Research Statement

Gabbrielle M Johnson

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.

My research program argues that there is a natural kind *social bias* that all three cases fall under and defends a theory of what that kind is. My theory explains how the cases are unified, how they differ, and why the differences between the cases matter. Within a computational theory of mind, the tasks of unification and differentiation can appear to be at odds with one another. The more we highlight differences among how Mr. T, Ms. O, and MatGPT were processing informational states, the harder it is to use those same computational resources to say what they have in common. My analysis reconciles these tasks within a cognitive science framework by shifting to a higher level of abstraction, arguing for an emergent, unifying nature of bias as a functional entity. This functional account of social bias specifies what bias is by the function that it serves, or how it transitions us from input to outputs. In all cases, biases guide inductive inferences by taking as an input some underdetermining evidential state about what social group a person belongs to and producing as an output some determinate "best guess" as to the environmental regularities that pair that social group with stereotypical properties.

By conceptualizing bias as a function that can be realized in myriad ways, my analysis liberates the concept from previous conceptual constraints and opens up new avenues for investigation and intervention. In particular, my functional account moves us away from the narrow focus of what mental states an individual harbors, towards a more expansive interest in how those states are manipulated. It delves into the diverse factors both within individuals and external to them that can underpin this function. This shift in perspective deepens our understanding of how we interact with others and helps us improve to become better moral agents. I aim to give accounts of social cognition that are both empirically and conceptually well-informed by working across the fields of social psychology, cognitive science, computer science, philosophy of mind and language, and ethics.

The Nature of Bias

I argue that social bias (hereafter, simply 'bias') is a functionally defined mental entity that takes structured mental states as inputs and returns structured mental states as outputs in a way that mimics inductions made on the basis of social kind membership. All three cases of bias relate the input that Mary is a woman to the output that she's not suitable for a mathematics program. Like functional analyses of other mental states, my analysis of bias entails that it is multiply realizable by a variety of computational systems and decision-making processes. For instance, biases could be realized by an explicit belief that women are ill-suited for mathematics (as Mr. T has) or 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 its training instances (as MatGPT does).

My analysis also reveals an additional, previously unappreciated way that bias can arise within cognitive architecture. The history of inquiries into the nature of bias have, to their detriment, tended toward over-intellectualization and over-individualization. These are theories that treat Mr. T's bias as the paradigmatic case and theorize other instances within its shadow. This gives rise to a theory that focuses on particular mental states—beliefs, stereotypes, prejudices—that an individual harbors, evidenced by their avowal of those states. If you want to know if someone is biased toward some group, you just ask them. In my article "The Psychology of Bias" (in *An Introduction to Implicit Bias: Knowledge, Justice, and the Social Mind*, Routledge), I critique this narrow perspective in the existing literature on psychological models of bias. In it, I argue that psychological and philosophical attempts to model social bias have centered on empirical data patterns displayed by individuals which provide different, often conflicting accounts. I provide an in-depth

introduction that addresses these tensions in data, method, and theory, arguing that traditional approaches lack the resources to account for bias's diversity.

Accordingly, my functional approach to bias flips the traditional script, treating the obvious and overt cases of bias as occurring in the margins. My article "The Structure of Bias" (*Mind*) puts forward my positive functional analysis. In it, I argue that a representationalist approach to bias, i.e., one that identifies having a bias with having representations of stereotypes, is a mistake because it conceptualizes social bias in ways that do not fully capture the phenomenon. Crucially, this view fails to capture a heretofore neglected possibility of bias: one that influences an individual's beliefs about and actions toward other people, but is, nevertheless, nowhere represented in that individual's cognitive repertoire. I then demonstrate how my functional analysis is uniquely suited to handle such cases. In the resulting view, biases are more like rules guiding inferences about others than like stand-alone belief states. No other account of bias has the ability to unify cases across them, while also locating theoretically important distinctions between them.

Varieties of Bias

This expansive understanding of what functional role diverse cases of bias have in common sheds light ultimately on neglected forms bias can take. I have thus extended this framework to explore the diversity of biases in human cognitive architectures, algorithms and artificial intelligence, and in society and scientific practice.

Cognitive Architecture

Whereas traditional theories of implicit bias simply *posit* the existence of unconscious biases that are difficult to correct, my functional proposal helps *explain* these features. If some content is conscious, 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 can explain why Ms. O is oblivious to her rationale for rejecting Mary. A second way, neglected in theories of bias until now, is if the content is an abstraction from personal-level, explicitly represented states; in this way it emerges "above" the level of 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. In my article, "Unconscious Perception and Unconscious Bias: Parallel Debates About Unconscious Content" (*Oxford Studies in Philosophy of Mind*), I compare and contrast two parallel debates surrounding unconscious content in philosophy and psychology: unconscious perception and unconscious bias. I use the debates surrounding both to highlight the significance of different theoretical assumptions for empirical investigations into conscious accessibility.

Likewise, in my article "The (Dis)Unity of Psychological (Social) Bias" (revise and resubmit at *Philosophical Psychology*, and recipient of the Eleventh Annual Essay Prize at the University of Antwerp's Center for Philosophical Psychology), I detail several examples of what I call the "system-dependence" of bias. Using contemporary empirical work, I argue that while visual social biases and cognitive social biases share a functional profile, their diverse instantiations render them importantly different. Visual biases are principally example-based; whereas social biases in central cognitive capacities are principally theory-based. For example, perceptual bias might rely on past examples pulled from misrepresentations in the media 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. This functional analysis likewise helps to address a longstanding question in the philosophical study of bias: why is combating social bias so difficult? Because bias is multiply realizable, combating one instance of a social bias can easily leave another intact. For example, 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.

Algorithms and Artificial Intelligence

Emergent bias—that is, one that emerges 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.

In my article "Algorithmic Bias" (*Synthese*), I explore in depth these biases that emerge out of seemingly innocuous patterns of information processing in artificial systems. This emergent bias, absent any human agent, obscures the existence of the bias itself, making it difficult to identify, mitigate, or evaluate using standard resources in epistemology and ethics. The insistence that human biases and computer biases are fundamentally distinct undermines our ability to extend lessons from one domain to the other. In this paper, I argue for the commonalities between the two under my functional model, ultimately paving way for this intellectual exchange.

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 machine learning program 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. My paper "The Hard Proxy Problem: Proxies Aren't Intentional, They're Intentional" (MS) confronts this conceptual problem head on by demonstrating how a theory of proxies is integral to a fully expansive understanding of social bias. This is because the functional account allows us to recognize the source of bias as entirely outside of individuals. This occurs when explanations of the functional transitions from the inputs to the outputs ultimately makes reference to social kinds in the external environment. 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 all current theories of bias, Dr. R's evaluation method wouldn't be considered biased against women. However, through my functional theory and understanding of proxy attributes, we can recognize the potential for bias if Dr. R's criteria indirectly reflect wider discriminatory practices against women in their environment. For example, I argue that if societal factors like unequal access to resources for test preparation ultimately explain the lower scores for women on standardized tests, then Dr. R's evaluation, though seemingly impartial, instantiates a form of social gender bias.

Societies and Scientific Practice

Now at its most expansive analysis, my account of bias moves beyond individuals to intellectual communities and the biases that they collectively instantiate. Firstly, we need to be mindful of how bias can emerge within mixed human-technical systems. In a co-authored paper "Uncanny Performance, Divergent Competence: Biases as Principled Barriers to Human-Machine Communication" (in *Communication with AI: Philosophical Perspectives,* under contract with OUP), Gabriel Dupre and I argue that the inherent biases that are encoded in humans and machines make it impossible for us to know that our concepts (with all of their built-in biased associations) are the same as those concepts in machines (which likely lack the full range of biased associations harbored by humans). This leads to critical miscommunications when, for example, we take an algorithm's use of some concept like "recidivism risk" to be the same as our own.

Additionally, we need to be mindful of how biases might emerge out of group-based inquiry. In my "Are Algorithms Value-Free" (*Journal of Moral Philosophy*), I extend familiar debates about values in science to the new domain of machine learning programs. The core insight of the paper is that the sources for values in science are the same sources that give rise to the need for bias in everyday thought and inquiry. In future work, I want to further explore how we can think of scientific practice as a model for evaluating intellectual communities for the biases they collectively harbor.

Additional Research Avenues

Recapturing Normativity

Once we have such an expansive notion of bias, we lose certain evaluative tools. By no longer centering individuals and the mental states they harbor, we can no longer rely on standard evaluative resources in ethics and epistemology. To address this issue, I have begun drafting a paper in which I aim to flesh out an

epistemic theory of psychological entitlements for social biases that is expansive enough to account for biases of all varieties, in particular the most expansive notion constituted by proxies.

Mitigation Techniques

Now that we've identified what bias is and how to evaluate it, we can get to the most pressing question of how we combat problematic biases. My principle contribution to this literature will argue that, given the multifaceted nature of bias, to achieve mitigation, we have two strategies: the first strategy focuses on bias's functional role, while the second focuses on individual differences in how biases operate. Because biases are internal mechanisms that function to mimic environmental regularities, we can either adopt strategies for changing the mechanism (manipulating training data, input filters, blocking inferences) or we can adopt strategies for changing the regularities they function to track (change the world). Both approaches have their virtues and vices: mechanistic interventions are easier to achieve, but short-lived; functional interventions are hard to bring about, but once successful, long-lasting. Thus, in future work I aim to outline concrete strategies (both mechanism-based and function-based) for ameliorating problematic biases across the various domains in which bias operates, alleviating group-based inequality in all of its manifestations.