Synthese (2016) 193:2659–2684
DOI 10.1007/s11229-015-0874-2

Why implicit attitudes are (probably) not beliefs

Alex Madva[1]

Received: 1 April 2015 / Accepted: 23 August 2015 / Published online: 14 September 2015

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. I appeal to two areas of research that seem to show that implicit attitudes fail *spectacularly* to satisfy this condition—although persistent gaps in the empirical literature leave matters inconclusive. I sketch an alternative account, according to which implicit attitudes are sensitive merely to spatiotemporal relations 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, 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 consciously but long enough to register subliminally. Given these subliminal perceptions, Madeleine acquires 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

✉ Alex Madva
alexmadva@gmail.com

[1] Vassar College, Box 93, 124 Raymond Ave, Poughkeepsie, NY 12604, USA

she would another candidate. Were she to read his résumé, she would dwell longer on his deficiencies than 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]

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. On the other, beliefs are thought to guide actions, together with desires and ends. In this case, 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 shares a common structure with more troubling cases. Many members of liberal democracies sincerely report anti-racist beliefs but harbor unwitting or unwilling racial biases. In these instances of "aversive racism" (Pearson et al. 2009), 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 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, implicit work-performance biases in Sweden predicted real-world hiring discrimination against Arab-Muslims (Rooth 2010) and obese individuals (Agerström and Rooth 2011). Employers who implicitly associated these groups with laziness and incompetence were less likely to contact job applicants from these groups for an interview. Implicit attitudes predicted discrimination over and above explicit attitudes.

What, if anything, can we do to combat these implicit biases? The answer depends in part on the nature of the underlying psychological states. Combating implicit biases requires knowing what we are up against. One question is how *belief-like* implicit attitudes are. If implicit attitudes are belief-like, perhaps we can combat them via rational argument. If not, the role of argument might be more circumscribed. Arguments might draw our attention to our biases, and motivate us to do better, even if arguments don't *themselves* reduce our biases.

Here I argue that, contrary to the views of many philosophers (Frankish forthcoming; Gertler 2011; Huebner 2009; Hunter 2011; Kwong 2012; Mandelbaum 2013, 2014; Rowbottom 2007; Schwitzgebel 2010; Webber forthcoming) and some psychologists (De Houwer 2011, 2014; Mitchell et al. 2009), these unreflective dispositions are likely not expressions of a belief-like attitude, but of an altogether different psychological kind. Implicit attitudes are responsive to an agent's thoughts, but, unlike beliefs, they seem *insensitive to the logical form* of those thoughts. Specifically, they seem insensitive to the logical constituents of mental content (e.g., operators like negation and conditional). I argue that belief-like cognitive states are, and implicit attitudes are probably not, sensitive to logical form.

[1] This case is based on Rydell et al. (2006), which measured the influence of subliminal conditioning on a timed association task (the Implicit Association Test) but not on more ecologically valid behaviors (e.g., Kawakami et al. 2007a, b).

In what follows, I explain sensitivity to logical form and argue that it is a necessary condition for belief (Sect. 2). I survey prominent arguments that implicit attitudes are belief-like, and explain how these arguments implicitly depend on sensitivity to logical form (Sect. 3). I then appeal to two areas of research to suggest that implicit attitudes are insensitive to logical form (Sects. 4, 5), although I emphasize gaps in the empirical literature that leave matters inconclusive. I conjecture that implicit attitudes are merely sensitive to experienced spatiotemporal relations, i.e., the orders in which people think, see, or hear things. I also consider the empirical evidence that tempts some to adopt a belief-based construal (or BBC) of implicit attitudes, and explain how little existing findings actually do to support that construal (Sect. 6).

2 Logical form and belief

One view holds that implicit attitudes are obviously beliefs, because they seem to meet certain very generic criteria, such as being "states of taking the world to be a certain way."[2] But such criteria are too permissive (not to mention vague), insofar as they fail to differentiate beliefs from other intentional states, from primitive perceptions to complex imaginings. Another view holds that implicit attitudes are obviously *not* beliefs, because they fail to meet certain sophisticated criteria, such as being readily revisable with the evidence, readily available for conscious reflection, or readily assimilable with other beliefs, desires, and intentions (Gendler 2008a, b; Levy 2014a; Zimmerman 2007). Such criteria are, however, too demanding, insofar as they rule out that infants and non-human animals ever have beliefs, and that human adults can have irrational, unconscious, or cognitively encapsulated beliefs. The correct criterion will fall between these two extremes. *Sensitivity to logical form* (or *form-sensitivity*) is a good fit for this purpose. I propose that beliefs are, and implicit attitudes are probably not, sensitive to the logical form of other mental states.

My interest is not primarily in deciding what 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. As I use the term, logical form is closely tied to semantic content, i.e., the truth conditions of cognitive states like belief and the satisfaction conditions of conative states like intention. I focus on logical form rather than the more general notion of semantic content because there may be ways in which implicit attitudes respond to meanings, e.g., of terms. Consider Deutsch and Strack's (2010, pp. 64–65) prediction that individuals might form an implicit attitude linking "Arab" with "terrorism" in response to media exposure, regardless whether they would reflectively agree that, "Most Arabs are terrorists" (2010, pp. 64–65). Individuals may consciously agree when they hear, "It is wrong to identify Arabs with terrorism," and "Most Arabs do not support terrorism." Yet simply hearing the conjunction of terms in these very claims may reinforce an implicit attitude associating Arabs and terrorism. Similarly, Gawronski et al. (2008, p. 376) predict that trying to reject a common stereotype by thinking, "it is not true that old people are bad drivers," reinforces rather than undermines a negative implicit attitude

[2] Sommers (2009) writes, "To believe is to take something to be so and so… animal and human belief is mainly… propositionless" (pp. 269, 270).

toward elderly drivers. These researchers seem to hypothesize that implicit attitudes are sensitive to certain linguistic tokens ("Arabs", "bad drivers"), but insensitive to the logical form of thoughts as a whole, and, specifically, insensitive to the logical constituents of the content (e.g., the "not" and perhaps even the "are" in "old people are not bad drivers"). Implicit attitudes might be insensitive to logical or predicative relations, and sensitive merely to experienced relations of spatiotemporal contiguity.

I try to elucidate logical form, and why sensitivity to logical form is necessary for belief, by reference to examples. To avoid incurring tangential commitments, I defend no particular view of logical form. Some take logical form to be the underlying (real or "deep") structure of thoughts or sentences (Harman 1970; Stanley 2000; Mandelbaum 2014).[3] Others, following Quine and Davidson, advocate abandoning this "reified" notion of logical form. Lepore and Ludwig (2002) take *sameness of logical form* to be basic, making it possible to say that "Snow is white" has the same logical form as "Schnee ist weiss," without committing to the existence of some third abstract entity in Platonic heaven which is the logical form that the two sentences share. Either view would be congenial to my argument.

The aim of sidestepping peripheral debates also lies behind my focus on whether these attitudes are *sensitive* to logical form, rather than whether they *have* logical form, i.e., are structured propositionally or linguistically. Some theorists point to the ways that implicit attitudes respond to other mental states in order to argue that they have the propositional structure characteristic of belief (Mandelbaum 2014; Levy 2014b), but this inference from dispositional profile to internal structure is contentious. Functionalists or dispositionalists may deny that beliefs as a class share any substantive internal-structural features, but accept that beliefs respond to logically structured information. They accept 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 content, many theorists will agree (to some version of the claim) that beliefs are sensitive to logical form. I therefore remain neutral about mental-state structure in what follows. Perhaps implicit attitudes are structured like generics or non-strict generalizations, or have an action-outcome or map-like representational structure.[4] I hazard my views about the content of implicit attitudes elsewhere (Brownstein and Madva 2012a, b; Madva 2012; Madva and Brownstein, under review). Here I claim that, whether and however implicit attitudes are structured, they are likely insensitive to logical form.

I believe this claim is categorically true. Implicit attitudes are not just "less" systematically sensitive to logical form than beliefs, but, as a class, wholly insensitive. Where some see conclusive evidence for partial sensitivity (e.g., Levy 2014b, p. 8), I see suggestive evidence for total insensitivity. In those specific studies where implicit attitudes seem form-insensitive (Sects. 4, 5), they seem sensitive only to spatiotemporal relations among contiguous stimuli in perception and thought. In Sect. 6, I sketch

[3] This "descriptive" approach differs from the "normative" understanding of logical form, as the idealized structure of sentences or thoughts. Here logical form refers to properties of concrete entities, not idealized abstractions.

[4] Thanks to two referees for noting these possibilities. See Brownstein (2015), Huebner (forthcoming), Gendler (2008a, b), Leslie (forthcoming) for various ways implicit attitudes might be structured.

how this "contiguity-sensitivity" can explain away the appearance of even partial form-sensitivity. However, my proposals about contiguity-sensitivity are speculative, and not meant to be the "whole story" about implicit attitudes. I intend these psychological proposals to be consistent with various accounts of the neural underpinnings of implicit attitudes (Huebner forthcoming, Madva and Brownstein, under review). My broader point is about the state of the evidence: how little it speaks *against* mere contiguity-sensitivity and *for* form-sensitivity. I gesture toward an array of studies to fill these gaps.

To get a better handle on form-sensitivity, 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, i.e., what she takes him to be saying. Now consider some variations of this example.

(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!" In this case, truth is preserved from prior, premise-like states to subsequent, conclusion-like states. Madeleine's prior belief *that Mason is one of her closest friends* is sensitive to the logical form of the thought *that Mason is a john*, and vice versa. One state does more than respond to the fact that the other also refers to Mason or includes the linguistic token "Mason." It responds 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.[5]

(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. Suppose that, in this case, Madeleine's belief *that Mason is one of her closest friends* 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 or may not be rational. Perhaps she knows Theo is trustworthy, and so should believe his testimony, but simply cannot bring herself to do it. Nevertheless, her failure to revise her attitudes in light of the evidence is entirely consistent with those attitudes being beliefs, because the operative states are appropriately sensitive to logical form. Here Madeleine responds by 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 leading her to respond one way rather than another

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

(I discuss examples in Sect. 7). Whether her reaction is fully rational depends on the quality of her reasons. Whether her reaction involves interactions between beliefs, however, depends on whether those states are sensitive to logical form.

Form-sensitivity is thus a substantially less demanding condition than evidence-sensitivity, i.e., the disposition to revise immediately in light of changes in evidence (Gendler 2008a, b). Many belief-like states—strong convictions, tacit knowledge, "habituated beliefs" (Webber forthcoming)—may not budge in response to contravening evidence. Nevertheless, becoming occurrently aware of contravening evidence 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 attitude perseverance require that the operative mental states be sensitive to logical form.[6] Form-sensitivity need not even be a matter of responding or failing to respond to evidence. It can manifest in practical reasoning (Sect. 5), idle daydreaming, or hypothetical deductions.

(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 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 (1). Here Madeleine succumbs to an isolated "performance" error, 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 form-sensitivity 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. 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?" Although Madeleine started out by thinking *that John is not a mason*, she subsequently acts as if John *is* a mason, and asks which sort of mason he is. These psychological transitions are not just responding to logical form in an objectionable way, as in (2) and (3), but failing to respond to logical form at all. Her responses are becoming unintelligible.

Concerned, Theo exclaims, "John is no mason of any kind!" But to no avail. Madeleine responds each time by asking him how John developed a propensity for masonry. In fact, the more times Theo tries to persuade her that John is not a mason, the stronger her John-is-a-mason dispositions become. It is as if she only hears the conjunction of "John" and "mason" in Theo's utterances, and cannot appreciate the relations being predicated of them. She is, for whatever reason,

[6] Critics of "wide-scope" interpretations of rationality point to asymmetries between ways of resolving inconsistency (e.g., Kolodny 2005). If Madeleine intends to drink a beer and believes beer is 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 reflect form-sensitivity.

> 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 effect on her beliefs, the intervening psychological transitions fail to respect the logical form of her initial, premise-like mental states. At least one operative mental state is not appropriately sensitive to logical form.

(4) is so bizarre that one might reasonably wonder whether something is wrong with *Madeleine*, rather than her mental states. This reflects a limitation in the analogy between my toy example, which envisages conscious sequences of belief-like thoughts unfolding in inference-like ways, and the studies I discuss below, where my point is precisely that we should *not* posit such inferential sequences. Participants in these 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. The relevant cognitive processes in these studies may largely operate unconsciously, but it should strike us as no less forced or far-fetched to envision these processes operating at an unconscious or subpersonal level as it does to imagine them unfolding consciously in Madeleine's mind, as in case (4).

The difference between (3) and (4) brings out an essential component of form-sensitivity. In (3), Madeleine's confused inference is corrected once brought to her attention. In (4), the potential for correction is lost. This difference would remain even if the error had gone undetected. Imagine (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. Nevertheless, in (3*), Madeleine's response was open to correction, while in (4*) and (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.[7]

(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., responding in the same way to "John is a mason" and "John is not a mason." It is not enough that the state happens fortuitously to respond appropriately in certain circumscribed contexts. It must be counterfactual-supporting: it *would* have responded appropriately in different contexts. A first 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.

(DD) is a weak condition. Form-sensitivity requires that a state respond to the content of the states with which it interacts, but (DD) does not specify how that responsiveness should be manifest on particular occasions. There may not be any uniquely best way to respond in a given case, but to respond in very similar ways to blatantly diverging contents is decidedly wrong. It is to fail (DD). Of course, (DD) only holds *ceteris paribus*: when thoughts are not especially complex, concepts are familiar, minds are unclouded by fatigue, drugs, or brain lesions, etc.

[7] 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 (Fodor 1990; Gendler 2008b).

Another condition for form-sensitivity is brought out by a different example. Madeleine is conversing with her granddaughter and detects a hint of sarcasm. She exclaims, "A comedian, my granddaughter!" She could just as well have said, "My granddaughter is a comedian!" These utterances differ in a trivial 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 as if they differed radically in cognitive significance would fail to be form-sensitive. A state fails form-sensitivity if it responds to states with the *same* logical form as if they *differ*. 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.[8]

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

Whether a type of psychological state meets these conditions is testable. A state fails (DD) if it responds to *differing* logical forms in *similar* ways (e.g., responding similarly to "John is a mason" and "John is not a mason"). A state fails (SS) if it responds to the *same* logical form in *different* ways (e.g., responding differently to "I'm not kidding you" and "I kid you not"). I describe studies that begin to test (DD) and (SS) in Sects. 4 and 5. First I situate my account of form-sensitivity in relation to prominent defenses of BBC.

3 Belief-based construals

Debates about whether implicit attitudes are belief-like have generally focused on the extent to which they are evidence-sensitive, i.e., update with the incoming evidence and, relatedly, the extent to which they are "inferentially promiscuous," i.e., involved in inferences with other mental states. I suggested in Sect. 2 that such criteria are too demanding to be necessary conditions for belief; these conditions are likely not met by the beliefs of infants and non-human animals, nor by many of the irrational, dogmatic, or unconscious beliefs of adults. Despite the apparent stringency of these criteria, defenders of BBC frequently argue that implicit attitudes meet them. They appeal to research (which I discuss in Sect. 6) ostensibly showing that implicit attitudes are, at least to some degree, evidence-sensitive and inferentially promiscuous (Frankish forthcoming; Levy 2014b; Mandelbaum 2014; Schwitzgebel 2010; Webber forthcoming). Perhaps evidence-sensitivity and inferential promiscuity jointly constitute sufficient, if not necessary, conditions for belief.

[8] 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 bad, all-too-easily intelligible poetry rearranges words in this sort of way. Garbled as his syntax may be, Master Yoda's sage advice is easy to understand ("Strong is Vader. Mind what you have learned. Save you it can!").

De Houwer's (2011, 2014) "propositional model" holds that implicit attitudes form and change as a result of inductive inferences based on observed environmental contingencies (see also Mitchell et al. 2009; Webber forthcoming). De Houwer hypothesizes that individuals are consciously aware of these contingencies, while Mandelbaum (2013, 2014) argues that implicit attitudes are non-conscious beliefs with a language-like compositional structure.

Defenders of BBC disagree over how *systematically* evidence-sensitive and inferentially promiscuous implicit attitudes are. Frankish (forthcoming) argues that implicit attitudes exert the systematic, cross-contextual influence on reasoning and behavior characteristic of full-fledged belief. Levy (2014b) argues that implicit attitudes are *somewhat* evidence-sensitive but "not sensitive enough… to qualify as beliefs." They only "respond to semantic contents in a patchy and fragmented way" (2). For similar reasons, Schwitzgebel (2010) concludes that implicit attitudes occupy a nebulous middle-ground between belief and non-belief.

I defend the harder-line stance, similar to Gendler's (2008a, b), that implicit attitudes differ fundamentally in kind, and not just in degree, from beliefs, although to some extent this essay can be read as a friendly refinement of Levy and Schwitzgebel's views. Suppose they are right that implicit attitudes are somehow "between" belief and non-belief. Can we be more precise here? Are implicit attitudes sensitive to some types of evidence, or certain aspects of semantic content, and not others?

Greater precision is afforded by focusing on sensitivity to logical form. My view is that implicit attitudes are sensitive to certain spatiotemporal relations in thought and perception, but insensitive to logical relations (e.g., the "not" and perhaps the "are" in "old people are not bad drivers"). Alternatively, if Levy or Schwitzgebel is right, implicit attitudes might be sensitive to *certain* logical relations and not others (e.g., perhaps sensitive to evidence for cause-and-effect relations among environmental contingencies, but insensitive to negation). If De Houwer, Mandelbaum, or Frankish is right, implicit attitudes might be sensitive to a much broader and more systematic range of relations.

Arguably, form-sensitivity is a necessary condition for the more sophisticated capacities of inferential promiscuity and evidence-sensitivity. To be even *capable* of engaging in inferences with other mental states—whether in a systematic or merely "patchy" fashion—implicit attitudes must be sensitive to the logical form of those states. If implicit attitudes categorically fail to be form-sensitive, then ipso facto they will not be inferentially promiscuous, and they will be sensitive only to a highly circumscribed range of "evidence," e.g., experienced spatiotemporal relations. Form-sensitivity may therefore constitute a significant cognitive benchmark separating primitive from sophisticated mental states. In any case, it is considerably less demanding than full-fledged evidence-sensitivity and systematic inferential promiscuity. 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 next describe research in which implicit attitudes seem not to meet even the comparatively minimal condition of form-sensitivity. They seem flagrantly insensitive to substantive differences in logical form (Sect. 4) and overly sensitive to trivial differences that are plainly irrelevant to logical form (Sect. 5). However, outstanding

gaps and underexplored conditions leave matters inconclusive. Much more research remains to be done on these questions. My hope is that form-sensitivity, (DD), and (SS) will be useful for pursuing them.

4 Treating different as same

Here I summarize a few studies suggesting that implicit attitudes fail (DD). I think they fail (DD) systematically and spectacularly: even in simple and straightforward cases, they *never* respond to differences in logical form per se. Insofar as they ever respond differently to states with differing logical forms, this is explained by some further feature, such as differences in spatiotemporal experience; if we hold that further feature fixed, we can manipulate logical form without having any influence on implicit attitudes. Conclusively demonstrating this null hypothesis—the non-relation between logical form and implicit attitudes—is doubtless a tall order, requiring more than a handful of studies, and I propose novel experiments that might disconfirm it. The first set of studies I summarize has received considerable attention in the social-psychological literature, but is, as I explain, fraught with complications that prevent straightforward inferences about (DD). The second set of studies avoids many of these complications.

First, implicit attitudes seem insensitive to negation, although research initially seemed to suggest otherwise. In Kawakami et al. (2000), participants repeatedly "negated" or "affirmed" stereotypical or counterstereotypical associations. They saw images of racially typical black and white male faces paired with potentially stereotypical traits. In the "Stereotype Negation Condition" (p. 881), participants pressed a button labeled "NO" whenever they saw a stereotypical pairing, e.g., a black face paired with the word "athletic," and a button labeled "YES" when they saw a counterstereotypical pairing, e.g., a white face paired with the word "athletic." In the "Stereotype Maintain Condition," participants affirmed (pressed "YES" in response to) stereotypical pairings and negated counterstereotypical pairings. Participants in the Stereotype Negation Condition became completely unbiased according to one measure, while those in the Maintain Condition (and those who underwent no training) continued to exhibit bias. "In short," the authors conclude, "practice does make perfect—or at least very good—stereotype negators" (p. 884). If it were true that, as the paper's title suggests, one could "Just Say No (to stereotyping)," implicit attitudes would seem to possess at least a *minimal* sensitivity to logical form. I say "minimal" because hundreds of stereotype negations were necessary, suggesting perhaps that implicit attitudes are "habituated" beliefs that change gradually after repeated involvement in inferences (Webber forthcoming), or some sort of non-strict or generic belief (Leslie forthcoming).

There were, however, four distinct tasks confounded in these studies: 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., affirming both stereotypes and counterstereotypes). (DD) predicts that affirming versus negating stereotypes, and affirming versus negating counterstereotypes, should have markedly different effects on implicit attitudes. Perhaps affirming counterstereotypes and negating stereotypes should have similar effects (namely, as Kawakami predicted, each reducing bias), but comparing these conditions

is less diagnostic. The two cognitive exercises are not obviously on a par: one asserts a less familiar correlation, e.g., whites are lazy, while another denies a familiar correlation, e.g., it is not that whites are industrious. Further predictions might be that affirming stereotypes and negating counterstereotypes should each enhance bias, but such predictions must be tempered by the fact that many adults are racially biased already. Ceiling effects may prevent them from becoming significantly more so. (Ceiling effects can be avoided by studying attitudes toward novel stimuli, such as I describe shortly).

Gawronski et al. (2008) began to tease apart these tasks by splitting participants into two groups, all of whom saw the same overall set of face-word pairings, but instructed some to simply affirm counterstereotypical pairings and others to simply negate stereotypical pairings. They found that while affirming counterstereotypes reduced implicit racial bias, negating stereotypes did not. In fact, negating stereotypes had the opposite effect: it *enhanced* bias. Evidently, implicit attitudes reflected the perceived contiguity of faces and words, regardless whether participants intended to *reject* or *affirm* the face-word pairings. Regrettably, Gawronski et al. 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 significant gap in the empirical literature. Barring ceiling effects, there is independent reason to suspect that affirming stereotypes will enhance bias just as negating stereotypes seems to do. The key condition to be tested 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 between whether something is being asserted or denied, and fail (DD). Meanwhile, the ironic effects of negation are not congenial to BBC, especially since the original findings were advertised precisely as demonstrating the efficacy of stereotype negation.

There are further obstacles to drawing conclusions about (DD). 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 suggest that pertinent processes of belief revision are afoot. Indeed, many participants form the belief that the researchers are trying to influence their social attitudes, and sometimes briefly try to resist the training (Kawakami et al. 2007a). But what exactly is the content of participants' thoughts? "That's a stereotype: negate it"? "It is false that blacks are athletic"? "I'm morally opposed to this"? Strategies for addressing this question in future research might include asking participants to report what their thoughts were (Briñol et al. 2009), or to affirm/negate statements rather than face-word pairings, or running separate conditions in which participants are instructed to type or recite (vocally or internally) various specific statements. The question is whether systematically varying the contents of these cognitive exercises differentially affects implicit attitudes, or in each case merely reinforces participants' tendency to associate whichever contiguous stimuli they are perceiving.

If the latter, then the best explanation for this finding may not make reference to belief revision but to an entirely different psychological mechanism. The perceived spatiotemporal contiguity of the words and faces may drive the effect, independently of the logical form of participants' thoughts and beliefs *about* those faces and words.

Specifically, the effect may be driven by increased attention to one rather than another type of contiguous face-word pairing, since both groups of participants saw the same set of faces and words (Gawronski et al. 2008, p. 375).

Prejudice reduction research is practically important, but, because of ceiling effects and the socially sensitive material, its implications for the underlying states and processes are not always clear. Many of these complications are absent 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 alien always appeared on the screen just before the onset of an unpleasant sound ("a horrifying human scream") while a second appeared before that unpleasant sound stopped. A third 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 relaxing melody over those who "stopped" the melody and over those who "started" the scream. They also preferred those who stopped the scream over those who started the scream and over those who stopped the melody. Like maximizers of self-interest, they learned to like those who increased pleasure or reduced suffering more than those who reduced pleasure or increased suffering. Forming these preferences requires that participants' thoughts *somehow* consciously or unconsciously tracked and compared the various alien-sound contingencies, e.g., "Lo, the detestable red alien! The scream is nigh," or, "The green alien makes the terrible scream stop."

Implicit attitudes, however, failed to respect the dramatic difference between starting and stopping valenced stimuli. Implicitly, participants preferred *both* aliens who appeared with the melody over *both* who appeared with the scream, regardless of who started or stopped the sounds. 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 one case, attending to racial stereotypes leads to less favorable implicit attitudes toward blacks, ostensibly regardless whether participants affirm or reject these stereotypes. In another, attending to images contiguous with unpleasant sounds leads to less favorable implicit attitudes, regardless whether participants judge that the images start or stop those sounds. Once again, mere spatiotemporal contiguity seems to drive the effect, in apparent independence of the logical form of participants' thoughts about the relations among the stimuli. Gawronski et al. (under review) found the same pattern of results when participants learned about drugs that caused versus prevented good versus bad outcomes (e.g., they explicitly liked but implicitly disliked medicine that prevented negative outcomes). Similarly (Sect. 1), Rydell et al. (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 rehash every study suggesting dissociations between form-sensitive beliefs and contiguity-sensitive implicit attitudes (see Gawronski and Bodenhausen 2006, 2011). Moran and Bar-Anan's studies are notable for eliciting this dissociation without any trickery or subliminal priming, and even without any overtly linguistic stimuli in the learning procedure, suggesting that implicit attitudes are not just insensitive to the form of natural-language sentences (e.g., experimenters' verbal instructions) but to the form of participants' own thoughts. We can, in this case, remain relatively agnostic about the precise content

of their thoughts (e.g., do they think the aliens "cause" or merely "signal" the stimulus changes? Do they believe that *all*, or *most*, or merely *some* green aliens stop the scream?). Whatever is going on in their minds to explain their reported preferences is not influencing their implicit attitudes.

This dissociation poses a problem for an influential BBC of evaluative conditioning: that, in the absence of additional information, participants who observe co-occurring stimuli form the belief that the stimuli co-occur, share similar valence, and so on De Houwer (2011, p. 411). In this case, participants might form the belief that green aliens co-occur with screams, and then judge that green aliens are unpleasant. Supporting this interpretation, Zanon et al. (2014, Study 2) found that simply telling participants that two stimuli co-occur had the same effect on measures of implicit attitudes as did exposing them to repeated co-occurrence. But Zanon et al. posit that such beliefs form only in the absence of additional, countervailing information. When participants were told that a novel stimulus would actually have the "opposite" meaning of a contiguous pleasant stimulus (e.g., an unfamiliar word paired with "happy" would mean "sad"), they implicitly disliked the novel stimulus (Study 1). Yet in Moran and Bar-Anan's studies, participants recognized not just that the green alien co-occurred with the scream, but that it co-occurred with the *end* of the scream, which is why they explicitly preferred it. Participants had exactly the additional information that should guide their implicit attitudes away from the putative "default" inference that contiguous stimuli share valence. (I will raise independent problems for the ostensibly BBC-friendly upshots of studies like Zanon's in Sect. 6).

Although the literature speaking to (DD) is expanding rapidly, there are many potentially relevant contrasts that have not 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 relations, 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. We might learn, against my predictions, that implicit attitudes respond to some differences in logical form and not others, which would suggest that they are "between" belief and non-belief after all—and point to precisely where along this continuum they lie.

5 Treating same as different

While the evidence that implicit attitudes treat different as same remains gappy, the possibility that they treat same as different is almost completely unexplored. 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")? What happens if we

hold the logical form of participants' thoughts as fixed as possible while manipulating the spatiotemporal ordering of their experiences? Some suggestive findings emerge from research on implicit "shooter" bias, which began in response to tragic cases of police shooting unarmed black men. Among the many causes behind such tragedies, one might be an implicit attitude associating blacks with weapons (e.g., Glaser and Knowles 2008). In one measure, participants are instructed to press a button labeled "shoot" when they see 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.

It initially seemed that trying to control shooter bias only made it worse. 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). While it is intuitive that concretizing our intentions could be helpful, the documented effects of implementation intentions on shooter bias are striking. I should note here that my aim is not to delve into the vast research on implementation intentions, which influence a wide range of mental life and behavior besides implicit attitudes, but to use a sample of this research as a vehicle for illustrating the kinds of experimental manipulations that could speak to whether implicit attitudes pass (SS), and for illustrating challenges that research on (SS) must navigate. I conclude this section by sketching ways of testing (SS) without implementation intentions. Heretofore almost no such research has been done. One virtue of considering research specifically on intentions is that it points to the role of form-sensitivity not just in assessments of evidence but also in practical rationality and agency, i.e., putting one's plans into action at the right time in the right way.

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 et al. 2010, p. 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, almost "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. The researchers say, "The observed results are striking, given that the basic instructions for completing the task were essentially the same for each condition" (p. 519). Somehow the sheer phrasing or word order of our plans can make the difference between going on to act in egalitarian or prejudiced ways.

Webb and Sheeran (2008) argue that implementation intentions work in part by making the specified cue more accessible.[9] They found that participants who formed an if-then plan to retrieve a coupon after the experiment were quicker to identify "if" components of the plan on an implicit measure. Perhaps (IF) similarly works by making one cue (the gun) more accessible, and making other cues (like race) less accessible. Webb and Sheeran further argue that implementation intentions create an automatic associative "link" between the cue and the planned action (Gollwitzer and Sheeran (2006) call them "instant habits," but this is exaggeration. For example, implementation intentions work best when they conform to participants' existing goals[10]). Webb and Sheeran's account of implementation intentions thus relies on associative mechanisms, in the traditional sense of laying down co-activating mental links, in this case creating a plan-like structure associating environmental cues with actions.

This account is very plausible, but incomplete. As traditionally understood, associative mechanisms are symmetrical. Thoughts of "salt" call up thoughts of "pepper," *and vice versa*. But this symmetry is lacking in the case of implementation intentions. While Webb and Sheeran (2008) found that cue-related words heightened the accessibility of action-related words (e.g., seeing "gun" would prime "shoot"), they found that action-related words did *not* prime cue-related words ("shoot" would not prime "gun"). The mental link between cues and actions is asymmetrical. A traditional associative account seems similarly ill-equipped to explain the differential effects of (SI) and (IF). Both (SI) and (IF) specify the same cue (guns) and the same planned action (shooting). Subjects in both conditions are presumably equally motivated to perform accurately and without bias (or, if they are differently motivated, this too would need explaining). Why doesn't (SI) heighten the accessibility of gun cues and create an automatic link between seeing guns and pressing "shoot" in just the way that (IF) does? That is, why doesn't (SI) have precisely the same associative effects as (IF)?

Part of the answer, perhaps hiding in plain sight, might be the order in which the words "gun" and "shoot" are thought, or the order in which representations of the cue and the action are tokened. 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 order in which (IF) participants *think* about

[9] See my (Madva, forthcoming) for further discussion of cognitive accessibility and the mechanisms underlying implementation intentions, from which this summary borrows.

[10] Thanks to an anonymous referee for emphasizing this point.

those steps. (IF) causes the participants to form an automatic association between the cue and the behavior *in that order*, while (SI) does not. The order of words or steps as they figure in the participants' cognition of the intention plausibly helps to explain their differential effects. (This also bears on Webb and Sheeran's (2008) finding that cue-related words prime action-related words, but not vice versa). If this is right, 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). Despite the extensive literature on implementation intentions, the effects of such grammatically impoverished "plans" are unknown (p.c., David Amodio).

Note that in both (SI) and (IF), the words "gun" and "shoot" are in the same spatiotemporal ballpark. The effect, in this case, may depend not on the mere fact that two stimuli are contiguous, but on more particular features of the spatiotemporal structure of experience. Perhaps implicit attitudes are sensitive to certain asymmetric spatiotemporal relations, e.g., guiding attention *first* to the cue and *second* to the response. If so, this might have implications for, say, Moran and Bar-Anan's studies on aliens and sounds. Although participants implicitly disliked both aliens who co-appeared with the scream, other asymmetries might emerge in their implicit dispositions toward the scream-starting red aliens versus scream-stopping green aliens. Measures of attention and accessibility, such as eye-tracking or priming, might demonstrate that seeing the red alien cues attention to the scream, but not vice versa, because hearing the scream cues attention only to the *green* alien (who had appeared at the end of the scream).

The different spatiotemporal orderings of (SI) and (IF) seem unrelated to their logical form. Again, although we can call (SI) a "simple" intention, it specifies precisely the same cue and action as (IF): to shoot in the condition when participants see a person with a gun. Both only fail to be fulfilled when participants see a person with a gun, but do not shoot. When participants in both groups come to believe that they will fulfill their intentions, their beliefs share truth conditions. 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 share logical form. Alternatively, some might want to include spatiotemporal ordering as part of logical form, so that "If A, B" differs from "B, if A." I think speaking that way is apt to be confusing, but if we do, there could still be significant differences between mental states that respond only to "the spatiotemporally-experienced aspects of logical form," and those that respond to predicates, quantifiers, and connectives.

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). (SI) does not explicitly contain the conditional "if" (although in both cases, the intended context for action is the same). (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 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.[11] These complications could be avoided in a follow-up study that employed a better semantic mirror of (IF), "I will shoot, if I see a person with a gun." But even if the underlying logical form of (SI) and (IF) differ somehow, it is mysterious *how* this difference could be relevant to the task. If there are reasonable ways of prying apart their logical forms, do these differences plausibly explain why only one intention was effective? A state that treated such clearly similar intentions as if they were utterly dissimilar would fail to be form-sensitive.

There are, however, alternative explanations for the differential effects of (IF) and (SI) worth exploring. One alternative hospitable to BBC is that (SI) is ineffective because it is more difficult to parse. The role of parsing difficulty could be investigated by testing if-then intentions with awkward constructions, 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 awkward constructions still improve performance relative to simple intentions (as I predict), then the temporal structure of if-then formulations may really drive the effect. If awkwardly constructed intentions do not improve performance (or harm it), then (SI) might simply be too difficult to think through in the moment. This study might, then, not furnish evidence that shooter bias fails (SS), but that cognitive load and time constraints prevent making the necessary inferences (cf. case (3) in Sect. 2 on performance errors).

More research into these questions is needed sorely. Mendoza et al.'s single study may not say much on its own, but it points toward further research, which can and should be pursued without implementation intentions. By holding fixed (as much as possible) the logical form of participants' thoughts, we can pinpoint more precisely which features of their external and internal environments influence implicit attitudes. Spatiotemporal manipulations may have an outsized impact. For example, due to halo effects (Asch 1946), the valence of temporally prior words in a sentence might influence implicit attitudes more than later words. Implicit attitudes might (and explicit attitudes might not) respond very differently to thinking *that p & q & r* versus *that r & q & p*. Reading a series of statements like "When Bob is happily 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 happily relaxing with friends" or "When Bob curses, yells, and tells vulgar jokes, Bob is happily relaxing with friends." Perhaps the differential effects of active versus passive constructions (e.g., Henley et al. 1995) depend in part on the sheer ordering of the words, in addition to (or even underlying) tacit implications of agency and blame. Active constructions ("Bob violently destroyed the beautiful jewel") might lead to more negative implicit attitudes toward Bob than passive constructions ("The beautiful jewel was violently destroyed by Bob"). A further area of investigation might be manipulating the spatiotemporal presentation of information, for example, contrasting the effects of reading left-to-right versus top-to-bottom, or by revealing sentences one word at a time versus displaying whole sentences at once. Such manipulations need not involve natural language. A slowly appearing picture might first

[11] Thanks to Katie Gasdaglis for this suggestion.

reveal Bob with positive stimuli and gradually reveal negative stimuli. For example, a scene where Bob is holding a beautiful bouquet in a dark, gloomy cemetery might have different effects on implicit attitudes if Bob's face is first seen with flowers or with tombstones.

6 Counterevidence?

Several studies are cited to support BBC, such as Briñol et al.'s (2009) finding that implicit racial biases decreased after participants read persuasive arguments for hiring more African-American professors at their university.[12] However, these studies, while independently interesting, currently do little to support BBC. Where the empirical case against BBC has gaps, the case for it has chasms.

Before explaining why, two 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 exhibit greater bias than alert individuals (e.g., Govorun and Payne 2006). A change in performance on these measures might reflect an effect on attitudes, or behavioral control, or both. Several process-dissociation models attempt to disentangle these possibilities, and it is commonplace to use them to analyze data. Process-dissociation modeling suggests that counter-attitude training (Calanchini et al. 2013) and implementation intentions (Mendoza et al. 2010) *both* reduce implicit biases *and* increase the capacity to control their expression. Second, measures of implicit attitudes are, like measures of blood pressure, susceptible to myriad contextual and motivational factors. Implicit biases increase after taking oxytocin (De Dreu et al. 2011) and decrease after taking beta blockers (Terbeck et al. 2012). They decrease in the mere presence of a black experimenter (Lowery et al. 2001). Nicotine-deprived smokers exhibit positive implicit attitudes toward smoking, but after smoking they exhibit negative attitudes—slightly more negative than nonsmokers (Sherman et al. 2003). Interpreting these short-lived effects is beyond this paper's scope, but clearly they do not portend genuine attitude change. No one would propose that smokers should smoke in order to reverse their implicit attitudes about smoking.

To distinguish genuine changes from context effects, experimenters delay the posttest and change the context, although truly longitudinal and context-general experiments remain scant. While studies suggest that long-term change is possible (Devine et al. 2012), the conditions are not sufficiently controlled to isolate precise causes. Possible exceptions include Wiers et al.'s (2011) research on patients recovering from alcoholism. Participants who repeatedly avoided images of alcohol (in four 15-min sessions) prior to three months of standard therapy were less likely to relapse at least one year after discharge. Eberl et al. (2013) replicated these effects, finding that alcohol-avoidance training generated negative implicit attitudes toward alcohol, and that this change mediated the improvement in long-term recovery. Standard therapy

[12] See also Horcajo et al.'s (2010) findings that persuasive arguments influenced implicit attitudes toward vegetables and brands. See Levy (2014a, b) and de Houwer (2011, 2014) for surveys of other BBC-relevant studies.

without training or with sham training had no effect on implicit attitudes (and relapse was more likely). These studies on addiction recovery do not speak to whether implicit attitudes are form-sensitive (although the failure of months of therapy to make even a *dent* in implicit attitudes is striking). They demonstrate that implicit attitude change can endure and generalize to "real-world" behavior. Less far-ranging studies find that the effects of counterstereotype training last at least 24–30 h, on a variety of measures (Forbes and Schmader 2010). If anything, the effects seem to grow in strength over that span (Kawakami et al. 2000) and after intervening tasks (Kawakami et al. 2007a). Implementation intentions influence implicit attitudes for at least three weeks (Webb et al. 2012) and have other effects lasting months (Chapman and Armitage 2010). Broad patterns of evidence suggest these interventions are more than momentary flukes.

By contrast, the studies cited to support BBC (such as Zanon et al. 2014, discussed in Sect. 4) have neither used process-dissociation models nor tested the effects after even a brief delay. For all we know, these manipulations only generate transient context effects—ways to briefly "fool" the measure rather than influence the intended object of measurement (cf. Han et al. 2010). Huebner (forthcoming) similarly speculates that argument-based interventions temporarily boost motivation or control, rather than affect attitudes. Some BBC supporters have recently acknowledged these concerns. Smith and Houwer (2014) found that a persuasive message influenced implicit attitudes on one measure, immediately after reading the message, but not a second measure, immediately after the first. Variability across measures is common, but they also consider that "the effects of the persuasive message might have dissipated" before the second measure. Perhaps the effects are especially fragile and short-lived. Given that process-dissociation models have not been applied to these manipulations (cf. Smith and Houwer 2014, p. 444), and given widespread evidence of fluky context effects, it is difficult to see how these studies provide any *distinctive* support for BBC at all, as opposed to just more context effects that temporarily "fool" the test. In other words, while the evidence against the form-sensitivity of implicit attitudes is admittedly gappy, clear evidence for it is almost nonexistent. The further evidence needed is straightforward and commonplace: test after delays and across contexts, and analyze data with process-dissociation models.

But suppose, for the sake of argument, that some argument-based interventions do prove to have durable, generalizable effects. Such findings might fall short of suggesting that implicit attitudes are sensitive to logical form per se, rather than to its "downstream" effects. The mere conveyance of logically structured information in a manipulation does not indicate that the effect occurs *by virtue* of form-sensitivity. Suppose I persuade you to stand up in order to reduce a measure of your blood pressure. Following my advice, your blood pressure drops. Should we conclude that blood pressure is sensitive to persuasive argument? Positing this direct connection would be absurd. Similarly, logical form might influence implicit attitudes *indirectly*.[13]

Take the ostensibly BBC-friendly finding that both explicit and implicit attitudes toward a person named Bob formed and subsequently reversed in response to reading

[13] Gawronski and Bodenhausen (2011, Sect. 2.1.2) summarize several potential pathways of indirect influence, but do not discuss the strategy I highlight below.

valenced statements, e.g., "Bob continually yells at his wife in public" (Rydell et al. 2007). Explicit attitudes formed after reading only 20 statements, whereas implicit attitudes formed much more slowly, after about 100 statements. BBC's supporters might infer that implicit attitudes are *somewhat less form-sensitive* than explicit beliefs, or perhaps that they are some sort of experience-based generalization. However, findings like these are consistent with implicit attitudes per se being categorically form-insensitive, if the relationship between logical form and implicit attitudes is mediated by some further variable.

The best evidence for a mediated, indirect relationship is when the causal connection can be severed or supplanted. As it happens, the effects of valenced statements on implicit attitudes can be entirely thwarted, by subliminal priming. When positive statements about Bob are paired with subliminal negative words, self-reported beliefs become positive while implicit attitudes become negative (Sect. 1, Rydell et al. 2006). Subliminal priming intervenes in the "normal" movement from reading information to forming implicit attitudes. What is the intermediate step? One plausible mediator is affect. In this case, subliminal perceptions of valenced words might activate 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. Implicit attitudes would then reflect the contiguity of Bob's face and affective responses. But these affective responses can presumably be induced in numerous ways, including reading valenced statements. Personally, when I read Rydell's example of a negative statement—"Bob continually yells at his wife in public"—I feel a visceral discomfort. After repeatedly reading such sentences while seeing Bob's face, eventually just seeing him activates subtly negative feelings. If so, the effect of logical form on implicit attitudes is mediated, roughly, by the contingent and interruptible effect of belief on affect. This could explain the intermittent appearance that implicit attitudes are form-sensitive despite being categorically form-insensitive.

If the relations sketched here between logical form, affect, and implicit attitudes seem ad hoc or mysterious, it bears mentioning that BBC is equally committed to them. BBC posits causal relations between beliefs and evaluative dispositions (e.g., "Bob is a jerk. Therefore, I dislike Bob. Therefore…" … negative affective dispositions toward Bob), while offering no illuminating explanation for why these relations obtain. As Walther et al. (2011, p. 193) succinctly put it, "it is not clear how propositional knowledge is translated into liking." This "translation" is simply stipulated. It is no less mysterious for BBC than rival theories.

Moreover, in some ostensibly BBC-supporting studies, researchers interpret their findings in precisely these mediated terms. Consider Briñol et al.'s (2009) research comparing strong versus weak arguments for hiring African-American professors. Among participants encouraged to think extensively about the arguments, those who read strong arguments showed less bias than those who read weak arguments. Is this evidence that implicit attitudes are form-sensitive? The researchers don't think so. They propose that the effect of argument quality on implicit attitudes is a function of the sheer quantity of positively versus negatively valenced thoughts that participants entertain:

> 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, emphasis added)

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 were indeed mediated by the net valence of reported thoughts, i.e., persuasive arguments reduced bias by inducing a greater number of "happy thoughts" about blacks. It is striking that many researchers—even in, as it were, the belly of the beast of BBC—take such studies 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 logically structured information from affective experience, 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. 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 statements, e.g., "Bob loves the taste of delicious food; Bob really likes his friends; Bob enjoys fun hobbies; Bob follows the advice of wise, trustworthy people," and even questionable statements like, "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 statements might influence implicit attitudes even if participants already have decisive reason to dislike him (e.g., because he is an unrepentant serial murderer). Participants should also come to implicitly like any attention-capturing sights, smells, or sounds (e.g., mock advertisements) spatiotemporally contiguous with the statements.

7 Objections

In Sects. 2 and 3, I claimed that form-sensitivity has the virtue of being a significantly less demanding condition on belief than evidence-sensitivity and inferential promiscuity. One might object, however, that form-sensitivity is still too strong, because it seems to require that a mental state responds to the content, the whole content, and nothing but the content of other states. (SS), which mandates responding 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. We 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!" They arguably 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 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 (Karg 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.

One might also worry that form-sensitivity is too *linguistic* to be a necessary condition for belief. Several studies discussed in Sects. 4–6 refer to negations and grammatical features in English. The cognitive states of non-human animals and infants, and many cognitively encapsulated belief-like states in adults, will be largely insensitive to these linguistic niceties. One might worry, then, that form-sensitivity, just like more sophisticated criteria such as evidence-sensitivity, rules out that such states are beliefs. However, my argument does not presuppose that logical form be cashed out in terms of natural language. Presumably, implicit attitude research predominantly involves language-dependent manipulations because they are more tractable, but it need not. Implicit attitudes can be changed merely by approaching or avoiding stimuli (Kawakami et al. 2007b; Wiers et al. 2011), a task that any being capable of associative learning could approximate.

For example, far from form-sensitivity's being too demanding to apply to animals, theoretical discussions of non-human cognition commonly address whether such cognition is marked by analogues of form-sensitivity. Whether a bit of animal behavior should be explained in terms of belief and desire, or exemplifies rationality, often turns 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 examples of form-sensitivity discussed in Sects. 4–6. In fact, the procedure in Moran and Bar-Anan's (2013) studies on learning about creatures who "start" and "stop" sounds included nothing but images and sounds, not explicit language. It would likely be adaptive for adults, infants, and non-human animals alike to discriminate among stimuli that signal the imminent increase versus decrease in pleasure versus suffering, and to instinctively prefer signals for reduced suffering over signals for reduced pleasure or increased suffering. These are ecologically meaningful, experience-based proto-inferences that one might predict even relatively

unsophisticated beings (or rudimentary cognitive processes) could make. Yet while adults' self-reported preferences tracked these 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.

8 Conclusion

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. In crucial respects, however, the empirical evidence remains inconclusive. I have indicated how further research could address the gaps.

Although they seem to differ from beliefs, implicit attitudes must also be distinguished from "mere associations." The effects in these studies are not completely indifferent to the meaning and spatiotemporal structure of agents' thoughts, perceptions, and feelings. 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 certain affect-laden social cues and gestures, such as subtle expressions of approach or avoidance. These features of implicit attitudes may be important for combating them. If we cannot simply dispense with implicit attitudes by reflectively rejecting them, what should we do? Emerging evidence points 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.

Acknowledgments This essay was revised during my Mellon Postdoctoral Fellowship at the University of California, Berkeley, and subsequently with institutional support from Vassar College. For insightful comments on earlier drafts or lively debates about the ideas in this paper, I am indebted to Alejandro Arango, Charles Michael Brent, Michael Brownstein, Taylor Carman, Guillermo del Pinal, Andrew Franklin-Hall, Katie Gasdaglis, Bertram Gawronski, Tamar Szabó Gendler, Lydia Goehr, Brian Kim, Patricia Kitcher, Felix Koch, Chloe Layman, Eric Mandelbaum, Christia Mercer, Nate Meyvis, John Morrison, Matthew Moss, Marco Nathan, Andreja Novakovic, Christiana Olfert, Katherine Rickus, David Rosenthal, Michael Seifried, Beau Shaw, Susanna Siegel, Virginia Valian, Anubav Vasudevan, Sebastian Watzl, and several anonymous referees. Thanks also to audiences at Columbia University and the Eastern Division meeting of the American Philosophical Association in Atlanta, December 2012.

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: Oxford University Press.

Bermúdez, J. L. (2006). Animal reasoning and proto-logic. In S. Hurley & M. Nudds (Eds.), *Rational animals?* (pp. 127–137). Oxford: Oxford University Press.

Briñol, P., Petty, R. E., & McCaslin, M. (2009). *Changing attitudes on implicit versus explicit measures: What is the difference* (pp. 285–326). Attitudes: Insights from the new implicit measures.

Brownstein, M. (2015). Implicit bias. In: Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/archives/spr2015/entries/implicit-bias/.

Brownstein, M., & Madva, A. (2012a). Ethical automaticity. *Philosophy of the Social Sciences, 42*(1), 67–97.

Brownstein, M., & Madva, A. (2012b). The normativity of automaticity. *Mind and Language, 27*(4), 410–434.

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 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 of the United States of America*, *108*(4), 1262–1266.

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

De Houwer, J. (2014). A propositional model of implicit evaluations. *Social and Personality Compass*, *8*(7), 342–353.

Deutsch, R., & Strack, F. (2010). Building blocks of social behavior: Reflective and impulsive processes. In B. Gawronski & B. K. Payne (Eds.), *Handbook of implicit social cognition: Measurement, theory, and applications* (pp. 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.

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.

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.

Frankish, K. (Forthcoming). Implicit bias, dual process, and metacognitive motivation. In Brownstein & Saul (Eds.), *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.

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., & 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.

Glaser, J., & Knowles, E. (2008). Implicit motivation to control prejudice. *Journal of Experimental Social Psychology, 44*, 164–172.

Gollwitzer, P. M., & Sheeran, P. (2006). Implementation intentions and goal achievement: A meta-analysis of effects and processes. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 69–119). New York: Academic Press.

Govorun, O., & Payne, B. K. (2006). Ego-depletion and prejudice: Separating automatic and controlled components. *Social Cognition*, *24*(2), 111–136.

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.
Huebner, B. (2009). Trouble with stereotypes for Spinozan minds. *Philosophy of the Social Sciences*, *39*, 63–92.
Huebner, B. (Forthcoming). Implicit bias, reinforcement learning, and scaffolded moral cognition. In Brownstein & Saul (Eds.), *Implicit bias & philosophy: Volume I, metaphysics and epistemology*. Oxford: Oxford University Press.
Hunter, D. (2011). Alienated belief. *Dialectica*, *65*(2), 221–240.
Karg, B., & 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., & 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., & 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.
Kolodny, N. (2005). Why be rational? *Mind*, *114*, 509–563.
Kwong, J. M. C. (2012). Resisting aliefs: Gendler on alief-discordant behaviors. *Philosophical Psychology*, *25*(1), 77–91.
Lepore, E., & Ludwig, K. (2002). What is logical form? *Logical Form and Language*, *54*, 90.
Leslie, S. (Forthcoming). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*.
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*. doi:10.1111/nous.12074.
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.
Madva, A. (2012). *The hidden mechanisms of prejudice: Implicit bias and interpersonal fluency*. PhD dissertation, Columbia University.
Madva, A. (Forthcoming). Virtue, social Knowledge, and implicit bias. In Brownstein & Saul (Eds.), *Implicit Bias & Philosophy: Volume I, Metaphysics and Epistemology*. Oxford: Oxford University Press.
Madva, A., & Brownstein, M. (under review). *The blurry boundary between stereotyping and evaluation in implicit cognition*.
Mandelbaum, E. (2013). Against alief. *Philosophical Studies*, *165*(1), 197–211.
Mandelbaum, E. (2014). *Attitude*. Inference, association: On the propositional structure of implicit bias. Noûs. doi:10.1111/nous.12089.
Mendoza, S. A., Gollwitzer, P. M., & 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., & Bar-Anan, Y. (2013). The effect of object–valence relations on automatic evaluation. *Cognition & Emotion*, *27*(4), 743–752.
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.
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. (2014). The impact of persuasive messages on IAT performance is moderated by source attractiveness and likeability. *Social Psychology*, *45*(6), 437–448.

Sommers, F. (2009). Dissonant beliefs. *Analysis*, *69*(2), 267–274.

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

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.

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. (2008). Mechanisms of implementation intention effects: The role of goal intentions, self-efficacy, and accessibility of plan components. *British Journal of Social Psychology*, *47*, 373–395.

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.

Webber, J. (Forthcoming). Habituation and first-person authority. In R. Altshuler & M. Sigrist (Eds.), *Time and the philosophy of action*. Routledge.

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.

Zanon, R., De Houwer, J., Gast, A., & Smith, C. T. (2014). When does relational information influence evaluative conditioning? *The Quarterly Journal of Experimental Psychology*, *67*(11), 2105–2122.

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