arguments (e.g., Briñol, Petty, and McCaslin

2008). If implicit attitudes are radically unlike beliefs, however, then arguments may fail to change them. Such a finding would not entail, of course, that rational argumentation had no role to play in the fight against implicit prejudice, but its role might be relatively circumscribed: perhaps it can serve to draw people's attention to their unnoticed biases and to motivate them to take steps to control or change them, but argumentation would not itself *reduce* these biases, or provide practical guidance about how to do so.

Here I argue that, contrary to the views of many philosophers and some psychologists,[2] these unreflective rogue dispositions are likely not expressions of a belief-like attitude, but of a different psychological kind altogether. Implicit attitudes are responsive to an agent's thoughts, but, unlike beliefs, they seem to be *insensitive to the logical form* of those thoughts. A state is sensitive to logical form only if it is sensitive to the content of the states with which it interacts. Specifically, sensitivity to logical form requires sensitivity to the logical constituents of the content (e.g., logical operators like negation and conditional).[3] I argue that belief-like cognitive states are, and implicit attitudes are probably not, sensitive to logical form.

In what follows, I survey prominent arguments that implicit attitudes are belief-like (§2). I then explain sensitivity to logical form and argue that it is a necessary condition for belief (§3). I appeal to two areas of research to show that implicit attitudes seem to fail *spectacularly* to satisfy this condition (§§4-5), although I note significant gaps in the empirical literature that leave matters inconclusive. I sketch an alternative account of how implicit attitudes respond to

[2] De Houwer (2011, 2014), Egan (2011), Gertler (2011), Huddleston (2011), Huebner (2009), Hunter (2011), Kwong (2012), Mandelbaum (2013, forthcoming), Mitchell and colleagues (2009), Muller and Bashour (2011), Rowbottom (2007), and Schwitzgebel (2010).

[3] For my purposes, it does not matter whether one takes the constituents of mental content to be properties and objects (a Russellian view) or modes of presentation of properties and objects (a Fregean view). For ease of exposition, I often assume in what follows that the contents of mental states like belief have some internal structure, but much of what I say could be reformulated to address views that do not attribute structure to mental content. Implicit attitudes are puzzling enough to require special explanation on just about any view, e.g., Schwitzgebel's (2010) dispositionalist treatment. See §2 for further discussion.

other mental states. Specifically, I propose that implicit attitudes are sensitive to relations of mere spatiotemporal contiguity in thought and perception, i.e., the spatial and temporal orders in which people think, see, or hear things. I briefly consider the empirical evidence that has tempted many philosophers and some psychologists to adopt a belief-based construal (or BBC) of implicit attitudes, and explain how little the findings in question actually do to support that construal (§6). I consider the broader implications of my account for cognitive science and the philosophy of mind (§7), and address outstanding objections (§8).

2. Belief-Based Construals of Implicit Attitudes

One view holds that implicit attitudes are obviously *not* beliefs, because they fail to meet certain sophisticated cognitive criteria, such as being readily revisable with the evidence, readily available for conscious reflection, or readily assimilable with other beliefs, desires, and intentions.[4] Such criteria are, however, too demanding, insofar as they rule out the possibility that infants and non-human animals ever have beliefs. Another view holds that implicit attitudes obviously *are* beliefs, because they seem to meet certain very generic criteria, such as being "states of taking the world to be a certain way."[5] But such criteria are too permissive (not to mention vague), insofar as they fail to differentiate beliefs from other sorts of intentional

[4] See, e.g., Gendler (2008a,b), Levy (2014a), and Zimmerman (2007). Another reason one might distinguish implicit attitudes from beliefs is that the former inherently include an affective or motivational component, whereas beliefs (whether conscious or otherwise) are putatively "cold" cognitive states (name removed, p.c.). However, the cognitive/conative distinction is somewhat orthogonal here. For example, de Houwer (2014) defends a "propositional" (belief-like) account of our implicit likes and dislikes, while Mandelbaum (2013) argues that all beliefs are affective and motivational. I disagree with de Houwer, but, as I discuss in §7, some motivational states, such as intentions, arguably are sensitive to logical form. I say more about the conative aspects of implicit attitudes and beliefs in (citations removed).

[5] E.g., Sommers (2009) writes, "To believe is to take something to be so and so" (269), and further claims that, "animal and human belief is mainly… propositionless" (270).

states, from the most primitive perceptions to complex imaginings. The correct criterion will fall between these two extremes, and, in the next section, I argue that *sensitivity to logical form* is a good fit for this purpose. Since a primary aim of this paper is to argue that implicit attitudes are not beliefs, I begin by surveying some prominent belief-based construals (or BBCs) of implicit attitudes.[6]

Debates about whether implicit attitudes are belief-like have generally focused on the extent to which they revise in the light of incoming evidence and, relatedly, the extent to which they are "inferentially promiscuous," that is, involved in inferences with other mental states. Much of the discussion of sensitivity to evidence was set in motion by Tamar Szabó Gendler's influential work on implicit attitudes, which she calls "aliefs." Gendler summarizes one of her central arguments for why implicit attitudes (or aliefs) differ from beliefs in the following way:

> Beliefs change in response to changes in evidence; aliefs change in response to changes in habit. If new evidence won't cause you to change your behavior in response to an apparent stimulus, then your reaction is due to alief rather than belief. (2008b, 566)

Gendler argues that aliefs and beliefs differ in kind because beliefs revise immediately in light of the evidence, while aliefs do not. Whither goes the evidence, thither your beliefs shall follow. However, Gendler acknowledges that, as stated, such claims are too strong, because there are many beliefs that fail to revise with the evidence. The argument rules out the possibility that an otherwise belief-like state could fail to meet the standards of ideal rationality and still qualify as a belief. Although this criterion therefore seems too demanding, defenders of BBC have nevertheless argued that implicit attitudes are, at least sometimes and to some degree, appropriately responsive to changes in evidence. They also make the related claim that implicit attitudes are involved, at least sometimes, in rational inferential processes. Evidence-

[6] I canvass a broader range of interpretations of implicit attitudes elsewhere (citation removed).

sensitivity and inferential promiscuity are related in a number of ways. Plausibly, evidence-sensitivity is itself often a matter of inference, e.g., making inductive inferences about the way the world is based on experience, inferring that the incoming information is inconsistent with one's prior beliefs, and so on. Neil Levy (2014b, 6) sketches further connections between evidence-sensitivity and inferential promiscuity this way:

> Whereas inferential promiscuity is a matter of how beliefs cause behavior and update other mental states, responsiveness to evidence is a matter of how the belief itself can be expected to update, given appropriate evidence. Inferential promiscuity and responsiveness to evidence are two sides of the same coin: beliefs are inferentially promiscuous, causing the update of other beliefs, because beliefs are responsive to evidence. Any state which is inferentially promiscuous and appropriately responsive to evidence is a belief…

In short, although evidence-sensitivity and inferential promiscuity may be too demanding as necessary conditions on belief, perhaps they jointly constitute sufficient conditions. Citing evidence (which I discuss in §6) that seems to suggest that implicit attitudes at least sometimes respond to evidence and integrate inferentially with other attitudes, some defenders of BBC conclude that implicit attitudes are full-fledged beliefs, while others suggest that the difference between implicit attitudes and beliefs is more a matter of degree than of kind.

Jan de Houwer (2011, 2014) outlines a "propositional model" of implicit attitudes, according to which implicit attitudes typically form and change as a result of inductive inferences that individuals make on the basis of observed environmental contingencies (see also Mitchell, de Houwer, and Lovibond 2009). Mandelbaum (2013, forthcoming) defends a similar view, arguing that implicit attitudes are non-conscious beliefs whose truth conditions are determined by their language-like compositional structure. (The most salient difference between de Houwer and Mandelbaum's views about implicit attitudes is that the former hypothesizes that individuals are consciously aware of the observed environmental

contingencies that shape their implicit attitudes, while the latter supposes that the brunt of the cognitive action is unconscious.) These BBCs can readily explain cases in which implicit attitudes seem to update in rational ways and integrate inferentially with other attitudes, but they may have a harder time explaining cases when implicit attitudes seem more evidence-recalcitrant and inferentially encapsulated than other beliefs. Views like de Houwer's, that is, may *overestimate* the rationality of implicit attitudes.[7]

Eric Schwitzgebel (2010), by contrast, argues that implicit attitudes and beliefs do differ, but only by degree, rather than by kind. Schwitzgebel agrees that evidence-sensitivity and inferential dispositions are relevant to belief attribution (2010, 532, 540-1, 549n5, 550n10), but claims that in many cases implicit attitudes occupy a nebulous middle-ground such that they are somewhat evidence-sensitive and inferentially promiscuous, but less so than paradigmatic beliefs. Schwitzgebel would interpret Madeleine's ambivalent dispositions toward Bob, and aversive racists' ambivalent dispositions toward blacks, as cases of "in-between belief," such that there is no determinate fact about, say, whether Madeleine *really* believes that Bob is agreeable. Also in contrast to Mandelbaum, Schwitzgebel adopts a dispositionalist approach to belief. Whereas Mandelbaum appeals to the internal structure of implicit attitudes to explain why they (are disposed to) behave as they do, Schwitzgebel takes their dispositional profile to be basic.

More recently, Levy (2014a,b) has developed a view of implicit attitudes that incorporates elements of each of Mandelbaum, Gendler, and Schwitzgebel's views. Levy (2014b, 2) suggests that implicit attitudes are *somewhat* sensitive "to evidence and to the

[7] A related worry is that collapsing implicit attitudes and other beliefs may *underestimate* the rationality of other beliefs. Mandelbaum (2014) suggests that the mechanisms of belief formation are radically less rational than philosophers ordinarily think.

semantic content of other attitudes," but insists that, "they are not sensitive enough… to qualify as beliefs." They only "respond to semantic contents in a patchy and fragmented way" (2). They are sometimes involved in inferential processes, and sometimes revise in light of the evidence, but not reliably or widely enough to qualify as full-fledged beliefs. In these respects, Levy seems to agree with Schwitzgebel that implicit attitudes are not sharply, categorically distinct from ordinary beliefs, but simply lie further along a continuum away from ideal evidence-sensitivity and inferential promiscuity. Nevertheless, Levy seems to think that the distance between full-fledged beliefs and implicit attitudes is sufficiently large to treat them as separate psychological categories. So, following Gendler and (citations removed), Levy proposes that implicit attitudes "are *sui generis* states for which we lack any term in our folk psychological vocabulary" (17). Levy introduces a new term (initially proposed by Susanna Siegel) for implicit attitudes: "patchy endorsements." But whereas Gendler takes an explicitly agnostic stance regarding how the representational content of implicit attitudes is structured ("perhaps propositionally, perhaps nonpropositionally, perhaps conceptually, perhaps nonconceptually," 2008a, 643), Levy sides with Mandelbaum in arguing that they are structured propositionally. Levy's primary argument for this claim is that they exhibit some middling degree of evidence-sensitivity and inferential promiscuity:

> implicit attitudes are involved in seemingly inferential processes: explaining this evidence requires us to attribute propositional structure to implicit attitudes. But evidence of propositional structure is not sufficient to establish that implicit attitudes are beliefs. The evidence that implicit attitudes have propositional structure consists in evidence that they respond to the semantic contents of other states, and beliefs are states that respond in this kind of way, but beliefs do this *systematically*. If implicit attitudes are states that respond to semantic contents in a patchy and fragmented way, they are neither associations nor beliefs. (2014b, 2)

In what follows, I defend the harder-line stance that implicit attitudes differ fundamentally in kind, and not just in degree, from beliefs (although I also suggest that the

empirical evidence remains inconclusive). I advance novel arguments for a view similar to Gendler's (2008a,b). To a certain extent, however, this essay can also be read as a friendly refinement of Levy and Schwitzgebel's views. Suppose that they are right that implicit attitudes are in an important sense "between" belief and non-belief. Can we say something more precise about why they occupy this middle-ground of sensitivity to evidence and semantic content? Is there an explanation for why they are involved in *these* inferences and not *those*, and respond to *this* evidence and not *that*? Are they perhaps sensitive to some types of evidence, or certain aspects of semantic content, and not others?

3. Logical Form and Belief

I propose that the more precise notion we are circling around is sensitivity to logical form: beliefs are, and implicit attitudes are not, sensitive to the logical form of other mental states. To convey the basic idea of sensitivity to logical form (or form-sensitivity), consider some examples drawn from influential theories of implicit attitudes in social psychology. Roland Deutsch and Fritz Strack (2010, 64-5) propose that individuals might form an implicit attitude linking "Arab" with "terror" in response to frequent media exposure, regardless whether they would reflectively agree that, "Most Arabs are terrorists" (2010, 64-65). Individuals may consciously agree when they hear, "It is wrong to identify Arabs with terrorism," and "Most Arabs do not support terrorism." At the same time, Deutsch and Strack predict that simply hearing the conjunction of the terms "Arabs" and "terrorism" in these very claims may reinforce an implicit attitude that associates Arabs with terror. Similarly, Bertram Gawronski and colleagues (2008, 376) suggest that trying to reject a common stereotype by thinking, "it is not true that old people are bad

drivers," may reinforce rather than undermine a negative implicit attitude toward elderly drivers. In these cases, the effects on implicit attitudes seem to reflect a sensitivity to certain linguistic tokens ("Arabs", "terrorism," "old people," "bad drivers") but an insensitivity to the logical form of the thought as a whole.[8] In these examples, the only "evidence-sensitivity" posited of implicit attitudes is sensitivity to experienced relations of spatiotemporal contiguity, i.e., the order in which individuals perceive, think, and feel things.

Giving a substantive account of "logical form," and what sensitivity to logical form requires, is no easy task. As I will use the term, logical form is closely tied to semantic content, which is to say, the truth conditions of cognitive states like belief and the satisfaction conditions of conative states like intention and desire. However, I focus on logical form rather than the more general notion of semantic content because, as the examples of the prior paragraph indicate, there may be some senses in which implicit attitudes are responding to the meanings of terms (e.g., "old people" and "terrorism"), while they are insensitive to the logical constituents of content (e.g., the "not" in "old people are not bad drivers").[9] Focusing on logical form allows greater precision than is afforded by Schwitzgebel and Levy's views. We can say more than that implicit attitudes are sometimes and to some degree sensitive to semantic content. More concretely and specifically, they are *insensitive* to logical operators like negation and conditional.

[8] One referee found it implausible to suggest that implicit attitudes are *completely* unresponsive to logical form, because a consequence would be that "the person with the pernicious implicit bias that 'blacks carry weapons' would not differentiate in his implicit attitudes between this and 'blacks don't carry weapons' or 'weapons carry blacks.'" While I recognize that this view may seem implausible, it is consistent with prominent social-psychological theories and with the evidence I discuss below. Another virtue of this hypothesis is its falsifiability. The hypothesis that implicit attitudes are categorically insensitive to logical form seems especially amenable to empirical disconfirmation, and I propose a number of further tests in what follows. I also endeavor to explain away the appearance that implicit attitudes are even *somewhat* or *sometimes* sensitive to logical form.

[9] Some doubt whether any principled distinction can be drawn between logical and non-logical terms (see Lepore and Ludwig 2001 for discussion). I will not argue here that logical terms are semantically or syntactically special, although it does seem to be the case that implicit attitudes respond differently to logical versus non-logical terms, so that may be one reason to distinguish them.

In this section, I try to elucidate logical form, and why sensitivity to logical form is necessary for belief, primarily by reference to examples. I do not offer jointly necessary and sufficient conditions, which are notoriously hard to come by for most interesting concepts. The primary reason I do not defend any particular notion of logical form is to avoid incurring theoretical commitments that are tangential to the central topics of this essay, and to address as broad a range of views as possible. On one view, logical form represents the underlying (real or "deep") structure of thoughts or sentences (Harman 1970; Stanley 2000; Mandelbaum forthcoming).[10] This view of logical form may be particularly congenial to the story I tell in what follows, but my story does not require it. For example, Lepore and Ludwig (2001), following Quine and Davidson, advocate abandoning the idea of a "reified" notion of logical form, according to which there is some distinctive entity which is *the* logical form of a sentence. They instead take *sameness of logical form* to be basic, which makes it possible to say that "Snow is white" has the same logical form as "Schnee ist weiss," without committing oneself to the existence of some third abstract entity in Platonic heaven which is the logical form that the two sentences share. Such a view would also be congenial to my approach.

The aim of sidestepping peripheral debates also lies behind my focus on whether these attitudes are *sensitive* to logical form, rather than whether they themselves *have* logical form, i.e., are propositionally structured, which is a point of emphasis for Mandelbaum and Levy. Levy infers that implicit attitudes are propositionally structured because they respond to semantic content. Perhaps he is right, but this is a further, contentious inference (from dispositional profile to internal structure), which needlessly narrows the range of views at play. Functionalists

[10] Such a "descriptive" view of logical form differs from another historically important "normative" understanding of logical form, as the idealized structure of sentences or thoughts, which are only imperfectly approximated by natural language. Logical form, as I will use the term here, refers to actual properties of concrete entities, rather than to idealized abstractions.

about belief, or dispositionalists like Schwitzgebel, may deny that beliefs are propositionally structured, or even that beliefs as a class share any substantive internal-structural features, but they do not deny that beliefs respond to logically structured information. They do not deny, for example, that beliefs are, other things equal, disposed to respond differently to "It is true that old people are bad drivers," and "It is not true that old people are bad drivers." Despite background differences in theories of mental content, all the defenders of BBC mentioned in the previous section should agree (to some version of the claim) that implicit attitudes are (at least to some degree) sensitive to logical form. I hazard my own views about the intentional content of implicit attitudes in (citations removed). Here I merely claim that, whether and however implicit attitudes are internally structured, they are insensitive to logical form.

I believe this claim is categorically true. Implicit attitudes are not just "less" sensitive to logical form than beliefs, but, as a class, wholly insensitive. Where some see conclusive evidence for partial sensitivity (e.g., Levy 2014b, 8), I see inconclusive evidence for total insensitivity. In §6, I attempt to show how sensitivity to relations of spatiotemporal contiguity can explain (away) much of the appearance that implicit attitudes are somewhat-but-far-from-ideally sensitive to evidence and logical form, that is, to explain how implicit attitudes can walk and talk like beliefs in certain circumscribed contexts while being nothing of the sort.

Ultimately, my interest is not primarily in what we choose to call "belief," but in carving the mind at its joints. Sensitivity to logical form marks an important distinction, and we would be remiss in grouping states that have and lack this sensitivity together. I hope to show that sensitivity to logical form constitutes a significant cognitive benchmark separating primitive from sophisticated mental states (§7). Arguably, form-sensitivity is a necessary condition for the more sophisticated conditions of evidence-sensitivity and inferential promiscuity that Gendler

and others attribute to belief. To be even *capable* of engaging in inferences with other mental states, implicit attitudes must be sensitive to the logical form of those states. If they are not, then they are not even in the ballpark of belief. Thus, while form-sensitivity may be a significant dividing line between more and less cognitively sophisticated states, it is a considerably less demanding condition than has been proposed by other writers. As I explain in this section, it is possible for a mental state, such as a belief, to be robustly sensitive to the logical form of other states while being extremely recalcitrant to changes in the incoming evidence and highly susceptible to a wide range of nonrational and even irrational influences. I return to this point in §8.

To get a better handle on form-sensitivity, as I will be employing the notion, let us return to Madeleine, who is daydreaming while her friend Theo tells her the latest gossip. Due to her distraction, Madeleine only recalls that Theo's utterance included the words "Mason" and "John." Without letting on that she wasn't really listening, she tries to piece together what he was saying: "Did he say *that John is a mason* or *that Mason is a john*?" What she comes to *believe* depends not just on the words passing through her "inner monologue," but also on the *logical form* of her thoughts about Theo's utterance, that is, what she takes him to be saying. Now consider some variations of this example. The first two cases involve interactions among psychological states that are sensitive to logical form, whereas the latter cases involve interactions among states that fail to respect logical form.

(1) Suppose Madeleine comes to think that Theo meant to break the bad news to her about Mason. Her mind starts reeling: "Mason is one my closest friends… Mason is a john?!… Ugh, one of my closest friends is a john!" This psychological transition is evidently rational, and the operative mental states are sensitive to logical form.

Take Madeleine's prior belief *that Mason is one of her closest friends*, which interacts with the thought *that Mason is a john*. The outcome of this interaction is the formation of the new belief *that one of her closest friends is a john*. In this case, truth is preserved from the prior, premise-like states to the subsequent conclusion-like states. In order for transitions of this sort to be reliably truth-preserving, Madeleine's prior belief *that Mason is one of her closest friends* must be sensitive to the logical form of the thought *that Mason is a john*, and vice versa. One state must do more than respond to the fact that the other state also refers to Mason or includes the linguistic token "Mason." It must respond to *what* the other state is *saying* about him. A mental state must respond this way, in very simple and straightforward cases like this, in order to be a belief.[11]

(2) Now imagine that Madeleine's attachment to Mason distorts her reasoning. Her thoughts continue: "… One of my closest friends is a john! Ugh, I can't believe it! I can't believe one of my friends does that. There is no way that *Mason* does that." Madeleine then jumps to Mason's defense and accuses Theo of spreading rumors.

Take Madeleine's prior belief *that Mason is one of her closest friends*. Suppose that, in this case, this state interacts with the belief *that none of her friends is a john*. The outcome is, inter alia, that Madeleine fails to adopt the belief *that Mason is a john*. Madeleine's response may not be fully rational. Perhaps she knows Theo to be trustworthy, and so should believe his testimony, but simply cannot bring herself to do it. Nevertheless, her failure to *revise* her attitudes toward Mason is entirely consistent with those attitudes *being beliefs*, because the operative states are appropriately sensitive to logical form. In this case, Madeleine responds by

[11] What role does the *agent* play in these sorts of psychological transitions? Which sensitivities and abilities must an agent, or a cognitive system, have for these transitions to take place? Here I remain neutral about these questions, which require a separate treatment. My focus is on the properties of certain states within the cognitive system.

rejecting a premise (that Mason is a john) instead of accepting the conclusion (that one of her closest friends is a john). There may be any number of rational, nonrational, or irrational factors that lead her to respond one way rather than another (I discuss some examples in §8). Whether her reaction is fully rational depends on the quality of her *reasons* for responding one way or another. Whether her reaction involves interactions between beliefs, however, depends on whether those states are sensitive to logical form.

Many states of belief, such as strong convictions or unconsciously persevering beliefs, do not revise immediately in light of contravening evidence. Nevertheless, becoming occurrently aware of such apparent inconsistencies disposes agents either to reject *other* inconsistent beliefs, to discredit the new evidence, or to consider ways in which the appearance of inconsistency is illusory. These cases of (potentially but not necessarily irrational) attitude perseverance require that the operative mental states be sensitive to logical form.[12]

(3) Next imagine that Madeleine responds to Theo's utterance by thinking, "Mason is a john. John is one of my friends. One of my friends is a *mason*." Now something has gone wrong. Madeleine replies by asking whether Theo meant that John is a Freemason or a masonry worker. "What?" Theo says, "John is not a mason!" Madeleine suddenly realizes that she has made a mistake, perhaps due to her distraction. She thinks through what he said again and her thoughts follow the original pattern of case (1): "Mason is one my closest friends. Mason is a john. One of my closest friends is a john."

[12] This relates to a criticism of "wide-scope" interpretations of rational requirements, raised by, e.g., Kolodny (2005), that there are often palpable "asymmetries" between different ways of resolving cognitive inconsistencies. For example, if Madeleine intends to drink a beer and believes that there is beer in the fridge, it seems better, rationally speaking, to resolve the situation by going to the fridge to get the beer than by abandoning her belief that there *is* beer in the fridge. But both responses are acceptable from the perspective of form-sensitivity.

In this case, Madeleine succumbs to an isolated "performance" error due to a momentary cognitive lapse, which is quickly corrected when she turns her full attention to the task. Such isolated departures from form-sensitivity are common and unremarkable. Her prior attitude *that Mason is one of her closest friends* displays sensitivity to the logical form of other states when she is undistracted. It is still clearly a belief.

(4) This case begins like (3). Theo breaks the bad news about Mason, then Madeleine puzzlingly asks what sort of mason John is. In this case, however, after Theo exclaims that John is *not* a mason, Madeleine thinks, "John is not a mason. John is one of my friends. One of my friends is a mason." She repeats, "But what sort of mason *is* he?" Madeleine started out by thinking *that John is not a mason,* but ends up wanting to know just what sort of mason he *is*. These psychological transitions are not just responding to logical form in an objectionable way, as in cases (2) and (3), but failing to respond to logical form at all.

Concerned, Theo exclaims, "John is no mason of any kind!" But to no avail. Madeleine responds each time by asking him how their friend John developed a propensity for masonry. She is, for whatever reason, systematically unable to properly think through what Theo is saying. Although she seems to be sensitive to some part of the meaning of Theo's assertions, and although there is some sort of effect on her beliefs, the intervening psychological transitions fail to respect the logical form of her initial, premise-like mental states. At least one of the operative mental states is not appropriately sensitive to the logical form of the others. As a result, her responses are becoming unintelligible.[13]

[13] Case (4) is so bizarre that one might reasonably wonder whether something is wrong with *Madeleine*, rather than her mental states (see also note #11). This reflects a limitation in the analogy between my toy example, in which we are envisaging a conscious sequence of belief-like thoughts unfolding in inference-like ways, and the research I discuss below, where my point is precisely that we should *not* posit such a belief-like inferential sequence. The

The difference between cases (3) and (4) brings out an essential component of form-sensitivity. In (3), Madeleine's confused inference is corrected once it is brought to her attention. In (4), the potential for correction is lost. Moreover, this difference would remain even if the error had *not* been brought to anyone's attention. Imagine cases (3*) and (4*) in which Theo happens to actually say, "John is a mason." Madeleine's response remains the same: "What kind of mason is he?" In (3*) and (4*), Madeleine's behavior would seem to indicate that she had made inferences in good logical standing, and Theo would not have thought anything amiss. Nevertheless, in (3*) Madeleine's response was open to potential correction in a way it was not in (4*). In (4*), as in (4), Madeleine would have responded in the same way whether Theo had said that John was a mason, that John was *not* a mason, or that Mason was a john. But even a broken clock tells the right time twice daily.[14]

Cases (4) and (4*) exemplify how a mental state fails to be form-sensitive if, in simple and unambiguous cases, it responds to states with *differing* logical form as if they were the *same*, e.g., responds in the same way to "John is a mason" and "John is not a mason." It is not enough that the state fortuitously happens to respond in the right way in certain circumscribed contexts. It must also be counterfactual-supporting: it *would* have responded in the right way to other states in different contexts. We have arrived at a first important condition for form-sensitivity:

> (DIFFERENT-DIFFERENT) (DD) A mental state is sensitive to logical form only if it responds to states with differing logical form in different ways.

---

participants in those studies are healthy, cognitively normal adults, but their behavior should seem just as bizarre as Madeleine's in (4)—so long as we foist BBC on the operative mental states.

[14] In (4) and (4*), Madeleine is a *little* better off than a broken clock, perhaps more like the frog who endlessly laps its tongue at things that look like flies and never learns any better. See Fodor (1990), Gendler (2008b), and McDowell (1998) for philosophical discussion of the incorrigibility of frog perception.

(DD) is a weak condition. Sensitivity to logical form requires that a state be sensitive to the content of the states with which it interacts, but (DD) does not specify how that sensitivity should be manifest on particular occasions. There may not be any uniquely best way for one mental state to respond to another in a given case. However, to respond in the same, or very similar, ways to states with blatantly diverging contents is decidedly *wrong*. It is to fail (DD). Of course, (DD) only holds *ceteris paribus*: the interaction between psychological states must take place in a mind unclouded by fatigue, drugs, or brain lesions; the concepts involved cannot be unfamiliar or difficult; the thoughts cannot be especially complex, etc.

A second important condition for form-sensitivity is brought out by a different sort of example. Imagine that Madeleine is having a conversation with her granddaughter and detects a hint of sarcasm. Madeleine exclaims, "A comedian, my granddaughter!" She could just as well have said, "My granddaughter is a comedian!" These utterances differ in a trivial grammatical way but express much the same thought. They share logical form. Similarly, her granddaughter will demonstrate sufficient understanding whether she replies by saying, "I'm not kidding you," or "Granny, I kid you not." Genuine form-sensitivity requires *ignoring* such grammatical superficialities and differences of word order. The mental states of an agent who putatively understood both expressions but responded to them as if they differed radically in cognitive significance would fail to be form-sensitive.

A state fails to be form-sensitive if it responds to states with the *same* logical form as if they were *different*. It thereby fails a second condition for form-sensitivity:

> (SAME-SIMILAR) (SS) A mental state is sensitive to logical form only if it responds to states with the same logical form in similar ways.[15]

[15] Satisfaction of this condition might require that the agent be equally familiar with the two distinct formulations, but it is not clear how much prior familiarity is necessary. A lot of very bad, all-too-easily intelligible poetry

In particular, I mean to rule out cases like those above, in which the ordering of words, concepts, or phrases can be rearranged without affecting the content of the state.

Form-sensitivity demands that a mental state responds to the content itself of the mental states with which it interacts, and not to grammatically superficial properties.[16] Whether a type of psychological state meets these conditions is testable. A state fails (DD) if it responds to states of *differing* logical form as if they were the *same* (e.g., responding in the same way to "John is a mason" and "John is not a mason"). A state fails (SS) if it responds to states with the *same* logical form as if they were *different* (e.g., responding in different ways to "I'm not kidding you" and "I kid you not").

In the next two sections I describe research in which implicit attitudes evidently fail to meet these conditions. In one context (§4), implicit attitudes seem flagrantly insensitive to differences in logical form. In the other (§5), they seem *overly* sensitive to trivial differences that are plainly irrelevant to logical form. In both, however, there are outstanding gaps and underexplored conditions that leave matters inconclusive. My broader aim, therefore, is to show how attending to form-sensitivity brings these gaps into sharpest relief, and throughout I gesture toward further studies that might speak more precisely to the underlying nature of implicit attitudes. Indeed, while there are admittedly a few holes in the studies that seem to speak *against* BBC, there are gaping chasms in the studies cited to speak *for* it (§6). Much more research

---

rearranges words in this sort of way. Garbled as his syntax may be, Master Yoda's sage advice is often *too* easy to understand ("Strong is Vader. Mind what you have learned. Save you it can!").

[16] A referee suggested that this demand is too strong, because beliefs are also susceptible to being influenced by such nonrational factors. I believe that most such cases will fall into either case (2) or case (3) above, but I say more about such cases in §8. If, ultimately, (SS) is less diagnostic for form-sensitivity because paradigmatically form-sensitive beliefs fail to meet it, then the empirical evidence and research proposals described in §5, on testing (SS), may be less decisive than the evidence and proposals described in §4, on testing (DD).

remains to be done on these questions, and my hope is that (DD) and (SS) will be useful for pursuing them. The literature is, of course, vast and what follows will perforce be selective.

4. Treating Different as Same

According to (DD), a state fails to be form-sensitive if it responds to states with differing logical form in similar ways. Here I summarize a few studies suggesting that implicit attitudes fail this test. First, like Madeleine's responses in cases (4) and (4*), implicit attitudes seem to be insufficiently sensitive to the logical operator of negation. In one study, participants either repeatedly "negated" or "affirmed" stereotypical associations (Gawronski and colleagues 2008). They saw images of racially typical white or black faces paired with potentially stereotypical traits. In one condition ("negation training"), participants pressed a button labeled "NO" whenever they saw a stereotypical pairing, e.g., a black face paired with the word "athletic." In the other condition ("affirmation training"), participants pressed a button labeled "YES" whenever they saw a counter-stereotypical pairing. Researchers found that, in isolation, affirmation training reduced racial bias on indirect measures, whereas negation training *enhanced* racial bias. Their implicit attitudes responded in the same way regardless whether they intended to *reject* or *affirm* the face-word pairings they perceived.

If implicit attitudes were belief-like, one would predict that they would respond differently to such dramatic differences in the content of the participants' intentions. Indeed, this study was a follow-up to an earlier one in which researchers made precisely this prediction (Kawakami and colleagues 2000). In the original study, participants performed both tasks: one group repeatedly negated stereotypes and affirmed counterstereotypes, while another group

repeatedly affirmed stereotypes and negated counterstereotypes. The training was sandwiched between a pretest and a posttest of implicit racial bias. Participants who negated stereotypes (and affirmed counterstereotypes) went from being significantly biased to completely unbiased on this measure. Those who affirmed stereotypes or underwent no training at all, however, continued to show significant racial bias at posttest. "In short," wrote Kawakami and colleagues (2000, 884), "practice does make perfect—or at least very good—stereotype negators." If it were true that, as they suggest in their paper's title, one could "Just Say No (to Stereotyping)," this finding would suggest that implicit attitudes possess a *minimal* sensitivity to logical form. It would indicate that they are belief-like.

There were, however, four distinct tasks confounded in the original study: affirming stereotypes, affirming counterstereotypes, negating stereotypes, and negating counterstereotypes. A better measure of form-sensitivity would test each separately (and mix and match conditions, e.g., by affirming both stereotypes and counterstereotypes). If implicit attitudes are form-sensitive, the core predictions of (DD) are that the effects of affirming versus negating stereotypes, and of affirming versus negating counterstereotypes, should differ markedly. If implicit attitudes are form-sensitive, one might also predict that affirming counterstereotypes and negating stereotypes should have similar effects (namely, as predicted by Kawakami et al. 2000, they should each reduce bias), although comparing these two conditions is less diagnostic, because the two cognitive exercises are not obviously on a par (the former asserts a less familiar correlation, e.g., that whites are lazy, while the latter denies the existence of a familiar correlation, e.g., it is not the case that whites are industrious[17]). Further complementary predictions might be that affirming stereotypes and negating counterstereotypes should each

[17] Thanks to anonymous referee for emphasizing this point.

enhance bias, but these further predictions must be tempered by the fact that most adults are racially biased already. Ceiling effects may often prevent them from becoming significantly more so. These sorts of ceiling effects can be avoided by studying attitudes toward novel stimuli, such as I will describe shortly.

Gawronski and colleagues' follow-up study suggests, however, that while affirming counterstereotypes does reduce bias, negating stereotypes per se does not, and in fact has the opposite effect. Regrettably, they did not test the isolated effects of affirming stereotypes or negating counterstereotypes, which prevents direct comparisons between affirming versus negating the same stimuli, and so prevents a direct test of (DD). This represents a gap in the empirical literature that could be extremely informative to fill. There is, of course, independent reason to suspect that affirming stereotypes will, barring ceiling effects, enhance bias in just the way that negating stereotypes seems to do. (If affirming stereotypes instead turns out to reduce bias, while negating stereotypes enhances bias, this would technically pass (DD), but it would be no victory for BBC—nor would it be easy to interpret on alternative theories.) The key condition that remains to be tested, then, is repeatedly negating counterstereotypes. If affirming and negating counterstereotypes both tend to reduce bias, while affirming and negating stereotypes both tend to enhance it, then implicit attitudes would fail to respect the dramatic difference in logical form between whether something is being asserted or denied, and thus fail (DD). In the meantime, the ironic effects of stereotype negation are certainly not congenial to BBC, especially since the original finding was advertised precisely as a demonstration of the efficacy of stereotype negation rather than of counterstereotype affirmation.

Of course, in all of these conditions, participants engage in quite a bit of high-level cognitive activity. They have to identify social group membership, recognize a stereotype or

counterstereotype, and act on that basis. The presence of all this cognitive activity might lead one to think that there are relevant processes of belief formation and revision afoot. Indeed, participants are well aware that they are negating stereotypes, and that the researchers are trying to influence their racial attitudes (Kawakami and colleagues 2007a). Yet while this cognitive activity does seem to influence participants' beliefs in certain ways, the full content of all this activity seems peripheral to the effects on their implicit attitudes. "That's a stereotype: negate it!" seems to have much the same effect as "That's not a stereotype: affirm it!" Both cognitive exercises reinforce participants' tendency to associate whichever face-word pairing they are perceiving.

The best explanation for this finding does not make reference to belief revision but to an entirely different psychological mechanism. The perceived spatiotemporal contiguity of the words and faces seems to drive the effect, independently of the logical form of participants' thoughts and beliefs *about* those faces and words. Specifically, the effect is likely driven by increased attention to one rather than another type of (spatiotemporally contiguous) face-word pairing, since both groups of participants saw the same set of faces and words (Gawronski and colleagues 2008, 375).

These studies on effective techniques for reducing prejudice are practically important, but, because of ceiling effects and the socially sensitive material, their implications for getting at the underlying mental states and processes are not always clear. Defenders of BBC may find many nits to pick in research on counterstereotype training. Many of these complications are absent, however, in two studies by Moran and Bar-Anan (2013). Participants learned about four types of alien creature, each characterized by a distinctive color and shape. One type of alien always appeared on the screen just before the onset of an unpleasant sound ("a horrifying human

scream") while a second type appeared just before that unpleasant sound stopped. A third type appeared before the onset of a pleasant sound ("a relaxing musical melody") while the fourth appeared before the pleasant sound stopped. Explicitly, participants reported a preference for the aliens who "started" the pleasant sound over those who "ended" the pleasant sound and over those who "started" the unpleasant sound. They also preferred those who ended the unpleasant sound over those who started the unpleasant sound and over those who ended the pleasant sound. That is, like maximizers of self-interest, they learned to like those who increased pleasure or relieved suffering more than those who reduced pleasure or increased suffering.

Their implicit attitudes, however, told a different story. Implicitly, participants preferred *both* types of aliens who appeared on the screen with the pleasant sound over *both* aliens who appeared with the unpleasant sound, regardless of who had started or ended those sounds. Unlike their explicit attitudes, their implicit attitudes failed to respect the dramatic difference between (signals for) the initiation and termination of valenced stimuli. In lieu of affirmations and negations, we have starting and stopping, and in lieu of stereotypes and counterstereotypes, we have pleasant and unpleasant sounds. The result is structurally the same. In the first case, evidence suggests that attending to racial stereotypes leads to less favorable implicit attitudes toward blacks, regardless whether participants intend to affirm or reject these stereotypes. In the second case, attending to images that appear with unpleasant sounds leads to less favorable implicit attitudes toward those images, regardless whether participants judge that the images are responsible for (or signal) the starting or stopping of those unpleasant sounds. Once again, the mere spatiotemporal contiguity of the images and sounds drives the effect on implicit attitudes, in apparent independence of the logical form of participants' thoughts and beliefs about the

relations among those stimuli.[18] Similarly, as I explained in §1, Rydell and colleagues (2006) found that self-reported attitudes toward a person named Bob tracked verbal descriptions of him while implicit attitudes tracked the valence of contiguous subliminal primes. I will not rehearse every study demonstrating dissociations between form-sensitive beliefs and contiguity-sensitive implicit attitudes here (for more extensive reviews, see Gawronski and Bodenhausen 2006, 2011). Moran and Bar-Anan's studies are notable for eliciting this dissociation without using any trickery or subliminal priming, and even without using any overtly linguistic stimuli in the learning procedure.

One fallback for defenders of BBC in the face of findings like Moran and Bar-Anan's is to propose that participants form the belief, say, that the green aliens co-occur with the unpleasant sound, which leads participants to judge that the green aliens are unpleasant and to dislike them (see e.g., de Houwer 2011, 411). If such gestures do not constitute washed-out, unfalsifiable trivializations of BBC, nothing does.[19] These interpretations are also difficult to make sense of. Suppose that, in isolation, beliefs about co-occurrence dispose individuals to judge that the green aliens are unpleasant, while, also in isolation, beliefs about *ending* the unpleasant sound dispose individuals to judge that the green aliens are not unpleasant. Why, when the two beliefs are combined, don't they cancel each other out and fail to create a significant implicit attitude at all? How is it even possible to acquire such diametrically opposed beliefs with respect to the very same stimuli? Above all, why wouldn't the green aliens' association with the reduction of suffering make them more automatically or implicitly likeable

[18] I assume we can be relatively agnostic about the precise content of participants' beliefs about these environmental contingencies (e.g., do they think the aliens "cause" or merely "signal" the stimulus changes?). Any reasonable construal of their beliefs would be hard-pressed to explain the dissociation between implicit and self-reported responses.

[19] But see, e.g., Zanon and colleagues (2014, Experiment 2) for putative evidence that co-occurrence is cognized by participants in this belief-like way.

than the blue aliens who reduce pleasure? If we stop trying to force implicit attitudes into a belief-shaped box, these puzzles vanish. There are stronger arguments and evidence in favor of the claim that implicit attitudes are sensitive to logical form, to which I turn in §6.

Although the literature speaking to (DD) is rapidly expanding, there are many potentially relevant contrasts that have not yet been studied, such as contrasting disjunction and conjunction; conditional and biconditional; possibility, actuality, and necessity; existential, universal, and generic quantifiers; past, present, and future tenses; obligation and permission; and propositional attitudes such as believing, knowing, pretending, and imagining. If implicit attitudes are primarily sensitive to spatiotemporal contiguity, then they should treat, e.g., conjunctions and disjunctions (whether inclusive or exclusive) as more or less on a par. "Either Bob is a mailman or a murderer" should lead to similarly negative implicit attitudes as does "Bob is a mailman and a murderer," perhaps even if participants subsequently rule out his being a murderer or rule in his being a mailman. "Bob is required to steal" should generate similar responses to "Bob is permitted to steal," and perhaps even "Bob is pretending to steal," and so on. This uncharted terrain could prove fertile.

5. Treating Same as Different

The evidence that implicit attitudes treat different as same is substantial, albeit still gappy in important ways. A comparatively underexplored possibility is whether implicit attitudes also treat same as different. Do they respond to states with the same logical form in different ways (e.g., responding differently to "I'm not kidding you" and "I kid you not")? Some suggestive findings emerge from research on a type of implicit attitude known as "shooter bias." Research

began in response to a series of tragic cases in which police shot unarmed black men. Among the many causes behind such tragedies, one is thought to be an implicit attitude associating black men with weapons.[20] In one laboratory measure, participants are instructed to press a button labeled "shoot" when they see an image of a person holding a gun, and to press "don't shoot" when they see a person holding a cell phone. Many participants, including African Americans, are faster and more likely to "shoot" unarmed blacks than unarmed whites, and faster and more likely to "not shoot" armed whites than armed blacks.

It initially seemed that trying to control shooter bias only made it worse. In particular, when participants consciously intend to "avoid race bias," their bias *increases*. However, one peculiar class of intentions, called "implementation" or "if-then" intentions, seems to effectively curb the expression of shooter bias. If-then intentions specify a concrete cue or situation in which the agent will perform an action, such as, "the next time I see Bob, I shall tell him how much I like him." Other examples are, "If I feel a craving for cigarettes, then I will chew gum," or "When I leave work, I will go to the gym." These contrast with "simple" intentions, which do not refer to any specific cue, such as, "I'll tell Bob how much I like him," "I'm planning to cut back on smoking," or "My New Year's resolution is to work out more." Research suggests that concrete intentions specifying when, where, or how an action will be performed are far more successful and efficient means for making good on our plans than just having abstract goals to perform some action some time (Gollwitzer and Sheeran 2006). The idea that making our intentions concrete will help us reach our goals sounds intuitive enough, but the documented effects of implementation intentions on shooter bias are striking.

[20] See, e.g., Glaser and Knowles' (2008) finding that shooter bias was predicted by an automatic association of black men and weapons.

In one study, participants were given additional instructions to help curb their shooter bias:

> You should be careful not to let other features of the targets affect the way you respond. In order to help you achieve this, research has shown it to be helpful for you to adopt the following strategy… (Mendoza and colleagues 2010, 515)

Some participants were instructed to rehearse a simple intention:

> (SI) I will always shoot a person I see with a gun.

Others rehearsed an if-then intention:

> (IF) If I see a person with a gun, then I will shoot.

Although the two intentions were, as the researchers noted, "semantically parallel," the results were strikingly different (518). Participants who rehearsed the simple intention (SI) performed no better than participants with no plan at all, while participants who rehearsed the if-then intentions (IF) were significantly more accurate. As the researchers say, "The observed results are striking, given that the basic instructions for completing the task were essentially the same for each condition" (519). Somehow the sheer phrasing or word order of one's plan can make the difference between going on to act in egalitarian or prejudiced ways.

The best explanation for the difference might be the order in which the words "gun" and "shoot" are thought. The shooter task involves (roughly) two steps: to perceptually identify a stimulus and to press a button, in that order. The temporal order of these steps corresponds to the temporal order in which participants in the (IF)-condition *think* about those steps. (IF) causes the participants to form an automatic association between the cue (guns) and the behavior (shooting),

in that order, while (SI) does not. The temporal order of words as they figure in the participants' cognition of the intention plausibly explains their differential effects.[21]

This difference in word order has nothing to do with the logical form of the rehearsed intentions. For these agents in this context, (SI) and (IF) likely share the same content and logical form. Although we can *call* (SI) a "simple" intention, it specifies precisely the same cue for action as (IF). Both (SI) and (IF) express the participants' intention *to shoot in the condition when they see a person with a gun*. Both only fail to be fulfilled when participants see a person with a gun, but do not shoot. Hence when participants in both groups come to believe that they will fulfill their intentions, their beliefs share truth conditions. Moreover, both intentions play the same inferential roles in practical syllogisms. Employing one rather than another intention in otherwise identical bits of practical reasoning, would, other things equal, make no difference to an agent's deliberation. Given the shared features of these intentions, it is plausible that they also share logical form.

Admittedly, (SI) and (IF) are not perfect mirrors of each other, which might suggest that they differ somewhat in logical form. (SI) could be more "off-putting" because it says to "shoot *a person*" whereas (IF) just says to "shoot" (although in both cases, the actually intended action is the same: to press a button labeled "shoot"). (SI) does not explicitly contain the conditional "if" (although in both cases, the intended context for action is the same: every time an image of a person with a gun is seen during the task). (SI) and (IF) contain potentially different temporal operators, "I will always" versus "I will" (although the global operator "always will" should if anything be stronger than the merely futural operator "will," whereas the opposite was

[21] One wonders how important the actual grammar of the rehearsed intention is. Perhaps rehearsing even more spare thoughts such as "if gun, then shoot" or "see gun, press shoot" or even just "gun—shoot" might be effective. The fewer the words, the lesser the tax on working memory (Baddeley 2007).

observed). (SI) might even be ambiguous between two readings: "I will always shoot [a person with a gun]" or "I will always shoot [a person] with a gun"—as opposed to shooting the person with a bow and arrow.[22] Of course, all of these complications could be avoided in a follow-up study that employed a better semantic mirror of (IF), e.g., "I will shoot, if I see a person with a gun." But even if the underlying logical form of (SI) and (IF) did differ in some subtle respect, it is mysterious *how* such a difference could be relevant to the task. The point is not that there is no reasonable way of prying apart their logical forms, but that none of these differences plausibly explains why one intention was effective and the other was not. A state, like shooter bias, that treats such clearly similar intentions as if they were utterly dissimilar fails to be form-sensitive.

There are, however, alternative explanations for the differential effects of (IF) and (SI) worth exploring. A particularly germane alternative, which might be more hospitable to a form-sensitive construal of the operative mental states, is that (SI) fails to influence performance because it is a more difficult construction for participants to parse. The role of parsing difficulty could be investigated by testing whether if-then intentions with awkward constructions are still effective, along the lines of: "If a person with a gun I see, then shoot will I!" or "If I see a gun with a person, then will I shoot!" If participants who rehearse these awkward constructions still outperform those with simple intentions (as I predict they would), that would suggest that the temporal structure of the if-then formulation really is driving the effect. If, however, these awkwardly constructed intentions do not influence performance (or harm it), then the upshot might simply be that (SI) is too complex for participants to think through in the moment. In that case, this study would not furnish evidence that shooter bias fails to be form-sensitive, but rather that participants who are cognitively taxed and time-constrained cannot properly think through

[22] Thanks to (name removed) for this suggestion.

the relevant inferences. This would then amount to a kind of exceptional case, like Madeleine's performance error in (3).

Ultimately, whether implicit attitudes are form-sensitive is an empirical question. And whether implicit attitudes do in fact meet this criterion, they *must* meet it in order to qualify as beliefs. I have discussed a small sample of research that seems to speak against it, but more research is sorely needed. Whether implicit attitudes treat different as same has received considerable attention, but the further question whether they treat same as different is relatively unexplored. Mendoza and colleagues' single study may not tell us much on its own, but it points toward further research that could. By holding fixed (as much as possible) the logical form of different plans and descriptions, we can pinpoint more precisely which features of participants' external (and internal) environments exert an influence on implicit attitudes. For example, due to halo effects (Asch 1946), the valence of temporally prior words in a sentence might exert a stronger influence on implicit attitudes than do later words. Reading "When Bob is relaxing with friends, Bob curses, yells, and tells vulgar jokes" might lead to more positive implicit attitudes toward Bob than does "Bob curses, yells, and tells vulgar jokes when Bob is relaxing with friends" or "When Bob curses, yells, and tells vulgar jokes, Bob is relaxing with friends." Perhaps the differential effects of constructions in the active versus passive voice (e.g., Henley and colleagues 1995) depend in part on the sheer ordering of the words, in addition to more sophisticated considerations regarding tacit implications of agency and blame. Reading a series of active constructions like "Bob violently destroyed the beautiful jewel" might lead to more negative implicit attitudes toward Bob than passive constructions like "The beautiful jewel was violently destroyed by Bob."

6. Evidence to the Contrary?

Several studies purport to show that implicit attitudes are sensitive to logical form in certain contexts, leading some to insist that they are full-blooded beliefs (e.g., de Houwer, 2014; Mandelbaum 2013, forthcoming) and others to suggest that they differ from beliefs only as a matter of degree (e.g., Schwitzgebel 2010 and arguably Levy 2014a,b). Even BBC's leading opponents have published findings that seem to support it (e.g., Peters and Gawronski 2011). In this section, I briefly review these findings, and argue that they actually provide very little support for BBC. Where the empirical case against BBC has gaps, the case for it has chasms.

Before discussing these findings, a number of caveats are in order. First, measures of implicit attitudes are not "process-pure." They reflect a mix of automatic and effortful processes. Cognitively depleted individuals tend to exhibit greater biases than alert individuals (e.g., Govorun and Payne 2006). As a result, a change in performance on these measures might reflect a change in implicit attitudes, or a change in behavioral control, or both. Several models have emerged to disentangle these possibilities, and it has become commonplace to use them to analyze experimental data. Research suggests, for example, that counterstereotype training (Calanchini and colleagues 2013) and implementation intentions (Mendoza and colleagues 2010) lead *both* to reductions in implicit bias *and* to increases in the capacity to control the expression of bias. Applying these process-dissociation models to the data strengthens our confidence that the interventions in question lead to genuine changes in implicit attitudes (and, evidently, to changes in the capacity to control them).

Second, measures of implicit attitudes are (like measures of blood pressure) susceptible to a wide variety of contextual and motivational factors. Implicit racial and ethnic biases

increase after taking oxytocin (De Dreu and colleagues 2011) and decrease after taking beta blockers (Terbeck and colleagues 2012). They decrease in the mere presence of a black experimenter (Lowery and colleagues 2001). Smokers who are deprived of nicotine exhibit positive implicit attitudes toward smoking, while smokers who have just finished a cigarette exhibit negative implicit attitudes—slightly more negative, in fact, than the attitudes of nonsmokers (Sherman and colleagues 2003). How exactly to interpret these temporary, context-specific effects is beyond the scope of this paper, but clearly they do not portend genuine implicit attitude change.

To assess whether a manipulation leads to genuine attitude change rather than some fleeting context effect, experimenters can delay the posttest, change the context, and so on. To this point, tests of long-term, context-general changes in implicit attitudes remain scant. While several exploratory studies suggest that long-term implicit attitude change is possible (e.g., Devine and colleagues 2012), the conditions are typically not sufficiently controlled to pin down precise causes. Possible exceptions include Reinout Wiers and colleagues' (2011) findings that retraining implicit attitudes can bolster clinical treatment for addiction. Patients recovering from alcoholism repeatedly avoided images of alcohol (in 4 sessions lasting 15 minutes each) prior to undergoing 3 months of standard therapeutic treatment. These participants were significantly less likely to relapse *one year* after being discharged from therapy, in comparison to those who underwent standard therapy with a sham training procedure or with no training at all. Carolin Eberl and colleagues (2013) replicated these effects, finding that alcohol-avoidance training generated negative implicit attitudes toward alcohol, and that this change in implicit attitudes mediated the improvement in long-term recovery. By contrast, participants who underwent 3 months of therapy without alcohol-avoidance training showed no changes in implicit attitudes

(and, again, were more likely to relapse). Of course, these studies on addiction recovery do not permit inferences about whether implicit attitudes are or are not sensitive to logical form (although the failure of standard therapy to make even a *dent* in implicit attitudes is striking). They demonstrate that implicit attitude change can be relatively durable, context-general, and relevant to "real-world" behavior. Less far-ranging studies find that, in isolation, the effects of counterstereotype training last at least 24-30 hours, on a variety of different measures (Forbes and Schmader 2010). If anything, the effects seem to grow in strength over that span (Kawakami and colleagues 2000) and after intervening tasks (Kawakami and colleagues 2007a). The effects of implementation intentions on measures of implicit attitudes can last at least 3 weeks (Webb and colleagues 2012) and can have broader behavioral effects lasting months (Chapman and Armitage 2010). There are, in other words, broad patterns of evidence that clearly suggest that these interventions are more than momentary flukes.

By contrast, as far as I can tell, the studies cited by defenders of BBC to show that implicit attitudes are sensitive to logical form have not tested the effects after even a brief delay.[23] Nor have these studies employed process-dissociation models to assess whether the effects reflect genuine attitude change rather than a temporary boost in motivation or control. I am skeptical, therefore, that the brunt of these studies reflects anything but context effects (for further discussion, see, e.g., Han and colleagues 2010). Leading defenders of BBC, such as Jan de Houwer, seem to have recently acknowledged these concerns. Smith and de Houwer (in press) found that that a persuasive message influenced implicit attitudes according to one measure, which was administered immediately after participants read the message, but not a second measure, which was administered immediately after the first measure. Variability across

[23] See Levy (2014a,b), Mandelbaum (2013, forthcoming), and de Houwer (2011, 2014) for surveys of relevant studies. I discuss specific examples in the remainder of this section.

different measures of attitudes is common, but Smith and de Houwer also consider that, by the time of the second measure, "the effects of the persuasive message might have dissipated." The effects of this persuasive message might, that is, be especially fragile or short-lived. Smith and de Houwer go on to note that future studies should use process-dissociation models to better analyze the effects of persuasive messages on measures of implicit attitudes. Given that these models have not yet been applied to these manipulations, and against a background of widespread evidence that these measures are susceptible to context effects, it is difficult to see why we should interpret the existing research on persuasive messages as anything *but* more evidence for context effects. It is difficult to see how these studies provide any *specific* or *distinctive* support for BBC at all. In other words, while the evidence against the form-sensitivity of implicit attitudes is admittedly gappy, clear evidence for it is almost nonexistent.

The mere fact that a temporarily effective intervention involves the conveying of logically structured information does little to suggest that the effect occurs *by virtue* of sensitivity to logical form. Suppose I persuade you to stand up quickly in order to make a measure of your blood pressure come out low. If you follow my advice, and your blood pressure indeed drops, must we conclude that blood pressure is sensitive to persuasive argument? Consider Maria's advocacy for the power of positive thinking to the von Trapp children: "When the dog bites, when the bee stings, when I am feeling sad, I simply remember my favorite things, and then I don't feel so bad." Suppose that the children find Maria's argument for this technique persuasive, and try it one time to relieve their fear during a thunderstorm. Perhaps they would, if tested immediately thereafter, exhibit significantly less negative implicit attitudes toward lighting and thunder. Obviously, this would not, just as such, be evidence that implicit attitudes about thunder are sensitive to logical form. First, it is implausible that simply thinking of pleasant

things during one frightening storm would genuinely change one's storm-related attitudes. More likely, the effects of the attitudes are momentarily muted. (Durable attitude change might come with practice.) Second, it is implausible to attribute this effect *directly* to the persuasiveness or logical form of Maria's argument. Her argument need not even pertain to whether thunderstorms warrant fear, rather than to whether a certain cognitive strategy can alleviate fear. Plausibly, the strength of her argument leads the children to *try* the strategy, the strategy leads them to think positively valenced thoughts, and the thoughts reduce their fear.

This type of case exemplifies one of many ways in which beliefs and implicit attitudes might interact, and specifically, a way in which logical form might indirectly influence implicit attitudes via directly influencing explicit beliefs. It represents one sort of general strategy for explaining (away) studies that ostensibly demonstrate the form-sensitivity of implicit attitudes.[24] While the durability and context-generality of these studies remain untested, I believe that some of them stand a fighting chance of being more than fluky context effects. Nevertheless, these studies fall far short of suggesting that implicit attitudes are sensitive to logical form per se, rather than to "downstream" effects of logically structured information.

In one such study (Rydell and colleagues 2007), participants first read varying numbers of positively valenced descriptions about a person named Bob, such as, "Bob fought against a discriminatory law that made renting difficult for minorities." On explicit measures, they reported that Bob was agreeable after reading only 20 positive descriptions. Their implicit attitudes also became favorable, but only after reading about 100 positive descriptions. Subsequently, the participants were exposed to negative descriptions, such as, "Bob continually yells at his wife in public," interspersed with neutral descriptions, such as, "Bob bought a shelf."

[24] For additional strategies for reinterpreting these studies, see Gawronski and Bodenhausen (2011). They say comparatively little about the strategy I highlight here.

Just 20 negative descriptions were sufficient for participants to withdraw their prior report that Bob was agreeable. By contrast, participants still held positive *implicit* attitudes toward Bob after having read 20 and even 40 negative descriptions. Gradually, however, as participants learned more and more negative information, their implicit attitudes toward Bob grew sourer and sourer. Given the slow linearity of implicit attitude change in this study, the manipulation strikes me as standing a reasonable chance of being durable. Moreover, the findings suggest that implicit attitudes can change in response to logically structured information (e.g., Mandelbaum forthcoming, n37). In this case, they just respond more slowly than do self-reported attitudes. Findings like this might seem to support the view that the difference between implicit attitudes and beliefs is a matter of degree. Perhaps implicit attitudes are simply less sensitive to logical form than are self-reported beliefs. However, this study is perfectly consistent with a different interpretation, according to which implicit attitudes per se are completely and categorically insensitive to logical form, and merely sensitive to experienced relations of spatiotemporal contiguity. On this alternative view, roughly, the effect of logical form on implicit attitudes is mediated by the effect of belief on affect.

A powerful reason to think the effect is mediated is that it can be interrupted. As I described in §1, the effects of reading valenced descriptions on implicit attitudes can be entirely overridden by subliminal priming. When positive descriptions are paired with subliminal negative words, self-reported beliefs become positive and implicit attitudes become negative (Rydell and colleagues 2006). Subliminal priming intervenes somehow to prevent or dampen the "normal" effect of valenced descriptions on implicit attitudes. What is the intermediate step? Plausibly, in the case of subliminal priming, the unconscious perception of valenced words activates subtle affective responses. Every time participants see Bob's face, they experience a

certain low-level feeling. Eventually, the mere sight of Bob activates the feeling. In this case, implicit attitudes respond to the experienced spatiotemporal contiguity of Bob's face, valenced words, and affective responses.

This sketch of subliminal conditioning suggests a natural interpretation of the intermediate step between reading logically structured sentences and changes in implicit attitudes. Reading negative descriptions of Bob leads participants to judge that he is disagreeable. These anti-Bob judgments, in turn, activate negative affective responses. Personally, when I read Rydell and colleagues' (2007) example of a negative description—"Bob continually yells at his wife in public"—I feel a palpable sense of discomfort, a visceral negative reaction. Reading such sentences may play the same role, then, that the mere perception of negative words does in the context of subliminal conditioning. After repeatedly experiencing negative affect while seeing Bob's face, eventually just seeing him activates those negative feelings. In such a case, the effect of logical form on implicit attitudes is indirect: the incoming information influences participants' beliefs, which leads them to rehearse negatively valenced thoughts, which activates negative feelings. In short, the judgments cause feelings and over time the feelings cause changes in attitude.

This sketch is, of course, speculative. It is not ad hoc. In broad strokes, it is sufficiently commonsensical that young children grasp the rationale behind "Raindrops on roses and whiskers on kittens." If the relations sketched here between logically structured information, affect, and attitude change nevertheless continue to seem ad hoc or mysterious, it bears mentioning that BBC is more or less committed to them as well. BBC posits causal relations between beliefs and evaluative dispositions (e.g., "Bob is a jerk. Therefore, I should dislike Bob. Therefore…" … negative affective dispositions toward Bob), while offering no illuminating

explanation for why these relations obtain. As Walther and colleagues (2011, 193) succinctly put it, "it is not clear how propositional knowledge is translated into liking." This "translation" is simply stipulated. It is no *less* miraculous or mysterious for BBC than for rival theories.

In fact, this sort of mediated account of how logical form influences implicit attitudes is continuous with how leading researchers interpret (their own) evidence on the effects of rationally persuasive arguments. For example, Briñol, Petty, and McCaslin (2008) describe a study in which participants' implicit racial biases were measured after they read either a strong or a weak argument in favor of hiring more African-American professors at their university.[25] Among participants who were motivated to think extensively about the arguments, those who read strong arguments showed less bias than those who read weak arguments. Is this powerful evidence that implicit attitudes are sensitive to logical form? The researchers don't quite interpret it that way. Instead, they argue that the effect of argument quality on implicit attitudes is ultimately a function of the sheer quantity of positively versus negatively valenced thoughts that participants are induced to think by the arguments:

> the strong message led to many favorable thoughts… the generation of each positive (negative) thought provides people with the opportunity to rehearse a favorable (unfavorable) evaluation of blacks, and it is the rehearsal of the evaluation allowed by the thoughts (not the thoughts directly) that are responsible for the effects on the implicit measure. (2009, 295)

Support for this interpretation came from a subsequent study in which participants were asked to list all their thoughts about the arguments. The effects of argument quality on implicit attitudes were indeed mediated by the net valence of reported thoughts. Maria von Trapp seems vindicated. It is striking that many of the very researchers involved in these studies—in, as it

[25] Apparently, the original study described here on implicit racial bias has not been published in a peer-reviewed journal, which makes it an odd point of emphasis for defenders of BBC (Mandelbaum forthcoming). Similar findings have, however, been published on attitudes toward vegetables and commercial brands (Horcajo, Briñol, and Petty 2010).

were, the belly of the beast of BBC—do not conclude that implicit attitudes are sensitive to logical form. Instead, these studies are taken to show how implicit attitudes can walk and talk like beliefs within a narrow range of contexts, while the underlying states and mechanisms aren't belief-like at all.

This indirect account of how logical form can influence implicit attitudes could be disconfirmed in numerous ways. If affect is a primary mediator, then in any manipulation that dissociates the valence of the logically structured information from the affective experiences of those considering that information, implicit attitudes should track the latter rather than the former. Participants who are in a bad mood while they read positive information about Bob might fail to form pro-Bob implicit attitudes.[26] Participants who read narratives with surprise endings, where long-trusted allies are revealed as traitors and long-hated enemies as allies, might fail to reverse their implicit attitudes. Participants might form pro-Bob implicit attitudes simply by reading a litany of positively valenced but uninformative descriptions of him, such as "Bob loves the taste of delicious food; Bob really likes his friends; Bob enjoys fun hobbies; Bob follows the advice of people he trusts," and even questionable descriptions, such as, "Bob loves to befriend wealthy people; Bob follows the advice of beautiful celebrities; Bob loans money to royal princes who email him." These non-substantive, positive descriptions might influence implicit attitudes even if participants already have decisive reason to dislike him (e.g., because he is an unrepentant serial murderer). Participants who read a litany of positively valenced descriptions should also come to implicitly like any sights, smells, or sounds that are spatiotemporally contiguous with the descriptions (e.g., mock advertisements that appear on the screen simultaneously).

[26] The literature on mood, information-processing, and attitude change is vast (Vogel and colleagues 2014, 98-102, 147-9), but the effects on *implicit* attitudes remain underexplored.

7. Implications for Philosophy of Mind and Empirical Research

While much more research remains to be done, implicit attitudes seem to respond to mental states with differing logical form in similar ways, and states with the same logical form in different ways. Beliefs would be sensitive to these similarities and differences in the content of the mental states with which they interact. Instead, the effects may depend on relations of spatiotemporal contiguity in the agents' thoughts and perceptions, which is to say, the spatial and temporal order in which people see, think, and feel things. In this section, I draw out consequences of the foregoing considerations for further empirical research and the philosophy of mind more generally. In §8, I address outstanding objections.

Form-sensitivity is a significantly less demanding condition than evidence-sensitivity, inferential promiscuity, or the other sophisticated cognitive dispositions discussed by Gendler, Levy, and others. It is possible for an agent's mental states to be less than ideally responsive to the incoming evidence, while being robustly responsive to the logical form of other mental states. This is just the difference between Madeleine's responses to Theo in cases (1) and (2) above. In the first, Madeleine's attitudes about her friend Mason adjust appropriately. In the second, they fail to adjust at all. Her attitudes in case (2) can still qualify as beliefs because they are responsive to the logical form of the states with which they interact.

However, form-sensitivity is necessary but not sufficient for belief. Other mental states, like intentions, are plausibly also sensitive to logical form. My view is strictly neutral regarding further conditions for belief. Perhaps beliefs must be readily available for conscious self-ascription and sensitive to the normative standard of truth. Arguably, form-sensitivity is a

necessary condition for these more sophisticated properties. If so, it constitutes an important cognitive benchmark *between* extremely primitive and sophisticated intentional states.

Countenancing such a cognitive benchmark would fill a persisting gap in theories of belief and intentionality. Take Fred Dretske's view, for example. Dretske sets the bar for intentionality rather low, such that primitive forms of associative learning meet it, but grants that there might be higher-order conditions necessary for a state to be a full-fledged belief. In particular, a state might have to be assimilated into a larger nexus of beliefs and desires, that is, be inferentially promiscuous. Precisely *how much* assimilation is necessary is, he insists, just a "terminological boundary dispute of negligible philosophical interest" (1988, 107). Even if Dretske is right that there is no definitive point at which a state becomes sufficiently belief-like, we can nevertheless say *something* informative about the intermediate conditions between primitive and sophisticated forms of intentionality. Form-sensitivity is, plausibly, necessary for a state to assimilate with other states, regardless whether it does or not. A state has to be at least sensitive to logical form in order to interact with sufficiently many other states in the right way. For the purposes of understanding implicit attitudes, we can set aside debates about precisely what it takes to be a belief. Implicit attitudes seem not even to be in the ballpark of belief. The mind is likely populated with many other similarly primitive psychological kinds.

A related way in which form-sensitivity makes for an attractively minimal condition is how little it requires in terms of cognitive consistency. Full rationality plausibly requires cognitive consistency (or closure), such that, e.g., an agent who believes that *P* and that *if P, then Q* also believes that *Q*. An agent who has standing beliefs that *P* and that *if P, then Q*, but who

fails to believe that *Q* would be less than fully rational.[27] But of course ordinary human agents often fail to, as the saying goes, put 2 and 2 together. Cognitive work is often required to make our standing beliefs internally consistent, and form-sensitivity is well-suited for capturing this. A standing mental state could be perfectly form-sensitive without ever, as an empirical matter, having interacted with other psychological states, and thereby never taking part in processes of belief formation or revision. Such a state would be *ready* or *available* to interact with other form-sensitive states, even if it happened to remain in splendid isolation.[28] This creates a significant wedge between form-sensitivity and full rationality. It also invites us to investigate hitherto underexplored questions: what *prompts* states to interact? Under what conditions do mental states, which may or may not be form-sensitive, interact? One familiar condition is when an apparent inconsistency is explicitly pointed out to an agent, but there are many others. The question of the rationality of our beliefs and other attitudes should be broken up into *different* questions: Is a given type of state form-sensitive? When does it interact with other states? Which other states? Is the state malleable in any way? And so on.

Form-sensitivity is an empirically tractable notion. We can inquire into whether a type of state meets this condition, and I have sketched a variety of ways of doing so (§§4-6). Form-sensitivity also has the virtue of leaving a great deal of room between ideal rationality and the actual rationality of finite agents. Form-sensitivity can assist in the empirical investigation of belief without making unrealistic demands on human cognitive capacities. It does, however, place something like a lower limit on human rationality. It is difficult to construe the relevant

[27] The failure I have in mind occurs when the inconsistency goes unrecognized, e.g., because it has not occupied the agent's attention. Such failures differ from the occurrent endorsement of inconsistency in Moorean-paradoxical expressions (e.g., "Bob is agreeable, but I don't believe it"), and from the non-inferences in case (4) in §3.
[28] See (citation removed) for further discussion of cognitive accessibility.

cognitive processes in §§4-5 in inferential terms. Implicit attitudes seem to respond to other mental states in ways that would be not just irrational but *unintelligible* if they were beliefs.

8. Objections

One might object that form-sensitivity is still too strong. Form-sensitivity seems to demand that a mental state responds to the content, the whole content, and nothing but the content of other states. The requirement of (SS), to respond to states with the same logical form in similar ways, might seem particularly strong. Of course, psychological responses to states that share content but differ in some other way can themselves differ. An agent might find one turn of phrase more lyrical or memorable than another. Compare "The spoils go to the victor!" and "To the victor go the spoils!" The phrases share logical form, but only the latter is in trochaic tetrameter. However, if Madeleine tries to persuade Theo that the winner of the next poker hand should get the whole pot, it does not much matter whether she says one phrase or the other. Theo is apt to make similar inferences and prepare similar replies, regardless whether "the spoils" or "the victor" crosses his mind first. Even if, for example, hearing the phrase "Spoils to the victor!" puts Theo in a bad mood because he associates it with political cronyism, it is not as if the activation of this negative association *disables* his capacity to think through the content and respond in an intelligible way.

Beliefs are not magically exempt from these associative connections, but neither do these associative connections truly prevent beliefs from responding to the logical form of other states. Recall case (3), in which a momentary cognitive lapse leads Madeleine to ask Theo what sort of mason John is, but the error is quickly corrected. For my purposes, the source of the error does

not matter, but suppose that some sort of idiosyncratic association is responsible. Perhaps, while Theo was talking, Madeleine was occupied trying to remember the lyrics to "Unforgettable" as sung by Nat King Cole, who, she recently learned, is alleged to have been a Freemason (Kang and Young 2009), leading her to wonder whether any of her acquaintances might secretly be Freemasons, too. In her state of distraction, merely hearing the word "Mason" reminded her of all this, leading her to wonder whether John might be a member of that fraternity. However, once her mind stops wandering, she can think through these inferences in the right way, and the operative states are still form-sensitive in the relevant sense. They still respond to states with the same logical form in sufficiently similar ways.

Moreover, if Madeleine's beliefs about Mason happened *not* to stand in any such associative connections with other states, they would not thereby be disqualified from beliefhood. Standing in these associative relations does not make beliefs what they are. Whatever *does* make beliefs what they are, it also makes it the case that they are sensitive to logical form. Implicit attitudes, by contrast, do not respond to the logical form of an agent's thoughts and perceptions. They are irredeemably yoked to relations of spatiotemporally contiguity.

One might also worry that form-sensitivity is too strong because it is overly *linguistic*. Several of the studies discussed in §§4-6 refer to negations and grammatical features in English. Obviously, the cognitive states of non-human animals and infants will be largely insensitive to many of these linguistic features. One might think that sensitivity to logical form rules out that such states are beliefs, much like the kinds of sophisticated cognitive conditions discussed in §2 and §7.

While many of the examples and studies rely heavily on linguistic phenomena, my argument does not presuppose that logical form must be cashed out in terms of natural language. Research on implicit attitudes predominantly involves language-dependent manipulations, presumably because it is much more tractable, but it need not. Implicit attitudes can be changed merely by moving a joystick back and forth in response to images of faces (Kawakami and colleagues 2007b) or beverages (Wiers and colleagues 2011), in a task that any being capable of associative learning could approximate.

Theoretical discussions of non-human cognition commonly address whether such cognition is marked by analogues of form-sensitivity. Arguments that a bit of animal behavior should be explained in terms of belief and desire, or exemplifies rationality, often turn on whether there is counterfactual-supporting evidence that the animal is engaging in "proto-inferences" (Bermúdez 2006). Such capacities are not a far cry from the more language-based form-sensitivity on which I focused. Consider again Moran and Bar-Anan's (2013) studies on learning about creatures who "started" and "stopped" valenced sounds. The learning procedure in these studies included nothing but images and sounds; no explicit language was involved. It would, presumably, be highly adaptive for human and non-human animals to discriminate among stimuli that signaled the imminent increase versus decrease in pleasure versus suffering, and, in particular, to automatically prefer those stimuli that predicted (or caused) the reduction of suffering over those that predicted the reduction of pleasure or the onset of suffering. These are proto-inferences worth making. Yet while adults' self-reported preferences tracked these environmental contingencies adaptively, their implicit attitudes failed spectacularly. Their immediate, intuitive dispositions reflected a simple liking for the stimuli that co-appeared with positive sounds over those that co-appeared with negative sounds.

As for infant cognition, children learn very early on what an isolated "No!" means (Cameron-Faulkner et al. 2007). (So, for that matter, do many other animals.) In the negation training study, the participants were doing precisely that: just saying "No!" They were repeating the most primitive kind of linguistic negation but were unable to negate or reject the association. Of course, the participants in this study *were* adults, who are otherwise capable of excising beliefs by negating them in this way. ("Do you believe that 2 + 2 = 5?" "No!") There is doubtless much that is automatic and unconscious in the operations of belief, but most theorists agree that adults are at least capable of excising beliefs when, for example, they are hit over the head with what they admit to be overwhelming counterevidence, and they have no special commitment to the truth or falsity of the proposition. Although further research is clearly warranted, implicit attitudes evidently cannot be excised in a similar fashion.

9. Conclusion

In this paper, I have urged that implicit attitudes seem to differ greatly from beliefs. Whereas beliefs (even irrational, evidence-recalcitrant beliefs) are sensitive to the logical form of other states, implicit attitudes seem to respond to states of differing logical form in similar ways, and perhaps to states of similar logical form in differing ways. However, I also emphasized that, in key respects, the empirical evidence surrounding implicit attitudes and sensitivity to logical form remains inconclusive. I indicated how further research could address the gaps. More broadly, I proposed that, however the empirical chips fall, sensitivity to logical form constitutes an important cognitive benchmark dividing primitive from sophisticated mental states.

Although they seem to differ from beliefs, implicit attitudes must also be distinguished from "mere associations" (Gawronski and Bodenhausen 2011). The effects in these studies were not completely indifferent to the meaning of agents' thoughts and perceptions. Implicit attitudes are, in some sense, sensitive to the meaning of words and images, if not to the content per se of an agent's conscious thoughts. They are also sensitive to the meaning of social cues and gestures, such as subtle expressions of approach or avoidance. These features of implicit attitudes may be important for understanding how to combat them. If we cannot simply dispense with implicit attitudes by reflectively rejecting them, what should we do about them? Emerging evidence points to a new way forward, beyond, say, arguing persuasively that stereotypes are illegitimate. Harmful implicit attitudes *can* be changed through practice, the formation of new psychological associations, and the transformation of old ones—genuine features of *training*, properly so called. Becoming a more egalitarian person may have less to do with acquiring a better appreciation of the facts and more to do with acquiring better habits.[29]

References

Agerström, J., & Rooth, D. O. (2011). The role of automatic obesity stereotypes in real hiring discrimination. *Journal of Applied Psychology*, *96*(4), 790.

Asch, S. E. (1946). Forming Impressions of Personality. *Journal of Abnormal and Social Psychology,* (41), 259-290.

Baddeley, A. (2007). *Working Memory, Thought, and Action*. Oxford University Press.

Bermúdez, J.L. (2006). Animal Reasoning and Proto-logic. In *Rational Animals?*, edited by S. Hurley and M. Nudds, 127-137. Oxford: Oxford University Press.

Brinol, P., Petty, R. E., & McCaslin, M. (2008). Changing attitudes on implicit versus explicit measures: What is the difference. *Attitudes: Insights from the new implicit measures*, 285-326.

[29] Acknowledgments.

Calanchini, J., Gonsalkorale, K., Sherman, J. W., & Klauer, K. C. (2013). Counter-prejudicial training reduces activation of biased associations and enhances response monitoring. *European Journal of Social Psychology*, *43*(5), 321-325.

Chapman, J., & Armitage, C. J. (2010). Evidence that boosters augment the long-term impact of implementation intentions on fruit and vegetable intake. *Psychology and Health*, *25*(3), 365-381.

De Houwer, J. (2011). Evaluative conditioning: A review of procedure knowledge and mental process theories. *Associative learning and conditioning theory: Human and non-human applications*, eds. T.R. Schachtman and S.S. Reilly, 399-416. Oxford, UK: Oxford University Press.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluations. *Social and Personality Compass*.

De Dreu, C. K., Greer, L. L., Van Kleef, G. A., Shalvi, S., & Handgraaf, M. J. (2011). Oxytocin promotes human ethnocentrism. *Proceedings of the National Academy of Sciences*, *108*(4), 1262-1266.

Deutsch, R., and Strack, F. (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.

Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of experimental social psychology*, *48*(6), 1267-1278.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. Cambridge, MA: MIT Press.

Eberl, C., Wiers, R. W., Pawelczack, S., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2013). Approach bias modification in alcohol dependence: Do clinical effects replicate and for whom does it work best?. *Developmental cognitive neuroscience*, *4*, 38-51.

Egan, A. (2011). Comments on Gendler's "The Epistemic Costs of Implicit Bias".

Fodor, J.A. 1990: *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Forbes, C. E., & Schmader, T. (2010). Retraining attitudes and stereotypes to affect motivation and cognitive capacity under stereotype threat. *Journal of personality and social psychology*, 99(5), 740.

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.

Gawronski, B., & Bodenhausen, G. V. (2011). The Associative-Propositional Evaluation Model: Theory, Evidence, and Open Questions. *Advances in experimental social psychology*, *44*, 59.

Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., and Strack, F. (2008). When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology*, 44, 370-377.

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

Gendler, T.S. 2008b: Alief in action (and reaction). *Mind and Language* 23 (5), 552-585

Gertler, B. 2011: Self-Knowledge and the Transparency of Belief. In A. Hatzimoysis (ed.), *Self-Knowledge*. Oxford: Oxford.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (Forthcoming). Statistically Small Effects of the Implicit Association Test Can Have Societally Large Effects. *Journal of Personality and Social Psychology*.

Gollwitzer, P.M., and Sheeran, P. 2006: Implementation intentions and goal achievement: A meta-analysis of effects and processes. In *Advances in experimental social psychology*, edited by M.P. Zanna, 69-119). US: Academic Press.

Han, H. A., Czellar, S., Olson, M. A., & Fazio, R. H. (2010). Malleability of attitudes or malleability of the IAT?. *Journal of experimental social psychology*,*46*(2), 286-298.

Harman, G. (1970). Deep structure as logical form. *Synthese*, *21*(3-4), 275-297.

Henley, N. M., Miller, M., & Beazley, J. A. (1995). Syntax, semantics, and sexual violence agency and the passive voice. *Journal of Language and Social Psychology*,*14*(1-2), 60-84.

Horcajo, J., Briñol, P., & Petty, R. E. (2010). Consumer persuasion: Indirect change and implicit balance. *Psychology & Marketing*, *27*(10), 938-963.

Huddleston, A. 2012: Naughty beliefs. *Philosophical studies*, 160 (2), 209-222.

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

Hunter, D. 2011: Alienated Belief. *Dialectica* 65 (2), 221–240.

Karg, B., and Young, J.K. 2009: *101 Secrets of the Freemason: The Truth Behind the World's Most Secret Society*. Avon, MA: Adams Media.

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

Kawakami, K., Dovidio, J. F., & van Kamp, S. (2007a). The impact of counterstereotypic training and related correction processes on the application of stereotypes. *Group processes & intergroup relations*, *10*(2), 139-156.

Kawakami, K., Phills, C.E., Steele, J.R., and Dovidio, J.F. 2007b: (Close) Distance Makes the Heart Grow Fonder: Improving Implicit Racial Attitudes and Interracial Interactions Through Approach Behaviors. *Journal of Personality and Social Psychology*, 92(6), 957–971.

Kwong, J.M.C. 2012: Resisting Aliefs: Gendler on Alief-Discordant Behaviors. *Philosophical Psychology*.

Kolodny, N. 2005: Why Be Rational? *Mind*, 114, 509-563.

Lepore, E., & Ludwig, K. (2002). What is logical form? *Logical Form and Language*, *54*, 90.

Levy, N. 2014a: Consciousness, Implicit Attitudes, and Moral Responsibility. *Noûs 48*(1), 21-40.

Levy, N. 2014b: Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*.

Lowery, B. S., Hardin, C. D., & Sinclair, S. (2001). Social influence effects on automatic racial prejudice. *Journal of personality and social psychology*, *81*(5), 842.

Mandelbaum, E. 2013: Against alief. *Philosophical studies*, *165*(1), 197-211.

Mandelbaum, E. (2014). Thinking is believing. *Inquiry*, *57*(1), 55-96.

Mandelbaum, E. Forthcoming. Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Noûs*.

McDowell, J.M. 1998b: The Content of Perceptual Experience. In *Mind, Value, & Reality*. Cambridge, MA: Harvard University Press.

Mendoza, S.A., Gollwitzer, P.M., and Amodio, D.M. 2010: Reducing the Expression of Implicit Stereotypes: Reflexive Control Through Implementation Intentions. *Personality and Social Psychology Bulletin*, 36(4), 512-523.

Mitchell, C. J., De Houwer, J., & Lovibond, P. F. (2009). The propositional nature of human associative learning. *Behavioral and Brain Sciences*, *32*(02), 183-198.

Moran, T., and Bar-Anan, Y. 2013: The effect of object–valence relations on automatic evaluation. *Cognition & emotion*, *27*(4), 743-752.

Muller, H., and Bashour, B. 2011: Why alief is not a legitimate psychological category. *Journal of Philosophical Research* 36, 371-389.

Nagel, J. (2012). Gendler on alief. *Analysis*, *72*(4), 774-788.

Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., and Tetlock, P. E. (2013). Predicting ethnic and racial discrimination: a meta-analysis of IAT criterion studies. *Journal of Personality and Social Psychology* 105: 171–192.

Payne, B.K. 2006: Weapon Bias: Split-Second Decisions and Unintended Stereotyping. *Current Directions in Psychological Science*, 15(6), 287-291.

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

Peters, K. R., & Gawronski, B. (2011). Are we puppets on a string? Comparing the impact of contingency and validity on implicit and explicit evaluations. *Personality and Social Psychology Bulletin*, *37*(4), 557-569.

Rooth, D. O. (2010). Automatic associations and discrimination in hiring: Real world evidence. *Labour Economics*, *17*(3), 523-534.

Rowbottom, D.P. 2007: 'In-between believing' and degrees of belief. *Teorema* 26, 131-137.

Rydell, R.J., McConnell, A.R., Mackie, D.M., Strain, L.M. 2006: Of Two Minds: Forming and Changing Valence-Inconsistent Implicit and Explicit Attitudes. *Psychological Science*, 17(11), 954-958.

Rydell, R. J., McConnell, A. R., Strain, L. M., Claypool, H. M., & Hugenberg, K. (2007). Implicit and explicit attitudes respond differently to increasing amounts of counterattitudinal information. *European Journal of Social Psychology*, *37*(5), 867-878.

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

Sherman, S. J., Rose, J. S., Koch, K., Presson, C. C., & Chassin, L. (2003). Implicit and explicit attitudes toward cigarette smoking: The effects of context and motivation. *Journal of Social and Clinical Psychology*, *22*(1), 13-39.

Smith, C. T., & De Houwer, J. (in press). The impact of persuasive messages on IAT performance is moderated by source attractiveness and likeability. *Social Psychology*.

Sommers, F. 2009: Dissonant Beliefs. *Analysis* 69 (2), 267-274.

Stanley, J. (2000). Context and logical form. *Linguistics and philosophy*, *23*(4), 391-434.

Stewart, B.D., and Payne, B.K. 2008: Bringing Automatic Stereotyping under Control: Implementation Intentions as Efficient Means of Thought Control. *Personality and Social Psychology Bulletin*, 34, 1332-1345.

Terbeck, S., Kahane, G., McTavish, S., Savulescu, J., Cowen, P. J., & Hewstone, M. (2012). Propranolol reduces implicit negative racial bias. *Psychopharmacology*, *222*(3), 419-424.

Unkelbach, C., Forgas, J.P., and Denson, T.F. 2008: The turban effect: The influence of Muslim headgear and induced affect on aggressive responses in the shooter bias paradigm. *Journal of Experimental Social Psychology*.

Valian, V. 1998: *Why So Slow? The Advancement of Women*. MIT Press.

Vogel, T., Bohner, G., and Wanke, M. (2014). *Attitudes and Attitude Change*. Psychology Press.

Walther, E., Weil, R., & Düsing, J. (2011). The role of evaluative conditioning in attitude formation. *Current Directions in Psychological Science*, *20*(3), 192-196.

Webb, T. L., Sheeran, P., & Pepper, J. (2012). Gaining control over responses to implicit attitude tests: Implementation intentions engender fast responses on attitude-incongruent trials. *British Journal of Social Psychology*, *51*(1), 13-32.

Wiers, R. W., Eberl, C., Rinck, M., Becker, E. S., & Lindenmeyer, J. (2011). Retraining automatic action tendencies changes alcoholic patients' approach bias for alcohol and improves treatment outcome. *Psychological Science*, *22*(4), 490-497.

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