The (Dis)Unity of Psychological (Social) Bias

Gabbrielle M Johnson

Forthcoming in *Philosophical Psychology*

Abstract

This paper explores the complex nature of social biases, arguing for a functional framework that recognizes their unity and diversity. The functional approach posits that all biases share a common functional role in overcoming underdetermination. This framework, I argue, provides a comprehensive understanding of how all psychological biases, including social biases, are unified. I then turn to the question of disunity, demonstrating how psychological social biases differ systematically in the mental states and processes that constitute them. These differences indicate that biases at various levels of the cognitive architecture require distinct treatment along at least two dimensions: epistemic evaluation and mitigation strategies. By examining social biases through this dual lens of unity and diversity, we can more effectively identify when and how to intervene on problematic biases. Ultimately, this approach provides a nuanced understanding of the nature of social bias, offering practical guidance for addressing existing biases and proactively managing emerging biases in both human and artificial minds.
Keywords: Implicit Bias, Perceptual Bias, Essentialism, Underdetermination

1 Introduction

The social biases studied by psychology manifest in diverse ways. Discussions on this diversity frequently adopt a *horizontal* perspective, suggesting that, e.g., the mechanisms underlying racial bias differ from those underlying age bias. For instance, racial bias may stem from emotional or associative factors, whereas age bias might rely more on statistical or inferential reasoning. This diversity cautions against viewing social bias as a singular, uniform phenomenon, implying that seeking commonalities could be misguided.[1] Where the

[1]For discussion of the dimensions of heterogeneity for implicit bias and their implications, see Holroyd and Sweetman 2016.

complexity of bias is acknowledged, it is often relegated to prefatory remarks and ultimately set aside to pursue other ends in theorizing about bias, with the assumption that as empirical research progresses, more nuanced distinctions within models will emerge.[2] This creates a theoretical tension: focusing on diversity raises doubts about overarching commonalities, while efforts to present a unified view may neglect the significance of diversity.

In this paper, I argue that a functional framework resolves the apparent conflict between unity and diversity in social biases. Empirical evidence already supports the recognition of bias heterogeneity interpreted *vertically*: social biases exist at various levels of the cognitive architecture, influenced by the specific psychological system in which they are embedded. This *system dependence* of social biases reflects a broader psychological principle that biases are tailored to their cognitive contexts.[3] Building on established lessons about the system dependence of bias, I demonstrate that different constitutions of social biases at different levels of the cognitive architecture often require distinct treatment in two key dimensions: epistemic evaluation and mitigation techniques. To illustrate this, I focus on two well-documented types of biases: essentialist social biases, which encode theoretical causal-explanatory assumptions involving category membership; and perceptual social biases, which are based primarily on statistical regularities. My choice to examine these specific social biases is strategic, grounded in substantial empirical research that elucidates their mechanisms and outlines strategies for evaluating and intervening in their operation. Specifically, I propose that essentialist social biases are warranted only when those assumptions are true; while perceptual social biases are warranted only when those regularities hold.[4] Additionally, I

[2]For an example the sort of prefatory remarks I have in mind, see Mandelbaum 2015, 3.

[3]In visual psychology, for example, the dominant approach to explaining how the visual perceptual system overcomes underdetermination involves positing that computational transitions within the system are guided by biases. These perceptual biases operate differently from those in higher-level cognitive transitions, such as inference (Burge 2005, 13). This paper extends this broader understanding of system-dependent biases to the particular domain of social bias.

[4]Here, I'm borrowing the notion of epistemic *warrant* from Burge (2003, 504), who regards *warrant* as a general species of epistemic evaluation, encompassing *justification* and *entitlement* as subspecies. Burge would resist applying the notion of epistemic warrant to perceptual states themselves, rather than to the perceptual beliefs formed on the basis of those states. I believe my main theses can be adapted to align with this more restricted understanding of the domain of warrant.

argue that some problematic cognitive social biases may be countered through reasoning, argumentation, and other logical interventions; while some problematic perceptual social biases may be addressed through perceptual learning and counter-stereotypical exemplar training. The existing empirical knowledge base, I contend, provides a solid foundation for exploring how social biases operate differently depending on their cognitive context.

To illustrate these differences initially, consider a social bias that assumes that (roughly) *men are dangerous*. This bias plausibly can arise in two distinct ways: through perception or cognition. In both cases, it seems the input-output profile will be the same: the bias will take as input the categorization of an individual as belonging to the social group *men* and produce the output that that individual possesses the stereotypical property of *being dangerous*. However, from here, biases instantiated in different systems operate differently. In cognition, this bias may be underwritten by a psychological assumption of essentialism, positing a shared gender essence that is causally-explanatorily responsible for outward properties like *being dangerous*. In perception, on the other hand, the bias may transition from input to output based on statistical regularities, without robust causal or explanatory assumptions. Crucially, while the two biases start and end similarly, their different transitional and inferential routes have important consequences for both our epistemic evaluation and the strategies we adopt to mitigate them.

Still, it's critical to recognize how perceptual and cognitive biases, among others, interact in complex patterns and create looping effects. Understanding precisely how these different biases combine and interact to perpetuate wider social stereotypes is essential for a comprehensive understanding of how bias globally operates. It is in recognizing this joint task of a theory of bias—both in unification and differentiation—that I aim to showcase the broader utility of a functional framework. This approach reconciles the similarities and differences among biases generally, underscoring the necessity of a unified framework that acknowledges both the shared functional roles of biases in addressing underdetermination and their distinct manifestations and implications at various cognitive levels. This nuanced

perspective not only clarifies the landscape of social biases, but also informs more targeted and effective strategies for epistemic evaluation and bias mitigation, emphasizing the critical role of context in shaping the nature and impact of biases within the cognitive architecture.

The paper will proceed as follows. In section 2, I present the functional account that unifies psychological social bias by highlighting its role in overcoming underdetermination, its having inputs involving social categories, and its having outputs involving stereotypical features. In section 3, I address the disunity of psychological social bias by contrasting biases in perception and cognition through case studies, illustrating how these biases, despite their distinct mechanisms, align with the functional approach. After establishing on both empirical and theoretical grounds that psychological social biases are multiply realized, I argue in section 4 that the differences between the mental constructs that give rise to bias are crucial for epistemic evaluation and mitigation strategies. Section 5 acknowledges the complex interplay between different types of biases and examines the opportunities this presents for mixed interventions. This section argues for a comprehensive approach that, while recognizing the limitations of singular strategies, reinforces the value of the functional model in understanding and managing social bias generally.

2 The Unity of Psychological (Social) Bias

I have previously argued (Johnson 2020a,b) that a unified account of social bias can be achieved through a functional characterization, which conceptualizes bias by its functional role in aiding induction and responding to the problem of underdetermination.[5] My model builds on work in epistemology, philosophy of science, and philosophy of perception that identifies the problem of underdetermination as the most significant challenge facing any psychological system that understands the world through the bottleneck of limited evidential

[5]Attempts to broadly unify (social) bias have become more popular recently. Notable examples include Munton (2021), who offers a unified account of prejudice as the misattribution of salience, and Kelly (2022), who offers a unified account of bias as a systematic departure from a norm.

experiences.[6] In what follows, I extend this account, understanding bias in general (and social bias specifically) as systematic solutions to the problem of underdetermination.[7]

The problem of underdetermination manifests in perception, where the collection of data in the form of sensory information registrations of light hitting our retina underdetermine the distal cause of the data. If the visual system relied solely on these data, then it would be immobilized by their limitless possible interpretations. However, the visual system is not immobilized by such possibilities, but produces a determinate representation of the world by relying on biases that structure the data and constrain the inferences it can make. These biases, refined through the evolutionary and learning histories of the visual system, limit possible interpretations in ways that tend to yield accurate results. The visual system tends to get things right.

Crucially, underdetermination occurs for any inductive (i.e., ampliative or non-deductive) decision procedure, including enumerative induction and abductive inference, or inference to the best explanation.[8] Thus, biases are the tools that guide systematic ampliative reasoning, and they occur wherever induction does.[9] This perspective renders some biases beneficial,

[6] Antony 2016, 161. See also Antony 2001 for a discussion of the role of bias in overcoming this epistemic challenge and its relation to historical debates in mainstream empiricism and philosophy of science. For more on underdetermination within psychological theories of perception, see Burge 2010, 90-92, 344-345, especially his footnote 41 on the use of the term 'bias'.

[7] See Kelly (2022, 145)'s discussion of bias as a violation of symmetry, which can be taken in similar spirit to bias as a response to underdetermination.

[8] For a thorough discussion on the relationship between underdetermination, inductive inference, and bias, see Johnson 2023, 27-38. There, I vacillate between the notions of "canons", "values", "virtues", and "biases", with all playing the same conceptual role. Likewise, Munton (2021, 6, 8, 11,13)'s account characterizes the explanatory import of bias (as salience structures) by its ability to facilitate abductive inference by making it computationally tractable (cf. Johnson 2020b, 1195 and Antony 2016, 161)

[9] One might worry that broadening the characterization of bias to encompass all non-deductive inferences might dilute its epistemic significance, as it would apply to inferences where our conclusion is, in some intuitive sense, strongly supported by evidence. To address this, one could restrict the view to only biases occurring in situations of "significant" underdetermination, however that is determined. Nonetheless, I maintain a broader perspective for several reasons. Firstly, long-standing philosophical debates concerning the nature of scientific confirmation highlight the challenge of clearly distinguishing between substantial and minimal underdetermination. Secondly, from a functional-as-etiological standpoint, the causal histories of insubstantial underdetermination problems likely make reference to more substantial cases. That is, even where there's not substantial underdetermination in a particular case of bias, I believe the basic cases that ground that bias's functioning are the cases in which there is a live problem of underdetermination. For instance, a computational transition in vision, like depth from convexity, might seem well-supported by multiple cues but is plausibly rooted in a general capacity developed to solve a substantial underdetermination problem. This broader, teleological view on the development of our cognitive capacities supports the

which contrasts with the common negative connotations associated with the notion. The everyday understanding of *bias* often implies something inherently negative, either from an epistemic or moral standpoint.[10] However, the concept of bias I employ here is more neutral, rooted in the idea of bias as simply a predisposition or tendency.[11] This means that not all biases are bad; some can be neutral or even beneficial. Moreover, evaluating a bias isn't just about recognizing it as a bias; it also involves looking at its effects and origins.[12] Ultimately, evaluating bias becomes, in part, an empirical question. Recognizing this, Louise Antony (2001, 2016) has argued that the epistemic evaluation of a bias should depend on its likelihood of leading us to the truth. This is a good start. However, I contend that this traditional account is incomplete. A more comprehensive understanding of what makes a bias epistemically problematic must also consider the psychological mechanisms behind it. This paper aims to develop a fuller account of what makes a bias bad and, when bad, how to combat it.

These same basic points about underdetermination apply to a wide range of knowledge-gathering domains, including scientific theorizing and, importantly for our interests, social interactions. Like visual psychology, social psychology is fundamentally concerned with veridicality conditions. As expressed by Bodenhausen and Morales (2013, 228, 241), one *core assumption* within social cognitive science is that "social behavior is a function of social cues as they are represented in the mind, rather than how they exist objectively in the social environment" raising the question of "in what ways are our social impressions firmly grounded in available informational cues versus derived from inferential processes that go

idea that biases, by guiding our inductive inferences, enhance our interaction with an inherently uncertain world. Thanks to Tom Kelly for pushing me to think more about the degree of underdetermination in our attributions of bias.

[10] Recent alternative philosophical views maintain this intuition, see Kelly 2022.

[11] Antony 2016, 161. See also Munton (2019a, 1)'s notion of a *formal bias* and Burge (2010, 92, fn. 41)'s notion of a *biasing principle*.

[12] Why not think bias is constitutively bad? Others have attempted to answer this question, at least in the social domain, as it relates to stereotyping more generally. (See Beeghly 2015 and Blum 2004.) However, my reasons for resisting this are grounded in the considerations discussed in footnote 9 and further elaborated by the discussion to follow, namely that I take the functional role of bias to be theoretically prior to our evaluation. We must know a bias's intended function to understand the circumstances in which it fails. For more discussion, see Antony 2001 and Johnson 2023.

beyond the given information." To put the point another way, the aim of social cognitive science is to explain how accurate or inaccurate social representations are formed and how those representations function in inference, dictating social interactions. The superficial properties we observe from others underdetermine both the social categories to which they belong and the properties we attribute to them on that basis. Thus, we rely on social biases to navigate our social environment, just as we rely on visual perceptual biases to navigate our physical environment.[13]

Bringing these points together, we arrive at a unifying definition of social bias that encompasses three constitutive features. First, social bias is a functional response to underdetermination. This is what solidifies its belonging to the general psychological kind *bias*. Second, social bias functions by taking in social-kind inputs. In paradigmatic cases, these inputs identify (accurately or inaccurately) individuals as belonging to certain social categories. And finally, social bias functions to systematically produce stereotypical-feature outputs. In paradigmatic cases, these outputs pair the individuals identified by the inputs with properties that are stereotypical of their presumed social category. Consider again the toy example of social bias with which the paper started that encodes roughly that *men are dangerous*. This bias takes as an input the categorization of an individual as a *man* and outputs that that individual has some property taken to be stereotypical of men, namely *being dangerous*. By maintaining these three conditions as constitutive, we can recognize a diversity of candidates that fulfill this functional role. One candidate might be an explicit stereotype belief that men are dangerous. Another might be an implicit association, manifesting as a cascade of conceptual activation where the concepts *man* and *dangerous* both light up. Crucially, while this functional account unifies bias by these constitutive features, it remains agnostic as to which combinations of states and processes bridges the inputs and outputs, allowing for multiple realizability.

In summary, psychological (social) biases are unified by their functional role as a response

[13]Of course, some visual perceptual biases will also be social. More on this in §3.2.

to underdetermination. They transition us from underdetermining inputs to determinate outputs in ways that aim to track features of the environment, getting us onto truth (as we'll see, these features can be statistical or explanatory). Social biases facilitate these transitions in the service of navigating the social environment and involve inputs and outputs that attribute distinctively social kinds and properties.[14] However, social biases can fulfill this unified role while differing dramatically in the states and processes that bridge inputs to outputs.

3 The Disunity of Psychological Social Bias

Research on psychological essentialism and social perception reveals how social biases can vary depending on the psychological system in which they are embedded. To reiterate, my focus on these specific examples is strategic: I aim to leverage both the wide consensus about the fundamental distinction between perception and cognition as well as the extensive empirical work concerning their operation and interventions. While I focus narrowly on these examples here, ultimately I believe there are many possible avenues for exploring the diversity of bias.[15]

[14] As much as possible, I'd like to proceed in the absence of a robust theory of what makes some psychological states distinctively social. It's not sufficient to label any state that underwrites a social interaction as social, since many mundane and non-distinctive psychological capacities contribute to such interactions. For instance, social interaction with a peer might begin with the perceptual capacity to identify them as a three-dimensional object with a cohesive, bounded shape (Burge 2010, 464). Considerations like these have led some, such as Beer and Ochsner (2006), to deny the existence of a social cognitive module, arguing that social functions are too wide-ranging to plausibly be traced back to dedicated neuro-cognitive resources. Therefore, I will focus on relatively uncontroversial examples, leaving borderline cases aside.

[15] A thorough theory of bias would detail the various types of bias that exist, though this paper will not cover all of them. Some alternatives worth mentioning include Del Pinal and Spaulding (2018) and Del Pinal et al. (2017)'s work, which emphasizes the importance of conceptual centrality and dependency networks in shaping implicit biases and extends their treatment beyond the focus on mere associative strength that is common in the literature. This approach shares important similarities with the essentialist construal of social bias discussed in this paper by highlighting the distinction between causal-explanatory vs statistical relationships between features. However, the conceptual centrality model avoids the additional assumption of a hidden essence found in the essentialist paradigm. Other models might type-individuate social bias based on different memory systems (e.g., Amodio and Mendoza 2010, Faucher 2014) or other well-known but contested distinctions in the literature on implicit bias (e.g., system one vs system two, automatic vs controlled, associative vs propositional). I'm grateful to two anonymous reviewers for suggesting these additional avenues for exploring the heterogeneity of bias.

3.1 Cognitive Social Bias

One canonical instance of cognitive-level social bias involves psychological essentialism. The literature on psychological essentialism is vast and rapidly evolving. For our purposes, it helps to focus on a standard model of psychological essentialism, according to which "people seem to assume that categories of things in the world have a true, underlying nature that imparts category identity [and] is thought to be the causal mechanism that results in those properties that we see."[16] For a kind to be essentialized in the psychological sense, there must be an assumption that members share some deep, hidden, internal feature that is both necessary and sufficient for belonging to the kind. This non-obvious essence is taken to be constitutive of the category and plays a causal-explanatory role in accounting for its observable characteristics.[17] For instance, we might think a tiger has stripes or a violent disposition due to its having a tiger essence, such as its specific DNA makeup. While possessing an essence is necessary and sufficient for belonging to the kind, the outward characteristics are neither necessary nor sufficient. We can imagine tigers without stripes or violent tendencies and non-tigers with all the outward traits of a tiger but lacking the essence. These are not conceptual impossibilities. Moreover, the tiger essence often features in our folk explanations for other characteristics, like sharp claws and loud roars. We think there's just something about being a tiger that disposes a creature toward these properties. The presumed sharing of an underlying essence thus has "inductive potential," allowing us to infer that if a creature belongs to the category *tiger*, it likely possesses stereotypical outward properties. This link to induction solidifies essentializing inferences as a form of psychological social bias, as discussed above.

Essentializing a kind involves assuming that a creature's essence causally-explanatorily

[16] Gelman et al. 1994, 344. Psychological essentialism differs from metaphysical essentialism in that the former pertains to how we represent the world as being, while the latter is a claim about what the world is actually like (for philosophical discussion, see Ritchie 2021 and Neufeld 2022).

[17] For canonical readings connecting essentialism and natural-kind concepts in philosophy of mind and language, see Putnam 1975, Burge 1979, and Kripke 1980. For canonical readings of psychological essentialism, see Gelman et al. 1994; Gelman 2004.

disposes it to have certain outward properties, even if those properties are not currently manifest. Indeed, being open to a distinction between how a thing veridically appears and what it actually is (or, more generally, a distinction between appearances and reality) is a key hallmark of harboring natural kind concepts.[18] Thus, while perceptual cues guide essentialized kind attributions, they do not determine them. Moreover, individuals need not be explicitly aware that they reason in this way about object categories, nor do they need to have robust notions of *essence* or theories about what the underlying essence of some group might be.[19] Psychological essentialism aligns with other psychological theories that posit human concepts are embedded in (or constituted by) rich conceptual webs or "theories," rather than manifesting in mere clusters of correlated properties.[20] Following this trend, I will regard essentialized-kind based inductions as a form of *theory-based social bias*.

Although originating primarily in the domain of how humans think and reason about biological kinds, the assumptions of psychological essentialism have been extended to other categories as well, including the social domain.[21] Evidence for psychological social essentialism dates back to Rothbart and Taylor (1992)'s work, which argued that people commonly treat social categories as natural kinds.[22] For example, a study by Taylor et al. (2009) found evidence that essentialist-based social-kind inductions about gender emerge early in development. In their study, children aged 5 to 10 were presented with stories about either a baby boy who went to live with his aunt on an island inhabited by girls and women, or a baby girl who went to live with her uncle on an island inhabited by boys and men. The children were then asked questions about the babies' future behavioral characteristics, such as whether they would like to sew or build things, play with tea sets or toy trucks, or aspire to

[18] Burge 2013, 237.

[19] Medin and Ortony (1989) use the idea of essence as a *placeholder notion*, wherein one is disposed to reason about a category as if it has an essence without knowing what the essence is or what an essence is.

[20] For an overview of the "theory-theory" view of concepts and how it differs from other theories of concepts, e.g., prototype and exemplar theories, see Murphy 2004. For helpful discussion of the causal structures underlying essentialized-kind attributions, see Neufeld (2019). For alternative causal models of concepts, again see the conceptual centrality literature cited in footnote 15.

[21] For more information on the extent of essentialized thinking in different domains, see Gelman 2004.

[22] For review, see Haslam et al. 2000, 113-117, Rhodes and Mandalaywala 2017, Pauker et al. 2010, and Neufeld 2019.

be nurses or firefighters. The study found that children "reliably made category-based predictions about behavioral properties (though less often than about physical properties)."[23] Moreover, research indicates that psychological essentialism regarding race is common and associated with increased endorsement of racial stereotypes and prejudices.[24] Thus, these essentializing mechanisms plausibly underwrite many real and empirically substantiated cases of social bias.

What combination of states and processes actually constitute an essentializing mechanism? Unfortunately, psychological models of essentialism are often agnostic about which aspects of their operation, specifically which states and processes posited within the models, are taken to be explicitly represented in some psychologically robust sense.[25] One how-possibly explanation of essentializing social bias is the following: from a very young age, humans possess an ingrained psychological mechanism built into the cognitive architecture that disposes them to, when faced with attributions of essentialized kinds, infer in ways that are indicative of kindhood being causally-explanatorily responsible for outward, superficial features. Here, there is no explicit representation of *essence*. Rather, the explicit representation of, say, *is a tiger* or *is a man* is enough to trigger the implicit essentializing mechanism. The relevance of which categories trigger the mechanism can also develop over time, and certain linguistic constructions can invite certain categories to be included.[26] Thus, while the basic essentializing mechanism is innate, children needn't begin with a complete repertoire of essentialized kinds. This view avoids over-intellectualizing psychological essentializing biases by not requiring explicit representations of *essence*.

[23] Taylor et al. 2009, 477. Curiously, this pattern does not hold for children's racial category-based predictions about behavioral properties in similar switched-at-birth paradigm experiments. For more on this, see Mandalaywala et al. 2018b and Mandalaywala et al. 2018a, as well as Williams and Eberhardt 2008 and Rhodes and Gelman 2009. This suggests that essentialized inference for behavioral properties as exhibited in racial stereotypes emerge closer to adulthood.

[24] Haslam et al. 2000, Keller 2005, Bastian and Haslam 2006, and Mandalaywala et al. 2018a. However, see also my previous footnote.

[25] See remarks in Strevens 2000, 150. This has led some including Strevens (2000) to posit more minimalist causal models underwriting essentializing biases.

[26] See Leslie (2017)'s, Neufeld (2019)'s, and Ritchie (2021)'s discussions of how generic and predicate-nominal constructions contribute to essentialized reasoning. We will return to this literature when discussing mitigation techniques in section 4.2.

Crucially, to repeat an important point, the essentializing mechanism operates largely independently of observable features used to identify social groups and those projected onto them based on their presumed shared essence. While attributions of social kinds such as *tiger* or *man* do depend partly on observable features, the mechanism itself is not entirely determined by these outward characteristics. Instead, it functions mainly through the intermediary attribution of relevant social kind attributives. This mechanism requires a capacity akin to inference to the best explanation (IBE), where one infers that the best explanation of superficial similarity is a shared essence. Therefore, while observable properties, such as being striped, play a non-determinative role in producing attributions of social kinds, such as being a tiger; it is the social kind attribution itself, not the observable property, that is constitutive of essentialized reasoning. This fundamentally individuates essentialized social biases from perceptual social biases, which I turn to next.

3.2 Perceptual Social Bias

In contrast to theory-based social biases that rely on tacit assumptions of essence, perceptual transitions are driven by transformation principles that function to track superficial statistical regularities found in the normal environment.[27] One example of a perceptual social bias is social-group-based perceptual expertise, exemplified by the well-documented "other-race effect" in race-based facial expertise. This phenomenon is characterized by subjects recognizing faces of their own race better than those of other races.[28] This ability emerges within the first six months of infancy and is likely part of a broader phenomenon known as perceptual narrowing, wherein children's perceptual window narrows as they are exposed to a vast amount of perceptual data. Through this process, they develop the capacity to

[27]It is worth noting that presenting clear-cut examples of perceptual transitions that constitute social biases is challenging, as discussions of them typically overlap with ongoing debates about the (rich vs. sparse) nature of perceptual content and the status of purported cases of cognitive penetration. To avoid these debates, I focus on cases of social-group-based perceptual expertise that utilize the attribution of low-level perceptual attributives considered both perceptual and social on a majority of the most conservative views of perceptual content.

[28]This case is also discussed at length by Munton (2019a, 5-8). For review, see Meissner and Brigham 2001.

better discriminate stimuli that are statistically prevalent in their local environments, such as features of the faces of their primary caretakers.[29] Similar findings have been observed for women, demonstrating an "own-gender bias" in facial recognition (Loven et al., 2011).

Social-group-based perceptual expertise can also influence social evaluation, as perceptual biases shape preferences for certain social groups. For example, research indicates that children as young as 2-3 years prefer same-gender peers (Kinzler et al., 2010). Interestingly, auditory cues influence social evaluation earlier than visual cues. From birth, infants show a preference for their native language over foreign languages; and by 10 months, they are more likely to accept toys from individuals who speak their native language (Kinzler et al., 2007). Even when race-based social evaluations emerge around the age of 4, accent remains a stronger modulator of preferential evaluation than visual racial cues (Kinzler et al., 2009). This preference for auditory over visual cues can be explained by the fact that many perceptual biases are evolutionarily endowed, making them more rigid than their cognitive-level counterparts. Throughout evolutionary history, relevant social out-groups likely did not look very different from us, but likely sounded very different.

Social-group-based perceptual expertise demonstrates the general statistical learning mechanisms that underwrite both perceptual capacities in general and perceptual social biases in particular. Perceptual systems encode previous environmental regularities and generalize them to new data encountered by the system. Bayesian approaches model these encoded regularities as perceptual priors. These priors *bias* the perceptual system to facilitate the interpretation of new data. As in the perceptual recognition tasks discussed so far, these environmental regularities are sometimes specific to social groups. Consequently, new perceptual data, such as faces or accents, are interpreted differently based on features that correlate with the individual's social category. Two key aspects of this perceptual process are that it relies primarily on superficial properties of stimuli and on statistical regularities.

The case studies of essentialized-kind social biases and perceptual social biases reveal a

[29] Kelly et al. 2007. For an overview and discussion of potential accounts of perceptual narrowing, see Nelson 2001.

significant disunity in psychological social bias. Firstly, theory-based social biases rely primarily on tacit theories of essentialism that resist straightforward statistical analysis, whereas perceptual social biases rely primarily on statistical analysis of environmental regularities.[30] Secondly, perceptual social bias relies primarily on the co-occurrences of superficial properties, while theory-based social bias relies primarily on the assumed presence of some deep, underlying essence distinct from superficial properties. These two points are related since the deep underlying essence allows theory-based inferences to persist even where superficial properties fail to correlate. Thus, the states and processes constituting perceptual social bias differ significantly from those constituting theory-based social bias.[31]

4 Evaluation and Mitigation

In this section, I turn to the implications of system-dependent social biases for epistemic evaluation and mitigation. I argue that social biases originating from different psychological systems necessitate distinct treatments for assessing their epistemic impact and devising effective mitigation strategies. Specifically, perceptual biases, which function primarily at the level of sensory processing and statistical learning, require different approaches compared to theory-based biases, which operate at the level of more abstract cognitive processes and are

[30] Consider that generic constructions promote essentialized reasoning, but generic statements resist analysis in terms of statistical regularities. For instance, the claim that mosquitoes carry West Nile is considered true even though a very small percentage actually do (Leslie 2014, 2017). Similarly, ongoing debates in psychology pertaining to what it would mean for social stereotypes to be "false" or "inaccurate" indicate again that straightforward statistical analyses are insufficient (Judd and Park 1993, Schneider 2004).

[31] Ultimately, I take these differences to stem from general functional differences that contribute to the separation of cognition and perception more broadly. As mentioned above, theory-based cognitive biases necessarily take the form of abductive inference, or inference to the best explanation (IBE): individuals notice some superficial properties hanging together in systematic, but not easily statistically predictable ways and infer to the best explanation of their relationship—having the common cause of an underlying essence. In contrast, perceptual biases rely predominantly on statistical correlations, without making IBE-like inferences to explain those correlations. While some perceptual transformations can be modeled as causal inference, ultimately I believe that the types of causal-theory-based explanations available to perception are far more restricted than those for cognition. A full discussion of these points relating to a general theory of what demarcates cognition and perception goes beyond the scope of this paper. For a plausible general theory, see Green (2020)'s dimension restriction hypothesis. For references probing the sophistication of perceptual capacities, see Leslie and Keeble 1987, Siegel 2011, Firestone and Scholl 2014, Helton 2016, and Burge 2022 (in particular, Chapter 12).

grounded in essentialist assumptions. Recognizing these differences is crucial for developing more nuanced and effective methods to combat social biases and promote justice.

4.1 Epistemic Evaluation

Let's revisit the idealized example from the start of the paper, which considers a social bias that assumes men are dangerous occurring at both the perceptual and cognitive levels.[32] These biases share similar content regarding their input-output profiles: each takes as an input a state pairing together an individual with the social kind attribute of *being a man*, and the tokening of that input for each causes the tokening of an output state that pairs that individual with the stereotype property of *being dangerous*.[33] However, one of these biases is perceptual, while the other involves essentialized-kind reasoning. If my analysis is correct, a perceptual bias suggests the bias tracks superficial properties. In contrast, the same bias at the cognitive level involves theory-based assumptions of causal-explanatory essence. How might these differences affect evaluation and mitigation?

To develop a theory of the epistemic evaluation of social biases, it's important to return to general remarks about the nature and purpose of biases. Recall that biases are mechanisms that help overcome underdetermination.[34] They occur primarily in domain-specific learning mechanisms honed throughout evolutionary history and the lifetime of the learner

[32]While there is no uncontroversial empirical evidence that connects the two relevant attributes of *being a man* and *being dangerous* in perception, existing literature on race-based and gender-based perceptual expertise, along with the acknowledgment of visual attributions of *danger* by even the most conservative views of perceptual content (Burge 2010, 280, 300, 324-325), suggests that this example is not empirically far-fetched. Ultimately, the connection between perceptions of race and gender and perceptions of threat or danger remains debated among psychologists (Payne 2001, Correll et al. 2002, Eberhardt et al. 2004, Trawalter et al. 2008, Cloutier et al. 2014, Wilson et al. 2017, among others).

[33]This is prescinding from debates about how that content is structured in the relevant psychological systems. It could be that the content is propositional, where the input is roughly *that person is a man* and the output is *that person is dangerous*; or the content could be in an iconic format, expressed best by a complex noun phrase of the rough form *that male person* and the output likewise takes the rough form *that dangerous person*. See Johnson 2020b, 1226, footnote 57 for discussion.

[34]Antony 2016, 161. Antony regards biases as "non-evidential tendencies". Insofar as the functional account allows for explicit beliefs to serve as biases, they can plausibly be epistemically evaluated in straightforward evidentialist ways, or by traditional accounts of justification and warrant for belief. However, biases that are not instantiated by beliefs, or that are built into the cognitive architecture of some computational system as the examples in this paper have illustrated, will resist this approach. It is for these reasons that a more general epistemic account, one having its roots in a theory of epistemic entitlements, is provided here.

to clear inductive gaps guiding us toward truth. They are tailored to their formative environment, and the assumptions they encode are specific to their mental systems. Therefore, the epistemic evaluation of a bias should consider its effectiveness in its intended operation and whether it performs well in the environment for which it was designed—Gigerenzer (2008) calls this *ecological rationality* and, drawing on Gigerenzer, Antony (2016) calls it *ecological validity*.

Antony (2016, 183) recommends the following recipe for identifying epistemically problematic biases: we start by (a) identifying the *markers* and *targets*—the targets are the properties we aim to track using the markers. For example, we might intend to track the target property of *being dangerous* using the marker property of *being a man*. Next, we (b) examine the *indication relation* between the two by determining whether the markers actually correlate with the target properties. Antony notes that many social biases will fail with respect to condition (b), because the purported markers do not actually track the intended targets. However, for the sake of argument, let's assume that *being a man* and *being dangerous* are correlated. Indeed, due to social biases and systematic oppression being as widespread and impactful as they are, we might find that many morally dubious correlations between social markers and targets are reliably found in the wider environment. Consider a nearby example of the connection between *being a man* and *being a philosopher*. In this case, the correlation of marker and target is empirically substantiated.[35] However, as the old adage goes: correlation does not equal causation. We needn't from this statistical result assume any robust causal-explanatory connection between the two, such as being a man causes one to be better at philosophy. It is likely that instead gender-oppressive histories of discrimination have resulted in this correlation. Likewise, in assuming a correlation between *being a man* and *being dangerous*, we are not thereby assuming any robust, causal-explanatory connections.

[35] According to 2003 data from the National Center for Education Statistics, women make up approximately 20% of instructional faculty members in the US. Statistics are provided by the APA Committee on the Status of Women's "Data on Women in Philosophy", https://csw.apaonline.org/data-on-women-in-philosophy.

Where the marker and target are correlated, Antony claims that "epistemic reform in such cases is not the issue [since] the justificatory connection between [marker and target] is completely proper."[36] She's adamant, however, that such biases remain problematic *from the perspective of justice*. This is because, as noted, upon examining the external *causal-explanatory mechanism* underwriting the indication relation (the third and final step (c) in her process), it becomes apparent that the correlation is driven by discriminatory practices deemed morally illegitimate. Thus, rather than eliminating such biases within cognitive systems, we should instead work to manipulate the regularities in the social environment that give rise to and epistemically legitimate such correlations. By doing so, we can "indulge our biases, without injustice."[37] Examples of morally repugnant social biases that appear epistemically non-problematic have long troubled research at the intersection of ethics and epistemology.[38] In summary, according to Antony, while such biases are morally problematic, they are epistemically praiseworthy, since they meet the minimal epistemic requirement put forward by her account: they often produce accurate outputs, thus guiding us toward truth.

I believe a case can still be made for the epistemic faults of some such biases. My strategy for this employs Antony's notion of ecological validity, but interprets its constraints more broadly.[39] This involves interpreting a bias's ecological validity with respect to its reliability and functional etiology, which can reveal deeper epistemic flaws.[40] Ultimately, I argue that while her account might be appropriate for perceptual biases, which function primarily to track statistical correlations among overt features, it is inadequate for theory-based social biases, which function to track deeper, causal-explanatory connections.

To begin, consider the case of theory-based social biases. If the story I've told about

[36] Antony 2016, 185.

[37] Antony 2016, 186.

[38] Reconciling the apparent tension between these two evaluative dimensions is a primary motivation for theories of moral encroachment, where moral shortcomings can undermine the epistemic warrant for some beliefs. (See, Basu and Schroeder 2019, Basu 2018, and Basu 2019.) Another example comes from Munton (2019b)'s analysis of what she calls *Technically Unimpeachable Perceptual Experiences*, or TUPES.

[39] Though, see her footnote 10, Antony 2016, 162.

[40] For discussion on the relationship between reliabilism, function, and environmental interaction in psychological theories of warrant, see Burge 2003, 538ff., as well as Plantinga 1993, Goldman 1976, and Klein 1996.

the relationship between many social-kind based inductions and psychological essentialism is right, then we can identify an avenue for epistemic criticism. Essentialized-kind inductions rely on a mechanism that tacitly assumes a shared, causally-efficacious essence. Given that social groups do not actually share such essences (e.g., there is no essence common to all and only men that causally explains their being dangerous), we have reason to think these social biases lack epistemic warrant.[41] The inadequacy of a theory that considers only the reliability of the inputs to systematically produce accurate outputs stems from a failure to recognize the error occurring in the transition between inputs and outputs. In this case, the cognitive system's mistake is assuming an underlying essence that causally explains the stereotypical property. This highlights an important lesson for any epistemic practice: it matters not only where we arrive, but also on the basis of what assumptions we got there.

The mismatch between the tacit essentializing assumption and the domain to which it is applied is a source of epistemic failure. It is also the result of a complex etiology. One argument for the developmental origin of essentialized biases is that the capacity to reason about social groups as sharing an essence has been "recruited" from our capacity to reason about biological groups.[42] Machery and Faucher (2001, 26-27) and Barrett (2001, 24-25) convincingly support this exaptation theory regarding the essentialization of race, with Barrett stating "essentialization of race may occur via cross-domain transfer of essentialist assumptions and inference procedures, from the domain of biological taxa to the domain of race."[43] In other words, our cognitive systems have adopted essentialized thinking from a domain where it was useful—biological kinds—to another domain where, usefulness aside, it is arguably unwarranted—social kinds.[44]

[41] Neufeld (2019) uses similar reasoning to argue that slurs, which she takes to encode attributions of social-kind essences causally responsible for specifically negative stereotypical properties, fail both morally and epistemically. These failures include having empty extensions and infringing on individual agency.

[42] This explanation was first suggested by Atran (1998) and made explicit by Gil-White (2001).

[43] For reasons to doubt this suggestion, see Hirschfeld and Gelman 1994, 210-211, Mandalaywala et al. 2018a, and Hochman 2013.

[44] I believe the evaluation of a bias's functional etiology, and consequently its "ecological validity," hinges on broader considerations than merely its impact on reproductive success or conferral of evolutionary advantage. This perspective suggests a departure from narrower traditional interpretations of ecological validity. For arguments that we should consider the functional evaluation of psychological capacities, see Burge 2010, 301-

This highlights theoretical resources for outlining the epistemic failure of many morally problematic social biases that, on their surface, appear epistemically warranted. Consider a social bias encoding roughly the content that *women apologize often*. Even at the level of cognitive bias, there are plausibly different etiologies for this bias. One etiology might be that the bias developed to track an internalized essence, and so the bias tacitly assumes that the internal, biological makeup of women causes them to be more submissive and deferential. Conversely, another etiology might be that the bias developed to track structural causal-explanatory features of the environment, and so the bias might instead assume that external, social-political forces have conditioned women to be deferential, perhaps as a form of self-preservation. In sum, the epistemic evaluation of social biases must consider not only the reliability of the inputs and outputs, as Antony does, but also the legitimacy of the underlying assumptions shaped through the etiology of the bias.

It's important to note that the theory of epistemic evaluation outlined above will differ depending on the states and processes that give rise to the bias. Theory-based social biases can lack epistemic warrant due to their reliance on a false, tacit assumption. In contrast, visual perceptual social biases, which aim to track mere statistical regularities in the physical environment, might seem epistemically warranted due to the nature of the visual system's function.

This explains the stubborn prevalence of some social biases, given that statistical regularities are easy enough to come by. Without the more robust causal-explanatory claims backing the correlations, biases aiming to track mere statistical regularities are constantly reaffirmed by an environment where such correlations obtain, even if spurious and capricious. This is significant for understanding the prevalence of biases in various domains, including the development of machine learning programs, which often exhibit so-called "algorithmic biases". These learning programs likewise operate on mere statistical correlations, with-

302. Ultimately, focusing on functional etiology (facilitated by adoption of the functional account generally) allows for a more nuanced exploration of the epistemic evaluation of biases, even if a comprehensive summary of these alternative evaluations falls outside this paper's scope.

out deeper, causal models linking features. To address these biases, it is crucial to move beyond mere statistical correlations and develop robust causal-explanatory models that provide deeper understanding of how various features are related. Incorporating these causal models into predictive algorithms can help proactively address emerging cases of social bias in machines and other contexts. However, building and maintaining these causal models requires substantial effort and resources, which is why many biases persist based on cheap and fortuitous statistical relationships. Despite these challenges, developing more sophisticated models of causality is essential for making progress in addressing social biases and their societal impact.[45]

Given that the visual system primarily functions to accurately track statistical regularities in the physical environment, it avoids the false assumption of an underlying causal structure linking social groups and stereotype features. The relevant condition underwriting the warrant in this case is the actual presence of regularities in the environment. However, as discussed, due to entrenched historical patterns of prejudice and discrimination, many input-output feature pairings might be superficially reliably correlated. Consequently, visual perceptual social biases that track these correlations can appear epistemically warranted within the context of the visual system's function. This makes sense when we consider the epistemic import of being able to recognize statistical patterns. While explanatorily shallow, such patterns can be significant practical guides in identifying problematic societal trends. As Madva (2016) notes, being aware of and sensitive to unjust societal patterns is crucial for identifying systematic patterns of oppression. Since some biases are responses to real patterns of injustices in society, our goal shouldn't always be to eliminate bias entirely, but to understand its origins and how it interacts and evolves within cognitive processes.[46]

However, it remains open for sophisticated theories of perceptual entitlements to put more

[45]See Johnson 2020a for discussion about the relationship between human and algorithmic biases. For a helpful survey of the challenges facing the adoption of causal models in alleviating algorithmic bias, see Hu 2019.

[46]Thanks to an anonymous referee for highlighting the epistemic importance of recognizing statistical patterns.

demanding epistemic constraints on the warrant of perceptual states. One example is Munton (2019b, 30)'s theory of perceptual skill, which requires that these perceptual states accurately encode not only the environmental regularities, but also the right *modal momentum*, i.e., the ability to "go beyond the actual sample encountered to predict what other instances of the kind will be like." This capacity is often missing in perceptual states that are the result of perceptual social biases.[47] The broader point is that epistemic evaluation depends on the states and processes that realize a bias, as it functions in the system in which it is embedded. Two biases with similar functional profiles might nevertheless differ in their epistemic warrant. Depending on the aims and function of their systems, some seemingly problematic social biases might be epistemically warranted, while others are not. The theory of epistemic evaluation presented here underscores the need to understand the states and processes that give rise to social biases, and evaluating them based on their functional etiology. By doing so, we better assess their epistemic and moral merits in the service of cultivating more just and reliable psychological practices.

4.2 Mitigation Techniques

Another consequence of system dependence is that biases at different levels will, when isolated, respond differently to mitigation techniques. The visual system, which relies principally on statistical regularities, will be plausibly counteracted using counter-stereotypical exemplar training. This method involves exposing subjects to numerous examples that contradict common stereotypes, such as showing subjects female philosophers.

Empirical studies support the idea that perceptual biases can be addressed through continuous exposure to counter-stereotypical examples (i.e., statistical learning interventions) and perceptual learning techniques. It is commonly believed that many perceptual biases are hardwired into the functional architecture of our perceptual system, suggesting they might

[47] I believe this starts to broach the more sophisticated functional capacities proprietary to cognition rather than perception, but I lack the space to fully develop the criticism here. For discussion, see footnote 31 above.

be heavily resistant to change.[48] However, research indicates that even our most fundamental perceptual assumptions can be modified. For example, recent findings demonstrate that adults' deeply engrained perceptual bias encoding roughly that *light comes from above* can be adjusted through training, highlighting that these biases are often products of statistical learning.[49] Indeed, there's recent evidence from Lall and Tanaka (2023) suggesting that the race-based perceptual expertise discussed above can be ameliorated in adults through sustained training. This aligns nicely with a general view about the efficacy of perceptual learning.[50] As summarized by Jenkin (2023, 5), "[b]ayesian models of perception state that perceptual learning consists of updating environmental priors in response to data from experience, in accordance with Bayes' Theorem."[51] This insight has led researchers like Munton (2019b) to investigate how alterations in the statistical landscape of our social environment could transform the social biases shaped by these statistical patterns. This line of reasoning bolsters the argument that counter-stereotypical exemplar training, which floods the perceptual system with counterexamples, could effectively diminish the impact of problematic social biases arising from these perceptual patterns.

On the other hand, theory-based baises, in isolation, are not effectively counteracted in this way, since essentialized inferences resist counterexamples. Instead, we need interventions that undermine the assumption of essentialism. Various empirical studies suggest effective strategies. One approach is to describe social categories in ways that avoid reinforcing essentialist thinking, such as using descriptions that focus on behavior or preferences rather than fixed identities. Research by Leslie (2017), Rhodes et al. (2018), Ritchie (2021), and Neufeld (2019) explores how predicate nominals can trigger essentializing. They suggest that avoiding predicate nominals can help reduce essentialized biases. For instance, instead of saying "a homosexual," one could adopt the construction "someone with homosexual preferences."

[48] Orlandi 2014; for this point applied to some social biases, see Johnson 2020b.

[49] Adams et al. 2004; for discussion, see Rescorla 2021, 6.

[50] For broad overviews of perceptual learning, see Connolly 2019 and Jenkin 2023. For a more conservative discussion about the efficacy of perceptual learning, see Burge 2022 (in particular, Chapter 18).

[51] See also Knill 2007.

This change can block the cognitive tendency to essentialize categories expressed in nominal form. In general, cognitive biases may be more susceptible to rational interventions, since those effectively disrupt the theory-like assumptions that underpin them.[52]

Perhaps the most promising intervention technique involves supplanting shallow essentialist explanatory schemas for more robust causal explanations. Recall that the standard model of psychological essentialism holds that there must be an assumption that members of the kind share some hidden essence that is necessary and sufficient for belonging to the kind. However, recent research in psychology from Ny Vasil and Tania Lombrozo suggests that individuals can be dissuaded from adopting essentialist construals of social groups by offering alternative explanations that emphasize the impact of societal structures and positions on group traits.[53] This "structural thinking" allows individuals to conceptually pivot from innate to external explanations of stereotypical features, facilitating a more robust form of causal reasoning. It effectively counters social biases by drawing attention to the varied external causes behind them.

Importantly, a theory recognizing the system dependence of bias anticipates the recalcitrance of social bias evident in empirical data concerning various mitigation strategies. Although different strategies have been somewhat effective, none have succeeded in fully eliminating an individual's biases.[54] This is precisely what we would expect if people have many biases that respond differently to various mitigation strategies. It moreover underscores the importance of tailoring interventions to specific psychological systems, considering both their functional etiology and normative evaluation.[55]

[52]For more on what we can infer about the structure of the mental states and processes underwriting bias from the differential counteracting strategies we employ, see Mandelbaum 2015 and Byrd 2019.

[53]Vasilyeva et al. 2018; Vasilyeva and Lombrozo 2020.

[54]For a meta-analysis, see Lai et al. 2014, Lai et al. 2016, and Forscher et al. 2019.

[55]However, the persistence of bias might also be explained by resistance to fully embracing effective mitigation techniques that are available. This resistance can stem from various sources, including ideology, indifference, ignorance, and individual unwillingness. For a comprehensive review of resistance to debiasing techniques and arguments for their adoption, see Madva 2017. Thanks to Alex Madva for drawing my attention to these alternative explanations for the recalcitrance of social bias.

5 Interaction

Ultimately, addressing social bias comprehensively requires strategies that mirror the complexity of how these biases at various levels inevitably interact.[56] Biases, like any aspect of our psychology, do not exist in isolation; they are part of a dynamic interplay between how we think and perceive. Reason influences attention, which impacts memory, which shapes anticipation, all of which, eventually, over time, affects perception. These interaction effects can often improve perception. However, they can also reinforce problematic social bias.[57]

Return once more to the example of the bias that *men are dangerous*. As an overarching cognitive and behavioral disposition, this bias likely encompasses both perceptual experiences and robust inferential assumptions about gender. Cognitive biases can heighten our awareness of statistical patterns by drawing our attention to them. For instance, we might pay increased attention to reports of crimes committed by men if influenced by gender-essentialist theories. This can create a vicious cycle wherein pre-existing beliefs draw individuals to certain visual evidence, reinforcing their biases at every level. As these patterns are noticed and rationalized, cognitive and perceptual biases continue to feed into each other, perpetuating the cycle. Recognizing how psychological social biases can globally emerge by being shaped and informed by biases at various levels of the cognitive architecture is essential to a complete theory of bias.

This suggests that evaluating and tackling our social biases requires some recognition of a combined approach. Of course, it's widely accepted that our beliefs are partially formed based on what we visually perceive, and likewise epistemic evaluations of belief consider how perception's proper functioning affects the warrant of a belief. Recent philosophical theories argue also that our perceptions are shaped by our beliefs, allowing for epistemic evaluation of perceptions based on their interactions with belief. For example, Siegel (2017) explores

[56] Thanks to two anonymous reviewers for pressing me to consider the complex interplay between perception and cognition.

[57] Crucially, however, recognizing these interactions should not push us to abandon the deep, principled divide between perception and cognition. For recent defenses of the border between perception and cognition, see Block 2022 and Burge 2022.

how our perceptions can be "hijacked" by our prior beliefs or fears, leading to prejudice and other sources of epistemic failure.[58] Thus, the epistemic dependency relation is bidirectional, suggesting a place for a holistic approach to the epistemic evaluation of bias that incorporates our theories of epistemic evaluation at each level.

Interaction between beliefs and perception likewise suggests a combined approach to mitigation. Empirical findings support the effectiveness of interventions targeting various psychological processes. These studies demonstrate that interventions aimed at traditionally perceived "associative", "automatic", or "perceptual" processes can both influence and be influenced by traditionally perceived "propositional", "deliberative", or "cognitive" processes.[59] Additionally, interventions targeting both kinds of processes simultaneously have shown efficacy.[60] Therefore, a holistic approach to the epistemic evaluation and mitigation of bias is supported by both philosophical theories and empirical evidence.

Crucially, the functional approach to understanding social biases offers a powerful framework for approaching biases both in isolation and in concert. This model is effective because it allows for the continual breakdown of biases into smaller components while recognizing how they integrate into larger effects. Functions can embed within them other functions, enabling a dynamic analysis of social bias. The states and processes that bridge some inputs to outputs can themselves be unfolded to reveal a variety of smaller functions from intermediary inputs to intermediary outputs. This nested structure recognizes that bias might initially activate via a perceptual bias, detecting statistical regularities among superficial features, which then feeds into a cognitive bias, identifying the target as belonging to an essentialized group, which ultimately directs attention back to superficial cues.

For instance, an input might begin with the superficial visual attribution of an individual as appearing to be a man, which thereby leads to the superficial attribution of danger, which

[58] See also Siegel 2012 and Siegel et al. 2014. For other accounts of epistemic evaluation that expand their scope beyond belief, see Burge 2003, 2020 and Jenkin 2020.

[59] For instance, see Kawakami et al. 2007 and López et al. 2016, among others.

[60] Calanchini et al. 2013, among others. For a comprehensive overview of mitigation approaches in implicit bias research, see Brownstein 2018 (in particular, pp. 167 ff. and pp. 182 ff.).

then contributes to their identification as belonging to the essentialized category of man, leading to the inference that they have the superficial properties that they do due to some underlying causal-explanatory essence. Here we have an unfolding chain of inputs to outputs that then serves as the input to other processes, eventuating in further outputs. This unique accordion-like quality of functional analysis positions the functional account to effectively manage the expanding and contracting evaluative and mitigative investigations essential to a comprehensive theory of social bias.

6 Conclusions

This paper has argued for both unity and disunity in psychological social bias. I have contended that biases are not inherently epistemically problematic, and that identifying and eliminating problematic biases will be more effective if we consider their functional roles within the cognitive systems that produce them. To achieve this, I have proposed a theory of epistemic evaluation that takes into account the ecological validity and functional etiology of biases, recognizing that biases differ in their epistemic warrant depending on the states and processes that give rise to them. It is also crucial to understand that biases can interact with each other in complex ways, sometimes reinforcing social stereotypes and prejudices.

The functional framework I advocate is uniquely situated to recognize both this unity and disunity of biases, providing a comprehensive tool for understanding and addressing them. This theory allows us to better comprehend the nature of biases and to develop more effective interventions tailored to the specific systems on which they depend. By anti-individualistically understanding biases in the context of the wider environment, we can develop more effective interventions that offset the ill effects of problematic social biases and cultivate a more just environment in which good biases can emerge.[61]

[61]A short version of this paper was joint winner of the Eleventh Annual Essay Prize at the Centre for Philosophical Psychology, University of Antwerp in 2023 on the topic of perceptual and cognitive biases. I am grateful for valuable feedback from Lavi Echeverria, Elli Neufeld, Kate Ritchie, Carolina Flores, the audience at the Social Identities and Cognition in the Desert workshop in 2023, and two anonymous referees.

References

Adams, W. J., Graf, E. W., and Ernst, M. O. (2004). Experience can change the 'light-from-above' prior. *Nature Neuroscience*, 7(10):1057–1058.

Amodio, D. M. and Mendoza, S. A. (2010). 19. Implicit intergroup bias: cognitive, affective, and motivational underpinnings. In *Handbook of implicit social cognition: Measurement, theory, and applications*, pages 353–374. Guilford Press.

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

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

Atran, S. (1998). Folk biology and the anthropology of science: Cognitive universals and cultural particulars. *Behavioral and Brain Sciences*, 21(4):547–569.

Barrett, H. C. (2001). On the functional orgins of essentialism. *Mind & Society*, 2(1):1–30.

Bastian, B. and Haslam, N. (2006). Psychological essentialism and stereotype endorsement. *Journal of Experimental Social Psychology*, 42(2):228–235.

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

Basu, R. (2019). Radical moral encroachment: The moral stakes of racist beliefs. *Philosophical Issues*, 29(1):9–23.

Basu, R. and Schroeder, M. (2019). Doxastic Wronging. In Kim, B. and McGrath, M., editors, *Pragmatic Encroachment in Epistemology*. Routledge.

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

Beer, J. S. and Ochsner, K. N. (2006). Social cognition: A multi level analysis. *Brain Research*, 1079(1):98–105.

Block, N. J. (2022). *The border between seeing and thinking*. Philosophy of mind series. Oxford University Press, New York, NY.

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

Bodenhausen, G. V. and Morales, J. R. (2013). Social cognition and perception. *Handbook of psychology*, 5:225–246.

Brownstein, M. (2018). *The implicit mind: cognitive architecture, the self, and ethics*. Oxford University Press, New York, NY.

Burge, T. (1979). Individualism and the Mental. *Midwest studies in philosophy*, 4:73–121.

Burge, T. (2003). Perceptual entitlement. *Philosophy and phenomenological research*, 67(3):503–548.

Burge, T. (2005). Disjunctivism and Perceptual Psychology. *Philosophical Topics*, 33(1,):1–78.

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

Burge, T. (2013). Some Remarks on Putnam's Contributions to Semantics: Some remarks on Putnam's contributions to semantics. *Theoria*, 79(3):229–241.

Burge, T. (2020). Entitlement: The Basis for Empirical Warrant. In Graham, P. J. and Pedersen, editors, *Epistemic Entitlement*, pages 37–142. Oxford University Press.

Burge, T. (2022). *Perception: first form of mind*. Oxford University Press, Oxford, United Kingdom. OCLC: 1319221816.

Byrd, N. (2019). What we can (and can't) infer about implicit bias from debiasing experiments. *Synthese*.

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

Cloutier, J., Li, T., and 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.

Connolly, K. J. (2019). *Perceptual learning: the flexibility of the senses*. Philosophy of mind series. Oxford university press, New York (N.Y.).

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

Del Pinal, G., Madva, A., and Reuter, K. (2017). Stereotypes, Conceptual Centrality and Gender Bias: An Empirical Investigation: Stereotypes, Conceptual Centrality and Gender Bias. *Ratio*, 30(4):384–410.

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

Eberhardt, J. L., Goff, P. A., Purdie, V. J., and Davies, P. G. (2004). Seeing Black: Race, Crime, and Visual Processing. *Journal of Personality and Social Psychology*, 87(6):876–893.

Faucher, L. (2014). Non-Reductive Integration in Social Cognitive Neuroscience: Multiple Systems Model and Situated Concepts. In *Brain Theory: Essays in Critical Neurophilosophy*, pages 217–240. Palgrave.

Firestone, C. and Scholl, B. J. (2014). "Please Tap the Shape, Anywhere You Like": Shape Skeletons in Human Vision Revealed by an Exceedingly Simple Measure. *Psychological Science*, 25(2):377–386.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., and Nosek, B. A. (2019). A Meta-Analysis of Change in Implicit Bias. *Journal of Personality and Social Psychology*.

Gelman, S. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, 8(9):404–409.

Gelman, S. A., Coley, J. D., and Gottfried, G. M. (1994). Essentialis Beliefs in Children. In Hirschfeld, L. A. and Gelman, S. A., editors, *Mapping the Mind: Domain Specificity in Cognition and Culture*, pages 341–365. Cambridge University Press.

Gigerenzer, G. (2008). Why heuristics work. *Perspectives on psychological science*, 3(1):20–29.

Gil-White, F. (2001). Are Ethnic Groups Biological "Species" to the Human Brain?: Essentialism in Our Cognition of Some Social Categories. *Current Anthropology*, 42(4):515–553.

Goldman, A. I. (1976). Discrimination and Perceptual Knoweldge. *The Journal of Philosophy*, 73:771–791.

Green, E. J. (2020). The Perception-Cognition Border: A Case for Architectural Division. *The Philosophical Review*, 129(3):323–393.

Haslam, N., Rothschild, L., and Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of Social Psychology*, 39(1):113–127.

Helton, G. (2016). Recent issues in high-level perception: High-Level Perception. *Philosophy Compass*, 11(12):851–862.

Hirschfeld, L. A. and Gelman, S. A. (1994). *Mapping the mind: Domain specificity in cognition and culture*. Cambridge University Press.

Hochman, A. (2013). Do We Need a Device to Acquire Ethnic Concepts? *Philosophy of Science*, 80(5):994–1005.

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

Hu, L. (2019). On the Hunt for the Correct Counterfactual. *The Phenomenal World*.

Jenkin, Z. (2020). The Epistemic Role of Core Cognition. *The Philosophical Review*, 129(2):251–298.

Jenkin, Z. (2023). Perceptual learning. *Philosophy Compass*, 18(6):e12932.

Johnson, G. M. (2020a). Algorithmic bias: on the implicit biases of social technology. *Synthese*.

Johnson, G. M. (2020b). The Structure of Bias. *Mind*, 129(516):1193–1236.

Johnson, G. M. (2023). Are Algorithms Value-Free?: Feminist Theoretical Virtues in Machine Learning. *Journal of Moral Philosophy*, 21(1-2):27–61.

Judd, C. M. and Park, B. (1993). Definition and Assessment of Accuracy in Social Stereotypes.

Kawakami, K., Dovidio, J. F., and Van Kamp, S. (2007). The Impact of Counterstereotypic Training and Related Correction Processes on the Application of Stereotypes. *Group Processes & Intergroup Relations*, 10(2):139–156.

Keller, J. (2005). In Genes We Trust: The Biological Component of Psychological Essentialism and Its Relationship to Mechanisms of Motivated Social Cognition. *Journal of Personality and Social Psychology*, 88(4):686–702.

Kelly, D. J., Quinn, P. C., Slater, A. M., Lee, K., Ge, L., and Pascalis, O. (2007). The other-race effect develops during infancy evidence of perceptual narrowing. *Psychological Science*, 18(12):1084–1089.

Kelly, T. (2022). Bias: A Philosophical Study. In *Bias*, pages 17–C1.P92. Oxford University PressOxford, 1 edition.

Kinzler, K. D., Dupoux, E., and Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences*, 104(30):12577–12580.

Kinzler, K. D., Shutts, K., and Correll, J. (2010). Priorities in social categories. *European Journal of Social Psychology*, 40(4):581–592.

Kinzler, K. D., Shutts, K., DeJesus, J., and Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social cognition*, 27(4):623.

Klein, P. (1996). Warrant, Proper Function, Reliabilism, and Defeasibility. In Kvanvig, J. L., editor, *Warrant in Contemporary Epistemology*. Rowman & Littlefield, Lanham, Maryland.

Knill, D. C. (2007). Learning Bayesian priors for depth perception. *Journal of Vision*, 7(8):13.

Kripke, S. (1980). *Naming and Necessity*. Basil Blackwell, Oxford.

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., Sartori, G., Dial, C. M., Sriram, N., Banaji, M. R., and Nosek, B. A. (2014). Reducing implicit racial preferences: I. A comparative investigation of 17 interventions. *Journal of Experimental Psychology: General*, 143(4):1765–1785.

Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., Calanchini, J., Xiao, Y. J., Pedram, C., Marhsburn, C. K., and others (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General (in press)*.

Lall, M. K. and Tanaka, J. W. (2023). The culture of perceptual expertise and the other-race effect. *British Journal of Psychology*, 114(S1):21–23.

Leslie, A. M. and Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, 25(3):265–288.

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

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

Loven, J., Herlitz, A., and Rehnman, J. (2011). Women's Own-Gender Bias in Face Recognition Memory: The Role of Attention at Encoding. *Experimental Psychology*, 58(4):333–340.

López, F. J., Alonso, R., and Luque, D. (2016). Rapid Top-Down Control of Behavior Due to Propositional Knowledge in Human Associative Learning. *PLOS ONE*, 11(11):e0167115.

Machery, E. and Faucher, L. (2001). Why Do We Think Racially? In *Handbook of Categorization in Cognitive Science*, pages 1009–33. Elsevier, Amsterdam.

Madva, A. (2016). Virtue, Social Knoweldge, and Implicit Bias. In Brownstein, M. and Saul, J., editors, *Implicit Bias and Philosophy, Volume 1: Metaphysics and Epistemology*, pages 191–215. Oxford University Press.

Madva, A. (2017). Biased against Debiasing: On the Role of (Institutionally Sponsored) Self-Transformation in the Struggle against Prejuice. *Ergo, an Open Access Journal of Philosophy*, 4(20201214).

Mandalaywala, T. M., Amodio, D. M., and Rhodes, M. (2018a). Essentialism Promotes Racial Prejudice by Increasing Endorsement of Social Hierarchies. *Social Psychological and Personality Science*, 9(4):461–469.

Mandalaywala, T. M., Ranger-Murdock, G., Amodio, D. M., and Rhodes, M. (2018b). The Nature and Consequences of Essentialist Beliefs About Race in Early Childhood. *Child Development*.

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

Medin, D. L. and Ortony, A. (1989). Psychological Essentialism. In Vosniadou, S. and Ortony, A., editors, *Similarity and Analogical Reasoning*, pages 179–195. Cambridge University Press, Cambridge ; New York.

Meissner, C. A. and Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1):3–35.

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

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

Munton, J. (2021). Prejudice as the misattribution of salience. *Analytic Philosophy*, page phib.12250.

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

Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development*, 10(1-2):3–18.

Neufeld, E. (2019). An Essentialist Theory of the Meaning of Slurs. *Philosophers' Imprint*, 19(35).

Neufeld, E. (2022). Psychological essentialism and the structure of concepts. *Philosophy Compass*, 17(5).

Orlandi, N. (2014). *The innocent eye: why vision is not a cognitive process*. Oxford University Press, Oxford.

Pauker, K., Ambady, N., and Apfelbaum, E. P. (2010). Race Salience and Essentialist Thinking in Racial Stereotype Development: Racial Stereotype Development. *Child Development*, 81(6):1799–1813.

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.

Plantinga, A. (1993). *Warrant and Proper Function*. Oxford University Press, New York.

Putnam, H. (1975). The Meaning of 'Meaning'. In *Mind, Language and Reality*, pages 215–271. Cambridge University Press.

Rescorla, M. (2021). Bayesian modeling of the mind: From norms to neurons. *WIREs Cognitive Science*, 12(1):e1540.

Rhodes, M. and Gelman, S. A. (2009). A developmental examination of the conceptual structure of animal, artifact, and human social categories across two cultural contexts. *Cognitive Psychology*, 59(3):244–274.

Rhodes, M., Leslie, S.-J., Bianchi, L., and Chalik, L. (2018). The Role of Generic Language in the Early Development of Social Categorization. *Child Development*, 89(1):148–155.

Rhodes, M. and Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism: Social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(4):e1437.

Ritchie, K. (2021). Essentializing Language and the Prospects for Ameliorative Projects. *Ethics*, 131(3):460–488.

Rothbart, M. and Taylor, M. (1992). Category Labels and Social Reality: Do we view social categories as natural kinds? In Semin, G. R. and Fiedler, K., editors, *Language, interaction and social cognition*, pages 11–36. SAGE Publications.

Schneider, D. J. (2004). *The psychology of stereotyping*. Distinguished contributions in psychology. Guilford Press, New York.

Siegel, S. (2011). *The Contents of Visual Experience*. Oxford University Press.

Siegel, S. (2012). Cognitive penetrability and perceptual justification. *Noûs*, 46(2):201–222.

Siegel, S. (2017). *The rationality of perception*. Oxford University Press, Oxford, United Kingdom, first edition edition.

Siegel, S., Silins, N., and Matthen, M. (2014). *The Epistemology of Perception*. Oxford University Press.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition*, 74(2):149–175.

Taylor, M. G., Rhodes, M., and Gelman, S. A. (2009). Boys Will Be Boys; Cows Will Be Cows: Children's Essentialist Reasoning About Gender Categories and Animal Species. *Child Development*, pages 461–481.

Trawalter, S., Todd, A. R., Baird, A. A., and Richeson, J. A. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology*, 44(5):1322–1327.

Vasilyeva, N., Gopnik, A., and Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology*, 54(9):1735–1744.

Vasilyeva, N. and Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition*, 204:104383.

Williams, M. J. and Eberhardt, J. L. (2008). Biological conceptions of race and the motivation to cross racial boundaries. *Journal of Personality and Social Psychology*, 94(6):1033–1047.

Wilson, J. P., Hugenberg, K., and 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.