Collective Implicit Attitudes: A Stakeholder Conception of Implicit Bias

Carole J. Lee (c3@uw.edu)
Department of Philosophy, Box 353350
Seattle, WA 98195 USA

Abstract

Psychologists and philosophers have not yet resolved what they take implicit attitudes to be; and, some, concerned about limitations in the psychometric evidence, have even challenged the predictive and theoretical value of positing implicit attitudes in explanations for social behavior. In the midst of this debate, prominent stakeholders in science have called for scientific communities to recognize and countenance implicit bias in STEM fields. In this paper, I stake out a stakeholder conception of implicit bias that responds to these challenges in ways that are responsive to the psychometric evidence, while also being resilient to the sorts of disagreements and scientific progress that would not undermine the soundness of this call. Along the way, my account advocates for attributing collective (group-level) implicit attitudes rather than individual-level implicit attitudes. This position raises new puzzles for future research on the relationship (metaphysical, epistemic, and ethical) between collective implicit attitudes and individual-level attitudes.

Keywords: collective implicit attitudes; implicit attitudes; implicit bias; science policy; dispositional attitudes; attitudes

Introduction

The American Association for the Advancement of Science and the National Academies of Science – non-profit organizations tasked with advancing science and providing science-based advice – have called for scientific communities to recognize and countenance implicit bias as an impediment on women and minority participation and advancement in science, technology, engineering, and mathematics (STEM) fields (Pinholster, 2016; National Academies of Science, 2015). How do we interpret their claims about implicit bias in the face of vociferous debate about what implicit attitudes are and real limitations to canonical methods for measuring them – all while creating evidence-based policies that are resilient to the sort of fine-grained empirical debate and progress that would not undermine the soundness of this call?

In this paper, I will stake out a stakeholder conception of implicit bias that finesses two major challenges: (current and future) disagreement about what implicit attitudes are, and the psychometric limits of the Implicit Association Test (IAT).[1] While responding to the second challenge, I will argue for the notion of collective implicit attitudes and identify some of the metaphysical, epistemic, and ethical questions they raise for future philosophical inquiry.

[1] Because there is more evidence on the discriminant validity of implicit attitudes as measured by the IAT as opposed to other measures for implicit attitudes, I focus on IAT-related evidence.

Challenge 1: Disagreement about what Implicit Attitudes Are

Ideally, stakeholders interested in articulating evidence-based policies do so in ways that create room for scientific disagreement about and continued discovery of finer-grained scientific details that do not impact the soundness of the policy. This practical constraint has interesting implications on whether stakeholder claims and policies should conceptualize implicit attitudes as mental states or as dispositions.

Implicit Attitudes as Mental States

Social psychologists advocating the mental state approach to conceptualizing implicit attitudes have imputed competing cognitive accounts of the hidden processes and representations that explain how social behavior is generated from stimuli in the environment (De Houwer, Gawronski, & Barnes-Holmes, 2013). For mental state view advocates, attitudes are evaluative judgments stored in long-term memory (Wilson, Lindsey, & Schooler, 2000) or constructed, "on the spot" (Schwarz, 2007, p. 650), in working memory (Gawronsky & Bodenhausen, 2006). Some mental state theorists advocate for just one kind of underlying representation to do this work (Fazio, 2007). Others advocate for the existence of two or more representations that generate judgments and behavior via different but interacting types of processes (Wilson, Lindsey, & Schooler, 2000; Wittenbrink, Judd, & Park, 2001). Yet others suggest that the representational base or "underlying ingredients" (Krosnick, Judd, & Wittenbrink 2005, p. 24) from which implicit and explicit attitudes are formed are shared, where observed dissociations between implicit and explicit attitudes result from different processes rather than from different, independently stored representations (Gawronski & Bodenhausen, 2007; Strack & Deutsch, 2004). In light of such disagreement, some mental state theorists have suggested that the term "attitude" could "be used as a general integrative label that subsumes *any* aspect of process that is responsible for positive or negative responses toward a given object" (Gawonski & Bodenhausen 2007, p. 709, italics mine). Among philosophers advocating a mental-state-like view, some suggest we add to beliefs a second type of content-laden attitude (Gendler, 2008; Levy, 2015; Mandelbaum, 2015; Holroyd, 2016). For others, "wheeling in the big gun of a new fundamental taxonomical category" (Egan, 2011, pp. 67-8) may not yet be merited (Kwong, 2012).

Which of these theories is the right one? The imputation of representations and processes is constrained by each other: mental representations can only be retrieved and generate behavior by means of mental processes (Anderson, 1976; De Houwer, Gawronski, & Barnes-Holmes, 2013; Machery, 2007); and, processes can only be triggered by and transform some representations (Gigerenzer & Hoffrage, 1995; Lee, 2007). However, because there isn't consensus about either the representations or processes involved, the field of possible, empirically permissible cognitive theories is large enough that "the same behavioral data" can be explained "as multiple processes operating on a single representation, one process operating on multiple representations, or any admixture of representations and processes" (Greenwald & Nosek, 2008, p. 80). Each of these cognitive theories – which posit different numbers and types of representations and processes – can be made and has been made consistent with the observed evidence. "[T]here are plausible arguments for any of these positions" (Gawronski & Bodenhausen, 2007, p. 708).

From a stakeholder perspective, formulating a policy that's conditioned on a prediction about the longevity or superior empirical adequacy of a particular mental state/process theory seems unwise. Mental state theories are so pervasively underdetermined (Bechtel, 2005) that disagreement among empirically adequate cognitive theories may be the norm rather than the exception (Greenwald, 2012). Even if we refrain from drawing a pessimistic induction over a longer history of unsettled debates among competing mental state theories for other kinds of cognitive capacities (Greenwald, 2012; Laudan, 1981), it is important to note that psychologists have voiced concerns that disagreement among competing cognitive theories of *implicit attitudes* in particular "will never end and should never end" (Eagly & Chaiken, 2007, pp. 585-6) and may be "impossible to resolve" (De Houwer, Gawronski, & Barnes-Holmes, 2013, p. 3), with analogies drawn to "the [unresolved] debate between abstractive and exemplar-based representations in the cognitive literature" (De Houwer, Gawronski, & Barnes-Holmes, 2013, p. 13).[2]

Implicit Attitudes as Dispositions

From a stakeholder perspective, conceptualizing a policy that's conditioned on a prediction about the longevity or superior empirical adequacy of any particular cognitive theory in explaining observed and accepted effects is unnecessary for their purposes if a dispositional approach to conceptualizing implicit attitudes is available instead.[3]

According to the dispositional approach, attitudes – and implicit attitudes more specifically – are tendencies to cognize and behave towards an object, where these tendencies can be (imperfectly) measured through various measurement procedures (Cronbach & Meehl, 1955; Krosnick, Judd, & Wittenbrink, 2005; Eagly & Chaiken, 2007; Greenwald & Nosek, 2008). Dispositions are kept conceptually separate from and remain agnostic about claims about the (number of) representations and processes underwriting them (Fazio, 2007; Borsboom, Mellenbergh, & van Heerden, 2004). Thus, stakeholders adopting a dispositional view would be committing to the idea that while there is *some* mental state(s)/process(es) underwriting implicit attitudes, making sense of their policies does not require theoretical pre-commitment to any *particular* cognitive theory at the level of mental states/processes.[4]

By adopting a dispositional approach, stakeholders would be adopting a stance of epistemic modesty. Such a stance would not be unique to the stakeholder perspective – as a matter of scientific practice, it is also a position that some psychological researchers adopt. For example, Alice Eagly and Shelly Chaiken propose characterizing attitudes as "evaluative tendencies" and "purposefully avoid further specification of the inner tendency" since "the description of this inner tendency inevitably changes as attitude research develops and different theoretical positions emerge, become popular, and then may erode" (Eagly & Chaiken, 2007, pp. 585-6). Anthony Greenwald and Brian Nosek (2008) adopt a dispositional approach because they take questions about the number of underlying representations and processes to be, at present, "empirically irresolvable" (Greenwald et al., 2009, p. 32). Going all the way back to 1935, when Gordon Allport declared the concept of attitude as "the keystone in the edifice of American social psychology" (Allport, 1935, p. 798), he noted that the only "common thread" running through diverging definitions of the concept "attitude" (Allport, 1935, p. 805) was the idea that attitudes involved a kind of disposition: "a mental and neural state of readiness, organized through experience, exerting a directive or dynamic influence upon the individual's response to all objects and situations with which it is related" (Allport, 1935, p. 810).

In general, characterizing attitudes as dispositions does not demote the insight or priority of research on the psychological basis of the attitude construct (Machery, 2016) or on psychologists' and philosophers' theories about the cognitive architecture underwriting those attitudes (for just one nice example of this genre at work, see Huebner, 2016). Nor does it deny the reality of progress in these domains of research. Indeed, the purpose of characterizing attitudes as dispositions is to provide an account that is broad enough to accommodate standard patterns of scientific disagreement and growth in these discussions. In the natural lifecycle of an interesting effect, second generation questions about representations and processes – studied by identifying boundary conditions and moderators

[2] *Pace* Machery (2009), De Houwer et al agree with Barsalou that "trying to determine whether people use exemplar or abstracted representations is futile" and "cannot be evaluated on the basis of behavioral data" (Barsalou, 1990, pp. 61-2).

[3] For more on the relative stability of and agreement about psychological effects versus their cognitive explanations, see Cummins (1983, 2000).

[4] Note that the dispositional approach does not construe implicit attitudes as behaviorist posits since implicit attitudes are underwritten by representation-rich processing.

for the effect (Spencer, Zanna, & Fong, 2005; Bechtel, 2005; Zanna & Fazio, 1982; Fischhoff, 1982) – are fruitful. They generate new evidence that informs and constrains future cognitive theories (De Houwer, Gawronski, & Barnes-Holmes, 2013; Jacoby & Sassenberg, 2011) and refine our understanding of the original effect (De Houwer, Geldof, & De Bruycker, 2005; Gawronski et al., 2008). Indeed, a mental state/process theory's generative role in such debate and progress is a critical part of evaluating the value of any given cognitive theory (De Houwer, Gawronski, & Barnes-Holmes, 2103). As such, stakeholders adopting a dispositional approach would not, by any means, be discounting the importance of second-generation research on community effects (Dasgupta, 2013), context effects (for a review see Gawronski & Bodenhausen, 2008), training effects (e.g., Brauer et al., 2012), or the influence of competing cognitive processes in conditions where there is sufficient cognitive capacity and motivation (e.g., Payne, 2005; Maddux et al., 2005). Such mental state research makes important strides towards elucidating the psychological basis of our attitudes – i.e., the representations and processes underwriting our evaluative tendencies – and the circumstances under which an individual's egalitarian convictions are more likely to be reflected in her cognition and behavior.

Overall, stakeholders interested in articulating evidence-based policies that can create room for scientific disagreement about and continued discovery of finer-grained scientific details can invoke a dispositional rather than mental state conceptualization of implicit attitudes. Doing so does not put stakeholders in the awkward position of betting on the longevity or superior empirical adequacy of any particular mental state/process theory. Moreover, such a position allows stakeholders to respect the ways in which debates at the level of cognitive theory advance our understanding of implicit attitudes.

Challenge 2: Psychometric Limits of the IAT

Stakeholders must address a second challenge. Some have invoked psychometric evidence to challenge the predictive and theoretical value of positing implicit attitudes (Machery 2016; Oswald et al., 2013). In response, I think stakeholders can draw on a fuller suite of the psychometric evidence to hold that implicit attitudes *are* legitimately posited – but, are best attributed to groups of cognizers rather than to individuals. As such, this position re-interprets the call to countenance implicit bias in STEM contexts as a call to countenance collective implicit bias. I will identify this view's methodological and meta-methodological rationale, its policy implications, and some puzzles it raises about implicit attitudes.

Psychometric Challenges

When it comes to the construct validity of implicit attitudes, the psychometric evidence is quite mixed. The IAT has a low test-retest reliability, which may indicate that much of the variation in its scores is attributable to random errors of measurement rather than to the presence of an underlying construct (for a contrary view, see Cunningham, Preacher, & Banaji, 2001). The IAT has only small-to-moderate predictive validity (Greenwald et al., 2009; Oswald et al., 2013; Greenwald, Banaji, & Nosek, 2015). When we look beyond the IAT to a fuller set of techniques for measuring implicit attitudes, we see that these have low convergent validity with each other (Olson & Fazio, 2003; Rudman & Kilianski, 2000), which may be interpreted as suggesting that there isn't a shared, underlying construct that they all measure.

On the basis of the evidence above, some have suggested that we should reconsider whether to posit implicit attitudes at all (Machery, 2016; Oswald et al., 2013) – a position standing in direct contrast to the overwhelming view among psychologists and philosophers that implicit attitudes exist and are sensibly attributed to individual cognizers (for an overview, see Gawronski & Bodenhausen, 2006; Brownstein & Saul, 2016a, 2016b).

In contrast to both these mainstream and radical views, I think that stakeholders can hold what may, at first, sound like an unusual position: when considering the fuller suite of psychometric evidence, stakeholders can support claims about the construct validity of implicit attitudes, but only when describing and predicting group-level behavior rather than individual behavior.

Discriminant Validity

The strongest evidence favoring the positing of implicit attitudes is evidence of its discriminant validity. This evidence tends to be ignored or discounted by those adverting to the psychometric evidence to challenge the legitimacy of implicit attitudes (Machery, 2016; Oswald et al., 2013). To propose a new construct, psychologists must bring to bear evidence that distinguishes it from constructs already in use (Campbell & Fiske, 1959). As such, much of the research on implicit attitudes has focused on their discriminant validity in relation to explicit attitudes. Such evidence includes low correlations between tests thought to measure these different constructs (Krosnick, Judd, & Wittenbrink, 2005): meta-analysis measures demonstrate that correlations between the IAT and explicit measures (designed to measure explicit attitudes) are only small-to-moderate (Greenwald et al., 2009). Further evidence for discriminant validity includes dissociations (Greenwald & Nosek, 2008): some factors affect implicit but not explicit attitudes (Karpinski & Hilton, 2001) while other factors have been shown to impact explicit but not implicit attitudes (Greenwald et al., 2009). Finally, IAT scores and measures for explicit attitudes each predict variance not predicted by the other: the predictive validity of IAT scores begins to catch up to measures for explicit attitudes when dealing with socially sensitive topics and then outperforms measures for explicit attitudes in predicting intergroup behavior (Greenwald et al., 2009).

Despite the discriminant validity of IAT scores, the IAT's low-to-moderate test-retest reliability and low-to-moderate

predictive validity measures mean that IAT scores cannot be used diagnostically to predict *individual* differences in the "propensity to discriminate" (Oswald et al., 2013, p. 187) – not without risking "undesirably high rates of erroneous classifications" (Greenwald, Banaji, & Nosek, 2015, p. 557). As such, some suggest that IAT scores should be used to characterize cognition and behavior at the group or societal level (Greenwald, Banaji, & Nosek, 2015).[5]

Collective Implicit Attitudes

Because of limitations in the psychometric evidence, stakeholders should adopt the view that implicit attitudes are best attributed at the group rather than at the individual level. This reinterprets calls to recognize implicit bias in STEM as calls to address biases embodied and exhibited by collectives. Under this lens, previous questions conceived in terms of individuals are reimagined at the explicitly collective level: how do collective implicit biases reflect unjust social structures – and, what should institutions (including stakeholders in STEM) do to address them? As such, this collective account of implicit attitude foregrounds that "broader conception of attitude that is elastic enough to apply. . . to broad patterns of culture" – the conceptual "meeting point for discussion and research" between psychologists and sociologists (Allport 1935, p. 798).

This view also raises a number of difficult philosophical puzzles. Previous work on collective intentionality has grappled with a number of important challenges associated with trying to characterize the relationship between collective attitudes versus the attitudes of the individuals belonging to those groups (Gilbert, 1989; Pettit, 2001, 2007; List & Pettit, 2011). The notion of collective implicit attitudes raises a number of analogous questions. How should we characterize the relationship between collective implicit biases versus the attitudes of the individuals belonging to those groups? And, what are the ramifications of these accounts on assessing the epistemic and moral responsibility of the collective versus its individuals?

Finally, there may be some who remain skeptical about whether the psychometric evidence is sufficient to begin thinking or talking about launching interventions at all. Note that, even if the IAT is not a diagnostically terrific screening tool for predicting which *individuals* will commit discriminatory acts, stakeholders can nevertheless adopt interventions designed to reduce the overall risk of biased behavior, even among groups who do not have high IAT scores; and, some interventions may be inexpensive and beneficial enough to merit their broad acceptance and adoption. To understand how this might be the case, consider this analogy to the case of blood pressure. From an epidemiological perspective, some take blood pressure readings to be a poor screening tool when it comes to predicting cardiovascular events such as heart attack or stroke (Law, Wald, & Morris, 2003). However, because lowering blood pressure is good for decreasing the risk of cardiovascular events even among groups without high blood pressure, and because the cost of interventions used to lower blood pressure are inexpensive and beneficial, it makes good sense to implement such interventions broadly (for those above a certain age) regardless of their blood pressure reading (Law, Wald, & Morris, 2003).

[5] Note that IAT scores can still inform the evidence base used to characterize individuals; however, they may be best used fruitfully towards attributing, not implicit attitudes, but courser-grained attitudes like aversive racism (Lee, forthcoming; Dovidio & Gaertner, 2004).

Conclusion

In this paper, I staked out a stakeholder conception of implicit bias. This position responds to two important challenges. First, it navigates debates about whether implicit attitudes should be conceptualized as mental states or as dispositions by appealing to a form of epistemic modesty that is savvy to disagreement and growth in psychological research. This allows stakeholder injunctions (to attend to implicit bias in scientific communities) to be resilient to finer-grained forms of scientific debate and progress that would not impact the ultimate soundness of this call. Second, it recognizes real strengths and weaknesses in the psychometric evidence for implicit attitudes by understanding tests for implicit attitudes (in particular, the IAT) as tools for characterizing group-level rather than individual-level behavior. This position raises rich and pressing philosophical questions about how we should understand the relationship – metaphysical, epistemic, and ethical – between collective implicit attitudes and individual-level attitudes.

Acknowledgments

I am grateful for Royalty Research Fund Grant #A79071 (University of Washington) which afforded research time to complete this article. For illuminating comments and/or conversations, many thanks to Liam Kofi Bright, Michael Brownstein, Anthony Greenwald, Edouard Machery, Jennifer Nagel, Eric Schwitzgebel, Yuchi Shoda, Samuel Wang, the reviewers, and the audience at the 2015 Central American Philosophical Association meeting where I first presented some of these ideas. I am not writing in my capacity as a contractor for the U.S. National Institutes of Health.

References

Allport, G. W. (1935). Attitudes. In C. Murchison (Ed.), *A Handbook of Social Psychology*. Worcester, MA: Clark University Press.

Anderson, J. R. (1976). *Language, Memory, and Thought*. Hillsdale, NJ: Erlbaum.

Barsalou, L. W. (1990). On the indistinguishability of exemplar memory and abstraction in category representation. In T. K. Srull & R. S. Wyer (Eds.), *Advances in Social Cognition, Volume III: Content and Process Specificity in the Effects of Prior Experiences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Bechtel, W. (2005). The challenge of characterizing operations in the mechanisms underlying behavior. *Journal of the Experimental Analysis of Behavior, 84*, 313-325.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061-1071.

Brauer, M., Er-rafiy, A., Kawakami, K, & Phills, C. E. (2012). Describing a group in positive terms reduces prejudice less effectively than describing it in positive *and* negative terms. *Journal of Experimental Social Psychology*, *48,* 757-761.

Brownstein, M., & Saul, J. (Eds.). (2016a). *Implicit Bias & Philosophy, Volume 1: Metaphysics and Epistemology.* Oxford, UK: Oxford University Press.

Brownstein, M., & Saul, J. (Eds.). (2016b). *Implicit Bias & Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics.* Oxford, UK: Oxford University Press.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*, 281-302.

Cummins, R. (1983). *The Nature of Psychological Explanation*. Cambridge, MA: The MIT Press.

Cummins, R. (2000). "How does it work?" vesus "What are the laws?" Two concepts of psychological explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and Cognition*. Cambridge, MA: The MIT Press.

Cunningham, W. A., Preacher, K. J., & Banaji, M. R. (2001). Implicit attitude measures: Consistency, stability, and convergent validity. *Psychological Science, 12,* 163-170.

Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. *Advances in Experimental Social Psychology, 47*, 233-279.

De Houwer, J., Gawronski, B., & Barnes-Holmes, D. (2013). A functional-cognitive framework for attitude research. *European Journal of Social Psychology, 24*, 252-287.

De Houwer, J., Geldof, T, & De Bruycker, E. (2005). The Implicit Association Test as a general test of similarity. *Canadian Journal of Experimental Psychology, 59,* 228-239.

Dovidio, J. F., & Gaertner, S. L. (2004). Aversive racism. *Advances in Experimental Social Psychology*, *36*, 1-52.

Egan, A. (2011). Comments on Gendler's "The epistemic costs of implicit bias." *Philosophical Studies*, *140*, 47-63.

Eagly, A. H., & Chaiken, S. (2007). The advantages of an inclusive definition of attitude. *Social Cognition*, *25,* 582-602.

Fazio, R. H. (2007). Attitudes as object-evaluation associations of varying strength. *Social Cognition*, *25,* 603-637.

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge, U.K.: Cambridge 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*, 692-731.

Gawronski, B., & Bodenhausen, G. V. (2007). Unraveling the processes underlying evaluation: Attitudes from the perspective of the APE model. *Social Cognition*, *25,* 687-717.

Gawronski, B., Deutsch, R., LeBel, E. P., & Peters, K. R. (2008). Response interference as a mechanism underlying implicit measures: Some traps and gaps in the assessment of mental associations with experimental paradigms. *European Journal of Psychological Assessment, 24,* 218-225.

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

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gilbert, M. (1989). *On Social Facts*. Princeton, NJ: Princeton University Press.

Greenwald, A. G. (2012). There is nothing so theoretical as a good method. *Perspectives on Psychological Science, 7,* 99-108.

Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology, 108*, 553-561.

Greenwald, A. G., & Nosek, B. A. (2008). Attitudinal dissociation: What does it mean? In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the New Implicit Measures*. Hillsdale, NJ: Lawrence Erlbaum.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, *97*, 17-41.

Holroyd, J. (2016). What do we want from a model of implicit cognition? *Proceedings of the Aristotelian Society*, *116*, 153-179.

Huebner, B. (2016). Implicit bias, reinforcement learning, and scaffolded moral cognition. In M. Brownstein & J. Saul (Eds.), *Implicit Bias & Philosophy, Volume 1: Metaphysics and Epistemology.* Oxford, UK: Oxford University Press.

Jacoby, J., & Sassenberg, K. (2011). Interactions do not only tell us *when*, but can also tell us *how*: Testing process hypotheses by interaction. *European Journal of Social Psychology*, *41*, 180-190.

Karpinski, A., & Hilton, J. L. (2001). Attitudes and the Implicit Association Test. *Journal of Personality and Social Psychology*, *81*, 774-788.

Krosnick, J. A., Judd, C. A., & Wittenbrink, B. (2005). The measurement of attitudes. In D. Albarracin, B. T. Johnson & M. P. Zanna (Eds.), *Handbook of Attitudes and Attitude Change*. Mahwah, NJ: Erlbaum.

Kwong, J. M. (2012). Resisting aliefs: Gendler on belief-discordant behaviors. *Philosophical Psychology, 25*, 77-91.

Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science, 48*, 19-49.

Law, M., Wald, N., & Morris, J. (2003). Lowering blood pressure to prevent myocardial infarcation and stroke: A new preventive strategy. *Health Technology Assessment*, *7* (31).

Lee, C. J. (2007). The Representation of Judgment Heuristics and the Generality Problem. *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1211-1216). Austin, TX: Cognitive Science Society.

Lee, C. J. (Forthcoming). A dispositional account of aversive racism. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*. Madison, WI: Cognitive Science Society.

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs, 49,* 800-823.

List, C., & Pettit, P. (2011). *Group Agency: The Possibility, Design, and Status of Corporate Agents*. New York, NY: Oxford University Press.

Machery, E. (2007). Concept empiricism: A methodological critique. *Cognition*, *104*, 19-46.

Machery, E. (2009). *Doing Without Concepts*. Oxford, U.K.: Oxford University Press.

Machery, E. (2016). De-Freuding implicit attitudes. In J. Saul & M. Brownstein (Eds.), *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*. Oxford, U.K.: Oxford University Press.

Maddux, W. W., Barden. J., Brewer, M. B., & Petty, R. E. (2005). Saying no to negativity: The effects of context and motivation to control prejudice on automatic evaluative responses. *Journal of Experimental Social Psychology*, *41*, 19-35.

Mandelbaum, E. (2013). Against alief. *Philosophical Studies, 165,* 197-211.

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *49,* pp. 629-658.

National Academies of Science. (2015). Implicit bias workshop with Anthony Greenwald and Brian Nosek. Retrieved from http://sites.nationalacademies.org/PGA/cwsem/PGA_173396.

Olson, M. A., & Fazio, R. H. (2003). Relations between implicit measures of prejudice: What are we measuring? *Psychological Science, 14*, 636-639.

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

Payne, B. K. (2005). Conceptualizing control in social cognition: How executive functioning modulates the expression of automatic stereotyping. *Journal of Personality and Social Psychology, 89*, 488-503.

Petit, P. (2007). Responsibility incorporated. *Ethics, 117*, 171-201.

Petit, P. (2011). *A Theory of Freedom: From the Psychology to the Politics of Agency*. New York, NY: Oxford University Press.

Pinholster, G. (2016). Journals and funders confront implicit bias in peer review. *Science*, *352*, 1067-1068.

Rudman, L. A., & Kilianski, S. E. (2000). Implicit and explicit attitudes toward female authority. *Personality and Social Psychology Bulletin, 26*, 1315-1328.

Saul, Schwarz, N. (2007). Attitude construction: Evaluation in context. *Social Cognition, 25,* 638-656.

Schwitzgebel, E. (2001). In-between believing. *The Philosophical Quarterly, 51*, 76-82.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Noûs, 36,* 249-275.

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

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In N. Nottelmann (Ed.), *New Essays on Belief: Constitution, Content, and Structure*. New York: Palgrave MacMillan.

Spencer, S. J., Zanna, M. P., & Fong, G. T. (2005). Establishing a causal chain: Why experiments are often more effective than mediational analyses in examining psychological processes. *Journal of Personality and Social Psychology, 89*, 845-851.

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Personality and Social Psychology Review, 8,* 220-247.

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

Wittenbrink, B., Judd, C. A., & Park, B. (2001). Evaluative versus conceptual judgments in automatic stereotyping and prejudice. *Journal of Experimental Social Psychology, 37*, 244-252.

Zanna, M. P., & Fazio, R. H. (1982). The attitude-behavior relation: Moving toward a third generation of research. In M. P. Zanna, E. T. Higgins, & C. P. Herman (Eds.), *Consistency in Social Behavior: The Ontario Symposium*. Hillsdale, NJ: Erlbaum.