The Structure of Bias

Gabbrielle M Johnson

New York University

Abstract

What is a bias? Standard philosophical views of both implicit and explicit bias focus this question on the representations one harbors, e.g., stereotypes or implicit attitudes, rather than the ways in which those representations (or other mental states) are manipulated. I call this approach *representationalism*. In this paper, I argue that representationalism taken as a general theory of psychological social bias is a mistake because it conceptualizes bias in ways that do not fully capture the phenomenon. Crucially, this view fails to capture a heretofore neglected possibility of bias, one that influences an individual's beliefs about or actions toward others, but is, nevertheless, nowhere represented in that individual's cognitive repertoire. In place of representationalism, I develop a functional account of psychological social bias that characterizes it as a mental entity that takes propositional mental states as inputs and returns propositional mental states as outputs in a way that instantiates social-kind inductions. This functional characterization leaves open which mental states and processes bridge the gap between the inputs and outputs, ultimately highlighting the diversity of candidates that can serve this role.

1 Introduction

Over the last half century, our concept of *social bias* has changed dramatically. In the early 1970s, it seemed obvious to many that psychological social biases were constituted by consciously accessible mental representations in the form of stereotypes.[1] This was an understandable assumption in an era when people wore many of their social prejudices on their sleeves. However, over the last twenty years, this standard assumption has been abandoned. Research methods have developed to reveal psychological sources of discriminatory behavior that are not obvious to either researchers or the individuals they study.[2] This research has paved the way for recognizing a new species of social bias, one that involves *unconscious* attitudes that give rise to discriminatory behaviors, without the awareness or control of the individual who harbors such attitudes. Researchers call these *implicit biases*, reserving the term *explicit biases* for those biases that are readily consciously accessible.

In this paper, I develop a concept of psychological social bias that is specially designed to incorporate not only these known kinds of bias, but also other forms of bias so far neglected in the literature. In order to make progress on a theory of a general notion of psychological social

[1] Banaji and Greenwald (2013, 170-184) credit this to the "widespread and overt racism before 1950" and its metamorphosis into a "covert, less detectable form" of racial bias. (See also Banaji and Greenwald 1994 and Dovidio et al. 2005.) Kovel (1970) presents one of the first accounts of this latter form of bias with his theory of *aversive racism*. (See Gaertner and Dovidio 1986 and Dovidio and Gaertner 2004 for discussion.)

[2] See Greenwald and Banaji 1995, 4 for a brief historical guide. Measures of unconscious bias include the Implicit Association Test (IAT) (Greenwald et al. (1998)), semantic priming (Banaji and Hardin 1996), evaluative priming (Fazio et al. 1995), the Go/No-Go Associations Task (Nosek and Banaji 2001), the Sorting Paired Features Task (Bar-Anan et al. 2009), the Weapons Identification Task (Correll et al. 2002), the Extrinsic Affective Simon Task (De Houwer 2003), and the Affect Misattribution Procedure (Payne et al. 2005).

bias (hereafter, simply 'bias'), I begin with some preliminary observations about explanations involving it. Uncontroversially, biases influence an individual's beliefs about or actions toward others on the basis of their presumed membership in a social group. Theories of explicit bias handle this explanatory datum by postulating *conscious stereotypes*—generalizations accessible via introspection and potentially offered as rationale for one's discriminatory behavior. However, such an analysis cannot account for cases of implicit bias. A parsimonious way to expand the analysis is to keep intact the idea that biases involve representational attitudes (e.g., stereotypes) and merely strip away the feature that now appears inessential: conscious accessibility. The result is a strategy that characterizes biases as constituted by mental representations (explicit or implicit)—I call this approach *representationalism about bias* because it equates bias with a type of mental representation.[3]

In the first half of this paper, I argue that representationalism about bias (hereafter, simply 'representationalism') is a mistake because it conceptualizes bias in ways that do not fully capture the phenomenon. Just as the concept of bias turned out not to require conscious access, so too, I contend, they need not be representational at all. In fact, there are arguably biases that influence an individual's beliefs about or actions toward other people, but are, nevertheless, nowhere represented in that individual's cognitive repertoire. I call this type of bias *truly implicit bias*, and it is a counterexample to representationalism.

In the second part of the paper, beginning with Section 4, I present my positive proposal: a functional theory that explicates the concept of bias. This general theory treats bias as consisting of mental entities that take propositional mental states as inputs and return propositional mental states as outputs in a way that instantiates social-kind inductions. This functional characterization leaves open which mental states and processes bridge the gap between the inputs and outputs, ultimately highlighting the diversity of candidates that can serve this role.

The focus of this paper is a theory of the nature of psychological social bias. It attempts to provide what Carnap (1950) called an *explication*, as opposed to a logical analysis of our ordinary notion of social bias that would likely evade necessary and sufficient conditions. I aim to introduce a new, precise concept that furthers our understanding of a genuine psychological kind. This explication makes clear the role that the concept *bias* plays in psychological explanations. This account is best suited to capture the properties of psychological social bias that I believe secure its place within a broader unified kind—psychological bias more generally. The primary unifying feature of this general notion of bias (which social bias shares) is that it is a necessary response to underdetermination. The problem of underdetermination is, as Louise Antony describes, "the largest epistemic challenge facing any finite knower."[4] Our access to the external world and the mind-independent realities that accompany it is mediated by our access to a finite amount of evidence. This evidence underdetermines the way the world actually is and, likewise, is consistent with an indefinite number of possible theories or conclusions we could draw about it. Psychological biases are our solution to this underdetermination problem. They serve as ways "to reduce hypothesis space to a tractable size."[5] They bridge the otherwise limitless inductive gap that exists between evidence and theory. This gap is well studied in psychological accounts of visual perception.[6] However, it exists in all domains in which we attempt to gain information about the world, including

[3] Proponents of representationalism about implicit bias include Mandelbaum (2015), Carruthers (2017), De Houwer (2014), and Levy (2015). Alternative to standard representationalist views among theories of implicit bias are *associationist views*, which characterize them merely in terms of associations between solitary conceptual representations. In §4.2.2, I demonstrate how the standard associationist view can be assimilated into my functional account.

[4] Antony 2016, 161. See also Antony 2000 and Antony 2001.

[5] Antony 2016, 161.

[6] Burge 2010a, 90-92, 344-345.

the social domain. Like visual psychology, social psychology is fundamentally concerned with veridicality conditions. The aim of social cognitive science is, at least in part, to explain how accurate social representations are formed when they are (i.e., what basic mechanisms underwrite accurate representation of our social environment) and how inaccurate social representations are formed when they are (i.e., how these basic mechanisms go awry). The superficial properties that we glean from observation vastly underdetermine the social categories to which individuals belong and the social properties they in fact exhibit. Just as we rely on visual perceptual heuristics to navigate our physical environment, so too we rely on social heuristics to navigate our social environment.[7]

Other theories of this psychological notion of bias, such as those from which I've extrapolated the representationalist strategy, are typically empirical theories of particular cases, and therefore have little to say about the general phenomenon. As I argue, the conceptual explication of the general kind psychological social bias is best undertaken at the functional level. Only by adopting a functionalist conception can we capture the real unity among the diversity that empirical theories of psychological bias evince.

2 Starting Points for a Theory of Bias

I begin by presenting a standard example of *explicit bias* that highlights the central fact about bias: biases influence beliefs about or actions toward an individual on the basis of that individual's presumed membership in a social group. I use this example as a core case in order to sketch a naive theory that treats explicit bias as the paradigmatic case. I then demonstrate how implicit bias prompts revision to this account, resulting in what I call *representationalism*.

2.1 Explicit Bias and Terminology

Imagine a fellow academic (call him 'E') attempts to help his colleague Jan join a Skype interview because he believes she is bad with computers. When asked why he believes this, he explains that it's because Jan is elderly and that elderly people are bad with computers. Here, the inference E is making is straightforward:

(i) Jan is elderly.
(ii) Elderly people are bad with computers.
$\therefore$ (iii) Jan is bad with computers.

This case clearly involves an *explicit bias*: E is completely aware that he is drawing conclusions about Jan based on his beliefs about the elderly and Jan's belonging to that group.[8]

Although it's clear there's a bias here, it's less clear which mental states and actions correspond to *the bias*. The term 'bias' is often ambiguous between a stereotype belief, the conclusion of some inference involving a stereotype belief, and the behavior based on the conclusion of some inference involving a stereotype belief.[9] To keep these components distinct, I call the belief about a particular person on the basis of which a discriminatory judgment is formed—in this case, E's belief that Jan is elderly—*the bias-input*. Next, I call the collection of states and processes that—in tandem with the bias-input—cause a discriminatory judgment *the bias-construct*; the bias-construct in E's case is his stereotype belief that elderly people are bad with computers (together with whatever inferential

[7] For interesting work on the relationship between social biases and visual perceptual biases, see Munton 2017, 2019a,b, and between social biases and cognitive biases more generally, see Antony 2001, 2016.

[8] Calling him 'E' helps track that his case involves an *explicit* bias.

[9] See Holroyd and Sweetman 2016, 81-82.

processes are necessary to derive the conclusion). I call the discriminatory judgment that bias-constructs and bias-inputs together cause—like E's belief that Jan is bad with computers—*the bias-output*. Finally, I call actions that are performed on the basis of bias-outputs—like E's trying to help Jan with the Skype interview—*bias-actions*. I use the notion of a *mental construct* to pick out an open-ended collection of mental states and processes.[10] Importantly, a mental construct in this sense need not be constituted by mental representations, although in the case of explicit bias, it is (more in §2.2).

Social psychologists and philosophers investigating bias-constructs standardly regard them as involving *stereotypes*. The definition of 'stereotype', however, is not a straightforward matter. For our purposes, we can stick with a standard textbook definition: a stereotype is "a set of cognitive generalizations (e.g., beliefs, expectations) about the qualities and characteristics of the members of a group or social category."[11] On this view, stereotypes are beliefs about members of social groups that take the form of generalizations. Following standard representationalist views about belief, I regard them as propositionally structured mental states. Since my focus is on the representational components of mental states, I set aside affective aspects of bias states (such as prejudices).[12]

Finally, it is necessary to introduce the concept of a *contrast social group*. A contrast social group is distinct from the relevant *target social group*, namely the social group referenced in the stereotype, but the two relate to each other along some salient dimension. In the case above, E wouldn't have a bias against the elderly if it turned out that he believed that the young, the elderly, and everyone in between were bad with computers and, on the basis of this, concluded that everyone he interacted with needed help with their computers. Instead, E needs to treat the target group (the elderly) differently from how he treats the contrast group (young people). This notion of *differential treatment* (in thought or behavior) is crucial to understanding the operation of social bias.[13]

With these theoretical distinctions at our disposal, here's a naive characterization of this core case of *psychological social bias*: a person P has a bias toward a target social group G if and only if on the basis of a consciously accessible stereotype that P harbors toward G, P forms different conclusions about individuals he or she regards as belonging to G than those conclusions he or she forms about individuals regarded as belonging to some contrast social group H. According to this first-pass analysis, social bias just is explicit bias. In what follows, I explore how this picture must be amended in light of the existence of implicit biases.

2.2 From Implicit Bias to Representationalism

Not all biases operate in ways that are obvious to the agents harboring them. Consider another core case similar to E's in which a different colleague, T, also considers Jan elderly.[14] Like E, T assumes that Jan needs assistance joining a Skype interview, and she does not make this assumption about younger colleagues. However, unlike E, T appears to lack any conscious stereotype that elderly

[10] This is roughly what Greenwald and Nosek (2008) and Machery (2017) have in mind with their uses of 'construct'. This is also the central mental entity that literature on implicit bias is generally about.

[11] VandenBos 2015, 1031. For more philosophical treatments, see Beeghly 2015 and Blum 2004.

[12] I'm forced to oversimplify a vast literature on these topics. For more on the distinction between stereotypes and prejudices in psychology and philosophy, see Allport 1979, Banaji and Greenwald 1994, Greenwald and Banaji 1995, Banaji and Hardin 1996, Madva and Brownstein 2018, and particularly Machery 2016, 105-110.

[13] The importance of this notion of differential treatment can be seen in the structure of tests for bias: both direct and indirect measures are designed to reveal differential treatment. It also allows for important structural similarities to models of discrimination in ethics and law (Lippert-Rasmussen 2014, 15-16). Thanks to Alex Madva and Gabe Dupre for drawing my attention to these points.

[14] 'T' will help highlight the possibility that her bias is *truly implicit*.

people are bad with computers. In fact, if you asked T, she would deny the claim and assert instead that elderly individuals are just as good with computers as anyone else is. (Let's assume we also lack any obvious alternative explanations for T's assumption, e.g., that Jan admitted to her that she's bad with computers.)

This is a standard picture of an implicit bias at play. Given the difference between the two cases, it becomes difficult to explain what prompted T's assumption. She seems to share the following mental representations (or beliefs) with E:

(i) Jan is elderly
(iii) Jan is bad with computers

It seems that, in order to explain T's behavior, we must posit the existence of a mental entity, which is non-obvious to T herself, and that plays the same role for her as (ii) does for E, i.e., mapping (i) onto (iii). The task for theories of implicit bias is to give a proper account of the relevant mental entity.

Perhaps the most natural way to deal with implicit biases is to say that they're exactly like explicit biases save that the stereotype they involve is not necessarily conscious, i.e., the bias-construct for an implicit bias is an *explicitly represented* but *consciously inaccessible* stereotype. Saying that a bias-construct is consciously accessible is typically taken to mean that the stereotype involved has the potential to be "drawn up" (to use an imprecise metaphor) to consciousness, causing the person harboring the state to use it in a variety of ways, such as in reasoning, in rationally guiding action, and, most notably, in reporting to others the content of the state.[15] Alternatively, if a mental state is not consciously accessible, then the individual cannot report on it. The lack of available reports is evidence against a state's being consciously accessible. Of course, it's also evidence that such a state is lacking altogether. Thus, we can't infer from a failure of reportability that a state exists unconsciously without some additional evidence for its existence.

In the case of bias, individuals sometimes exhibit behaviors that serve as evidence for the existence of a bias, but they are unable to report on the bias or appear surprised when confronted with the evidence. Greenwald and Banaji (2013, 56-58), for example, discuss the "disturbed" feeling subjects report when confronted with evidence from the Implicit Association Test (IAT). Some of the most compelling examples they present include a gay activist who finds out he harbors negative associations toward the gay community and a writer whose mother is Jamaican finding out he harbors pro-White biases, who state that the revelations were "creepy," "dispiriting," and "devastating."[16] Unfortunately, data alleging the conscious inaccessibility of bias are often anecdotal, and competing notions of conscious accessibility within philosophy of psychology make it especially difficult to adjudicate these and other data either in favor or against the claim that implicit biases are in principle consciously inaccessible.[17] This issue is made especially fraught by ambiguity and confusion with respect to the two central concepts this claim involves: *bias* and *conscious accessibility*. This paper is an attempt to resolve some ambiguities regarding the first notion, by distinguishing between different aspects of a bias's operation and drawing theoretical attention to the bias-construct rather than various bias-outputs, e.g., "spontaneous affective reactions."[18] Ambiguities regarding the second notion (conscious accessibility) are well documented

[15] See Block (1995)'s notion of *access consciousness*.

[16] For empirical evidence of the surprised or defensive responses subjects often exhibit when presented with their results on indirect measures, see Hillard et al. 2013, Howell et al. 2015, and Howell and Ratliff 2017.

[17] There has been a recent surge in literature challenging the idea that the mental constructs that give rise to results on indirect measures are in fact consciously inaccessible, most notably Gawronski and Bodenhausen 2006, 2014, Monteith et al. 2001 Hahn et al. 2014, Hahn and Gawronski 2014, Hahn and Gawronski 2019, and Rivers and Hahn 2018.

[18] Hahn and Gawronski 2019.

in contemporary philosophy of mind.[19] Fortunately, debates about the conscious accessibility of bias need not be resolved here.[20] The explication I provide concerns the representational nature of bias, and thus we can remain agnostic as to whether implicit biases are ultimately in principle consciously inaccessible.

Ultimately, the strategy of regarding implicit biases as exactly like explicit biases, save for implicit biases involve stereotype beliefs that are not necessarily consciously accessible, reflects a background commitment to *representationalism* in the theory of bias. On this view, we assume that—despite her assertions to the contrary—T has a mental state with the content (*ii*) elderly people are bad with computers.[21] This follows the now commonplace thought that implicit and explicit biases differ principally along the dimension of conscious accessibility.[22] However, as I argue in the next section, representationalism fails to capture an additional, largely neglected dimension of difference: that of the bias's *representational status*.

3 Against Representationalism

In this section, I argue for the possibility of biases that are representationally implicit, with no belief-like representation at the core of the bias-construct. I call these *truly implicit biases*.[23]

3.1 What Is Truly Implicit Bias?

The notion of *representationally implicit* or *merely encoded* content is not new to psychological theories. In visual psychology, for example, theorists often describe the operation of various transformations within the visual system as abiding by rules or principles that are not ascribed to or represented in the individual.[24] One example is the visual system's ability to ascertain the location (that is, the distance and direction) of an object in the environment based on the convergence of the two lines of sight from the eyes. To do this, the visual system utilizes facts about the geometry of binocular vision to determine an approximate location. When an individual focuses on a distal stimulus, proprioceptive receptors in the muscles surrounding the eye provide information about the vergence angle, or the angle at which the two lines of sight converge. The visual system then uses this information, coupled with the constant distance between the two eyes, to compute the distance

[19] Block 1995, 231. See also debates about unconscious perception in visual psychology (Peters et al. 2017, Lau 2008) and philosophy (Phillips 2015, Phillips and Block 2017).

[20] Doing so will likely require more rigorous empirical work than has been attempted in social psychology by proponents on either side of the issue. Ultimately, such matters require more extensive treatment than I can offer here, and so are taken up thoroughly in my paper 'Are Implicit Biases Implicit Biases?' MS.

[21] Mandelbaum (2015, 7) himself takes up such a view in his *Structured Belief Hypothesis* "that implicit bias is underwritten by unconscious beliefs. These beliefs . . . are honest-to-god propositionally structured mental representations [of the form] BLACK MALES ARE DANGEROUS." Other defenders of representationalist views about implicit bias include Carruthers (2017), De Houwer (2014), and Levy (2015), though they do not endorse drawing a distinction between implicit and explicit along the lines of conscious accessibility.

[22] In what follows, I will not take conscious inaccessibility as settled; however, I will assume that it represents an important and appealing position within theories of implicit bias because this best captures the historical progression in the literature. For these reasons, I adopt the mainstream tendency to refer to "implicit biases" as those that differ from "explicit biases" with respect to their conscious accessibility, but I flag where this assumption might matter for the arguments I present. Thanks to an anonymous referee for pushing me to be more explicit about my commitments regarding conscious accessibility and their impact on my explication of bias.

[23] Holroyd et al. (2017, 3-7) survey the standard interpretations of 'implicit' in the moniker 'implicit bias'. Among these include implicit as unconscious, as beyond control, as dissonant/unendorsed, as accessed by certain kinds of measure, and as discursively useful. None of these interpretations deal exclusively or directly with the interpretation of it as truly implicit, i.e., non-representational.

[24] See Burge 2010a and, for socially-laden visual transformations, Munton 2019a.

and direction of the distal stimulus from the midpoint between the pupils, or the cyclopean eye. In the case of convergence, the calculation of distance and direction follows a complex geometric algorithm. This algorithm is carried out without the representation of the algorithm itself. Instead, the algorithm is *merely encoded* in the "functional architecture" of the visual system.[25]

States or processes belonging to the *functional architecture* are non-representational, though they might perform operations on representational states.[26] They are embodied in the structure of the processing system, and they correspond to a system's capacity to process information in certain ways and to achieve certain goals. The convergence algorithm is an example: it corresponds to a capacity to produce veridical representational outputs concerning location on the basis of (underdetermining) informational inputs.

Theories according to which some types of content are representationally implicit have become familiar in non-perceptual cases as well. Carroll (1895)'s famous discussion of what the tortoise said to Achilles demonstrates that some rules of inference must be representationally implicit, otherwise we run the risk of an infinite regress.[27]

I propose that certain stereotype-like contents can also be encoded by embedding them in non-representational transformation rules. These contents are representationally implicit, i.e., truly implicit. The claim that there can exist content without representation may appear to run afoul of the computationalist orthodoxy that there are no mental contents without explicit tokenings of mental representations.[28] However, given that the relevant content, on my theory, is always embedded in a *rule*, it's more accurate to consider it an implicit aspect of the relevant rule and not the content of a self-standing state. Importantly, these implicit aspects of rules are standardly regarded as being inaccessible to other mental operations. Thus, the possibility that implicit biases are truly implicit has explanatory purchase: it explains why implicit biases are consciously inaccessible, automatic, and difficult to revise, since conscious access, cognitive control, and deliberative capacities presumably all operate on states that are independently accessible to multiple computational processes.[29]

Here's another example. In computer programing, values that are built directly into the computational rules are called *literals*. Literals built into one computational procedure are inaccessible to others. To adapt an example from Gallistel and King (2009, 151), a Turing machine that computes the doubling function need not have the value *2* explicitly represented on the tape, i.e., in memory. Typically, such symbols recorded on the tape are regarded as "vehicles" of representation, and a content's corresponding to a vehicle is what constitutes its being explicitly represented. In this case, the value *2* needn't have a corresponding vehicle. Rather, this value is encoded in the operation of the machine itself. Therefore, "the machine's 'knowledge' of this number is of the opaque implicit

[25] Devitt (2006, 50) presents another example of a baseball player running to catch a fly ball. Of this, he says "[players] surely don't manage [the task] by *representing* the algorithm for being in the right place and applying it to a series of representations of the acceleration of the tangent of the angle" (emphasis in original).

[26] See, for example, Pylyshyn 1991, 14. Lande (2018) makes the argument that these architectural features account for the perspectival character of perception.

[27] Three other examples include moral inference (Horgan and Timmons 2007), hypothesized *Bare Inferential Transitions* (Quilty-Dunn and Mandelbaum 2017b), and internalized generative grammar rules (Stabler 1983; Chomsky 1965).

[28] Fodor 1987, 25. More precisely, Fodor states that tokenings of attitudes that occur as "episodes" in mental processes "*must* correspond to tokenings of mental representations" (emphasis in original). For reasons to follow, my view isn't obviously in tension with this more precise dictum.

[29] Of course, it's always possible that a sufficiently reflective thinker might be able to deduce the computational rules a system abides by. For these reasons, even truly implicit content (or truly implicit bias) can—in some indirect way—come to be consciously accessible. Thus, this theoretical point isn't directly undermined by recent empirical evidence for the conscious accessibility of bias (see the discussion of conscious accessibility in §2.2 above and the discussion of Smith and Zarate 1990 in footnote 45 below).

kind; the machine cannot gain computational access to this information outside of the state into which it is built."[30] The computationalist orthodoxy requires only that content available for further processing be explicitly tokened. Thus, my view, which treats truly implicit content as being built into the procedure in the same way the value *2* is built into the doubling machine, doesn't commit itself to any truly implicit content being available for independent processing. Therefore, it not only leaves the computationalist orthodoxy intact, but also renders expected the empirical observations regarding conscious accessibility, automaticity, and recalcitrance.

Contents that are representationally implicit can be identified by their *emergent* qualities.[31] Although the contents are never explicitly tokened, the rules they instantiate still seem present in some abstract way: they *emerge* from the combination of the other explicitly represented and non-represented principles in operation along with the functional hardware of the machine.[32] This highlights one possible strategy for characterizing truly implicit content: truly implicit content is that which would be captured in a description of the operation of the machine in describing the transformation principles between particular contents.[33]

Up to this point, I've been characterizing the explicit-implicit distinction in terms of representational vehicles, so as to accord with contemporary computational theories of mind. But the distinction generalizes. Suppose one is a functionalist about all mental states. An instance of such a functionalism would be one that denies explicit representation requires *vehicles*, but instead only corresponds to a particular kind of role. One might worry that on such a view the implicit-explicit distinction collapses: the sorts of architectural facts I've just isolated *just are* (or fulfill) the functional conditions for being an explicit representational content. If so, this would eliminate my alleged difference between explicit stereotype beliefs and truly implicit bias. However, even adopting this form of functionalism, there is reason to capture an implicit-explicit distinction in the putative functional roles we assign to each kind of state. One canonical role for explicit representational states has already been discussed: they exhibit a kind of informational promiscuity that allows their content to be available for wide and systematic use in other computational processes.[34] As I've discussed, this causal role is absent in putative cases of implicit content. In sum, even on a functionalist picture such as this, truly implicit biases will not be assimilated to explicit representational states like beliefs, because their functional role will differ in important respects.[35]

My claim that truly implicit social biases are possible implies that the stereotype-like principles that guide our reasoning about others—such as the "rule" that elderly people are bad with computers—could fail to be explicitly represented, but nonetheless accurately describe the operation of an individual's reasoning about others. At this point, we've seen how distinctions between explicitly represented and implicitly represented contents can be made, and how those distinctions apply in the case of social bias, demonstrating the generality of the view I'm advancing. One might

[30] Gallistel and King 2009, 151.

[31] In this way, they're similar to Dennett (1981, 107)'s notion of *emergent content*, which he demonstrates with his famous example of a chess-playing program that abides by the rule "get my queen out early."

[32] Dennett would never identify his emergent content with content computational theory of mind proponents posit within the functional architecture. More on this in §4.2.1.

[33] Crucially, it's never the case that these computational principles have to be explicitly represented. Fodor (1987, 25) concedes that certain principles including "get my queen out early" that guide the transformations of mental contents can *either* be fully represented or not fully represented. See also Cummins 1982 and Pylyshyn 1980.

[34] Importantly, this is a claim about the capacities that correspond to states like beliefs: that they be capable of entering into inferences more generally. It needn't be the case that any particular belief content always or even often exercise this capacity.

[35] Of course, biases *can* be explicit. I merely claim that they need not be. My view is ultimately congenial to a functionalist analysis of what bias-constructs—explicit and implicit—have in common. Such an analysis must occupy a higher level of abstraction than the analysis provided by the functionalist about explicit representation, as discussed in §4.1.

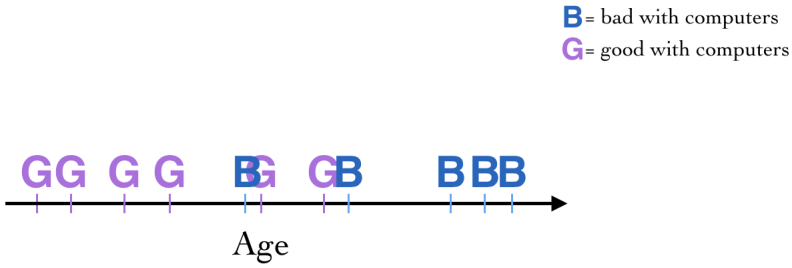
wonder, however, how a truly implicit bias could come to be naturally instantiated in human cognition or whether we have reason to think they in fact do. In the next section, I address both of these concerns by presenting examples of classification tasks that both exhibit truly implicit bias and provide a model of how humans arguably harbor such biases.

3.2 Truly Implicit Bias: An Illustrative Example

The point that biases could be truly implicit loses its explanatory force without an example of how this sort of bias might plausibly be instantiated in humans.[36] In what follows, I present a case borrowed from computer science that demonstrates a classification model whose operations involve truly implicit biases. The case is primarily intended as a proof of concept. However, given that the classification algorithm outlined below is widely adopted in cognitive science and empirical psychology in the modeling of exemplar-based classification, it also serves as an empirical argument for the existence of truly implicit biases. I turn to this point at the end of the section, ultimately demonstrating that truly implicit bias is neither conceptually nor empirically far-fetched.

The task of a machine learning classification program is to categorize novel inputs correctly. It does this by being exposed to many pre-categorized objects and "learning" the relationship between the features of those objects and the category to which they belong. Once it has adequately refined a predictive model that assigns category estimates to various combinations of features, it gets applied to new data. For each novel input, it looks for the familiar combinations of features and categorizes the input appropriately.

For example, imagine that an engineer is creating a program that classifies individuals as good or bad with computers. Let's say she thinks one relevant feature for determining this classification is a person's age.[37] She begins stage one by training her program on pre-labeled *training data*. These include many instances of individuals already categorized with respect to being good or bad with computers. We can represent the relationships between the relevant properties and classifications by plotting the data on a one-dimensional *features space* as follows:



Each data-point specifies two things: the *feature value*, i.e., value corresponding to the specific observed property of the data-point, in this case, age; and a *class label*, i.e., a label denoting the class of objects to which that data-point belongs, in this case, being good or bad with computers.

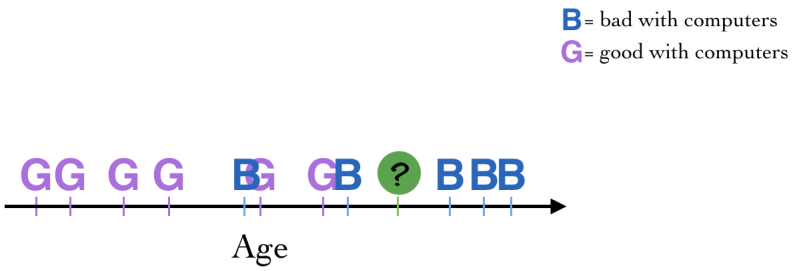
As we see from the graph above, the data in this example are skewed: those individuals that are bad with computers are clustered near the end of the line, while those who are good with computers are near the beginning. There are many ways for the test data to come to be patterned like this. For example, maybe the programmer pulled training data from the library on the same day it was

[36] See, for example, Carruthers (2017)'s footnote 8.

[37] The problem of which properties, i.e., features, are most relevant for some classification test is an interesting and complicated question from a computer science perspective that is unfortunately beyond the purview of this paper. A similar case using a classification of objects as skis or snowboards based on two features, length and width, is discussed by Daumé III (2015, 30-32).

hosting an after-school event for tech-savvy high-schoolers to provide social media training to elderly individuals who struggle with computer technology. In this case, we would expect her sample to be disproportionately filled with young people who are good with computers and elderly individuals who are bad with computers compared to the general population.[38] This example demonstrates an important lesson: a machine learning program is only as good as the data on which it is trained, giving rise to the oft-cited motto "garbage in, garbage out." If the data going into the training period are biased, then we can expect the generalizations the program makes based on those data to be biased as well.

This is precisely what we see in the next phase of the algorithm when it is applied to new, unclassified *test data* and aims to classify each datum as either good or bad with computers on the basis of the values of its relevant feature. One simple way to perform this task is to classify new instances on the basis of their proximity in the feature space to known classifications. For example, say the programmer had an individual who she doesn't know is good or bad with computers, but whose age she does know. The program could plot this new datum on the feature space based on the age of the individual as follows:



The program then predicts the classification based on the datum's relationship to the other data points. Many different methods are used to make this prediction, but one standard method is to simply classify based on a majority "vote" of its k-nearest neighbors (kNN). If we set k to 5, then the program will decide on the basis of the five nearest neighbors in the feature space. In this case, the five nearest neighbors comprise four data labeled as bad with computers and one datum labeled as good with computers. Thus, the new individual will, on the basis of their age, be labeled as bad with computers.

With those individuals who are older patterning roughly with those who are bad with computers (and vice versa for young individuals and those good with computers), we could think of this entire data-set as "reflecting" the social generalization (or, loosely, the stereotype) that elderly people are bad with computers. But crucially this content is never explicitly represented as a rule in the program's code. At this point, one might demur: why not think that the content *elderly people are bad with computers* in fact is explicitly represented by the data set, say as the decision boundary for the algorithm, despite there not being any syntactic elements in the program's code that correspond to that content? My analysis of kNN is an instance of my claim that explicit representational content corresponds to a particular functional role.[39] The lack of *availability* illustrates the way in which the state (in this case, a certain clustering of data points or a decision boundary) fails to

[38] The data can come to reflect biases in a variety of other ways too, e.g., if the person labeling them was—due to her own biases and prejudices—more likely to label elderly individuals as bad with computers even when they weren't or if the relevant patterns were ubiquitous in the environment, making even a representative sample reflective of biases. This begins to highlight the dynamic relationship between machine bias and human cognitive biases, resulting in what computer scientists call *algorithmic bias*.

[39] This claim is discussed at the end of §3.1 when generalizing the implicit-explicit distinction to functionalism about all mental states.

satisfy the functional conditions concomitant with explicit representational content. The functional role of this state is limited in ways associated with implicitly represented contents, which are not available for further processing. Notice that were we to include in the program's code an explicit rule with content *it's not the case that elderly people are bad with computers*, there isn't obviously a contradiction anywhere in the program, e.g., as there might be if we also programmed as a rule *elderly people are bad with computers*. That there would be no breakdown of the machine demonstrates that the implicitly represented content cannot enter into causal relations with other states or processes, and so no contradiction is ever arrived at nor detected. Thus, the stereotype-like content merely implicitly emerges out of the distribution of training instances in the feature space, i.e., it is truly implicit. Notice though that the results are roughly *as if* the algorithm represented such a rule: it still classifies individuals as being good or bad with computers on the basis of their age, with elderly individuals more likely to be labeled as bad.[40]

Nothing in principle rules out that humans can operate with a similar cognitive makeup, one that stores representations only of individuals they've encountered and those individuals' feature values, and not the generalizations that emerge from the relations of those feature values. This picture also allows for an expanded understanding of the origins of socially relevant stereotypes and bias-constructs. We can imagine a corollary of "garbage in, garbage out" in human cases of bias. Individuals raised in environments that are historically shaped by discriminatory practices will receive many "data points" that reinforce common stereotypes. Often these data points will grossly misrepresent actual patterns in the environment, such as when media, entertainment, or propaganda depict individuals from marginalized groups in offensive stereotypical roles, creating a mass production of inaccurate inputs. Thus, the classifications an individual makes on the basis of previous examples (including both actual examples and inaccurate portrayals) will reflect the stereotypes that shaped the production of the examples themselves, producing new examples that continue the cycle.

Crucially, the idea that some aspects of social cognition consist in tendencies to store representations only of individuals encountered and not generalizations is precisely the thesis of exemplar-based psychological theories, which often utilize the kNN algorithm as a model.[41] Historically, questions about categorization have centered around a rivalry between four views: the classical view, the prototype view, the exemplar view and, most recently, the theory-theory view.[42] Most relevant for our discussion is the contrast between prototype-based and exemplar-based categorization theories. According to the prototype view, categorization of new instances is based on a *summary representation* prototypical of some category. Thus, my categorization of a new animal as a bird will depend on its similarity to a stored representation I have of a prototypical bird, which has a summary representation that is an abstraction from features I take take to be typical of birds. (Otherwise often regarded as a "stereotype".) If this new animal meets some threshold of similarity

[40] Another way that a computer might fail to represent the relevant stereotype is if it relies on so-called 'proxy attributes': attributes that correlate with some other attribute in an environment such that the former can serve (either purposely or accidentally) as a proxy for the latter, e.g., when redlining practices substituted zip code as a proxy for race (Massey and Denton 1993, 51 ff.). Machine learning programmers have long struggled with eliminating biases that are based on proxy attributes (Adler et al. 2016). Moreover, the notion of a proxy discrimination is familiar in discussions of discrimination in ethics and law (Alexander 1992, 167-173). The possibility that cognitive biases might also rely on proxy attributes gives reason to resist representationalist views that require strict representational contents of the relevant stereotypes. I discuss this issue in greater length in my paper 'Algorithmic Bias: On the Implicit Biases of Social Technology', MS.

[41] See Murphy 2004, 49-60 and Reed 2006, 198-203.

[42] The notion of prototype was popularized by Rosch (1978). For a psychological overview of the historic rivalry between it and other views, see Smith and Medin 1981 and Medin and Smith 1984. For a philosophical perspectives, see Rey 1983 and Burge 1993. For exchanges between the two, see Smith et al. 1984, Rey 1985, and Carey 2009, especially chapter 13.

for that prototype, then it will be classified as a bird; if not, then it won't. The exemplar view, on the other hand, holds that classification is based on a set of stored exemplar representations, comprising past instances of group members. Thus, my categorization of a new animal as a bird will depend on its similarity to a set of representations I have of birds I've encountered in the past. (Just as classification occurs in the kNN algorithm, making it an instantiation of the exemplar model.)

The history of literature attempting to adjudicate between these two theories is vast and a recap is well beyond the scope of this paper. I focus instead on studies demonstrating the superiority of the exemplar model over the prototype model in at least some contexts.[43] Crucially, such studies have been replicated in the domain of social judgements specifically. For example, Smith and Zarate (1990) devised a study wherein subjects are provided descriptions of nine people, five of whom were in group A and four of whom were in group B. They are then asked to categorize these and new individuals on the basis of descriptions. In one version of the test, subjects were provided with summary information (akin to a stereotype-like generalization) about each group before they were introduced to the descriptions of the nine members. In the second task, individuals were not given summary information beforehand. The test was able to adjudicate whether an individual was categorizing on the basis of prototypes or exemplars by having some descriptions that (1) closely matched one group's summary representation but not any of its individual members and (2) closely matched particular members of the other group but not its summary representation. The idea was that if subjects categorized these individuals into the first group, then they were relying on prototypes, but if they classified them into the second group, then they were relying on exemplars. What they found was that "in the absence of prior stereotype knowledge, classification is based on similarity to known individuals most of the time—regardless of similarity to the group average or prototype," demonstrating that there are at least some contexts in which social judgements are formed on the basis of stored representations of past examples and not summarizing information, i.e., stereotypes.[44] It stands to reason that these results would generalize to a large number of cases in which individuals lack stereotype knowledge—either because they've never been exposed to stereotype generalization or because they have avowed egalitarian beliefs denying stereotype generalizations that they have been exposed to.[45] This suggests that exemplar-based models (such as the kNN algorithm above) are the right account of many examples of implicit bias; as a consequence, these are cases of truly implicit bias.

Finally, exemplar-based social reasoning exemplified by the kNN algorithm can also helpfully demonstrate the aforementioned explanatory benefits of postulating truly implicit content concerning conscious accessibility, automaticity, and recalcitrance. Focusing on recalcitrance, we see that since it's possible that some implicit bias-constructs operate with truly implicit stereotype-like rules, mitigation techniques that aim to negate those stereotypes by introducing reasons to infer

[43] Medin et al. 1984, Linville et al. 1989, and Andersen and Cole 1990.

[44] Smith and Zarate 1990, 257. See also Smith and Zarate 1992 and Mastro and Tukachinsky 2011.

[45] An anonymous referee helpfully points out that this empirical evidence does not adjudicate issues related to accessibility and self report. This is true, and thinking about the import of this case to those discussions demonstrates why evidence for conscious accessibility is difficult to interpret. Should subjects that classify based on similarity to known individuals be asked on the basis of what they categorized, one can imagine that a sufficiently reflective subject could recognize the pattern their responses instantiate and, on the basis of this, articulate some post-hoc generalizing rule for sorting on the basis of similarity. The same could be said of a programmer carefully attending to the results of the kNN algorithm's categorization of elderly individuals as bad with computers. But this ability to report on the basis of attending to the outputs of some computational procedure (i.e., the bias-outputs) is not the same as having direct access to the mechanism that gives rise to those outputs (i.e., the bias-contruct). Thus, mere reportability is not sufficient for establishing direct conscious access to bias-constructs nor for undermining the existence of truly implicit contents serving as those bias-constructs.

the opposite of the stereotype will be generally ineffective. For example, if someone has a truly implicit bias that merely encodes the content P, then causing them to token a belief with the explicitly represented content *not-P* will not obviously work to counteract the truly implicit bias. Building this rule into the computer would not change the distribution of known examples and, thus, wouldn't alter the operations of the Euclidean distance measure on those known examples in the feature space. Moreover, when a person with a truly implicit bias reflects on their explicitly represented mental states, there's no direct contradiction, just as there's not obvious breakdown of the machine when we insert the explicit rule that's at odds with the truly implicit content. Everything continues to operate as before.

The above provides a clear example of how truly implicit bias can emerge in social reasoning. These exemplar models serve as both a proof of concept for truly implicit human biases and an empirically substantiated model for many real-world social judgements. Since representationalist views of bias entail that an individual has a bias if and only if they represent a stereotype (which may be conscious or unconscious), and in cases of truly implicit bias, no such representation exists, representationalism fails to capture all cases of bias.

4 A Functional Account

I've demonstrated why representationalism about bias is a mistake. I now turn to presenting an alternative functionalist account that is flexible enough to cover the diversity of forms bias can take.

4.1 Having a Bias

In the cases of both E and T, the mental states that led them to their conclusions about Jan have a variable element that can change from context to context: namely, Jan. The fact that they drew conclusions about Jan is incidental to their having a bias; they could have just as easily formed similar conclusions about Jill, Joe, or Jack, assuming they consider each of them elderly. However, they couldn't have formed any of these conclusions if it weren't for the stability of the generalizing content in (*ii*). In the functional account I advance here, I regard a bias-construct as *whatever plays the role* of being the stable component that takes us from one variable aspect to the other, just as a function systematically relates variable inputs to variable outputs.

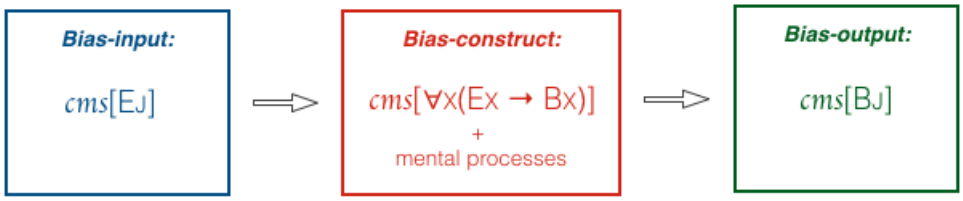
Importantly, the procedure instantiated by the bias-construct isn't wholly captured by the content of the second premise. The first premise, or the bias-input, states that Jan is elderly (Ej), and the conclusion, or the bias-output, states that Jan is bad with computers (Bj). Let's assume for the time being that the second premise takes the form of a universal generalization ($\forall x(Ex \rightarrow Bx)$). What makes for a valid inference from the first to the conclusion isn't merely the major premise, but rather the major premise together with inference rules (universal instantiation and *modus ponens*). Thus, these rules will need to be incorporated into the bias-construct. Moreover, it's possible that the second premise should be modeled by some content other than a universal generalization, since it seems many related contents can serve the role of taking us from variable inputs to variable outputs. Indeed, one of the lessons of this explication is that this intermediate content can take many forms. For example, it may be that this content ultimately takes the form of a generic rather than universal quantification.[46] Another possibility is that the bias-construct's operation is probabilistic, either because it is a statistical generalization (e.g., "most elderly are bad

[46]This is reasonably how we might model the stereotype generalization "elderly people are bad with computers." For specifics, see generic formalizations presented by Nickel (2016) and Leslie (2015, 2017).

with computers"), because it provides a confidence measure that any particular individual fits the generalization, or because "bad-with-computers" is itself a modal concept. Since the bias-construct is *whatever* bridges the gap between the two, it can account for any of these potentialities, and the bias-construct encodes a combination of the stereotype-like content and the rules that together instantiate the right dispositional profile.[47]

It's also important to note that mental states aren't standardly characterized merely as contents. Instead, they're properly regarded as constituted by a *content* together with a *mode*, or a way of relating to that content. For example, the belief that Jan is elderly is the content Ej together with the belief mode *bel*[ ]: *bel*[Ej]. Here, I remain agnostic as to whether the relevant mental states are bona fide beliefs. Instead, I assume only that the relevant components of the bias are *committal mental states* (CMSs), i.e., comprise some propositional content together with some committal mode *cms*[ ].

Tying it all together, we can model E's inference and the mental states and transitions it involves as follows:



Here, the arrows indicate that a bias-input "goes into" the bias-construct, "out of" which comes the bias-output. This functional account generalizes to any attribute, individual, or stereotype. Regarding T's implicit bias, her mental transitions will be similar in that they will involve the same inputs and outputs. Importantly, though, her bias-construct won't necessarily involve a CMS with the explicit representational content, roughly ∀x(Ex → Bx).

I thus propose a functional explication of the concept of psychological social bias exemplified by core cases, where each bias-construct is defined in terms of a systematic relation between bias-inputs and bias-outputs that mimic social-kind inductions. In all cases, I assume that the input and output are representational CMSs with propositional structure, an assumption I return to in §4.2.2.[48] In the case of *explicit bias*, the constructs involve consciously accessible, explicitly represented stereotypes. To repeat an important point, it's possible that in the case of implicit bias, the bias-constructs are representationally analogous, i.e., they too involve a fully represented (though not necessarily consciously accessible) stereotype. However, it's also possible that the content of the stereotype is *truly implicit*, meaning it's not represented at all. To account for this case, it's possible to collapse everything involved in the transition from the input to the output into one truly implicit content, so that we're left with a transformation principle that encodes *both* the content of the stereotype and the rule together. In all cases—even those in which the bias-construct is truly implicit—inputs and outputs are always explicitly represented. I regard these as *core cases* of bias because they illustrate the basic intuitive form that biases toward individuals take, namely inductions made on the basis of social kind membership.

Putting all of these functional components together, I propose the following definition of what it means for an individual P to *have a bias*, where $\boxed{G\text{·}T}$ is a label for a bias that results in attributions of a stereotypical property T to members of social group G.

[47] Thanks to an anonymous referee for pushing away from the categorical nature of universal instantiation.

[48] In this way, I build in more representational specificity than functional characterizations of implicit bias that leave the causal relata more open-ended. See, for example, Saul 2013, 40 and Holroyd 2016, 173.

Definition of Having a Bias:

For any social group G and attribute T, some person P *has the bias* $\boxed{G{\cdot}T}$ iff
there exists a contrast group H such that for any individual x
(1) P's having a CMS with the content Gx
reliably causes P to have a CMS with the content Tx, and
(2) P's having a CMS with the content Hx
does not reliably cause P to have a CMS with the content Tx.

Applying this definition to explicit bias as demonstrated by E, we can see that he instantiates the bias $\boxed{E{\cdot}B}$ since his tokenings of *cms*[EJ] reliably cause him to token *cms*[BJ], and his tokenings of *cms*[YJ] do not reliably cause him to token *cms*[BJ]. Moreover, this functional account is flexible enough to handle both implicit and truly implicit bias, for the definition of bias I've offered is neutral as to the nature of the mental construct. This allows us to say that E and T both share a bias toward the elderly, whether that bias be consciously accessible, consciously inaccessible, representationally explicit, or truly implicit.

This definition of *having a bias* incorporates the features fundamental to core cases of social bias—namely, the operation of systematically relating bias-inputs to bias-outputs in ways that mimic social-kind inductions—while remaining agnostic about certain phenomenal and representational features of the bias-construct itself. It also demonstrates an important point about the level of analysis that is most useful in characterizing bias: bias is just not the sort of thing that is constitutively characterized at the level of individual mental states, but rather something that emerges out of how certain mental states are related. It helps to contrast the functionalism I'm advocating for here with a functionalism about representational contents more generally. Crucially, my advocating for functionalism about a general characterization of bias involves moving to a higher level of abstraction from functionalism about individual states. The functional role I highlight here occurs at a level above the putative functional roles that can be identified for either truly implicit contents or explicitly represented stereotype beliefs. This is what allows it to be instantiated by both kinds of biases, whether representational or non-representational and regardless of whether we cash out representational contents themselves in further functional terms.

To put the point in familiar Marrian terms, bias is not a phenomenon best characterized at the *representation and algorithm level* of analysis. Rather, bias is a *computational-level* phenomenon, which could be instantiated in multiple different ways at the underlying levels.[49] Although bias can exist at the representational level, it's at the computational level that all cases of bias are unified. Other accounts of bias (such as the representationalist approaches to identifying particular instances of implicit bias) may be thought of as analyses of biases that manifest at these lower levels, namely the representational and algorithmic level. Such views are non-problematic to the extent that they're interpreted in this way, because these accounts might be right about the different states and processes that underlie many actual and empirically substantiated cases of bias. The theoretical lacuna in empirical theories of psychological social bias is a story about what unifies all cases of bias. My explication addresses this blind spot in the literature by giving a precise account of what unifies the heterogeneous instantiations of bias identified so far. According to my analysis, what unifies cases of bias—i.e., what makes them all instances of the natural kind *bias*—is that they all play the same functional role in overcoming underdetermination, which they achieve by sharing the same functional profile.[50]

[49] Marr 2010, 24-27.

[50] This model additionally finds unity among representationalist, dispositionalist, and associationist views of implicit bias, reconciling claims about heterogeneity (Holroyd and Sweetman 2016; Del Pinal and Spaulding 2018), as well as unifying these accounts with others gaining prominence in the literature, for example, those that postulate bias-

This basic definition of *having a bias* is likely too simple to account for many real-world instances of bias, but can still serve as an important foundation. Cases of intersectionality, familiarity, and context-effects can all affect the predication of an attribute to an individual despite that individual's supposed belonging to a target social category. Someone with the conscious belief that elderly people are good with computers and the implicit bias $\boxed{E{\cdot}B}$ will, on my view, have both biases $\boxed{E{\cdot}G}$ and $\boxed{E{\cdot}B}$; it's just that the operation of each will vary from context to context. This is expected in cases where a person's implicit biases conflict with their avowed egalitarian commitments.[51] How these two mental entities interact will be context sensitive. Moreover, my focus on "core cases" of social bias has been limited to those in which biases cause us to think about or behave toward individuals differently on the basis of social-kind membership (i.e., take the form of social-kind inductions). However, nothing in principle rules out that this theory might be extended to account for more peripheral forms that social bias can take, e.g., drawing conclusions about social groups as a whole on the basis of stereotypes about those groups. For example, we might go from a bias-construct encoding the content that elderly people are bad with computers to a bias-output that elderly people are ineligible for IT jobs.[52]

Notice also that my notion of having a bias leaves out important epistemic and ethical constraints, rendering this initial account normatively agnostic.[53] Such constraints will eventually be critical to a complete account of our ordinary notion of social bias, but are inessential to the psychological kind I'm explicating.[54] This way, the functional kind *bias* I employ shares more in common with cognitive notions of bias more generally that make their way into psychological explanations, say in reasoning about transformation principles within the visual perceptual system

constructs comprise imaginings (Sullivan-Bissett 2019, Welpinghus 2019) or social schemas (Soon 2019).

[51] Although my account unifies bias at a higher level of abstraction, there will still be cases where important distinctions still take place at lower levels, e.g., in distinguishing cases of explicit and truly implicit bias. This movement between levels of abstraction allows my view to meet the challenge presented by Holroyd (2016, 159, 169) of accounting for cases where a person harbors both an explicit bias $\boxed{E{\cdot}B}$ and an implicit bias $\boxed{E{\cdot}B}$. On my view, such individuation can occur at the lower functional level, say by differences in these biases' first-order functional roles that correspond to explicitly and implicitly represented contents, respectively.

[52] In this case, we would need to clarify that the individuals in the model of *having a bias* can refer to groups of people as well as individual members. That said, I find it hard to imagine that a social bias could fail to bottom out eventually in attributions of stereotypical properties to individuals within social groups. For example, if someone had a bias-construct that encoded the content that elderly people are bad with computers, but failed to ever believe of any particular elderly individual that they are bad with computers on the basis of this, then this begins to strain my intuitive understanding of how biases operate. It seems to me fundamental that a social bias ultimately affect how we think about and interact with individual members of social groups; however, nothing critical hinges on this intuition.

[53] For a detailed discussion of arguments in favor of and against including normative and accuracy conditions in the definition of bias and the related notion *stereotype*, see Antony 2016, 2001, Munton 2019a, Beeghly 2015, and Blum 2004.

[54] One might worry that my definition of bias is too inclusive because it applies to inferences that broach epistemic certainty. For example, according to my definition, a person's inference from an individual's being a bachelor to the conclusion that they're a man will count as a social bias. Although I admit this is an undesirable outcome, circumventing it is difficult. The easiest maneuver would be to add a stipulation to the definition that overtly excludes conclusions arrived at through logical deduction or conceptual analysis. This move can be independently motivated (and, thus, not regarded as ad-hoc) by reflecting on my earlier remarks about the nature of psychological bias more generally: psychological biases are unified by their being a response to underdetermination (see §1). It's only in circumstances where we're not certain of the conclusions that we rely on heuristics and biases, in order to guide us over inductive gaps. However, even this stipulation is difficult to employ in practice since, for any particular transition that is truly implicit, it's not obvious whether the processes involved are encoding deductive rules and universal generalizations rather than mere inferential rules and statistical/generic/normality generalizations (e.g., a transformation principle that encodes merely an assumption that *most* bachelors are unmarried men). For this reason, I think it difficult to build into the definition principles that definitively cordon bias off from logically certain social deductions.

or more general cognitive heuristics.[55] These constraints can, however, easily be added to the functional analysis to more closely approximate our everyday notion of social bias.[56]

It's worth taking time to reiterate points made above about how this view highlights important etiological facts about bias. The view of bias I'm putting forward allows for the possibility of biases that are a result of low-level mental states and processes that merely reflect problematic patterns that exist in our social environments. This reflects their fundamental role in overcoming underdetermination. Just as certain visual perceptual heuristics are explained by reference to the physical environment in which the perceptual system emerged (e.g., that the visual system assumes light comes from above because, in most situations, light does come from above), so too certain social psychological heuristics are explained by reference to the social environment in which the socio-cognitive system emerged (e.g., we might assume an individual with long hair and makeup is a woman because, in most situations, such individuals are women). Of course, the story about social biases will be complicated by the fact that the assumptions it makes are shaped not just by objective physical features, but various unreliable influences prevalent in society, including inaccurate portrayals in media, propaganda, and popular culture more generally.[57] The social psychological mechanisms that give rise to biases in my sense will exist in complex interactive networks with narratives (both accurate and inaccurate) that circulate in the wider culture. Thus, the view I'm advancing is deeply anti-individualistic and, by contrast, I believe representationalist views that require individuals with biases to harbor belief-like states result in overly individualistic and overly intellectual accounts of social bias.[58] Adequate causal stories of where some biases originate arguably requires extending our focus beyond individual agents to the social environments that provide these patterns.[59] A critical insight of the functional account is that our most effective explications of bias will look to how psychological systems globally operate, bolstering the claim that we must widen our analysis beyond individual mental states and agents.[60]

At this point, we've seen a variety of forms psychological social biases can take and a functional account that unifies them. One might next wonder if I've taken my theoretical strategy of stripping away the inessential features of bias far enough. It's possible, one might think, to eliminate further the representational status and propositional structure of the input and output states I've posited. However, as I argue below, one cannot strip away these aspects of the model without threatening our core explanatory aim of accounting for the fact that biases influence thoughts about and actions toward others. To reduce the structure any further would create an explanatorily flawed explication. Thus, the model I have proposed is one that is both general enough to cover the core case of bias (unlike the representationalist model) while also being able to explain the full range of phenomena (unlike the dispositional and associationist models, which I discuss next).

[55] See Helmholtz 1925 and Tversky and Kahneman 1974, respectively.

[56] For work investigating the epistemic and normative evaluation of mental states involving social group membership, see Basu 2018 and Bolinger 2018.

[57] See, for example, Valian 2005 for a discussion of how "cold" social-cognitive biases against women can manifest as a result of misogynistic environments, even when individuals in those environments express egalitarian beliefs.

[58] This is anti-individualism about bias in the sense used in theories of the natures of mental states, which claim that such natures depend constitutively on relations between the individual and the wider environment (Burge, 2010a, 61-82). In the case of social bias, the wider environment naturally includes an individual's social environment. This anti-individualism should be contrasted with its use in debates between so-called *structural prioritizers* and *anti-anti-indivisualists* in literature on implicit bias, which is more general (Madva 2016). However, I believe adopting anti-individualism in the former sense helps to resolve disputes regarding anti-individualism in the latter sense.

[59] As Anderson (2010, 6) argues, any account of race-based injustice must begin "from a structural account of the systematic disadvantages imposed on people because of their race in our society." See also Alexander 2012.

[60] Thanks to an anonymous referee for encouraging me to say more about the etiology of social biases.

4.2 Alternative Proposals

I now explore two options for further eliminating the features we take to be essential to the structure of bias. First, I address views that do away with all representational components—as inputs, outputs, or aspects of the bias-construct. The result is a form of *global dispositionalism* (akin to radical behaviorism), where biases just are dispositions to behave in certain ways when faced with environmental stimuli of certain sorts. After addressing global dispositionalism, I consider an approach that keeps representational mental states as inputs and outputs, but strips away their propositional structure. This amounts to the standard associationist view. The weakness of both is that they cannot provide a genuinely explanatory account of why biases influence beliefs about or actions toward an individual on the basis of that individual's presumed membership in a social group. That said, both views have the potential to—once supplemented by the right representational states—capture the relevant explanatory datum. The results are dispositional and associationist views that are consistent with the functional account I've presented.

4.2.1 Against Global Dispositionalism

I'll begin with Dennett's notion of emergent content, which I raised in my discussion of truly implicit content. Dennett holds a sort of *global dispositionalism* wherein attributions of mental states are grounded in dispositions to behave in certain ways when faced with certain environmental stimuli. Behavioral dispositionalist views have historically faced two primary problem cases: (1) cases where the behavioral dispositions are mixed, making it unclear whether to attribute a belief; and (2) cases where the relevant behavioral dispositions are lacking, but attributions of belief still seem apt. Likewise, problem cases of both varieties plague global dispositionalism about bias.

Problem cases in category (1) that apply to attributions of social bias are related to arguments some theorists have made for rejecting implicit biases as bona fide beliefs: to treat biases as beliefs would require that we attribute to individuals contradictory beliefs. To be clear, I'm not opposed to the idea that individuals can (and do) have contradictory beliefs. But to do so would violate the assumption of rationality adopted within standard global dispositionalist accounts. Consider, for example, Dennett's recipe for the attribution of belief, which requires the attribution of belief only where it lends to accurate predictions of behavior *under the assumption that the individual is rational*.[61] In the case where an individual has contradictory implicit and explicit biases, we can't attribute one without leading to a failed prediction of behavior or rationality. For example, although T performs some actions that make it seem like she believes elderly people are bad with computers, she explicitly (and we can assume honestly) asserts that she doesn't believe elderly people are bad with computers. Thus, if I attribute to T that she has the belief that elderly people are bad with computers, then I fail to predict that she will deny the claim that elderly people are bad with computers. If I attribute to T that she has the belief that elderly people are not bad with computers, then I fail to predict that she will attempt to help Jan with her computer. If I attribute to T *both* beliefs, then I violate the assumption that she's rational.[62]

Problem cases of the second type (2) are more straightforward: we can imagine an individual who has a bias toward the elderly, but who fails to ever behave in ways that reveal that bias. Imagine a scenario in which E is again preparing for a Skype interview, and he goes through his inference about Jan just as he does in the scenario above. But in this version, he has an acute awareness of the social stigma related to ageist beliefs. E is so good at hiding this fact that he *never* admits to or acts in ways that reveal having a bias. In other words, E is a super-spartan

[61] See, for example, Dennett (1987, 15)'s discussion of the *intentional strategy*.

[62] Similar criticisms apply to Newell (1994, 1988)'s "knowledge level" of analysis.

about bias.[63] In this case, it seems undeniable that E has a bias against the elderly, it's just that he's very good at keeping that bias private.

Liberal dispositionalist views about bias have arisen that promise to overcome these difficulties. For example, Schwitzgebel (2002, 2013) proposes that beliefs are best characterized as bundles of dispositions to behave *and cognize* in particular ways. However, this view, unlike my view, privileges phenomenal (as opposed to representational) properties of the mental states within the dispositional relata. As Schwitzgebel (2002, 250) notes, his view is similar to behavioral dispositionalism, however, "unlike dispositional accounts as typically conceived, it gives a central role to conscious experience, or 'phenomenology'."

By allowing reference to phenomenal experiences, we can see how Schwitzgebel's dispositionalism avoids the second set of problem cases (2). Although E doesn't have any *behavioral* dispositions associated with a bias toward the elderly, he undeniably has certain phenomenal dispositions, e.g., the disposition to experience thinking that elderly individuals are bad with computers.

But even this sort of approach will still run into difficulties with problem cases of the first sort (1) when posed with cases of implicit bias. Just as before, although T performs some actions that make it seem as though she believes elderly people are just as good with computers as anyone else is, she often behaves in ways that are at odds with this belief, like helping Jan join the Skype interview. Thus, she seems disposed to a mixed and complicated collection of behavioral dispositions. Schwitzgebel (2002, 260) attempts to handle these cases of implicit bias by introducing what he calls 'in-between cases of believing': cases where "it seems not quite appropriate to describe the subject as either fully believing or not believing the proposition in question."[64] Since T is sometimes disposed to behave in ways that are consistent with believing elderly people are bad with computers and sometimes not, her case qualifies as one of "in-between" believing.

The troubling aspect of Schwitzgebel's responses to the two sorts of cases is that they fail to provide a unified explanation of both E's and T's behaviors. In one case, we say E has a bias against the elderly because he has phenomenal experiences of the right sort, despite failing to act in the appropriate ways; in the other case, we say that T has a bias against the elderly because she acts in the appropriate ways, despite failing to have phenomenal experiences of the right sort. Schwitzgebel regards these points as demonstrating that a unified account is a mistake. This rebuttal is compelling only insofar as there's no plausible unifying alternative. My functional account of bias makes clear that such an alternative is available. Thus, following an assumption that, all else equal, unifying accounts are better than non-unifying accounts, we should not settle for a disjunctive explanation in the case of bias.

This points to a deeper difficulty concerning any global dispositionalist view that departs from the computational orthodoxy within cognitive science having to do with psychological explanations: global dispositionalism, like classical behaviorism, struggles to explain in virtue of what we're disposed to act and experience in the ways that we do.[65] An approach to bias based in the representational theory of mind, which specifies dispositions in terms of representational contents as mine does, avoids this problem by supplying a common explanation for why E and T are prone to behave and experience as they do: both have a bias-construct that causes them to token the CMS that Jan is bad with computers on the basis of the CMS that Jan is elderly. These are the minimal elements needed for a psychological explanation consonant with contemporary cognitive theory.[66]

[63] Putnam 1980, 29.

[64] See also Schwitzgebel 2013, 85.

[65] For examples, see Putnam 1980, 26 ff. and Fodor 1981, 6 ff.. For a contemporary arguments, see Quilty-Dunn and Mandelbaum 2017a.

[66] Similar criticisms apply to both the trait approach presented by Machery (2016, 2017) and the indeterminate

There are, however, aspects of the dispositionalist theory that seem crucial to the story of bias. As my functional account highlights, bias is fundamentally a dispositional phenomenon: by its nature, bias is characterized by our being disposed to think and act toward others in particular ways on the basis of group membership. However, also needed are the unifying representations that go along with recognizing someone as being a member of a social group (bias-inputs) and the unifying representations that pair that individual together with some attribute stereotypical of the social group (the bias-output). Thus, the functional account of bias improves on the dispositional account by grounding attributions of bias in dispositions to systematically relate precisely these representational inputs to these representational outputs, irrespective of any behavioral or phenomenal components.

4.2.2 Propositional Inputs and Outputs

A notable aspect of the application of my functional model to implicit bias is that the input and output types of my model are propositional, whereas many psychologists and philosophers take implicit biases to be the product of mental associations between solitary concepts.[67] I'll call this view 'the simple associationist view'. According to it, the implicit bias operative in the case of T above would be the product of a mental association between two concepts—the solitary concept *elderly* and the solitary (but complex) concept *bad with computers*—rather than any propositions these concepts combine to form.

Most individuals are familiar with the experience of mental associations. Combinations like *salt/pepper*, *Lois/Clark*, and *peanut-butter/jelly* are all prime candidates for mental associations. Thinking of salt likely causes you to think also of pepper. The standard explanation for this is that you often see one with the other, and so your mind naturally pairs the occurrence of the concept representing one to the concept representing the other. To be precise, let's say that a concept X is associated with concept Y iff mental tokenings of X reliably cause the tokening of Y.

In order to demonstrate why propositional structures are necessary for bias-inputs and bias-outputs, I consider first why solitary concepts won't do. Firstly, although associations connect explicitly represented concepts, the connections between concepts don't themselves represent anything. Associative connections such as these are what Fodor (2003, 90) calls 'semantically transparent', that is, "the content of *A-associated-with-B* is just *the content of A* associated with *the content of B*." In other words, there's nothing to the content of an association other than the content of the solitary concepts the associations relate. Crucially, the associations themselves don't represent relations between the two concepts. Thus, an association between, e.g., *salt* and *pepper*, is just that: an association and nothing more. The connection between the two concepts doesn't itself carry content, nor does it compose that content with the contents of the concepts it connects. Thus, an association between *salt* and *pepper* cannot represent a determinate proposition, e.g., *salt is good with pepper* or *salt is often with pepper* or *salt is identical to pepper* or *salt and pepper share similar properties* or even *I see salt and pepper*. All associationist accounts that attempt to build into their models claims about *how* the concepts are related, beyond mere tokenings of one causing tokenings of another, outstrip the theoretic resources the simple associationist view affords.

With that in mind, consider that many theorists believe that implicit biases are associations that (i) represent relations between the associated concepts and (ii) are responsible for a wide range of discriminatory behaviors. Both of these effects render the simple associationist view insufficient.

content approach presented by Yumusak (2017).

[67] Proponents of this view arguably include Greenwald et al. (2002), Banaji and Greenwald (2013), Gendler (2008), Holroyd (2016), Madva and Brownstein (2018), and Gawronski and Bodenhausen (2014).

Regarding (i), many proponents of the associative model believe that when a social kind concept, e.g., *elderly*, is associated with a particular attribute concept, e.g., *bad with computers*, that association represents propositions reflecting certain relations between the objects those concepts pick out, e.g., *elderly individuals are bad with computers*. Alternatively, such an association might be taken to reflect a different, potentially contradictory relation, e.g., *elderly individuals are not bad with computers*. In fact, none of these are possibilities for the same reason *salt* being associated with *pepper* doesn't represent any of the possible ways salt and pepper might be related: associative connections are semantically transparent. They don't represent anything beyond just *salt* and *pepper*, or in the bias case, *elderly* and *bad with computers*.

Assumption (ii) also rules out the simple associationist view. The vast amount of empirical work on implicit biases undermines the claim that they have no influence on behavior whatsoever. In order for implicit biases to have any of these effects, however small, we need to posit committal propositional structure to the input and output representational states.

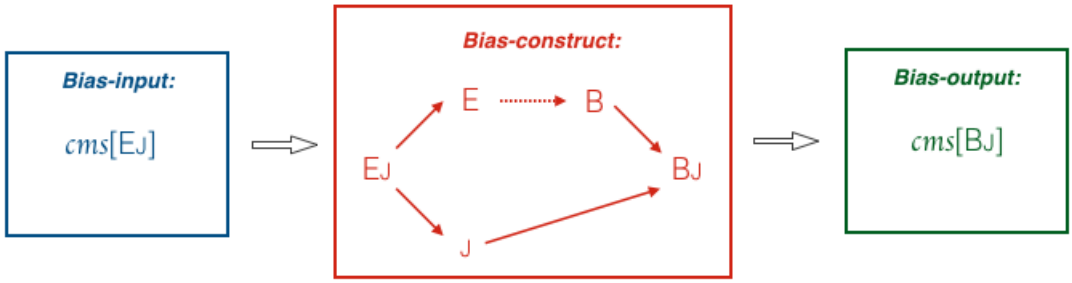
This point is demonstrated by what Mandelbaum (2013, 204) calls 'the binding argument'.[68] The reason discriminatory behaviors on the basis of biases necessitate structure beyond solitary concepts in the bias-outputs that cause them is because each of these behaviors has a distinct *target*, i.e., an individual in the social environment. And, in order for implicit biases to affect the behavior toward unique targets, the relevant mental representations that prompt those behaviors must include some attribute concept and some way of "binding" that attribute concept to the concept of the target. Consider again our example above. In order to explain why T treats *Jan* in a certain way, we need some mental state structure that binds the concept *Jan* with the relevant attribute concepts *elderly* and *bad with computers*. The tokening of the individual concept *Jan* might prompt the tokening of the attribute concept *bad with computers*, but without structure, there would be nothing that tells *how* those concepts are related. That is, there couldn't be content of the form *Jan is bad with computers* or any other content representing *who* is bad with computers. However, it's clear from T's behavior that there is a particular target in mind: Jan. T doesn't help everyone join the Skype interview, just Jan, and she does this *because* she thinks Jan is elderly. Thus, the relevant mental states that prompt the behavior must, at minimum, pair together the attributes with the target. Additionally, the behavior is explained only if we take T to regard that content as true; if she merely entertained or even doubted that content, without committing herself to it, this wouldn't explain why she behaves in ways that are consistent with her taking it to be true. Thus, the minimal structure needed for the inputs and outputs is the logical structure of propositions—*Ej* and *Bj*—together with a committal mode.[69] The relevant mental states that produce T's behavior must involve those modes and contents. Thus, the application of the functional model provided above to implicit bias—according to which the input-output types are committal propositions—is vindicated.

Again, however, there are aspects of the associationist picture that are consistent with the functional analysis I've provided. Although it's true that no combination of associative connections and solitary concepts will form a proposition and, therefore, that such combinations will never be sufficient to account for the core explanatory cases mentioned above, it's possible to "beef up" the

[68] Mandelbaum poses this argument as a challenge to Gendler (2008)'s associative approach to *aliefs*. Other adoptions of it include Fodor 2003, De Houwer 2014, and Levy 2015.

[69] The notion of *proposition* I invoke here is intended to be liberal enough to include any content minimally pairing individuals with attributes and is not intended to take a stand on more complex issues regarding the format of particular mental states. For example, Burge (2010b, 2018) argues that perceptual content is iconic and is best modeled not as a proposition, but as a complex noun phrase, e.g., (roughly) *that$_j$ elderly person*. This content would count as propositional in my liberal sense, since it still pairs the individual Jan (picked out by the demonstrative *that$_j$*) with an attribute (picked out by *elderly person*). Indeed, I take it that perceptual states can and do often serve as bias-inputs.

simple associationist view to have it avoid these issues.[70] In this case, the inputs and outputs to the bias-constructs would still need to be propositions, for the reasons stated above. However, we could find a place for associations within the bias-construct itself. For example, we could have some combination of processes that first takes a bias-input and systematically separates the targets from the social group predicates, next associates those predicates with attribute predicates, and then systematically recombines those new attribute predicates with the targets from before, as shown here:



In this case, the basic associationist view is present (represented by the dotted arrow), but it's folded into the bias-construct and supplemented by other processes so as to mimic the inferential form discussed before. So long as these supplementary processes are systematically (de)compositional, the sets of states and processes would instantiate a bias-construct, and the whole process will be an instantiation of the functional account I've provided.

The above discussion illustrates the strengths and weaknesses of the dispositional and associationist pictures. I've highlighted aspects of each that prevent either view from being regarded as a complete account of bias. However, we shouldn't for these reasons reject either view wholesale. Instead, I've demonstrated how both dispositionalist and associative views can be supplemented so as to meet the functional analysis I've provided and thereby account for the core explanatory datum of social bias. Thus, a functionalist approach can incorporate the positive aspects of each view while avoiding their explanatory misgivings.

5 Conclusion

In this paper, I have pursued two main goals. First, I've provided reasons to think that a representationalist approach to bias, which characterizes bias in terms of stereotype-like representational contents, is insufficient. This point was made by the possible (and plausible) existence of truly implicit bias. Second, I've presented a positive functional account of bias that meets the necessary explanatory and empirical needs of a computational-level theory of bias while also accounting for the diversity of forms that bias can take.

Although the notion of social bias that I explicate here is intended to narrowly characterize a core account of social bias, namely one that makes good on the role social bias plays in psychological explanations of how we think and act toward others on the basis of social-kind membership, it necessarily connects to a general theory of social bias construed more broadly. Individual psychologies reflect and contribute to larger social structures, and so a complete explanation of why they operate the way that they do will need to make reference to the wider social environment in which they're embedded.

[70] Following De Houwer (2014, 346), we can regard some such models as "specific instantiations of propositional models rather than as rivals."

For these reasons, optimistically, this account has the functional resources to be eventually extended to a variety of other domains, for example, cases of bias in the criminal justice system, the media, corporate hiring practices, and machine learning. For now, it lays the groundwork for better understanding these other kinds of biases by exploring the nature of social biases inherent in the minds of individuals.[71]

References

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. (2016). Auditing black-box models for indirect influence. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1–10. IEEE.

Alexander, L. (1992). What makes wrongful discrimination wrong? Biases, preferences, stereotypes, and proxies. *University of Pennsylvania Law Review*, 141(1):149.

Alexander, M. (2012). *The new Jim Crow: mass incarceration in the age of colorblindness*. New Press, New York, revised edition edition. OCLC: ocn656451603.

Allport, G. (1979). *The nature of prejudice*. Basic Book.

Andersen, S. M. and Cole, S. W. (1990). "Do I Know You?": The Role of Significant Others in General Social Perception. *Journal of Personality and Social Psychology*, 59(3):384–399.

Anderson, E. (2010). *The Imperative of Integration:*. Princeton University Press, Princeton.

Antony, L. (2001). Quine as Feminist: The Radical Import of Naturalized Epistemology. In Antony, L. and Witt, C. E., editors, *A Mind Of One's Own: Feminist Essays on Reason and Objectivity*, pages 110–153. Westview Press.

Antony, L. (2016). Bias: Friend or Foe? In Brownstein, M. and Saul, J., editors, *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, pages 157–190. Oxford University Press.

Antony, L. M. (2000). Naturalized Epistemology, Morality, and the Real World. *Canadian Journal of Philosophy*, 30(sup1):103–137.

Banaji, M. R. and Greenwald, A. G. (1994). Implicit stereotyping and prejudice. *The psychology of prejudice: The Ontario symposium*, 7:55–76.

Banaji, M. R. and Greenwald, A. G. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press, New York.

Banaji, M. R. and Hardin, C. D. (1996). Automatic stereotyping. *Psychological Science*, 7(3):136–141.

[71] I am grateful for help and critical comments from Josh Armstrong, Erin Beeghly, Ned Block, Michael Brownstein, Tyler Burge, Elisabeth Camp, David Chalmers, Sam Cumming, Guillermo Del Pinal, Jenna Donohue, Tina Eliassi-Rad, Branden Fitelson, Amber Kavka-Warren, Amy Kind, Alex Madva, Eric Mandelbaum, Annette Martin, Allison McCann, Jessie Munton, Nico Orlandi, Carlotta Pavese, Michael Rescorla, Ayana Samuel, Susanna Schellenberg, Seana Shiffrin, Susanna Siegel, and Ege Yumusak, and for comments on multiple drafts from Rima Basu, Gabe Dupre, Katie Elliott, Gabriel Greenberg, Bill Kowalsky, and Kevin Lande. Previous versions of this paper were presented at the Second California Philosophy Conference, the Implicit Bias workshop at the University of Antwerp and the Mind Discussion Group at New York University. Finally, I want to acknowledge the detailed and helpful comments received from the two anonymous referees and the *Mind* editors, which greatly improved the paper.

Bar-Anan, Y., Nosek, B. A., and Vianello, M. (2009). The Sorting Paired Features Task: A Measure of Association Strengths. *Experimental Psychology*, 56(5):329–343.

Basu, R. (2018). The Wrongs of Racist Beliefs. *Philosophical Studies*.

Beeghly, E. (2015). What is a Stereotype? What is Stereotyping? *Hypatia*, 30(4):675–691.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(02):227.

Blum, L. (2004). Stereotypes And Stereotyping: A Moral Analysis. *Philosophical Papers*, 33(3):251–289.

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese*.

Burge, T. (1993). Concepts, Definitions, and Meaning. *Metaphilosophy*, 24(4):309–325.

Burge, T. (2010a). *Origins of objectivity*. Oxford University Press, Oxford.

Burge, T. (2010b). Origins of Perception. *Disputatio*, 4(29):1–38.

Burge, T. (2018). Iconic Representation: Maps, Pictures, and Perception. In Wuppuluri, S. and Doria, F. A., editors, *The Map and the Territory*, pages 79–100. Springer International Publishing, Cham.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carnap, R. (1950). *The Logical Foundations of Probability*. Chicago: University of Chicago.

Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, 4(14):278–280.

Carruthers, P. (2017). Implicit versus Explicit Attitudes: Differing Manifestations of the Same Representational Structures? *Review of Philosophy and Psychology*, 9(1):51–72.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. MIT Press.

Correll, J., Park, B., Judd, C. M., and Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, 83(6):1314–1329.

Cummins, R. (1982). The Internal Manual Model of Psychological Explanation. *Cognition and Brain Theory*, 5(3):257–268.

Daumé III, H. (2015). *A Course in Machine Learning*. https://ciml.info/.

De Houwer, J. (2003). The extrinsic affective Simon task. *Experimental psychology*, 50(2):77.

De Houwer, J. (2014). A Propositional Model of Implicit Evaluation: Implicit evaluation. *Social and Personality Psychology Compass*, 8(7):342–353.

Del Pinal, G. and Spaulding, S. (2018). Conceptual centrality and implicit bias. *Mind & Language*, 33(1):95–111.

Dennett, D. C. (1981). A Cure for the Common Code. In *Brainstorms: philosophical essays on mind and psychology*, pages 90–108. MIT Press, Cambridge, Mass.

Dennett, D. C. (1987). *The intentional stance*. MIT Press, Cambridge, MA.

Devitt, M. (2006). *Ignorance of Language*. Oxford University Press.

Dovidio, J. F. and Gaertner, S. L. (2004). Aversive Racism. In Zanna, M. P., editor, *Advances in experimental social psychology*, volume 36, pages 1–52. Elsevier Academic Press, San Diego, CA.

Dovidio, J. F., Glick, P. S., and Rudman, L. A., editors (2005). *On the nature of prejudice: fifty years after Allport*. Blackwell Pub, Malden, MA.

Dupre, G. (2019). Linguistics and the explanatory economy. *Synthese*, pages 1–43.

Fazio, R. H. (1990). Multiple Processes by which Attitudes Guide Behavior: The Mode Model as an Integrative Framework. In *Advances in Experimental Social Psychology*, volume 23, pages 75–109. Elsevier.

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

Fazio, R. H. and Olson, M. A. (2003). Implicit Measures in Social Cognition Research: Their Meaning and Use. *Annual Review of Psychology*, 54(1):297–327.

Fodor, J. A. (1981). Introduction: Something of the State of the Art. In *Representations*, pages 1–31. MIT Press.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. MIT Press.

Fodor, J. A. (2003). *Hume variations*. Lines of thought. Clarendon Press. OCLC: ocm52485691.

Gaertner, S. L. and Dovidio, J. F. (1986). The Aversive Form of Racism. In Gaertner, S. L. and Dovidio, J. F., editors, *Prejudice, Discrimination, and Racism*, pages 61–89. Academic Press, San Diego.

Gallistel, C. R. and King, A. P. (2009). *Memory and the computational brain: Why cognitive science will transform neuroscience*. Wiley-Blackwell.

Gawronski, B. and Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, 132(5):692–731.

Gawronski, B. and Bodenhausen, G. V. (2014). Implicit and Explicit Evaluation: A Brief Review of the Associative-Propositional Evaluation Model: APE Model. *Social and Personality Psychology Compass*, 8(8):448–462.

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

Greenwald, A. G. and Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4–27.

Greenwald, A. G. and Banaji, M. R. (2013). *Blindspot: Hidden biases of good people*. Delacorte Press.

Greenwald, A. G., Banaji, M. R., Rudman, L. A., Farnham, S. D., Nosek, B. A., and Mellott, D. S. (2002). A unified theory of implicit attitudes, stereotypes, self-esteem, and self-concept. *Psychological Review*, 109(1):3–25.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Greenwald, A. G. and Nosek, B. A. (2008). Attitudinal dissociation: What does it mean. *Attitudes: Insights from the new implicit measures*, pages 65–82.

Hahn, A. and Gawronski, B. (2014). Do implicit evaluations reflect unconscious attitudes? *Behavioral and Brain Sciences*, 37:28–29.

Hahn, A. and Gawronski, B. (2019). Facing One's Implicit Biases: From Awareness to Acknowledgment. *Journal of Personality and Social Psychology*, 115(5):769–794.

Hahn, A., Judd, C. M., Hirsh, H. K., and Blair, I. V. (2014). Awareness of implicit attitudes. *Journal of Experimental Psychology: General*, 143(3):1369–1392.

Helmholtz, H. (1925). *Treatise on physiological optics*, volume 3. Optical Society of America, Rochester, NY. (Original work published 1910).

Hillard, A. L., Ryan, C. S., and Gervais, S. J. (2013). Reactions to the implicit association test as an educational tool: A mixed methods study. *Social Psychology of Education*, 16(3):495–516.

Holroyd, J. (2016). VIII—What Do We Want from a Model of Implicit Cognition? *Proceedings of the Aristotelian Society*, 116(2):153–179.

Holroyd, J., Scaife, R., and Stafford, T. (2017). What is implicit bias? *Philosophy Compass*, 12(10):e12437.

Holroyd, J. and Sweetman, J. (2016). The Heterogeneity of Implicit Bias. In Brownstein, M. and Saul, J., editors, *Implicit Bias and Philosophy Volume 1: Metaphysics and Epistemology*, pages 80–103. Oxford University Press.

Horgan, T. and Timmons, M. (2007). Morphological Rationalism and the Psychology of Moral Judgement. *Ethical Theory and Moral Practice*, 10(3):279–295.

Howell, J. L., Gaither, S. E., and Ratliff, K. A. (2015). Caught in the Middle: Defensive Responses to IAT Feedback Among Whites, Blacks, and Biracial Black/Whites. *Social Psychological and Personality Science*, 6(4):373–381.

Howell, J. L. and Ratliff, K. A. (2017). Not your average bigot: The better-than-average effect and defensive responding to Implicit Association Test feedback. *British Journal of Social Psychology*, 56(1):125–145.

Kovel, J. (1970). *White Racism: A Psychohistory*. Patheon, New York.

Lande, K. (2018). The Perspectival Character of Perception. *The Journal of Philosophy*, 115(4):187–214.

Lau, H. C. (2008). Are We Studying Consciousness Yet? In Weiskrantz, L. and Davies, M., editors, *Frontiers of Consciousness: Chichele Lectures*, pages 245–258. Oxford University Press.

Leslie, S.-J. (2015). Generics Oversimplified. *Nous*, 49(1):28–54.

Leslie, S.-J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, 114(8):393–421.

Levy, N. (2015). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous*, 49(4):800–823.

Linville, P. W., Salovey, P., and Fischer, G. W. (1989). Perceived Distributions of the Characteristics of In-Group and Out- Group Members: Empirical Evidence and a Computer Simulation. *Journal of Personality and Social Psychology*, 57(2):165–188.

Lippert-Rasmussen, K. (2014). *Born free and equal? A philosophical inquiry into the nature of discrimination*. Oxford University Press, Oxford, New York.

Machery, E. (2016). De-Freuding implicit attitudes. In *Implicit Bias & Philosophy: Metaphysics and Epistemology*, volume 1, pages 104–129. Oxford University Press.

Machery, E. (2017). Do Indirect Measures of Biases Measure Traits or Situations? *Psychological Inquiry*, 28(4):288–291.

Madva, A. (2016). A plea for Anti-Anti-Individualism: how oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, 3.

Madva, A. and Brownstein, M. (2018). Stereotypes, Prejudice, and the Taxonomy of the Implicit Social Mind. *Noûs*, 52(3):611–644.

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

Mandelbaum, E. (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous*, 50(3):1–30.

Marr, D. (2010). *Vision: a computational investigation into the human representation and processing of visual information*. MIT Press, Cambridge, Mass. OCLC: ocn472791457.

Massey, D. S. and Denton, N. A. (1993). *American apartheid: segregation and the making of the underclass*. Harvard University Press, Cambridge, Mass.

Mastro, D. and Tukachinsky, R. (2011). The Influence of Exemplar Versus Prototype-Based Media Primes on Racial/Ethnic Evaluations. *Journal of Communication*, 61(5):916–937.

Medin, D. L., Altom, M. W., and Murphy, T. (1984). Given Versus Induced Category Representations: Use of Prototype and Exemplar Information in Classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(3):333–352.

Medin, D. L. and Smith, E. E. (1984). Concepts and Concept Formation. *Annual Review of Psychology*, (35):113–138.

Monteith, M. J., Voils, C. I., and Ashburn-Nardo, L. (2001). Taking a look underground: Detecting, interpreting, and reacting to implicit racial biases. *Social Cognition*, 19(4):395–417.

Munton, J. (2017). The Eye's Mind: Perceptual Process and Epistemic Norms. *Philosophical Perspectives*, 31(1):317–347.

Munton, J. (2019a). Bias in a Biased System: Visual Perceptual Prejudice. In *Bias, Reason and Enquiry: New Perspectives from the Crossroads of Epistemology and Psychology*. Oxford University Press.

Munton, J. (2019b). Perceptual Skill And Social Structure. *Philosophy and Phenomenological Research*.

Murphy, G. (2004). *The big book of concepts*. MIT Press.

Newell, A. (1988). The intentional stance and the knowledge level. *Behavioral and Brain Sciences*, 11(03):520.

Newell, A. (1994). *Unified Theories of Cognition*. Harvard University Press.

Nickel, B. (2016). *Between Logic and the World*. Oxford University Press.

Nosek, B. A. and Banaji, M. R. (2001). The Go/No-Go Association Task. *Social Cognition*, 19(6):625–666.

Payne, B. K., Cheng, C. M., Govorun, O., and Stewart, B. D. (2005). An inkblot for attitudes: Affect misattribution as implicit measurement. *Journal of Personality and Social Psychology*, 89(3):277–293.

Peters, M. A. K., Kentridge, R. W., Phillips, I., and Block, N. (2017). Does unconscious perception really exist? Continuing the ASSC20 debate. *Neuroscience of Consciousness*, 2017(1).

Phillips, I. (2015). Consciousness and Criterion: On Block's Case for Unconscious Seeing. *Philosophy and Phenomenological Research*, DOI:10.1111/phpr.12224.

Phillips, I. and Block, N. (2017). Debate on Unconscious Perception. In Nanay, B., editor, *Current Controversies in Philosophy of Perception*, pages 165–192. Routledge.

Putnam, H. (1980). Brains and Behavior. In Block, N., editor, *Readings in Philosophy of Psychology*, pages 24–36. Harvard University Press.

Pylyshyn, Z. (1991). Rules and representations: Chomsky and representational realism. *The Chomskian Turn. Oxford: Basil Blackwell Limited*.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1):111–132.

Quilty-Dunn, J. and Mandelbaum, E. (2017a). Against dispositionalism: belief in cognitive science. *Philosophical Studies*.

Quilty-Dunn, J. and Mandelbaum, E. (2017b). Inferential Transitions. *Australasian Journal of Philosophy*, pages 1–16.

Reed, S. (2006). *Cognition: Theory and Appplications*. Cengage Learning, 7 edition.

Rey, G. (1983). Concepts and Stereotypes. *Cognition*, 15:237–262.

Rey, G. (1985). Concepts and Conceptions: A reply to Smith, Medin and Rips. *Cognition*, 19:297–303.

Rivers, A. M. and Hahn, A. (2018). What Cognitive Mechanisms Do People Reflect on When They Predict IAT Scores? *Personality and Social Psychology Bulletin*, page 15.

Rosch, E. (1978). Principles of categorization. In Rosch, E. and Lloyd, B. L., editors, *Cognition and Categorization*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Saul, J. (2013). Implicit bias, stereotype threat and women in philosophy. In Hutchison, K. and Jenkins, F., editors, *Women in Philosophy: What Needs to Change?*, pages 39–60. Oxford University Press.

Schwitzgebel, E. (2002). A phenomenal, dispositional account of belief. *Nous*, 36(2):249–275.

Schwitzgebel, E. (2013). A dispositional approach to attitudes: Thinking outside of the belief box. In Nottelmann, N., editor, *New Essays on Belief: Constitution, Content and Structure*, pages 75–99. Palgrave Macmillan UK, London.

Smith, E., Medin, D. L., and Rips, L. J. (1984). A psychological approach to concepts: Comments on Rey's "Concepts and stereotypes". *Cognition*, 17(3):265–274.

Smith, E. E. and Medin, D. L. (1981). *Categories and Concepts*. Harvard University Press.

Smith, E. R. and Zarate, M. A. (1990). Exemplar and Prototype Use in Social Categorization. *Social Cognition*, 8(3):243–262.

Smith, E. R. and Zarate, M. A. (1992). Exemplar-Based Model of Social Judgment. *Psychological Review*, 99(1):3–21.

Soon, V. (2019). Implicit bias and social schema: a transactive memory approach. *Philosophical Studies*.

Stabler, E. (1983). How are grammars represented. *Behavioral and Brain Sciences*, 6(3):391–421.

Sullivan-Bissett, E. (2019). Biased by our imaginings. *Mind & Language*, 34(5):627–647.

Tversky, A. and Kahneman, D. (1974). Judgement under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131.

Valian, V. (2005). Beyond Gender Schemas: Improving the Advancement of Women in Academia. *Hypatia*, 20(3):198–213.

VandenBos, G. R., editor (2015). *APA dictionary of psychology (2nd ed.)*. American Psychological Association, Washington.

Welpinghus, A. (2019). The imagination model of implicit bias. *Philosophical Studies*.

Yumusak, E. (2017). "Implicit Bias and the Unconscious". In *The 2017 Minds Online Conference*. The Brains Blog.