Research Statement

How do you know your own mind, how do others know your mind, and what is the importance of this knowledge? A great deal of my research explores two epistemologically puzzling phenomena in connection with these questions. The first phenomenon is our *first-person authority*. Roughly, to have first-person authority is to be owed (and typically receive) deference from your listeners when ascribing mental states to yourself. The second phenomenon is the *privileged and peculiar self-knowledge* we tend have of our own mental states. Roughly, in having privileged self-knowledge, one tends to know one's own mental states better than anyone else, and in having peculiar self-knowledge one tends to know these mental states in a way that no one else can use to gain knowledge of one's mind. I offer explanations of each of these phenomena, and I explain how they jointly ground different forms of social-epistemic cognition, such as interpersonal reasoning and linguistic interpretation.

While you can learn more about my published works at benwinokur.com/research, here I will describe my not-yet-published projects. I begin with two papers currently under review. The first has recently been revised and resubmitted. In it, I respond to recent skepticism about first-person authority. The skeptic argues that there is either (1) nothing philosophically puzzling about first-person authority, or (2) no such thing as first-person authority. I respond to this dilemma by proposing several refined specifications of first-person authority, each of which are plausible and philosophically puzzling.

In the second paper under review, I argue that Donald Davidson's philosophical corpus harbors the resources for an interesting but hitherto underdeveloped transcendental account of peculiar self-knowledge. On this account, we necessarily have peculiar self-knowledge insofar as we are capable of interpreting the speech and thoughts of others. Along the way, I critically address a reading of Davidson according to which he was a kind of deflationist or quietist about self-knowledge.

In a recently completed paper, I evaluate different "expressivist" accounts of first-person authority. Expressivist explanations of first-person authority share the idea that listeners are entitled to (and do) defer to a speaker's self-ascriptions insofar as those self-ascriptions directly express something about the speaker. But they differ in terms of what expressed feature is relevant. *Neo-expressivists* argue that the relevant expressed feature is the very mental state self-ascribed, whereas *agency-based expressivists* argue that the relevant expressed feature is the agent's cognitive agency with respect to the mental state self-ascribed. I defend neo-expressivism against criticisms from the agency-based expressivist camp, after which I argue against agency-based expressivism directly. I will soon publish this paper in a special issue of the journal *Philosophies*, tentatively titled "Expression and Self-Knowledge", that I am co-editing with Dorit Bar-On.

I will now describe a few projects in development. In one paper, I develop a "constitutivist" account of self-knowledge of propositional attitudes. On this account, a first-order propositional attitude is, in normal cognitive conditions, itself a *part* of one's second-order belief to the effect that one has that attitude, rather than an ontologically independent mental state that one introspectively *detects*. I defend my constitutivist account against a battery of criticisms, such as the criticism that constitutivist views are susceptible to a vicious regress, and that constitutivist views do not allow us to see self-knowledge as a cognitive achievement. Finally, I defend my constitutivist account against a recently proposed non-constitutivist alternative, one that is supposed to better explain the fallibility of our self-knowledge.

A final project within my primary research program is in its earlier stages. This is a monograph titled *The Functions of Self-Knowledge*. This is the first monograph dedicated solely to the question of what functional roles privileged and peculiar self-knowledge plays in our psychological economies. The monograph will contain six chapters: one introducing key intuitions about the privilege and peculiarity of much of our self-knowledge, one describing various skeptical responses to these intuitions, three comprising a critical survey of extant accounts of the functional roles of privileged and peculiar self-knowledge, and one offering my own account of the indispensability of privileged and peculiar self-knowledge for various social-epistemic ends, as originally argued in my doctoral thesis. A more detailed précis of this monograph (approximately 2200 words) is available upon request.

Here are a few other topics that I intend to broach soon:

1) The metaphysical differences between 'brute' and 'rational' desires, and the implications of these differences for how we acquire self-knowledge of them
2) The differences, if there be any, between self-ascriptive speech acts that *express* one's mental states and self-ascriptive speech acts that *testify* as to one's mental states
3) The question of whether our tendencies to 'confabulate' post-hoc rationalizations for our judgements impedes our first-person authority with respect to those judgements

Moving beyond my primary research program, I have been deepening my interests in social and digital epistemology. I am currently writing a paper about the epistemic injustice one suffers when one's sincere, competent online comments are ignored due to the assumption that they have been posted by a bot. This is increasingly prevalent in places like North America and the UK in light of recent revelations about the prevalence of (primarily Russian) bots on platforms like Reddit and Twitter. Of course, because many such bots are real, "accusations of bothood" are often true. But some are false. The false accusations are interesting, I argue, because they generate epistemic injustices that look a lot like 'testimonial injustices', except that they are not rooted in the hearer's prejudices about the speaker's identity. After all, as testimonial injustice is typically understood, a speaker's testimony is not taken *sufficiently seriously*, owing to the hearer's prejudices against speakers of a given sex, class, ethnicity, and so on. However, when someone assumes that a post has been made by a bot, it is being assumed that *no testimony has really been issued*, since bots are not genuine speakers. I explore some consequences of this fact for thinking about how to combat epistemic injustices produced by accusations of bothood.