Dissertation Abstract: *Cognition and the Structure of Bias*
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 Hal, an automated program that is trained on past admittance data about which students, when accepted, have gone on to successful careers in the field.

My dissertation 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 Hal 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.

I argue that social bias (hereafter, simply 'bias') is a functionally defined mental entity that takes propositional mental states as inputs and returns propositional 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).

My analysis also reveals an additional, previously unappreciated way that unconscious bias, i.e., *implicit bias*, can arise within cognitive architecture. 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.

This emergent bias—that is, one that emerges out of patterns of information processing—is similar to Hal's case, and indeed cases of it are primarily studied within computer science and machine learning. I argue that humans too can instantiate this kind of bias and that there's a perfectly reasonable sense in which human biases consist merely in patterns of how mental states are organized, even when those other mental states are not specifically about the values or stereotypes that the bias reflects. Such socially biased patterns reflect systematic regularities in how our society is organized. Thus, a second commitment of my general theory of bias is that biases need not be explicitly represented at all. I call these *truly implicit biases*.

My analysis allows bias to manifest in a wide range of ways. Different instantiations of bias will depend on the psychological system in which the bias is embedded. This wider system will determine the sorts of states and processes that bridge the gap between the functional inputs and outputs. For example, I conjecture that some social biases manifest in the visual system, where they will rely primarily on surface-level features. This can explain our tendency to identify ambiguous objects as weapons depending on the perceived skin tone of the individual holding the object. Other forms of bias manifest in a theory-laden belief system —where a bias will paradigmatically rely on below-the-surface features. This can explain our tendency to assume that members of marginalized demographics share an underlying essence that is causally responsible for stereotypical properties. These different psychological entities would nonetheless produce similar discriminatory patterns of social interaction; they are both biases.

My theory of bias has an important practical consequence: since the states and processes biases comprise are system-dependent, no one mitigation technique will be universally effective. Thus, a final insight of the dissertation is that our most effective debiasing techniques will be tailored to how mental systems globally operate, again bolstering my claim that we must widen our analysis beyond individual mental states.