Research Statement

Thalia Vrantsidis

My research investigates a fundamental aspect of human psychology: our ability to understand. When we understand something – say, another person – we do not simply memorize a set of facts, or have a "heap" of unrelated information about them. Instead, we *organize* and *structure* this information, for example, inferring causal relationships, or finding patterns (e.g., inferring general traits from behaviors). In studying understanding, my work thus examines our ability to form, represent, and use this type of structured knowledge. This ability is central to much of human cognition (e.g., our explanations, schemas, impressions), and is essential for solving problems, achieving goals, and navigating the world. Moreover, improving one's understanding is arguably one of the things people find most intrinsically valuable (as in the satisfaction of an "aha" moment when we discover a new connection). Given the widespread value of understanding, my work ultimately aims to help people understand better, so that we can have representations that better reflect, or are more likely to reflect, the true structure of the world.

In doing this research, I use a fundamentally interdisciplinary approach, which integrates social, cognitive, computational, and philosophical perspectives. With this approach, I address two interrelated questions. The first is **how understanding can go wrong.** Focusing on the domain of stereotyping, I examine the causes and consequences of errors in people's social understanding. The second question my work addresses is **how understanding can improve.** I examine an essential component of this process: how people select better ways of understanding something. To answer these questions, my work interrogates human behavior using a combination of behavioral studies and Bayesian cognitive modelling. By understanding the fundamental principles that guide human cognition, this work has broad implications for everything from basic cognitive and neural processes, to our interactions with the social world, to the core human values involved in living a meaningful life.

When understanding goes wrong: Causes and consequences of errors in people's social understanding

One line of my work investigates how understanding goes wrong, particularly in the domain of stereotyping. One of the key contributions of my work in this area has been to use **powerful theoretical and methodological tools to distinguish different types of wrongs involved in stereotyping**. For example, if we think a case of stereotyping is wrong, is it wrong because of errors somewhere in the relevant belief structures (e.g., beliefs about groups, individuals, and their relationships)? Because of errors in how we *use* these beliefs to make inferences? Or because, even if these are correct, it is still morally problematic? Distinguishing these errors involves overcoming two key challenges in the field: 1) the need for clear normative standards to define what counts as an error, and 2) the need to specify the full set of relevant beliefs and relationships, to pinpoint where in these structures errors may be occurring. Much of my work has overcome these challenges using Bayesian cognitive modelling, a powerful computational approach just beginning to be applied to stereotyping. This work has provided novel insights into the causes and consequences of stereotyping – providing some of the first tests of classic theories of stereotyping, identifying new forms of errors, and new ways these errors can be counteracted. In turn, these insights can be used to develop targeted interventions for reducing problematic stereotyping, by addressing the specific errors involved in each case.

In one example of this approach, my work has provided **one of the first clear tests of a classic idea within social psychology: that certain impression formation heuristics can lead people to over-rely on social categories such as race, gender, or occupation**, to the exclusion of other information. For example, these heuristics might lead people to over-rely on someone's race, and thus mis-judge a Black person as aggressive, despite the person's smiling expression and unaggressive behavior. These heuristics could thus increase the chance of misunderstanding individuals whenever they differ from stereotypes of their groups. Importantly, my work provided one of the first clear tests of this idea, by formalizing a normative Bayesian model of this process. This model allowed for precisely defining over-

reliance on categories, by contrasting this to valid Bayesian inference, given one's other beliefs. Category over-reliance could then be identified by comparing people's inferences to this model, while measuring and controlling for people's other beliefs. So far, this work has found little evidence for this form of heuristic-driven category over-reliance (**Vrantsidis**, & Cunningham, submitted), though ongoing work is examining other conditions in which it might occur (**Vrantsidis**, & Cunningham, in prep). These results suggest that, at least in the cases examined here, stereotype-based misunderstandings do not necessarily stem from errors in people's inference processes. Instead, they may stem from errors in people's beliefs about social groups (e.g., about how aggressive Black people are in general). This work also provides a powerful approach for identifying other errors in people's inferences. For example, my ongoing work tests whether people might over-rely on social categories because they fail to consider relevant causal relationships (e.g., that racial differences are often caused by differences in poverty levels; **Vrantsidis**, Buchsbaum & Cunningham, in prep).

In other work, I used a Bayesian approach to **more fully specify the relevant belief structures, and thus identify novel sources of errors in people's social understanding, and as well as novel ways these errors can be counteracted** (**Vrantsidis** & Cunningham, 2021). In particular, this work focused on the potentially biased information people receive about social groups, and the downstream errors this may cause. In a series of studies, I examined how the amount and source of information people had about social groups was related to people's beliefs about those groups, and their inferences about group members. By using a Bayesian-inspired approach, I was able to more fully specify the belief structures that could be affected by biased information (e.g., beliefs about a group's average, its variability, individual group members, as well as confidence in these beliefs). While previous work only studied these components in isolation, studying this structure as a whole allowed me to identify how errors propagate through this system. This approach revealed evidence for a novel form of error. Specifically, these studies suggest that having more second-hand information about a group biases people to underrepresent the groups' variability, which, in turn, leads to greater confidence when applying stereotypes to individuals. So, for example, someone who has only learned about Black people from the media might come to see them as overly uniform in their aggressiveness – thus misunderstanding the group – and then over-confidently apply this stereotype to every Black person they meet – thus misunderstanding the large portion of Black people who do not fit this stereotype. Moreover, considering the full belief structure highlighted novel ways that these errors might be counteracted. For example, my work suggested that having limited personal experience with a group also led to underrepresenting the group's variability. However, contrary to previous predictions, this did *not* lead to more confident inferences about group members. Examining the full set of relevant beliefs allowed for explaining this: people appeared to recognize the unreliability of their limited information, and thus appropriately reduced confidence in their other beliefs (here, about the group average). This suggests that highlighting the unreliability of one's information could also help reduce the impacts of other errors in people's social understanding.

In addition to investigating causes of errors in our social understanding, I have also examined the consequences of these errors. My work has **highlighted an under-appreciated route to stereotyping, that can lead even well-intentioned people to intentionally use problematic stereotypes** (**Vrantsidis** & Cunningham, submitted)**.** Specifically, when people have errors in their social understanding (e.g., thinking a stereotype is accurate, or not seeing the harms it causes), this can lead people to freely use problematic stereotypes, precisely because they view these as merely acceptable group-based generalizations. This process can produce a form of "bias blind-spot", where people tend to see stereotyping as something not done by themselves, but by other people – and especially by those whose social understanding differs from one's own (e.g., those with differing political views). In providing evidence for this view, my work highlights an important, yet under-appreciated, factor that can drive stereotype use: people's moral evaluations of their stereotypes, and the beliefs that feed info this. This

view can help better explain cases of stereotyping that are not well accounted for by major previous theories: e.g., cases where people intentionally use problematic stereotypes, without competing or ill-intentions. Practically, this work suggests that addressing people's underlying misunderstandings may be a valuable strategy for confronting stereotype use. This strategy may especially help reduce the backlash often caused by these confrontations – particularly important in today's increasingly antagonistic social and political environments.

My ongoing work has delved further into people's moral evaluations of stereotyping, and how this relates to correct or incorrect understanding (**Vrantsidis**, Feinberg, & Cunningham, in prep). In particular, I am examining cases where stereotyping may be viewed as immoral, even though it seems to reflect correct understanding of a social group (e.g., real differences between racial groups). Investigating these cases can shed light on the complex interplay between moral values and the value of understanding, and the consequences when these values conflict.

When understanding improves: How people select better ways of understanding

Given that understanding often goes wrong, my work also examines how understanding can improve. Just as evolutionary fitness can improve through variation and selection, understanding can improve through generating various ways of understanding something, and selecting the better ones. My work has focused on this second process: how people evaluate and select better ways of understanding.

My work sheds light on the relationship between two broad approaches to selecting better understanding, a topic of ongoing debate within philosophy and psychology. The first approach is probabilistic, and says that people prefer ways of understanding that are more likely to be true, given their observations, and the rules of Bayesian or probabilistic inference. The second approach focuses on "explanatory virtues", and says that people prefer ways of understanding that possess virtues such as greater simplicity, breadth, goodness-of-fit, etc. My work attempts to reconcile these views, by suggesting that these two approaches may play different psychological roles, and by using carefully controlled experiments to determine the precise nature of these roles.

This work has provided **novel insights into how Bayesian inference and explanatory virtues form complementary strategies for evaluating the probability of different ways of understanding something** (**Vrantsidis** & Lombrozo, 2022). Focusing on the virtue of simplicity, I examined how people evaluate simpler and more complex ways of understanding a set of observations (e.g., using one vs. two diseases to explain a set of symptoms). Participants in these studies evaluated the outputs of Bayesian inference for each explanation (i.e., its probability, given the observations), while either estimating or being provided with the inputs to Bayesian inference (e.g., baserates for having one vs. two diseases, in general). These studies showed that simplicity may help people overcome multiple challenges faced when using Bayesian inference to evaluate explanations. I found that people use simplicity as a cue to the inputs of Bayesian inference – e.g., using simplicity to infer higher baserates for one vs. two diseases. This may help deal with the uncertainty people often have about these input values. Furthermore, I found the first evidence that simplicity's role goes beyond this, serving as a direct cue to the outputs of Bayesian inference. That is, even when controlling for the different inputs to Bayesian inference, simpler explanations were still judged as more probable, given the observations. This occurred especially when there was more uncertainty and effort associated with these input values (because they were estimated, rather than provided), suggesting that simplicity may also help deal with the difficulty of mentally doing probabilistic computations. Building on these findings, my ongoing work investigates how flexible people's use of simplicity is (**Vrantsidis** & Lombrozo, in prep). Specifically, I examine whether people adjust their use of simplicity in appropriately context-sensitive ways, thus generally increasing the accuracy of their probability estimates, or whether they use it as a relatively inflexible and less-accurate heuristic. Overall, this line of work integrates psychological, philosophical,

and computational perspectives to provide insights into how understanding can improve, given the challenges we face as limited cognitive agents.

Future directions

My research program centers on human ability to understand: how this can go wrong, and how it can improve, with the ultimate goal of helping people do this better. Over the next several years I plan to extend and integrate my current lines of research. One of the major questions I plan to investigate is how the probabilistic view of understanding, as used in my current work, relates to alternative views based on cognitive efficiency. For example, other work has framed the goal of understanding as maximizing the cognitive efficiency of one's representations – e.g., so that we prefer simpler representations not because they are seen as more probable, because they allow us to more efficiently compress or re-represent known data. Indeed, the question of how these two views are related is important to much of cognitive science, as maximizing probability and maximizing cognitive efficiency have both been viewed as fundamental goals of the human mind. Yet, the relationship between these goals remains unclear: they have been variously viewed as equivalent, competing, derivative, or unrelated. And theoretical arguments about their relationship depend crucially on the nature of people's cognitive machinery, thus making this an important empirical question.

I therefore plan to investigate the relationship between probability and cognitive efficiency goals as applied to understanding. The first step in doing this will be to dissociate the consequences of these different goals (e.g., they may lead people to value different forms of simplicity). This in turn can be used to examine the relationship between these goals (e.g., do cognitively simpler representations increase or decrease people's probability of being accurate?). It can also allow for examining whether one of these goals might be more fundamental, in terms of being more strongly tied to the subjective (and perhaps intrinsic/innate) sense of satisfaction that often accompanies good understanding.

I plan to address these types of questions through a combined social-cognitive approach. To focus on the basic cognitive processes involved, I will build off the types of experiments used in my previous research (e.g., where symptoms could be explained by either one vs. two diseases). I also plan to extend these studies to the perceptual domain (e.g., where perceptual inputs could be caused by one vs. two external sources). In these perceptual cases, understanding often occurs more automatically, and thus may involve more basic cognitive mechanisms, which may be more closely linked to the potentially innate value of understanding. I will also study the implications of these processes for how we understand other people. For example, building off my work on impression formation heuristics, I plan to investigate other ways in which these different goals might lead people to simplify their impressions (e.g., representing fewer causes for a person's behavior, or simpler relationships between these causes), and thus perhaps sometimes over-rely on stereotypes. This work can thus provide insights into the fundamental principles of human cognition, as well as questions of broad social importance, such as why we use stereotypes, and how to reduce problematic stereotyping and promote better understanding of other people.

As a whole, my work uses a fundamentally interdisciplinary approach – integrating social, cognitive, computational, and philosophical perspectives – to shed light on a central aspect of human psychology: the ability to understand. My long-term research vision involves expanding this work, though my own lab and in collaboration with others, to develop an even richer and more interdisciplinary view of understanding. Ideally, this would bridge all the way from basic neural and cognitive processes to the role of understanding in the most valued human experiences, such as a sense of meaning, and a sense connection to other people and to the world around us.