Received: 28 September 2023 | Revised: 4 June 2024 | Accepted: 9 July 2024

DOI: 10.1111/phc3.13011

ARTICLE

WILEY

Varieties of Bias

Gabbrielle M. Johnson

Philosophy, Claremont McKenna College, Claremont, California, USA

Correspondence
Gabbrielle M. Johnson.
Email: Gabbrielle.Johnson@ClaremontMcKenna.edu

Abstract

The concept of *bias* is pervasive in both popular discourse and empirical theorizing within philosophy, cognitive science, and artificial intelligence. This widespread application threatens to render the concept too heterogeneous and unwieldy for systematic investigation. This article explores recent philosophical literature attempting to identify a single theoretical category—termed 'bias'—that could be unified across different contexts. To achieve this aim, the article provides a comprehensive review of theories of bias that are significant in the fields of philosophy of mind, cognitive science, machine learning, and epistemology. It focuses on key examples such as perceptual bias, implicit bias, explicit bias, and algorithmic bias, scrutinizing their similarities and differences. Although these explorations may not conclusively establish the existence of a natural theoretical kind, pursuing the possibility offers valuable insights into how bias is conceptualized and deployed across diverse domains, thus deepening our understanding of its complexities across a wide range of cognitive and computational processes.

1 | INTRODUCTION

Consider three structurally similar cases of social bias. Mary's application for graduate school in mathematics is rejected by the traditionalist Mr. T, an evaluator who has written a series of books arguing that women have a natural disposition toward being worse at abstract, logical thinking than men. Her application for a different program is rejected by oblivious Ms. O, an evaluator who avows egalitarian principles but finds that Mary *just seems*

less suitable for the program, for reasons that go unarticulated and would not pan out under pressure. Her application for a third program is rejected by MatGPT, an automated program that is trained on past admittance data about which students, when accepted, have gone on to successful careers in mathematics.

As technological innovations amplify human biases, we see how biases in one domain can influence those in others, highlighting the need for a unifying theory to understand and address bias comprehensively. This challenge grows with the variety of contexts in which bias manifests. Researchers often discuss instances of bias beyond the social domain, with specialized sciences focusing on biases within their respective fields, such as the human visual perceptual system or customized learning algorithms. Moreover, these discussions often challenge the intuition that bias is always negative, acknowledging "good biases" that contribute positively to system operation. This widespread application threatens to render the concept too heterogeneous and unwieldy for systematic investigation.

This article, situated at the intersection of philosophy of science, psychology, machine learning, and ethics, integrates recent interdisciplinary insights to explore the possibility of a unified concept of *bias* that can do requisite work of providing a foundational theory for understanding and addressing biases across various domains. It examines contemporary philosophical theories about this concept and evaluates their effectiveness in explaining how different cases of bias are unified, how they differ, and why these similarities and differences matter.

In the past decade, two prominent philosophical theories have emerged to address the nature of bias: the functional view and the norm-theoretic view. The functional view, adopting a normatively-neutral approach, understands bias through its function, focusing on its teleological origins and its purposeful transitions from underdetermining inputs to determinant outputs. In contrast, the norm-theoretic view characterizes bias as systematic deviations from genuine norms, assigning an inherently pejorative status to bias. This article examines both theories. It begins with the functional view, which seeks to capture a broader concept of bias that is applicable to purported cases of both good and bad biases. It then examines the norm-theoretic view, which complements the functional view by aiming to distinguish the good from the bad.

Following dominant trends in the literature, the functional analysis approaches bias from a computational theory of mind perspective. Here, the tasks of unification and differentiation can seem at odds. For example, the more we highlight differences among how Mr. T, Ms. O, and MatGPT process information, the harder it is to use those same computational resources to say what they have in common. The functional account reconciles these tasks by shifting to a higher level of abstraction. From this perspective, it posits that bias is a functional entity with an emergent, unifying nature. Bias arises to overcome uncertainty, taking underdetermining evidential states as inputs and producing a determinate "best guess" about reality. In Mary's case, each system—Mr. T, Ms. O, and MatGPT—attempts to reason from underdetermining information about Mary's social group to a conclusion about her suitability for the program. The functional account generalizes these insights to provide a comprehensive understanding of bias. On this view, bias is that which functions to resolve the uncertainty created by the problem of underdetermination. It does this by guiding us toward some conclusions over others when they all equally fit the evidence.

The success of identifying a unified concept of bias can be measured by its benefits for both theoretical investigation and practical intervention. Accordingly, this article explores the utility of the functional theory of bias through these benefits. Conceptualizing bias as a functional entity that can be realized in various ways liberates the concept from previous conceptual constraints and opens new avenues for exploration. Historically, inquiries into the nature of bias have tended toward over-intellectualization and over-individualism, often treating Mr. T's bias as the paradigmatic case and theorizing other instances within its shadow.[1] This gives rise to a theory that focuses on particular mental states—beliefs, stereotypes, prejudices—that an individual harbors, as evidenced by their avowal of those states. Put bluntly, if you want to know if someone is biased toward some object or group, you just ask them. The functional account, however, shifts the focus from individual mental states to a broader interest in how those states are manipulated and why they're manipulated the way that they are. It explores the diverse factors, both within individuals and external to them, that underpin this function. This shift in perspective aims to deepen our understanding of how we interact with others and helps us improve to become better moral agents.

The article begins in Section 2 by presenting the two philosophical accounts that aim to give a general theory of bias, though each makes a different choice about the centrality of normativity in the analysis. It first introduces the

functional account, championed by Gabbrielle Johnson and Louise Antony, and then compares it to the norm-theoretic account, introduced and defended by Thomas Kelly. After reviewing the strengths and weaknesses of these two approaches, in Section 3, it explores benefits of the functional account's expansive interpretation of bias, showing how it extends bias to various cognitive levels, artificial intelligence, and even entire intellectual communities. It likewise compares this approach to other theories of intelligent agent inquiry, touching on stereotypes, prejudice, generics, frames, salience structures, and patterns of attention. Section 4 concludes by outlining future research avenues illuminated by the discussion, particularly for expanding the normative analysis of bias toward combating problematic bias and addressing group-based inequality in all of its forms.

2 | THE NATURE OF BIAS

Many different things are said to be biased. People, of course, are biased; but so too are groups of people, such as organizations, news programs, and political parties; as well as parts of people, like visual systems or reasoning capacities. Inanimate objects, like coins and, more recently, algorithms, are also said to be biased. What, if anything, do all these forms of bias have in common? This section explores various approaches to understanding the nature of bias across diverse contexts and systems.

Defining the concept of *bias* presents the classic challenge of conceptual analysis: providing a definition that accurately encompasses all and only relevant cases. Given the messy application of the term 'bias' across various domains, it seems unlikely that we'll be able to provide anything like necessary and sufficient conditions that capture every use of the term. Acknowledging this challenge, the theories this article sets out to examine aim instead to provide what's called a "Carnapian explication." This method transforms a messy and ambiguous concept into a clearer, more rigorous one, making it apt for more systematic theoretical and explanatory application.[1] The goal is not merely to analyze the notion of *bias*, but to offer a reformulated version better suited for specific purposes. Thus, the path this explication takes and its effectiveness as an explication will depend on what we and the theorists under discussion take our aims to be and how well the concept helps us to fulfill those aims.

One relevant consideration in this endeavor is whether bias should be considered inherently negative. Does every instance of bias warrant concern, either morally or epistemically? Disagreement over the answer leads to two distinct approaches: the functional account, which removes normativity from the concept to uncover a deeper unifying nature; and the norm-theoretic account, which maintains that all bias is inherently bad, baking normativity into the very nature of bias itself. This section explores both approaches in turn.[2]

2.1 | Bias as Function

The functional account of bias takes as its starting point the problem of underdetermination. At the account's core is the idea that bias serves an essential function in overcoming the uncertainty inherent to this problem. Louise Antony characterizes underdetermination as "the largest epistemic challenge facing any finite knower."[3] Our engagement with the empirical world is possible only through first gathering limited evidence via our senses. Crucially, this evidence in principle underdetermines the way the world actually is and can always support numerous alternative theories or conclusions. Biases act as our solution to this problem. They systematically tend us toward some hypothesis over others. As a rough general definition, we can say that *biases are non-evidential assumptions that systematically limit the inductive hypothesis space to a tractable size*.[4] Biases systematically guide ampliative inference.[5] In slogan form: bias exists anywhere induction does.[6]

The vast import of the concept arises from recognizing the ubiquity of induction as our paradigmatic gateway to empirical knowledge. Take perception as a key example. In visual perception, our eyes capture incoming light on the retina, and our visual system computes its best educated guess about the distance source of those data. The

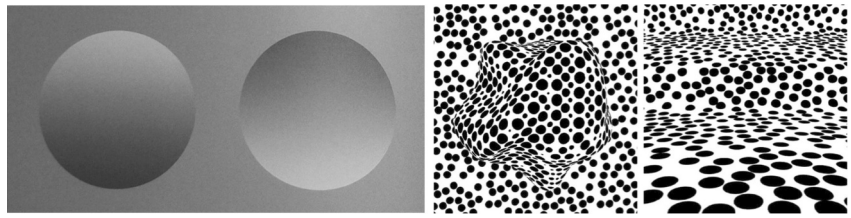

FIGURE 1 Examples of visual depth.

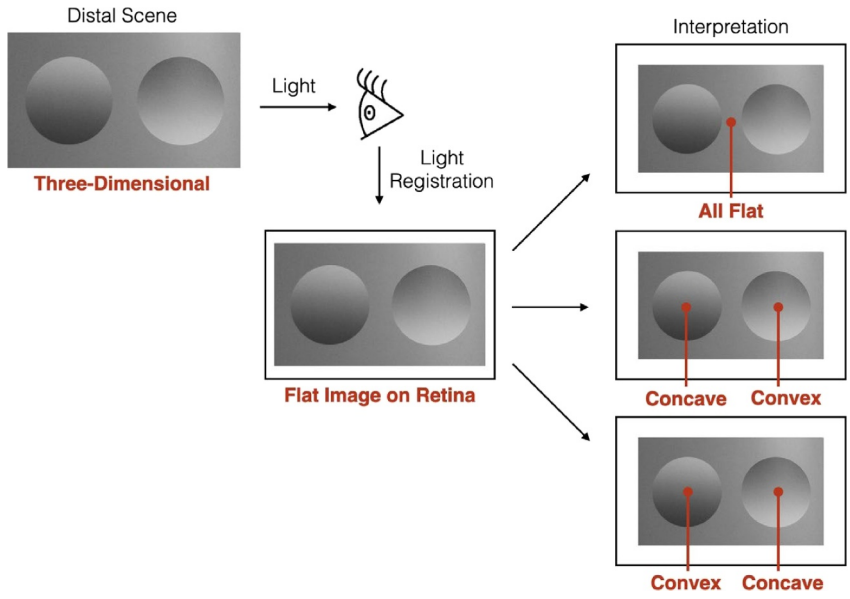

FIGURE 2 The underdetermination problem in vision.

retina receives a two-dimensional projection of a three-dimensional scene, resulting in an informational bottleneck of data that alone are insufficient for determining the exact cause. In other words, because the projection involves a dimensional reduction in information, it is always consistent with many possible scenes. Therefore, our visual system needs *something more* to settle on one interpretation among many. Bias serves as that something.

Consider the two images in Figure 1 that are similar to those often found in introductory vision science textbooks.[7]

A typically functioning visual system will usually interpret both images as having depth. For instance, in the image on the left, it generally perceives a convex circle to the left and a concave one to the right. However, this is just one interpretation among a multitude of possible ones. Let's consider three alternative interpretations presented in Figure 2.

Curiously, the "all flat" interpretation at the top is the accurate one here when simply looking at the image on the page. Strictly speaking, then, perception of this image involves an optical illusion, wherein the visual system adopts the inaccurate interpretation at the bottom. How does the visual system decide on just one interpretation when all three are compatible with the data received by the retina? Why does it settle on the one that it does, given its inaccuracy in this context? The answer lies in a built-in bias of our visual system: the visual system assumes that light is coming from above. This assumption, combined with the shading present in the flat image on the retina, systematically leads to the bottom interpretation being selected. In this way, the example illustrates the functional

role of bias described before: the evidence is the data on the retina, the possible conclusions of the inductive inference include the three possible interpretations (and countless others), and the assumption of light coming from above is the bias, which systematically narrows the inductive hypothesis space to its selected interpretation.

These visual examples illustrate two key points about the operation of bias that the functional approach generalizes to other cases. First, biases often operate outside of our awareness. Many people are familiar with the "light comes from above" bias in visual perception, but this knowledge comes from empirical investigation in vision science, not mere introspection. As forcefully argued by Orlandi (2014), visual perceptual biases are likely built into the cognitive architecture of the visual system and not available to introspection or cognition.[8] This is evident in the less-familiar example on the right in Figure 1, where the underlying assumption used by the visual system to interpret the spotted image as having depth is less obvious. While one might infer the bias through indirectly reasoning about their own visual system as a vision scientist would, it seems the natures of psychological biases are rarely obvious or available via direct introspection.[9] So, one important point about biases generally is that they are often non-obvious even to those harboring them. This suggests the de-intellectualization of our general concept of bias, as we move away from conscious access and subjective report as common criteria for some system's exhibiting bias.

Second, these examples demonstrate that biases are constitutively influenced by the wider environment, suggesting a broadly externalist (i.e., anti-individualistic) approach to bias investigation. Our visual system assumes that light comes from above because this is typical in the environment where the system evolved—for humans, light usually does come from above, courtesy of the sun. If we were instead deep-sea creatures, for whom light often emanates from below (through cracks in the Earth's mantle, say), then our visual systems would operate on an opposite assumption of the direction of light. If so, we would likely perceive the circles according to the middle interpretation in Figure 2, where perception of depth is inverted. Indeed, we get things wrong exactly in cases where the environment is abnormal, like an image cleverly disguised to exploit our tacit assumptions, leading to visual illusions.[10]

In the functional framework of bias, as presented and supported by Johnson (2020a, 2020b, 2023a, 2023b), social bias functions like these other psychological and computational biases. Specifically, social bias is a functionally-defined mental entity that takes structured mental states as inputs and yields structured mental states as outputs in ways that mimic inductions made on the basis of social-group membership.[12] For instance, the three cases of social bias mentioned at the beginning of the article all connect the input—that Mary is a woman—to the output—that she is unsuitable for a mathematics program. In this context, social bias serves a similar inductive role to that of the 'light comes from above' bias in visual perception. It navigates us from underdetermining inputs related to social-group membership to a definitive best guess about someone's traits, and it does so based on stereotypes associated with that group that aim to track regularities in the broader social environment.

Like functional analyses of other mental states, this analysis of (social) bias entails that it is multiply realizable by a variety of computational systems and decision-making processes. For instance, social biases could be realized by an explicit stereotype belief that women are ill-suited for mathematics (as Mr. T has), by an unconscious, automatic association between women and the stereotypical property of being bad at math (as Ms. O has), or by a general-purpose learning algorithm that computes a simple similarity metric among a set of training instances disproportionately comprising successful male exemplars (as MatGPT does). In the resulting view, biases are more like rules guiding inductive inferences than like stand-alone belief states. The approach again de-intellectualizes bias by extending it beyond traditional belief-like representations of stereotypes.

The functional approach to bias thereby flips the traditional script, treating the obvious and overt cases of bias as occurring in the margins. The traditional approach to bias, i.e., one that equates bias with harboring belief-like representations, captures only a modest fraction of biases responsible for how we think about and interact with the wider environment. Crucially, the traditional view fails to capture non-obvious influences on an individual's psychological capacities, like the tacit assumptions of the visual system or an unconscious transition between kinds and

properties. It is also ill-equipped to recognize biases that reside outside human agents, like those exhibited by MatGPT. The functional analysis, on the other hand, is especially well-suited to address these diverse cases.

In further departures from tradition, in treating social biases as akin to psychological and computational biases more generally, the functional account strips away negative connotations often associated with the concept *bias*. This normatively-neutral notion of bias is grounded in a wider definition, implying "a tendency; an inclination of temperament or outlook."[12] Accordingly, labeling something as *a bias* does not automatically determine its normative status. While some biases (like the visual bias of the direction of light) can be beneficial or even integral to gaining empirical knowledge, others (like those involving Mary) are obviously quite harmful and detrimental to our ability to accurately reason and learn about the world. Thus, even a normatively-neutral approach to bias in general will need an additional theory about what differentiates the good biases from the bad biases in specific cases. Toward an answer, we can explore alternative approaches that treat bias's badness as a constitutive feature of bias generally.

2.2 | The Norm-Theoretic Approach

Thomas Kelly (2022) introduces what he calls "the norm-theoretic account of bias," according to which bias is best understood as a "systematic departure from a [genuine] norm or standard of correctness."[13] The central idea behind the norm-theoretic account is to unify the sprawling concept of *bias* by recognizing how typical examples of biased reasoning involve a departure from what we regard as norms for successful decision-making.

At its core, the characterization requires that bias involves a deviation from a norm. As in the functional account, the deviation must be *systematic*, rather than accidental, arbitrary, or random. Norms here include a wide array of principles or standards in various domains of life, including the moral, epistemic, and practical.[14] For example, there likely exists an epistemic norm that we believe in accordance with our evidence. Deviating from this norm through, for example, motivated reasoning or willful ignorance would therefore be considered a form of bias.[15] Likewise, since there plausibly exists a moral norm to treat people with respect, failing to do so would also count as a bias.[16]

It's clear then how the norm-theoretic account captures the paradigmatic cases of bias with which the article started: if those evaluating Mary's application disregard evidence, e.g., her exceptional math grades, or if they refuse to consider her for admission because they don't respect women, then they are displaying bias, just as our intuitions suggest. In each case, the decision-maker departs from established moral and epistemic norms for decision-making regarding others. This shows that, on the norm-theoretic account, each instance of bias inherently involves a normative aspect: when we fall short of reasoning as we should (either morally, epistemically, or pragmatically), we have erred, making the corresponding bias inherently negative.

There are several strengths to the norm-theoretic account that Kelly outlines that highlight its role as an effective explication. At first glance, the notion that all biases are bad aligns well with our intuitions that, by and large, being biased is undesirable, and that accusations of bias are essentially accusations of error or wrongdoing. The account also sheds light on why we often disagree about whether something is biased and the perspectival way in which we label those who disagree with us as biased.[17] If bias is fundamentally a deviation from a particular norm, disagreements often emerge from contrasting perspectives on what underlying norms are operative or genuine. Accusing someone of bias communicates that they are not adhering to what the accuser perceives as the "correct" or "acceptable" standard. And finally, this account offers an explanation for what Kelly, drawing from psychological research, refers to as "the bias blindspot."[18] This is the tendency for individuals to consider themselves generally less prone to bias and more objective than others. According to the norm-theoretic view, this makes sense because each person operates based on their own set of norms, which they often regard as universally valid or objective. When people assess their own actions or beliefs through introspection, they are likely to find alignment with their personal norms, leading them to conclude that they are unbiased. Conversely, when evaluating

others, people tend to use their own norms as the standard, making it easier to perceive biases in those who don't conform to these norms.

Difficulties with the norm-theoretic account again resemble those of classic conceptual analysis: we sometimes have systematic deviations from genuine norms that we don't regard as biases (such as when I consistently mispronounce 'Gödel' as 'girdle') and we sometimes have biases that don't seem to be systematic deviations from genuine norms. For instance, as previously discussed, we often invoke the idea of good biases, like the "light comes from above" bias in vision science.[19] One might argue that these discrepancies are non-problematic since the objective was never to outline necessary and sufficient conditions for the concept *bias*, but rather to provide a Carnapian explication. To his credit, Kelly resists taking the easy way out by dismissing these discrepancies as irrelevant to the theory; instead, he takes seriously that the variations in how we use the term 'bias' are not merely accidental, but tracking something significant for the explication.[20] So, just as the normatively-neutral functional account needed a story about what makes bad biases bad, so too the norm-theoretical account needs a story about what's in common between negative and positive instances of bias.

According to Kelly (2022, p. 145), the commonality among both positive and negative instances of bias involves violation of a presumed 'symmetry norm'. This norm, roughly, stems from what we might regard as the popular wisdom that one ought to treat like cases alike. Crucially, however, Kelly rejects this popular wisdom as accurate: sometimes, he recognizes, we must treat like cases differently. He argues this point with reference to Buridan's ass, the donkey who must choose between two identical bales of hay that are equidistant from it. Were the animal to genuinely treat like cases alike, it would be immobilized by indecision, eventually starving itself to death. Thus, there must be legitimate reason for choosing between the two, treating one differently from the other. Were the donkey to make this choice in some systematic fashion, it would seem to be exhibiting a bias, insofar as we regard the symmetry norm as genuine. Whether we ultimately regard the donkey as biased likely depends on our sympathy to a normatively-neutral notion. Regardless, the sentiment that violations of a purported symmetry norm get at something fundamental about our attributions of bias aligns nicely with the functional account. Symmetry is just another instance of underdetermination—features of the hay bales underdetermine possible choices between them. Systematic violations of symmetry are themselves systematic responses to underdetermination. Thus, in both theories, bias serves as a mechanism that enables choice among equal options and, taken together, the two accounts fill in where the other was lacking, converging on the idea that biases emerge when we're confronted with situations where a neutral stance between options is untenable.[21]

Before advancing, it's useful to compare these philosophical perspectives on bias with key empirical debates in the sciences of bias over the latter part of the 20th Century. Daniel Kahneman and Amos Tversky's seminal work laid the foundation for understanding cognitive biases as systematic errors in judgment that often contradict traditional notions of rationality.[22] They highlighted that these errors are molded by heuristics, like the "availability heuristic" (Tversky & Kahneman, 1974, p. 1127) that shapes our probability assessments based on the ease with which instances come to mind. Contrastingly, Gerd Gigerenzer posited that heuristics aren't necessarily flawed but can be adaptive tools for decision-making in complex environments.[23] According to Gigerenzer, these mental shortcuts have evolved to be "fast and frugal," often yielding outcomes as favorable as more deliberative methods. In this context, the functional account of bias aligns more with Gigerenzer's adaptive view, celebrating the utility of biases and heuristics. On the other hand, the norm-theoretic account echoes Kahneman and Tversky's focus on biases as systematic deviations from rationality.[24]

Perhaps the closest historical analogue to the functional view of bias comes not from empirical psychology, but computer science. Tom Mitchell, a leading figure in the field of machine learning, is known for his seminal contributions to the understanding of learning algorithms. Mitchell's work, particularly in the 1980s, laid the groundwork for the formalization of machine learning problems. One of his most influential contributions to the field is the concept of *inductive biases*: assumptions that a learning algorithm makes to predict outputs for new, unseen data based on the data it has already encountered. Crucially, as argued by Mitchell (1980), these

assumptions are essential for the algorithm's ability to generalize. Without some form of inductive bias, an algorithm would be immobilized like Buridan's ass, unable to make any useful predictions and rendering learning impossible.

This discussion leaves us with an emerging view according to which, for both cognition and machine learning, the role of bias is essential. Biases play a critical function in facilitating our interactions with evidence, allowing for the handling of symmetries, and helping to resolve underdetermination. While the functional account posits that biases aren't inherently negative, it allows that individual instances can be either a bad thing (aligning with the norm-theoretic account) or a good thing, depending on the specific context of their emergence and application. This nuanced understanding of bias subsequently enables us to appreciate its diverse roles across various contexts, which we turn to next.

3 | VARIETIES OF BIAS

Revisiting the factors that guide a Carnapian explication of *bias*, one key consideration is how expansive our bias concept ought be. As discussed, traditional theories have predominantly focused on the mental states located within individual agents, portraying them as representations akin to beliefs, thus both individualizing and intellectualizing our conception of bias. This section explores the benefits provided by contemporary theories that extend our exploration beyond these conventional limits.

At the onset, this expansion project aligns well with ameliorative projects in philosophy. Ameliorative projects, sometimes called "conceptual engineering" projects, likewise aim to provide Carnapian explications; however, they are motivated by considerations not just of what aids in scientific explanation or logical analysis, but also by considerations of what advances political, ethical, or social justice efforts.[25]

A more encompassing definition of 'bias' beyond overt prejudices to include subtler, more insidious forms will likely contribute to such ameliorative projects. While explicit, individualized forms of bias are certainly damaging and warrant continued attention, they are now largely understood, recognized, and stigmatized, paving the way for an exploration into subtler, equally harmful forms of social injustice that might otherwise escape notice. In what follows, we'll explore how a more expansive philosophical understanding of (social) bias can likewise contribute to such projects by shedding light on a longstanding practical question: why is combating bias so challenging? To summarize, de-intellectualizing bias shows how it emerges in ways different from and less familiar than overt belief, explaining why our normal approaches to mitigation are now ill-suited for the job. Whereas belief is often thought of as consciously accessible and rationally revisable, some forms of bias seem to be neither. Likewise, de-individualizing bias shows how it persists through sources external to agents, and thus in spite of individuals' best efforts to adopt more egalitarian viewpoints. Fulfilling these ameliorative roles thus arguably contributes to any theory aiming to provide effective Carnapian explications.

In what follows, we'll explore how an expanded notion reveals neglected forms bias can take by extending the functional analysis to explorations in human cognitive architectures, algorithms and artificial intelligence, and in society and scientific practice.

3.1 | Cognitive Architecture

According to the functional view, bias is multiply realizable.[26] This explains why both recognizing bias and combating it are so difficult. Recognizing bias is difficult because it can be realized in ways that are both like and unlike traditional conscious beliefs. If some belief content is consciously accessible, then it must be explicitly represented—that is, there has to be a state to "bring up" to conscious introspection, as Mr. T is able to do with his sexist belief. However, there are multiple ways some content can be unconscious. One way, which is standard in

cognitive science, is if it is explicitly represented in some encapsulated subpersonal system, putting it "below" the purview of consciousness. Theories like this are popular in empirical studies of unconscious perception and implicit bias, and are often invoked to describe cases like Ms. O's being oblivious to her rationale for rejecting Mary.[27] However, whereas traditional theories of implicit bias simply posit the existence of unconscious biases that are difficult to correct, the functional account helps explain these features. It highlights a second way, neglected in theories of social bias until now, that some content can be unconscious: if the content is a functional abstraction from personal-level, explicitly represented states. In this way, it emerges "above" the level of conscious introspection and explicit representation. This is similar to how one might consciously represent each individual move they make in a chess game, but be unaware of the pattern they instantiate with their moves, making them surprised and confused when their opponent complains that they're always scheming to get their queen out early.[28]

Of course, this difficulty in recognizing biases will itself have implications for our ability to combat problematic biases. But even setting the question of conscious accessibility aside, the multiple realizability entailed by the functional account of bias presents independent complications for mitigation. According to the functional account, the same functional bias may manifest redundantly within an individual's psychology, occurring at different levels within cognitive architecture. Thus, the states and processes that constitute them can vary depending on whether it originates in, e.g., perception or cognition. Following Johnson (2024), consider a social bias that can be described roughly as an assumption that 'men are dangerous'. The functional account allows that this same functional bias can occur at both levels of perception and cognition. Functionally, all instances of this bias share a common input-output profile: identifying an individual as a man triggers an output that ascribes to them the stereotypical property of *being dangerous*. However, the underlying computational processes can vary for each instance. In cognition, empirical research suggests that this bias might be underwritten by a psychological assumption of essentialism: the input that a person belongs to the gender category *men* will trigger a tacit assumption of a shared gendered essence (whatever that may be) that is causally and explanatorily responsible for outward superficial properties, in particular, *being dangerous*.[30] In perception, on the other hand, this same social bias might transition from input to output simply according to a probability calculus that merely takes the having of one feature (*being male*) to raise the likelihood of having another feature (*being dangerous*), without any more robust assumptions about their causal or explanatory relationship.[31]

According to Johnson (2024), these differences subsequently have implications for how the respective biases operate, and how we intervene on each. Perceptual bias might, for example, rely on superficial past exposure to misrepresentations in the media in its tendency to readily identify ambiguous objects as weapons when held by a man. Whereas gender-essentialist assumptions might reinforce the same inference at the level of thought through an unconscious belief that men have a dangerous essence. Although the biases start and end in the same place, the differences in the inferential routes traveled will have important consequences for the strategies we adopt to mitigate them.[32] Johnson (2024) suggests that problematic cognitive social biases will be counteracted through reasoning, argumentation, and other logical interventions; whereas, problematic perceptual social biases will be counteracted through counter-stereotypical exemplar training and perceptual learning. We are then in a good position to address the long-standing question on how to combat bias: because bias is multiply realizable, combating one instance of a bias can easily leave another intact. Eliminating someone's perceptual bias might leave intact their unconscious essentialist beliefs. In fact, since a functional view recognizes that we have (social) biases springing up at various levels of some computational architecture, this theory renders the recalcitrance of social bias empirically predictable.[33]

3.2 | Algorithms and Artificial Intelligence

Emergent bias—that is, one that emerges "above" explicit representations, out of patterns of information processing—is similar to MatGPT's case, and indeed cases of it are primarily studied within computer science and machine

learning. Such socially biased patterns reflect systematic regularities in how our society is organized. Thus, they will extend to any computational system that aims to track those regularities, evidenced now by the growing concern about algorithmic bias.[33]

Standardly, 'algorithmic bias' refers to when machine learning programs trained on real-world data manifest biases found in the data themselves. This can occur in a variety of ways: biased engineers might explicitly write biases into the algorithmic code, poor data collection practices might result in unrepresentative or inaccurately labeled data, or systematic injustices might shape data to reflect unfair treatment between different social groups. Among these, the third is likely the most persistent and pervasive form of bias, since it emerges out of what seem to the engineer to be innocent approaches to data analysis (following roughly an individualist and internalist approach to system evaluation). According to Johnson (2020a, p. 9941), this often "obscures the existence of the bias itself, making it difficult to identify, mitigate, or evaluate using standard resources in epistemology and ethics." Moreover, the traditional approach that treats human biases and computer biases as fundamentally distinct will further complicate our ability to extend evaluative and mitigative resources from one domain to the other.

Progress in understanding bias can thus be facilitated through the functional model, which highlights similarities between how bias operates across different domains and encourages intellectual exchange. For instance, as already demonstrated in the case of Mary's admission to graduate school, we can identify functional similarities between MatGPT's algorithmic bias and the human biases via their inputs and outputs. Additionally, different instantiations of algorithmic biases within a single machine learning system can be recognized in much the same way as different social biases at varying levels of cognitive architecture. This allows for the recognition of how the same bias can be instantiated in different ways, such as through an explicit rule written by a biased programmer or by an emergent bias that is the result of innocuous rules together with problematic data. The same explanatory resources can then be extended to explain why mitigation and evaluation are so hard in machine learning cases.

Interestingly, this intellectual exchange between human and algorithmic bias operates in both directions: studying algorithmic bias introduces yet another way that problematic biases can emerge out of what, on the surface, seems like innocent information processing. This occurs when a system makes decisions on the basis of seemingly innocuous features that correlate with socially sensitive features, allowing those innocuous features to serve as "proxies" for the socially sensitive attributes themselves. Johnson (2020a, p. 9942) calls this "the proxy problem," and recognizing it in machine learning has prompted theorists to revise considerations of what counts as unbiased decision-making in humans.

For example, imagine now a fourth evaluator considering Mary's application. Reasonable, reliable Dr. R evaluates Mary's academic application solely based on standardized test scores and intentionally avoids any gender-identifying information. Since Mary has not performed well on these tests, her application is denied. According to highly individualizing theories of bias, Dr. R's evaluation method wouldn't be considered biased against women. However, recall that according to the functional account, biases are explained partly by their relation to the external environment they aim to track.[35] This provides the potential for identifying bias in this case, if Dr. R's criteria indirectly reflect wider discriminatory practices against women in their environment. For example, we could argue that if societal factors like unequal access to resources for test preparation ultimately explain the lower scores for women on standardized tests, and this wider pattern likewise explains why Dr. R is prone to transition from inputs to outputs in ways that exclude women, then Dr. R's evaluation, though seemingly impartial, instantiates a form of social gender bias.[36]

3.3 | Intellectual Communities

In its most expansive analysis, a functional account of bias moves beyond individuals to the wider environment. This includes intellectual communities and the biases that they collectively instantiate.[36] One relevant kind of intellectual community occurs through the exchange between humans and machines. Through its application to each

individually, the functional account provides a common framework with which to approach these mixed socio-technical systems. As always, the functional account allows us to recognize both similarities and differences in how biases manifest in humans and machines, highlighting when those differences matter. In the case of human-machine interactions, although inputs and outputs might be the same, the functional account recognizes also that important differences might exist between them. This puts us in a good position to recognize when surface-level alignment might obscure underlying differences, leading to serious misunderstandings. Jonson and Dupre (n.d.) explore this through the example of recidivism-risk algorithms. According to them, while a machine learning program might output a pairing of some individual with a property like "recidivism risk," there's no guarantee that the system's transition is based on the same kind of reasoning as when human judges make the same judgment. According to the functional account, the superficial level similarities are an invitation to investigate deeper differences, rather than simply take them for granted.

Additionally, the functional view lays a path for recognizing how biases emerge out of group-based inquiry. Johnson (2023a) extends comparisons of biases in algorithms to the adoption of values in scientific inference.[38] Drawing on the functional account of bias, these comparisons can be extended further, recognizing that the sources for values in machine learning programs and scientific practice more generally are the same sources that give rise to the need for bias in everyday thought and inquiry. Scientific practice can thus be viewed as a model for evaluating intellectual communities for the biases they collectively harbor.

We can thus expand lessons from feminist philosophy of science to the evaluation of biases in society more broadly, providing useful frameworks for addressing biases across society's intellectual communities. For instance, Heather Douglas's (2000) work shows the importance of integrating moral and societal values in our evaluation of inductive risk, likewise providing a risk-based model for forming everyday beliefs in inferences about others.[38] Sandra Harding (1992) argues for the distinctive epistemic advantages gained through the occupation of marginalized standpoints.[39] Such advantages are further validated through Helen Longino's (1995) emphasis on the role of diversity and communal interaction in knowledge production. By fostering intellectual communities that both recognize the indispensable role of marginalized perspectives and themselves better reflect the complexities of our diverse world, we can achieve a more balanced approach to group-based decision-making that aims to offset historical and individualized biases within groups.

3.4 | Fellow Travelers

The mechanisms intelligent agents employ to learn about and interact with their environments have been explored through countless theoretical frameworks. This section concludes by examining some of these parallel theoretical frameworks, or 'Fellow Travelers,' whose work aligns with the themes discussed so far. These theories investigate how information is arranged, perceived, constructed, and expressed, offering a variety of insights into how humans gather information about the world around them. Each of these perspectives contributes to a comprehensive understanding of how bias operates. By integrating the functional theory of bias with this broad range of theories, we likely move closer to a more complete understanding of how intelligent systems interact with and investigate the world.

To begin, no theory of bias would be complete without a comprehensive understanding of parallel work on stereotypes and prejudices, both within empirical psychology and philosophy. For example, the functional account's adoption of a normatively-neutral approach to bias is complemented by similar work by Erin Beeghly (2015) on a descriptive theory of stereotypes.[40] In similar spirit, Endre Begby's (2021) theory of prejudice carves space for morally unacceptable but epistemically justifiable treatment of others, again highlighting the complex interactions between moral and epistemic values in bias manifestation.

It's also important to understand how bias is facilitated in language and other forms of expression. For example, Elisabeth Camp's (2023) theory of perspectives and frames investigates how representational vehicles, like

metaphors and slurs, mold bias-like structures of thought. Crucially, such structures pre-empt the functional account's transition away from stand-alone belief states, so as to emphasize the importance of how frames subtly shape our transitions between attitudes, beliefs, and judgments. Ultimately such frames help to understand how patterns of communication act to reinforce societal biases.[41] Another research program lending to our understanding of how language naturally facilitates and perpetuates biased thinking comes in the form of work on generics. For example, Sarah-Jane Leslie (2017) provides a theory of what she calls "striking property generalizations" that she believes contribute to the essentialization of social kinds and perpetuate social stereotypes.[42] Ultimately this theory suggests that by changing the way we express certain generic formulations involving social kinds, we can combat the spread of social prejudices. In similar spirit, Eleonore Neufeld (2019, 2020) and Kate Ritchie (2021a, 2021b) argue that the mere adoption of predicate nominals can have essentializing effects, further indicating how language and cognitive structures embed and perpetuate biases systematically.

And finally, there's a growing interest in exploring how bias can manifest as patterns of attention and salience. According to Wayne Wu, attention results from our needing to solve what he calls 'the selection problem', similar in kind to the underdetermination and symmetry cases discussed above. According to Wu (2023a, p. 79), bias will be integral to the story of how attention achieves this, since "attentional phenomena necessarily reflect biases that solve the Selection Problem."[43] As before, we can generalize this fundamental understanding of the role of bias and attention in cognitive psychology to better understand how attention and salience influence us at the level of social interaction, guiding what we would regard as better or worse engagement with others. For example, Susanna Siegel (2017), in her investigation into the potential rationality of perception, gives a theory of how patterns of attention in perception can influence agents in ways that undermine empirical warrant for belief based on that perception. She then extends these points about attention and salience to better understand how journalistic practices influence and bias what the wider public deems as important, like when newspapers choose which stories to highlight on the front page.[44] Jessie Munton (2021) argues for a theory of prejudice as the misattribution of salience, wherein we unduly organize our prioritizing of information around social categories. Like the functional account of bias and Camp's theory of frames, this view emphasizes the importance of moving our theories of bias beyond individualized mental states, to how information is organized, prioritized, and accessed when we reason about the world around us. Likewise, Ella Whiteley (2022) gives a theory of harmful salience perspectives according to which individuals can experience harm if contingent features of their identity, i.e., those not reflecting personhood, are excessively emphasized in the minds of others, undermining their agency.[45]

This only scratches the surface of connections to the wider philosophical literature that will ultimately be important for a general theory of bias. Uniting perspectives across these domains will thereby inevitably deepen our holistic understanding of bias.

4 | CONCLUSION AND FUTURE RESEARCH

This article has explored recent philosophical literature positing a unified theory of bias through a Carnapian explication, aiming to enhance both theoretical understanding and practical application. It focused on two recent advancements: a functional account, recognizing bias's role in overcoming underdetermination across various contexts; and a norm-theoretic view, which characterizes bias as systematic deviations from genuine norms. By integrating these perspectives, the discussion highlighted previously neglected forms of bias and elucidated practical obstacles to eliminating problematic bias. To conclude, we can explore future research avenues that this comprehensive framework illuminates.

One major research avenue involves expanding the normative analysis. By no longer centering individuals and their high-level mental states, we must move beyond traditional evaluative resources in ethics and epistemology that took these as starting points. Following the norm-theoretic account, the project would be to provide a taxonomy of genuine norms—epistemic, moral, and beyond—that apply to systems beyond high-level belief and

reasoning in human agents. This includes artificial and subpersonal systems, as well as intellectual communities, expanding the scope of our analysis to encompass a diverse array of systems and the norms governing them.

Once we've identified what bias is and how to evaluate it, we can then get to the most pressing question of how we combat it. Thus, a second major research avenue further explores how we mitigate problematic biases. If one adopts the functional account as outlined here, then given the multifaceted nature of bias, two possible mitigation strategies emerge: one focuses on bias's general functional role, while the other addresses individual differences in how biases fulfill that function. Biases function generally as internal mechanisms that aim to mimic external environmental regularities. Therefore, it seems we can adopt a range of internal, external, or combined strategies to reshape that functional operation. Internal strategies aim to change individualized mechanisms that realize bias, such as manipulating training data, adjusting input filters, or blocking inferences. External strategies seek to alter the regularities that biases aim track, essentially changing the world. Each approach has its strengths and weaknesses. Internal mechanistic interventions are typically easier to implement but often provide short-term solutions. In contrast, external environmental interventions are more challenging to achieve but offer long-lasting effects once successful. Ultimately, given the feedback loops between the internal and external factors underpinning bias's function, a mixed strategy is likely necessary for complete amelioration. Future research should focus on developing concrete, function-based strategies that leverage these insights to mitigate problematic biases across domains, ultimately reducing group-based inequality in its many forms.

ACKNOWLEDGEMENTS

I am grateful for help and valuable feedback from Carolina Flores, Gabriel Greenberg, Katie Elliott, Susanna Siegel, Jessie Munton, and Elli Neufeld, as well as editor Nico Orlandi and the anonymous referee for Philosophy Compass.

ORCID

Gabbrielle M. Johnson https://orcid.org/0000-0003-1463-4496

ENDNOTES

[1] A theory is *intellectualizing* in the sense used throughout this article to the extent that it regards intellectual abilities (e.g., sentence-like representations, conscious accessibility, ability to self-report) as pre-conditions for bias. For example, if it turns out that some biases are unconscious, then any theory that defines 'bias' as consciously accessible would be overly intellectualizing. A theory is *individualizing* in the sense used throughout this article to the extent that it regards features of individuals (e.g., the states and computational processes internal to them) as solely constitutive of bias. For example, if semantic externalism is true, then any theory that says Oscar on Earth and Twin-Oscar on Twin-Earth have the same biases toward what they each call 'water' would be overly individualizing.

[2] Carnap (1950); for discussion of Carnapian explication in the two views of bias to follow, see Johnson (2020b), p. 1195 and Kelly (2022), p. 151, respectively.

[3] For an alternative analysis of bias as a natural kind, see Khalidi and Mugg's (2023) discussion on the plausibility of biases and heuristics as natural cognitive kinds. They distinguish cognitive heuristics and biases from social or perceptual instances, focusing on reasoning, inference, and decision-making in humans (omitting also consideration of artificial systems). Their notion of *heuristic* aligns more closely with what this article describes as a normatively neutral concept of bias that underlies normatively problematic instances. Like the functional account in this article, Khalidi and Mugg consider the normatively neutral notion to be a better candidate for a natural kind than the norm-theoretic concept, which incorporates normativity into its definition. However, they ultimately argue that neither notion is likely a good candidate for a cognitive kind, due to their emphasis on common causal origins. Their analysis restricts the concept of cognitive kind to those with unified causal processes, a more limited scope than the natural kind concept explored here, which focuses instead on "the role the concept *bias* plays in psychological [i.e., computational] explanations," not all of which are taken to be causal (Johnson, 2020b, p. 1195).

[4] Antony (2016), p. 161. See also Antony (2000, 2001).

[5] Antony (2016), p. 161.

[6] Following Johnson (2023a, p. 4), we can consider an inference to be inductive if it is ampliative, i.e., non-deductive. This includes enumerative induction and abduction, i.e., inference to the best explanation.

[7] For a thorough discussion of the relationship between underdetermination, induction, and bias, see Johnson (2023a). Johnson uses the terms 'canons', 'values', 'virtues', or 'biases' interchangeably to denote the same natural kind.

[8] Right image borrowed from Todd and Oomes (2002), p. 847. For more on the problem of underdetermination in vision, see Palmer (1999), p. 5 ff and Burge (2010), p. 89.

[9] For more discussion of this point as it relates to a functional account of social bias, see Johnson (2020b).

[10] Roughly, the visual system assumes that patterns are homogenous. If all the circles are assumed to be approximately the same size and shape, then the differences in their elliptical retinal projections are taken to result from a slanted surface, leading to the perception of a slant.

[11] This focus on how the environment can be more or less hospitable to cognitive biases in the face of underdetermination has recently been taken up by C. Thi Nguyen (2023) in his exploration of what he calls "hostile epistemology," extending the idea to include high-level threats to cognitive biases and heuristics, such as the misleading allure of clarity (elaborated in Nguyen, 2021).

[12] Johnson (2020b, p. 1226, fn. 57) defines these inputs as propositions, but clarifies that her notion of proposition includes non-sentence-like structures, e.g., pictorial representations found in the visual system. Minimally what is required is structure that is sufficient to support attribution of properties to individuals.

[13] Antony (2016), p. 161. See also Munton's (2019b, p. 1) notion of a *formal bias*. For detailed discussion of a normatively-neutral approach to the related notion of *stereotype*, see Beeghly (2015). I discuss the relationship between theories of bias and stereotypes further in Section 3.4.

[14] Kelly (2022), p. 63.

[15] More generally, we can follow Burge (2020, p. 38) in regarding *norms* as "standards for success, or for contributing to success, in fulfilling a function, purpose, or goal."

[16] See Kelly (2022), pp. 143–144 for discussion. For insightful work on ignorance, see Rosen (2002), Moody-Adams (1994), Mills (2007), Martín (2021), Miller Larsen (2023).

[17] It's possible also that moral, epistemic, and practical norms intersect with one another in interesting ways. For more discussion, see canonical work in literature on pragmatic and moral encroachment: Stanley (2005), Fantl and Mcgrath (2007), Moss (2018), Basu (2018, 2019), Bolinger (2018), Gardiner (2018); Munton (2017, 2019a, 2019c), among others.

[18] Kelly (2022), pp. 68–77.

[19] Kelly (2022), p. 88 ff.; elaborated also in Kelly (2023).

[20] For example, Burge (2005, p. 14) refers to the principles guiding the visual system in overcoming underdetermination "biasing principles". For further discussion on this choice of terminology, see Burge (2010, p. 92, fn. 41). Kelly (2023, pp. 19, 146) likewise cites Burge (1979, p. 116) as using 'bias' in a positive sense.

[21] See Kelly's (2022, p. 147) contrast of different uses of 'bias' and different uses of 'bank'.

[22] In a similar spirit, Wayne Wu invokes bias as central to his characterization of attention in order to solve what he calls 'the Selection Problem', which he sometimes demonstrates also with an appeal to Buridan's ass (Wu, 2023b, p. 3). I discuss the relationship between bias, attention, and priority further in Section 3.4. For Wu's discussion of the functional account of bias, see Wu (2023a, p. 162 ff).

[23] For canonical work, see Tversky and Kahneman (1974, 1983), Kahneman (2011).

[24] For canonical work, see Gigerenzer and Todd (2001); Gigerenzer (2002, 2008, 2010).

[25] I regret not being able to give a more nuanced discussion to this comparison. I will however note one complication: one might wonder whether all biases studied by Kahneman and Tversky can be framed as solutions to underdetermination problems. Some, like the status quo or confirmation bias, might seem odd characterized as responses to equally viable alternatives, but rather signify a disregard for whether conclusions are strongly supported or refuted by the evidence. A proponent of the functional approach could respond by appealing to bias's teleological function. Even if a specific instance of a psychological bias is (over)determined, they could argue, it likely evolved from capacities for solving substantial underdetermination. However, given the complexity of many cognitive biases and their likely mixed-process origins, pinpointing the exact underdetermination problems they evolved to tackle remains a challenging endeavor. I thank Tom Kelly for raising this concern.

[26] Haslanger (2000, 2012), Burgess and Plunkett (2013a, 2013b), Cappelen (2018), Chalmers (2020), Burgess et al. (2020).

[27] This fits well with claims about the heterogeneity of (implicit) bias, see Holroyd and Sweetman (2016), and views on which the content of bias is indeterminate, see Madva (2017).

[28] For a broad sampling of empirical and philosophical approaches to implicit bias, see Banaji and Greenwald (2013), Dovidio et al. (2005), Gaertner and Dovidio (1986), Fazio (1990), Gawronski and Bodenhausen (2006); Gawronski and Bodenhausen (2014) on the one hand; and Holroyd et al. (2017), Del Pinal and Spaulding (2018), Levy (2015), Mandelbaum (2015), Machery (2016), Soon (2019), Welpinghus (2019), Sullivan-Bissett (2019), Nanay (2021), Karlan (2021), Toribio (2021) on the other.

[29] This example comes originally from Dennett (1981, p. 107). See Johnson (2023b) for a comprehensive exploration of how different views about the structure of bias can intersect in interesting ways with questions of conscious accessibility.

[30] For canonical work on psychological essentialism, see Rothbart and Taylor (1992); Gelman (2004). For empirical work suggesting that social kinds are psychologically essentialized, see Haslam et al. (2000), Rhodes and Mandalaywala (2017); Pauker et al. (2010). For philosophical discussion of the implications of psychological essentialism for how we reason and talk about others, see Strevens (2000), Leslie (2017), Neufeld (2019, 2020, 2022), Ritchie (2021a, 2021b), among others. I discuss the relationship between bias and psychological essentialism further in Section 3.4.

[31] While empirical evidence directly linking *being a man* and *being dangerous* in perception is currently unavailable, existing literature does suggest that this example is empirically plausible (Correll et al., 2002; Cloutier et al., 2014; Eberhardt et al., 2004; Payne, 2001; Trawalter et al., 2008; Wilson et al., 2017, among others). See Johnson (2023a), footnote 32 for discussion.

[32] Johnson (2023a) additionally argues that these differences might likewise be relevant for differential epistemic evaluation. To summarize, cognitive social biases that encode theoretical assumptions about category membership are justified only when those assumptions are true. Since essentialist assumptions are rarely true for social kinds, it is clear why many social biases are problematic. Perceptual social biases, based on statistical regularities, seem justified when those regularities hold. However, Munton (2019c) uses a skill-based approach to perception to argue compellingly for a diminishment of epistemic evaluation of perceptual social biases due to considerations of counterfactual stability.

[33] For meta-analyses of empirical evidence for recalcitrance, see Lai et al. (2014), Lai et al. (2016), Forscher et al. (2019).

[34] For a broad sampling of empirical and philosophical approaches to algorithmic bias, see Dwork et al. (2011), O'Neil (2016), Caliskan et al. (2017, 2022), Noble (2018) on the one hand; and Barocas and Selbst (2016), Johnson (2020a), Fazelpour and Danks (2021) on the other.

[35] In this way, the functional account is tied to other externalist commitments in philosophy of mind, language, and epistemology. By recognizing the influence of discriminatory patterns in the wider environment, the theory takes seriously structural accounts of bias and stereotyping (Ayala Lopez & Beeghly, 2020; Haslanger, 2015; Vasilyeva & Lombrozo, 2020; Vasilyeva et al., 2018) as integral to a general theory of bias, while trying to reconcile so-called "individualist" and "structuralist" perspectives (Haslanger, 2016; Madva, 2016; Saul, 2018b) by denying one can fully individuate the bias from the wider structures in which it operates (for further discussion, see Johnson (2020b), p. 1219, fn. 48; compare Madva et al. (2023)).

[36] This case mimics roughly the distinction in the legal literature between theories of direct and indirect discrimination. For more discussion of the proxy problem and its relation to these legal theories, see Prince and Schwarcz (2020), Hellman (2023a, 2023b), Hu and Kohler-Hausmann (2020), Hu (2023), and Johnson (n.d.). Insofar as proxies involve problematic content being conveyed in implicit, seemingly innocuous ways, the proxy problem shares important theoretical similarities with so-called "trojan horse propaganda" and dogwhistles (see especially Taiwo, 2017, 2018; Saul, 2018a).

[37] Liao and Huebner (2021) extend the functional account of bias even further, recognizing bias as existing wholly outside of individuals, and giving a theory of "oppressive objects".

[38] For further discussion of the role of values in algorithms and scientific practice, see Rudner (1953), Levi (1960), Douglas (2000, 2009, 2016), Rooney (1992), Longino (1995, 1996), Dotan (2020), and Kasirzadeh and Gabriel (2023), among others.

[39] This sentiment has been largely taken up in the recent surge in work on pragmatic and moral encroachment. See citations in note 17.

[40] For helpful introductions and summaries of standpoint epistemology, see Toole (2019, 2023).

[41] Beeghly likewise provides a comprehensive review of the empirical work in psychology on stereotypying and prejudice. For a broad sampling of empirical and philosophical approaches to stereotypes and prejudice, see Rosch (1978), Allport (1979), Hilton and Von Hippel (1996), Blum (2004), Schneider (2004), Dovidio et al. (2010), Fiske et al. (2010), VandenBos (2015) on the one hand, and Lippert-Rasmussen (2014), Del Pinal et al. (2017), Puddifoot (2021), on the other.

[42] See also Flores and Camp (2023) and Flores and Woodard (2023).

[43] See also Leslie (2014, 2015).

[44] Wu is here discussing relevant work by Watzl (2017) on priority.

[45] Siegel (2022).

[46] For more insightful work on how patterns of attention can inform our normative evaluation of inquiry, see Watzl (2017), Gardiner (2022), Saint-Croix (2022); Irving (2016); Irving and Glasser (2020), among others.

REFERENCES

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

Antony, L. (2001). Quine as feminist: The radical import of naturalized epistemology. In L. Antony & C. E. Witt (Eds.), *A mind of one's own: Feminist essays on reason and objectivity* (pp. 110–153). Westview Press.

Antony, L. (2016). Bias: Friend or foe? In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy, volume 1: Metaphysics and epistemology* (pp. 157–190). Oxford University Press.

Antony, L. M. (2000). Naturalized epistemology, morality, and the real world. *Canadian Journal of Philosophy*, *30*(sup1), 103–137. https://doi.org/10.1080/00455091.2000.10717550

Ayala Lopez, S., & Beeghly, E. (2020). Explaining injustice: Structural analysis, bias, and individuals. In E. Beeghly & A. Madva (Eds.), *Introduction to implicit bias: Knowledge, justice, and the social mind*. Routledge.

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

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*.

Basu, R. (2018). *The wrongs of racist beliefs*. Philosophical Studies.

Basu, R. (2019). What we epistemically owe to each other. *Philosophical Studies*, *176*(4), 915–931. https://doi.org/10.1007/s11098-018-1219-z

Beeghly, E. (2015). What is a stereotype? What is stereotyping? *Hypatia*, *30*(4), 675–691. https://doi.org/10.1111/hypa.12170

Begby, E. (2021). *Prejudice: a study in non-ideal epistemology*. Oxford University Press.

Blum, L. (2004). Stereotypes and stereotyping: A moral analysis. *Philosophical Papers*, *33*(3), 251–289. https://doi.org/10.1080/05568640409485143

Bolinger, R. J. (2018). The rational impermissibility of accepting (some) racial generalizations. *Synthese*, *197*(6), 2415–2431. https://doi.org/10.1007/s11229-018-1809-5

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy*, *4*, 73–121. https://doi.org/10.1111/j.1475-4975.1979.tb00374.x

Burge, T. (2005). Disjunctivism and perceptual psychology. *Philosophical Topics*, *33*(1), 1–78. https://doi.org/10.5840/philtopics20053311

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

Burge, T. (2020). Entitlement: The basis for empirical warrant. In P. J. Graham & Pedersen (Eds.), *Epistemic entitlement* (pp. 37–142). Oxford University Press.

Burgess, A., Cappelen, H., and Plunkett, D. (Eds.) (2020). *Conceptual engineering and conceptual ethics* (1st ed.). Oxford University Press. OCLC: on1103006363.

Burgess, A., & Plunkett, D. (2013a). Conceptual ethics I: conceptual ethics I. *Philosophy Compass*, *8*(12), 1091–1101. https://doi.org/10.1111/phc3.12086

Burgess, A., & Plunkett, D. (2013b). Conceptual ethics II: conceptual ethics II. *Philosophy Compass*, *8*(12), 1102–1110. https://doi.org/10.1111/phc3.12085

Caliskan, A., Ajay, P. P., Charlesworth, T., Wolfe, R., & Banaji, M. R. (2022). Gender bias in word embeddings: A comprehensive analysis of frequency, syntax, and semantics. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 156–170). ACM.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Camp, E. (2023). Perspectives in imaginative engagement with fiction. In W. Riggs & N. Snow (Eds.), *Open-Mindedness and Pespective*. OUP.

Cappelen, H. (2018). *Fixing language: an essay on conceptual engineering* (1st ed.). Oxford University Press. OCLC: on1005880448.
Carnap, R. (1950). *Logical Foundations of Probability*. Chicago University Press.
Chalmers, D. J. (2020). What is conceptual engineering and what should it be? *Inquiry*, 1–18. https://doi.org/10.1080/0020174x.2020.1817141
Cloutier, J., Li, T., & Correll, J. (2014). The Impact of Childhood Experience on Amygdala Response to Perceptually Familiar Black and White Faces. *Journal of Cognitive Neuroscience*, *26*(9), 1992–2004. https://doi.org/10.1162/jocn_a_00605
Correll, J., Park, B., Judd, C. M., & 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. https://doi.org/10.1037//0022-3514.83.6.1314
Del Pinal, G., Madva, A., & Reuter, K. (2017). Stereotypes, conceptual centrality and gender bias: An empirical investigation: Stereotypes, conceptual centrality and gender bias. *Ratio*, *30*(4), 384–410. https://doi.org/10.1111/rati.12170
Del Pinal, G., & Spaulding, S. (2018). Conceptual centrality and implicit bias. *Mind & Language*, *33*(1), 95–111. https://doi.org/10.1111/mila.12166
Dennett, D. C. (1981). A cure for the common code. In *Brainstorms: Philosophical essays on mind and psychology* (pp. 90–108). MIT Press.
Dotan, R. (2020). Theory choice, non-epistemic values, and machine learning. *Synthese*, *198*(11), 11081–11101. https://doi.org/10.1007/s11229-020-02773-2
Douglas, H. (2000). Inductive risk and values in science. *Philosophy of Science*, *67*(4), 559–579. https://doi.org/10.1086/392855
Douglas, H. (2016). Values in science. In P. Humphreys (Ed.), *Oxford Handbook in the Philosophy of Science* (pp. 609–630). Oxford University Press.
Douglas, H. E. (2009). *Science, policy, and the value-free ideal*. University of Pittsburgh Press. OCLC: ocn297144848.
Dovidio, J. F., Glick, P. S., & Rudman, L. A. (Eds.) (2005). *On the nature of prejudice: Fifty years after Allport*. Blackwell Pub.
Dovidio, J. F., Hewstone, M., Glick, P., & Esses, V. (2010). Stereotyping and discrimination: Theoretical and empirical overview. In *The SAGE handbook of prejudice, stereotyping, and discrimination* (pp. 3–23). SAGE Publications.
Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2011). Fairness through awareness. *arXiv:1104.3913 [cs]*.
Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, *87*(6), 876–893. https://doi.org/10.1037/0022-3514.87.6.876
Fantl, J., & Mcgrath, M. (2007). On pragmatic encroachment in epistemology. *Philosophy and Phenomenological Research*, *75*(3), 558–589. https://doi.org/10.1111/j.1933-1592.2007.00093.x
Fazelpour, S., & Danks, D. (2021). Algorithmic bias: Senses, sources, solutions. *Philosophy Compass*, *16*(8). https://doi.org/10.1111/phc3.12760
Fazio, R. H. (1990). Multiple processes by which attitudes guide behavior: The Mode model as an integrative framework. *Advances in Experimental Social Psychology*, *23*, 75–109. https://doi.org/10.1016/s0065-2601(08)60318-4
Fiske, S. T., Gilbert, D. T., & Lindzey, G. (2010). *Handbook of social psychology* (Vol. 2). John Wiley & Sons.
Flores, C., & Camp, E. (2023). "That's All You Really Are": Centering social idenities and essentialist beliefs. In S. Haslanger (Ed.), *Mind, Language and Social Hierarchy*. OUP.
Flores, C., & Woodard, E. (2023). Epistemic norms on evidence-gathering. *Philosophical Studies*, *180*(9), 2547–2571. https://doi.org/10.1007/s11098-023-01978-8
Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of change in implicit bias. *Journal of Personality and Social Psychology*, *117*(3), 522–559. https://doi.org/10.1037/pspa0000160
Gaertner, S. L., & Dovidio, J. F. (1986). The aversive form of racism. In S. L. Gaertner & J. F. Dovidio (Eds.), *Prejudice, discrimination, and racism* (pp. 61–89). Academic Press.
Gardiner, G. (2018). Evidentialism and moral encroachment. In K. McCain (Ed.), *Believing in accordance with the evidence: new essays on evidentialism* (pp. 169–195). Springer.
Gardiner, G. (2022). Attunement: on the cognitive virtues of attention. In M. Alfano, J. d. Ridder, & C. Klein (Eds.), *Social virtue epistemology* (pp. 48–72). Routledge.
Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin*, *132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692
Gawronski, B., & 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. https://doi.org/10.1111/spc3.12124
Gelman, S. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, *8*(9), 404–409. https://doi.org/10.1016/j.tics.2004.07.001

Gigerenzer, G. (2002). *Adaptive thinking: Rationality in the real world*. Evolution and cognition. Oxford University Press.
Gigerenzer, G. (2008). Why heuristics work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x
Gigerenzer, G. (2010). *Rationality for mortals: How people cope with uncertainty*. Evolution and cognition. Oxford University Press.
Gigerenzer, G., & Todd, P. M. (2001). *Simple heuristics that make us smart. Evolution and cognition*. Oxford University Press. 1. Issued as an Oxford Univ. press paperback edition.
Harding, S. (1992). Rethinking standpoint epistemology: What is 'Strong Objectivity'? *Centennial Review*.
Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, *39*(1), 113–127. https://doi.org/10.1348/014466600164363
Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Nouˆs*, *34*(1), 31–55. https://doi.org/10.1111/0029-4624.00201
Haslanger, S. (2015). Distinguished lecture: Social structure, narrative and explanation. *Canadian Journal of Philosophy*, *45*(1), 1–15. https://doi.org/10.1080/00455091.2015.1019176
Haslanger, S. (2016). What is a (social) structural explanation? *Philosophical Studies*, *173*(1), 113–130. https://doi.org/10.1007/s11098-014-0434-5
Haslanger, S. A. (2012). *Resisting reality: Social construction and social critique*. Oxford University Press.
Hellman, D. (2023a). Big data and compounding injustice. *Journal of Moral Philosophy*, *21*(1–2), 62–83. https://doi.org/10.1163/17455243-20234373
Hellman, D. (2023b). Defining disparate treatment: A research agenda for our times. *University of Virginia School of Law*, *99*, 205.
Hilton, J. L., & Von Hippel, W. (1996). Stereotypes. *Annual Review of Psychology*, *47*(1), 237–271. https://doi.org/10.1146/annurev.psych.47.1.237
Holroyd, J., Scaife, R., & Stafford, T. (2017). What is implicit bias? *Philosophy Compass*, *12*(10), e12437. https://doi.org/10.1111/phc3.12437
Holroyd, J., & Sweetman, J. (2016). The heterogeneity of implicit bias. In M. Brownstein & J. Saul (Eds.), *Implicit bias and philosophy volume 1: Metaphysics and epistemology* (pp. 80–103). Oxford University Press.
Hu, L. (2023). What is "Race" in algorithmic discrimination on the basis of race? *Journal of Moral Philosophy*, *21*(1–2), 1–26. https://doi.org/10.1163/17455243-20234369
Hu, L. and Kohler-Hausmann, I. (2020). What's sex got to do with fair machine learning? *arXiv preprint arXiv:2006.01770*. 11.
Irving, Z. C. (2016). Mind-wandering is unguided attention: accounting for the "purposeful" wanderer. *Philosophical Studies*, *173*(2), 547–571. https://doi.org/10.1007/s11098-015-0506-1
Irving, Z. C., & Glasser, A. (2020). Mind-wandering: A philosophical guide. *Philosophy Compass*, *15*(1), e12644. https://doi.org/10.1111/phc3.12644
Johnson, G. M. (n.d.). Proxies aren't intentional; they're intentional. Unpublished Manuscript.
Johnson, G. M. (2020a). Algorithmic bias: On the implicit biases of social technology. *Synthese*, *198*(10), 9941–9961. https://doi.org/10.1007/s11229-020-02696-y
Johnson, G. M. (2020b). The structure of bias. *Mind*, *129*(516), 1193–1236. https://doi.org/10.1093/mind/fzaa011
Johnson, G. M. (2023a). Are algorithms value-free? Feminist theoretical virtues in machine learning. *Journal of Moral Philosophy*, *21*(1–2), 27–61. https://doi.org/10.1163/17455243-20234372
Johnson, G. M. (2023b). Unconscious perception and unconscious bias: Parallel debates about unconscious content. In *Oxford studies in philosophy of mind* (Vol. 3, pp. 87–130). OUP. https://doi.org/10.1093/oso/9780198879466.003.0004
Johnson, G. M. (2024). The (dis)unity of psychological (social) bias. *Philosophical Psychology*, 1–29. https://doi.org/10.1080/09515089.2024.2366418
Jonson, G. M., & Dupre, G. (n.d.). Uncanny performance, divergent competence. Unpublished Manuscript.
Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.
Karlan, B. (2021). The rational dynamics of implicit thought. *Australasian Journal of Philosophy*, *100*(4), 1–15. https://doi.org/10.1080/00048402.2021.1936581
Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, *36*(2), 27. https://doi.org/10.1007/s13347-023-00606-x
Kelly, T. (2022). Bias: A philosophical study. In *Bias* (1 ed., pp. 17–C1.P92). Oxford University Press.
Kelly, T. (2023). Bias, norms, introspection, and the bias blind spot 1. *Philosophy and Phenomenological Research*, phpr.12953.
Khalidi, M. A., & Mugg, J. (2023). Cognitive heuristics and biases. In *Cognitive ontology: Taxonomic practices in the mind-brain sciences* (1 ed., pp. 181–209). Cambridge University Press.
Lai, C. K., Marini, M., Lehr, S. A., Cerruti, C., Shin, J.-E. L., Joy-Gaba, J. A., Ho, A. K., Teachman, B. A., Wojcik, S. P., Koleva, S. P., Frazier, R. S., Heiphetz, L., Chen, E. E., Turner, R. N., Haidt, J., Kesebir, S., Hawkins, C. B., Schaefer, H. S., Rubichi,

S., … Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, *143*(4), 1765–1785. https://doi.org/10.1037/a0036260

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marhsburn, C. K., Simon, S., Blanchar, J. C., Joy-Gaba, J. A., Conway, J., Redford, L., Klein, R. A., Roussos, G., Schellhaas, F. M. H., Burns, M., …, & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*(8), 1001–1016. https://doi.org/10.1037/xge0000179

Leslie, S.-J. (2014). Carving up the social world with generics. *Oxford Studies in Experimental Philosophy*, *1*, 208–232.

Leslie, S.-J. (2015). Generics oversimplified. *Noûs*, *49*(1), 28–54. https://doi.org/10.1111/nous.12039

Leslie, S.-J. (2017). The original sin of cognition: Fear, prejudice, and generalization. *The Journal of Philosophy*, *114*(8), 393–421. https://doi.org/10.5840/jphil2017114828

Levi, I. (1960). Must the scientist make value judgments? *The Journal of Philosophy*, *57*(11), 345. https://doi.org/10.2307/2023504

Levy, N. (2015). Neither fish nor fowl: Implicit attitudes as patchy endorsements. *Noûs*, *49*(4), 800–823. https://doi.org/10.1111/nous.12074

Liao, S., & Huebner, B. (2021). Oppressive things*. *Philosophy and Phenomenological Research*, *103*(1), 92–113. https://doi.org/10.1111/phpr.12701

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

Longino, H. E. (1995). Gender, politics, and the theoretical virtues. *Synthese*, *104*(3), 383–397. https://doi.org/10.1007/bf01064506

Longino, H. E. (1996). Cognitive and non-cognitive values in science: Rethinking the dichotomy. In L. Hankinson Nelson & J. Nelson (Eds.), *Feminism, science, and the philosophy of science* (pp. 39–58). Kluwer. OCLC: 801321444.

Machery, E. (2016). De-Freuding implicit attitudes. *Implicit Bias & Philosophy: Metaphysics and Epistemology*, *1*, 104–129. https://doi.org/10.1093/acprof:oso/9780198713241.003.0005

Madva, A. (2016). A plea for anti-anti-individualism: How oversimple psychology misleads social policy. *Ergo, an Open Access Journal of Philosophy*, *3*. https://doi.org/10.3998/ergo.12405314.0003.027

Madva, A. (2017). Social psychology, phenomenology, & the indeterminate content of unreflective racial bias. In E. S. Lee (Ed.), *Race as phenomena: Between phenomenology and philosophy of race* (pp. 87–106). Rowman & Littlefield International.

Madva, A., Kelly, D., & Brownstein, M. (2023). Change the people or change the policy? On the moral education of anti-racists. *Ethical Theory & Moral Practice*, *27*(1), 91–110. https://doi.org/10.1007/s10677-023-10363-7

Mandelbaum, E. (2015). Attitude, inference, association: On the propositional structure of implicit bias. *Noûs*, *50*(3), 1–30. https://doi.org/10.1111/nous.12089

Martín, A. (2021). What is white ignorance? *The Philosophical Quarterly*, *71*(4), pqaa073. https://doi.org/10.1093/pq/pqaa073

Miller Larsen, E. (2023). *The ethics of ignorance*. Dissertation. Harvard University.

Mills, C. W. (2007). White ignorance. In S. Sullivan & N. Tuana (Eds.), *Race and epistemologies of ignorance* (pp. 11–38). University of New York Press.

Mitchell, T. M. (1980). The need for biases in learning generalizations. Rutgers Computer Science Department Technical Report CBM-TR-117 (pp. 3).

Moody-Adams, M. M. (1994). Culture, responsibility, and affected ignorance. *Ethics*, *104*(2), 291–309. https://doi.org/10.1086/293601

Moss, S. (2018). *Probabilistic knowledge*. Oxford University Press.

Munton, J. (2017). The Eye's mind: Perceptual process and epistemic norms. *Philosophical Perspectives*, *31*(1), 317–347. https://doi.org/10.1111/phpe.12105

Munton, J. (2019a). Beyond accuracy: Epistemic flaws with statistical generalizations. *Philosophical Issues*, *29*(1), 228–240. https://doi.org/10.1111/phis.12150

Munton, J. (2019b). 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. (2019c). Perceptual skill and social structure. *Philosophy and Phenomenological Research*, *99*(1), 131–161. https://doi.org/10.1111/phpr.12478

Munton, J. (2021). Prejudice as the misattribution of salience. *Analytic Philosophy*, *64*(1), phib.12250–19. https://doi.org/10.1111/phib.12250

Nanay, B. (2021). Implicit bias as mental imagery. *Journal of the American Philosophical Association*, *7*(3), 329–347. https://doi.org/10.1017/apa.2020.29

Neufeld, E. (2019). An essentialist theory of the meaning of slurs. *Philosophers' Imprint*, *19*(35).

Neufeld, E. (2020). Pornography and dehumanization: The essentialist dimension. *Australasian Journal of Philosophy*, *98*(4), 1–15. https://doi.org/10.1080/00048402.2019.1700291
Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, *17*(5). https://doi.org/10.1111/phc3.12823
Nguyen, C. T. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplement*, *89*, 227–255. https://doi.org/10.1017/s1358246121000035
Nguyen, C. T. (2023). Hostile epistemology. *Social Philosophy Today*, *39*, 9–32. https://doi.org/10.5840/socphiltoday2023391
Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishing Group.
Orlandi, N. (2014). *The innocent eye: Why vision is not a cognitive process*. Oxford University Press.
Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.
Pauker, K., Ambady, N., & Apfelbaum, E. P. (2010). Race salience and essentialist thinking in racial stereotype development: Racial stereotype development. *Child Development*, *81*(6), 1799–1813. https://doi.org/10.1111/j.1467-8624.2010.01511.x
Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of personality and social psychology*, *81*(2), 181–192. https://doi.org/10.1037//0022-3514.81.2.181
Prince, A. E. R., & Schwarcz, D. (2020). Proxy DISCRIMINATION IN THE AGE OF ARTIFICIAL INTELLIGENCE AND Big Data. *Iowa Law Review*, *105*, 1257–1318.
Puddifoot, K. (2021). *How stereotypes deceive us* (1st ed.). Oxford University Press. OCLC: 1265465206.
Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism: Social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*, *8*(4), e1437. https://doi.org/10.1002/wcs.1437
Ritchie, K. (2021a). Essentializing inferences. *Mind & Language*, *36*(4), 570–591. https://doi.org/10.1111/mila.12360
Ritchie, K. (2021b). Essentializing language and the prospects for ameliorative projects. *Ethics*, *131*(3), 460–488. https://doi.org/10.1086/712576
Rooney, P. (1992). On values in science: Is the epistemic/non-epistemic distinction useful? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1992*(1), 13–22. https://doi.org/10.1086/psaprocbienmeetp.1992.1.192740
Rosch, E. (1978). Principles of categorization. In E. Rosch & B. L. Lloyd (Eds.), *Cognition and Categorization*. Lawrence Erlbaum Associates.
Rosen, G. (2002). Culpability and ignorance. *Proceedings of the Aristotelian Society*, *103*(1), 61–84. https://doi.org/10.1111/1467-9264.00128
Rothbart, M., & Taylor, M. (1992). Category labels and social reality: Do we view social categories as natural kinds? In G. R. Semin & K. Fiedler (Eds.), *Language, interaction and social cognition* (pp. 11–36). SAGE Publications.
Rudner, R. (1953). The scientist qua scientist makes value judgments. *Philosophy of Science*, *20*(1), 1–6. https://doi.org/10.1086/287231
Saint-Croix, C. (2022). Rumination and wronging: The role of attention in epistemic morality. *Episteme*, *19*(4), 491–514. https://doi.org/10.1017/epi.2022.37
Saul, J. (2018a). Dogwhistles, political manipulation, and philosophy of language. In D. Fogal, D. Harris, & M. Moss (Eds.), *New work on speech acts* (pp. 360–383). Oxford University Press.
Saul, J. (2018b). (How) should we tell implicit bias stories? *Disputatio*, *10*(50), 217–244. https://doi.org/10.2478/disp-2018-0014
Schneider, D. J. (2004). *The psychology of stereotyping*. Distinguished contributions in psychology. Guilford Press.
Siegel, S. (2017). *The rationality of perception* (1st ed.). Oxford University Press.
Siegel, S. (2022). Salience principles for democracy. In S. Archer (Ed.), *Salience* (pp. 235–266). Routledge.
Soon, V. (2019). *Implicit bias and social schema: A transactive memory approach*. Philosophical Studies.
Stanley, J. (2005). *Knowledge and practical interest*. Oxford University Press.
Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, *74*(2), 149–175. https://doi.org/10.1016/s0010-0277(99)00071-2
Sullivan-Bissett, E. (2019). Biased by our imaginings. *Mind & Language*, *34*(5), 627–647. https://doi.org/10.1111/mila.12225
Taiwo, O. (2017). Beware of schools bearing gifts. *Public Affairs Quarterly*, *31*(1), 1–18. https://doi.org/10.2307/26897014
Taiwo, O. (2018). The empire has no clothes. *Disputatio*, *10*(51), 305–330. https://doi.org/10.2478/disp-2018-0007
Todd, J. T., & Oomes, A. H. (2002). Generic and non-generic conditions for the perception of surface shape from texture. *Vision Research*, *42*(7), 837–850. https://doi.org/10.1016/s0042-6989(01)00234-6
Toole, B. (2019). From standpoint epistemology to epistemic oppression. *Hypatia*, *34*(4), 598–618. https://doi.org/10.1111/hypa.12496

Toole, B. (2023). Standpoint epistemology and epistemic peerhood: A defense of epistemic privilege. *Journal of the American Philosophical Association*, 1–18. https://doi.org/10.1017/apa.2023.6

Toribio, J. (2021). Accessibility, implicit bias, and epistemic justification. *Synthese*, *198*(S7), 1529–1547. https://doi.org/10.1007/s11229-018-1795-7

Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Racebased patterns of selective attention. *Journal of Experimental Social Psychology*, *44*(5), 1322–1327. https://doi.org/10.1016/j.jesp.2008.03.006

Tversky, A., & Kahneman, D. (1974). Judgement under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037//0033-295x.90.4.293

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

Vasilyeva, N., Gopnik, A., & Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology*, *54*(9), 1735–1744. https://doi.org/10.1037/dev0000555

Vasilyeva, N., & Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition*, *204*, 104383. https://doi.org/10.1016/j.cognition.2020.104383

Watzl, S. (2017). *Structuring mind: The nature of attention and how it shapes consciousness* (1st ed.). Oxford University Press. OCLC: ocn964379343.

Welpinghus, A. (2019). The imagination model of implicit bias. *Philosophical Studies*, *177*(6), 1611–1633. https://doi.org/10.1007/s11098-019-01277-1

Whiteley, E. (2022). Harmful salience perspectives. In S. Archer (Ed.), *Salience.* Routledge.

Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology*, *113*(1), 59–80. https://doi.org/10.1037/pspi0000092

Wu, W. (2023a). *Movements of the mind: a theory of attention, intention and action*. Oxford University Press.

Wu, W. (2023b). On attention and norms: An opinionated review of recent work.

AUTHOR BIOGRAPHY

Gabbrielle M. Johnson is an Assistant Professor in the Department of Philosophy at Claremont McKenna College. She received her PhD in Philosophy at the University of California, Los Angeles in 2019. Her research interests are in philosophy of psychology (particularly perception and social cognition), philosophy of science, and philosophy of technology.

How to cite this article: Johnson, G. M. (2024). Varieties of bias. *Philosophy Compass*, e13011. https://doi.org/10.1111/phc3.13011