Mind Misreading

Shannon Spaulding

Invited contribution to Philosophical Issues, annual supplement to Nous

Abstract:
Most people think of themselves as pretty good at understanding others' beliefs, desires, emotions, and intentions. Accurate mindreading is an impressive cognitive feat, and for this reason the philosophical literature on mindreading has focused exclusively on explaining such successes. However, as it turns out, we regularly make mindreading mistakes. Understanding when and how mind *misreading* occurs is crucial for a complete account of mindreading. In this paper, I examine the conditions under which mind misreading occurs. I argue that these patterns of mind misreading shed light on the limits of mindreading, reveal new perspectives on how mindreading works, and have implications for social epistemology.

1. Introduction

In common parlance, mindreading is the telepathic ability to see into another person's mind and discern what they are thinking. Telepathic mindreading is exotic and intriguing – at least judging by the number of movies, TV shows, and novels about it – and not real. In philosophy and the cognitive sciences, there is another sense of mindreading that is less exotic but just as interesting. This kind of mindreading is the ability to attribute mental states to agents in order to interpret their behavior and anticipate what they will do next. It is a central, ubiquitous feature of our social lives. When we are driving on a busy freeway, taking care of our children, teaching, fielding questions at a talk, we attribute beliefs, desires, emotions, and other mental states to people in order to make sense of their behavior and interact successfully with them.

The social world is incredibly complex. Our unique experiences, physiological, behavioral, and psychological factors influence our mental states. Moreover, our mental states are dynamically related to others' mental states. What we think, feel, and intend depends on what others are thinking, feeling, or intending. Figuring out why another person behaved as she did and anticipating what she will do next involves grasping, at some level, how all of these factors influence her mental states.

The dynamics of real-world social interactions are so complex that it is amazing that we accurately mindread at all. Mindreading *seems* to come very easily to most of us. We often have little difficulty understanding others' mental states. We generally can tell what other drivers are trying to do and why, what our child wants and why, when our students are bored or interested, and whether the audience member understands our answers. Given how impressive this feat is, it is natural to frame the study of mindreading around the cognitive processes that make successful mindreading possible. Indeed, this has been focus of the mindreading literature since its inception in the late 1970s.

My focus here, however, will be on mindreading failures, i.e., mind *mis*reading. Most people think of themselves as pretty good at understanding others' beliefs, desires, emotions, and intentions. However, social psychologists have discovered that we are significantly worse at mindreading than we think we are (Ames & Kammrath, 2004; Epley, 2008; Hall, Andrzejewski, & Yopchick, 2009; Realo et al., 2003). We consistently and substantially overrate our ability to accurately judge others' mental states and interpret social interactions. This may be due to a lack of interest in correcting our mindreading mistakes, a lack of feedback on errors,

or an extreme instance of the Dunning-Kruger effect.[1] Whatever the cause, the consensus from the empirical literature is that mind misreading is *very* common.

The philosophical literature on mindreading does not study mind misreading in any systematic way.[2] This is unfortunate because there is a high theoretical payoff for examining our mindreading errors. Specifically, patterns of mind misreading shed light on our various mindreading strategies and the conditions under which we use (and misuse) these strategies. In this way, the investigation of mind misreading reveals the limits of our mindreading abilities, which are not apparent when one focuses solely on successful mindreading. In addition, the examination of mind misreading suggests novel hypotheses about how we understand others that are not evident simply from studying mindreading successes.

In this paper, I explore this divergence between our subjective sense of our mindreading abilities and the objective evaluation of our mindreading abilities. In the next Section, I briefly review the two main accounts of mindreading. In Section 3, I discuss the empirical literature on the varieties of mind misreading. I consider the distinctive errors that arise for accuracy-oriented mindreading and efficiency-oriented mindreading. In Section 4, I discuss the implications of mind misreading. I argue that these patterns of mind misreading indicate specific limits on our

[1] The Dunning-Kruger effect is a cognitive bias wherein poor performers in social and intellectual domains are unaware of their ignorance (Kruger & Dunning, 1999). Their deficiency is invisible to them perhaps because recognizing their deficiency requires the very competency they lack.

[2] This literature extensively discusses mindreading failures in chimpanzees and children in the context of establishing a phylogenetic and ontogenetic timeline for mature mindreading. It also examines the mindreading failures of individuals with autism. However, there is no systematic discussion of neurotypical adults' mindreading errors, which is what I focus on here.

mindreading abilities. Furthermore, I argue that these patterns of mind misreading suggest that self-reflection plays an important factor in mindreading accuracy. In Section 5, I discuss the implications for social epistemology.

2. Theories of Mindreading

Two competing accounts have dominated the mindreading literature: the Theory Theory (TT) and the Simulation Theory (ST). The TT holds that we explain and predict behavior by employing a tacit folk psychological theory about how mental states inform behavior. With our folk psychological theory, we infer from a target's behavior what his or her mental states probably are. From these inferences, plus the psychological principles in the theory connecting mental states to behavior, we predict the target's behavior.

On this view, interpreting a person's behavior and anticipating what they will do next fundamentally is the same as explaining and predicting the position of the electrons in a cloud chamber. In both cases, we rely on a rich body of domain-specific information about the target, which we use to infer causal states, and on the basis of this we make predictions about the behavior of the target. Our theory of mind is tacit and less formalized than our scientific theories, but, it is argued, the ability to understand others is best understood as the application of a theory.

The ST, in contrast, holds that we explain and predict a target's behavior by using our own minds as a simulation of the other person's mind. To explain a target's behavior, we put ourselves in another's shoes, so to speak, and imagine

what our mental states would be and how we would behave if we were that agent in that particular situation. To predict a target's behavior, we take the attributed mental states as input and simulate the target's decision about what to do next.

Simulation theorists reject the idea that mindreading consists in theorizing. According to ST, we do not require a large body of folk psychological information about how mental states inform behavior in order to mindread. On this view, all we need is the ability to imagine oneself in a different situation, figure out what one would think, feel, and do in that situation, and attribute those imagined mental states to another person. This simply requires us to use our ordinary cognitive mechanisms in an offline way for the purpose of mindreading. Thus, the ST is regarded as an information-poor theory, whereas the TT is regarded as an information-rich theory.

In addition to what we might call *pure TT* and *pure ST* are hybrid accounts that incorporate elements of TT and ST. These hybrid accounts aim to capture the theoretical advantages of ST and TT while avoiding the problems with both theories. Shaun Nichols and Stephen Stich (2003) have developed a TT-centric hybrid account, and Alvin Goldman (2006) has developed a ST-centric hybrid account. These two innovative accounts have served as pillars for the mindreading literature.

3. Mind Misreading

Studying both successful and unsuccessful processes is a common methodology in philosophy and the cognitive sciences. Consider, for example, the study of vision,

memory, and self-knowledge. In each of these cases, researchers study how the capacity works *and* how it breaks down. To learn how vision works, we study veridical perception but also misperception, visual hallucinations, and visual illusions. Memory researchers study how and when we have accurate memories, but this is paired with investigation of false memories, misremembering, and amnesia. Similarly, research on self-knowledge covers successful introspection, confabulation, and self-deception. These three cases are representative of the study of cognition in general. The underlying rationale is that to understand a process, you must understand when and how it fails.

The debate between the TT, the ST, and various hybrid accounts primarily focuses on explaining successful mindreading. Though it is important to study successful mindreading, for several reasons this discussion should be paired with an examination of mind misreading. First, as noted in the introduction, we are not nearly as good at mindreading as we think we are. Focusing purely on successful mindreading presents a misleading picture of our actual abilities. Second, patterns of mind misreading reveal the limits of our mindreading abilities that are not apparent when focusing solely on successful mindreading. Third, an examination of mind misreading suggests a novel perspective on what it takes to mindread successfully.

In this Section, I shall examine several prevalent but underexplored errors that arise for two types of mindreading. In the philosophical literature on mindreading, many theories tacitly assume that the primary aim of mindreading is accuracy. That is, when we attribute mental states to others in order to interpret

and anticipate their behavior, the most important goal is to attribute the correct mental states. Although this certainly is true in some cases, accuracy is not always the *primary* concern in mindreading. Sometimes we are not motivated to or simply cannot engage in a thorough deliberation about a target's mental states, and in these cases efficiency trumps accuracy. When efficiency is the primary goal in mindreading, we use various mindreading heuristics, which are cognitively less demanding and reliable when used appropriately.

In discussing the varieties of mind misreading, I shall distinguish between the errors that arise for accuracy-oriented mindreading and the errors that arise for efficiency-oriented mindreading. This is not a hard and fast distinction. Some processes will not fit cleanly into the accuracy-seeking or efficiency-seeking categories. Some efficient strategies may play a role in deliberative mindreading, and deliberative processes may influence efficient strategies. Despite these complications, the distinction between efficiency-oriented and accuracy-oriented mindreading is helpful in this context, and I will use it to illustrate the types of mind misreading.[3]

3.1 Mind Misreading: Aiming for Accuracy

[3] I distinguish between deliberative and efficient mindreading processes, but I remain neutral on the kind of cognitive system that underlies these processes. It could be that there are two separate systems – system 1 and system 2 – that realize each type of process. Alternatively, there may be one system that realizes all mindreading processes but is modulated by executive function or some other factor. There may be other options, as well. My arguments are neutral with respect to these different hypotheses.

In some social interactions, our primary aim is accurate mindreading. This usually occurs when something important depends on getting it right, when it matters to us personally, when we will be held responsible for our interpretation of the interaction, or when the situation is unusual or unexpected (Fiske & Neuberg, 1990; Kelley, 1973; Tetlock, 1992). When our primary aim is accuracy, we tend to search for relevant information in a controlled and deliberative fashion. Consider, for example, what it is like to go on a first date. You are trying to figure out whether the person is interested in you romantically, shares your beliefs and values, has a good personality, will not cheat, wants to be in a long-term relationship, etc. The stakes are relatively high for you; you do not want to invest time, energy, and emotion in someone who will turn out to be a poor match for you. Thus, you will want to consider all the relevant evidence and make sure your judgments are not based on merely superficial cues.

When we aim for accurate mindreading, errors can occur under three conditions: when we are under cognitive load and thus cannot engage in a thorough search for information; when we apply an inappropriate model to the situation; when our information search is skewed by other motivations.

Table 1: Types of mindreading errors for accuracy-oriented mindreading.

Accuracy-oriented Mindreading	**Types of Mind Misreading**		
Deliberation	Cognitive load interferes with information search	Apply the wrong model	Self-interest biases information search

The first sort of error that arises for deliberative mindreading occurs when one lacks the cognitive resources to engage in a thorough, objective information search. Deliberative mindreading is effortful and cognitively taxing, and it is difficult if one is under cognitive load or not well practiced in this kind of reflective reasoning (Gilbert, Krull, & Pelham, 1988). In such cases, the result is that our social inferences are biased toward the most readily accessible information, which may lead to error.

Consider again the first date example. Suppose that the evening you go on the date, you are tired, stressed about work, and distracted during the date. A thorough, objective deliberation about your date is doubly difficult for you: Not only must you try to make a good impression on your date by being personable and witty, you also must listen to and interpret what your date is telling you, figure out what food choice, clothes, questions and answers tell you about your date's mindset, and you must do all this while physiologically and cognitively taxed. A careful, deliberative information search requires going beyond just the salient cues. However, you are too cognitively taxed to do this with much care, and as a result your deliberation is guided by superficial but potentially misleading cues.

So, what are the salient cues in social interactions? For all of us, the most salient features of a person tend to be their age, race, and gender (Ito, Thompson, & Cacioppo, 2004; Liu, Harris, & Kanwisher, 2002). We rapidly sort people by age, race, and gender and other social categories, depending on the content. On the basis of this categorization, we spontaneously attribute personality traits such as

trustworthiness, competence, aggressiveness, and dominance (Olivola & Todorov, 2010; Rule, Ambady, & Adams Jr, 2009). Although the speed of spontaneous trait inferences is a matter of dispute, it occurs very rapidly: between 100 milliseconds and 1400 to 1600 milliseconds, even when we are under cognitive load (Malle & Holbrook, 2012; Todorov & Uleman, 2003).

In addition, we spontaneously and implicitly associate these social categories with specific characteristics. For example, we associate old and incompetent, female and warm, baby-face and unthreatening. These associations are the sort of thing tested by the Implicit Association Task, which measure the strength of a person's implicit associations (Greenwald, McGhee, & Schwartz, 1998; Greenwald, Poehlman, Uhlmann, & Banaji, 2009).

Putting all of this together, in ordinary social interactions the most accessible information about another person tends to be an individual's social category, spontaneously inferred personality traits, and implicit associations. It is possible in deliberation to override the implicit associations and spontaneously inferred traits if one is motivated and has the cognitive resources to do so. However, if one is busy, stressed, and tired, overriding these inferences and associations is extremely difficult, and they may bias one's deliberation. Thus, it is difficult for you on your hypothetical first date to deliberate objectively about whether your date is committed to being in a serious relationship, shares your values, is loyal, etc. Your deliberation is influenced by implicit associations and trait inferences, which under ideal circumstances you would reflect on and possibly reject. However, because you are under cognitive load you lack the ability to override these salient features in

favor of less salient but potentially more accurate features. Thus, errors arise for deliberative mindreading when we are cognitively taxed and cannot deliberate carefully.

The second kind of error that may occur when we are aiming for accurate mindreading concerns the framework we employ to make sense of a social interaction. Even in good deliberative mindreading, we do not consider *all* of the available information. That would be impossible because there is far too much information for human beings to process. Instead, we search for the most relevant information and base mindreading judgments on that information.[4] The situational context and one's past experiences determine what is taken to be relevant information. They shape expectations in social interactions, and they make certain interpretations more accessible to us, i.e., our attention is primed for these interpretations (Wittenbrink, Judd, & Park, 2001).

Consider the following simple example. Having spent much of my life on university campuses, I generally know what to expect when I visit a university campus, even one that is unfamiliar to me. I understand the general institutional structure, social roles, and typical behavior of administrators, faculty, and students. I have a model that guides my interpretation and expectations of what happens on campuses. Someone who has never attended a university and has no experience with life on a university campus may not have the same interpretations and expectations as I do. They will use a different, less appropriate framework to

[4] Some errors occur because the information we attend to includes statistical outliers or our information sample is small and/or biased. These statistical errors are common to every type of reasoning, so I will not devote special attention to them in discussing mind misreading.

understand and anticipate behavior they encounter on a university campus and thus are likely to misunderstand some of the idiosyncratic behaviors on university campuses.

The theory-ladenness of social observation is key to the second type of error in deliberative mindreading. We are likely to attend to irrelevant or misleading information when the framework that guides our information search is faulty in some respect. If, for example, the framework does not apply to the situation, or if the framework itself inaccurate, then we are likely to misinterpret others' behavior. Consider again my model for university campuses. It is useful and appropriate for most American and European universities, but despite some superficial similarities it is not appropriate for contemporary technology campuses like Googleplex or Microsoft Campus. If I apply my university model to Googleplex, I am likely to misunderstand the institutional and social dynamics, and I am likely to misunderstand the behavior and motivations of people in this environment.

In general, applying an inappropriate or faulty model to a situation can lead us to misinterpret social interactions, which paves the way for mind misreading. This is especially likely to happen when we are under cognitive load because we may fail to notice that our model does not fit the situation. This kind of error also is likely to occur when we are overly confident in our social interpretation, which is common when a situation seems very familiar to us. In such cases, because we are confident we understand the social dynamics, we do not reflect on our interpretation or consider the possibility that we are employing an inappropriate or faulty model of the social interaction.

A third sort of error in accuracy-oriented mindreading arises when the mindreading process is skewed by self-interest. In many social interactions, our social interpretations are shaped by the need for anxiety reduction, self-esteem preservation, and confirmation of one's worldview. In these cases, our mindreading inferences, in one way or another, serve self-interested purposes (Dunning, 1999; Kunda, 1990). These motivations lead to several specific mindreading errors.

Consider first the Self-Serving Attributional Bias, which describes our tendency to take credit for success and deny responsibility for failure. (Miller & Ross, 1975). We often attribute our successes to some internal factor, e.g., diligence or talent, and attribute our failures to external mitigating factors, e.g., bad luck or bias. In this way, we come to feel good about our successes and brush off our failures.

This pattern is found for judging in-group and out-group behaviors, as well. This is called the Group-Serving Attributional Bias (Brewer & Brown, 1998; Pettigrew, 1979). One tends to judge the success of an out-group to be the result of external, mitigating situational factors and the failure of an out-group as the result of internal factors, whereas one judges the success of one's in-group to be the result of internal factors and the failure of one's in-group to be the result of situational factors. One sees this pattern of reasoning very clearly in sports fans. When the Badgers win it is because they are talented and hard working, but when the Badgers lose it is because they were off their game that day, the other team got lucky a few times, and the referees were biased against the Badgers.

The Self- and Group-Serving Attributional Biases tend to occur in a context of threat or competition. In such contexts, we employ different types of explanations

depending on whose behavior we are explaining. Whether we cite situational factors or mental states depends on our perceived similarity to the target, not whether situational factors or mental states actually caused the behavior. Thus, these biases distort our judgments about our own and others' behavior.

Naïve Realism is another sort of mind misreading generated by self-interest. It describes the tendency to regard others as more susceptible to bias and misperception than oneself (Pronin, Lin, & Ross, 2002). We think that we simply see things as they are but others suffer from bias. This tendency is prevalent in interactions in which people disagree. For example, one regards those of a different political party as misguided and biased by their personal motivations, whereas one regards oneself (and to some extent other members of one's political party) simply as correct. We assume that we simply see things as they really are. Naïve Realism influences the mental states we attribute to ourselves and to others. This bias is entrenched in our reasoning, but it is especially common when we are overly confident. In those cases, we fail to consider seriously the idea that we are the ones who are biased and misperceiving.

Finally, confirmation bias describes a general tendency to seek only information that confirms one's preconceived ideas and interpret ambiguous information in light of these preconceived ideas. With respect to social cognition, we have preconceived ideas about other individuals and groups, and we tend to interpret social interactions in terms of those preconceived ideas. For example, racists notice when individuals behave in ways that confirm their racist beliefs but they often do not attend to the many cases where individuals act in ways that

disconfirm their racist beliefs. Confirmation bias occurs regardless of how the preconceived idea originated, how likely it is to be true, and whether accuracy is incentivized (Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992; Snyder, Campbell, & Preston, 1982).

3.2 Mind Misreading: Aiming for Efficiency

Section 3.1 explains three ways in which thoughtful deliberation about others' mental states can go awry. This Section explains the types of errors in efficient mindreading. Although sometimes our primary aim is accurate mindreading, this is not always the case. Often there are constraints on our motivation, time, and attention that prohibit even attempting to engage in a thorough search for information. In such cases, accuracy is a secondary aim and efficiency is the primary aim. When the social interaction seems ordinary and familiar, when not much hangs on it, or when we are otherwise cognitively taxed, we use cognitive shortcuts.

When our primary goal is efficient mindreading, several strategies are available. The strategies we use depend on whether or not the individual we are mindreading is part of our in-group. We identify people as part of our in-group or part of an out-group on the basis of perceived similarity (Ames, 2004a, 2004b; Ames, Weber, & Zou, 2012). That is, those who we perceive to be like us are categorized as part of in our in-group, and those who we perceive to be unlike us are categorized as part of an out-group. One tends to identify people who share one's age, race, gender, religion, or nationality as part of one's in-group. However, because people have

multiple, overlapping identities, and perceived similarity is relative to a context, social categorization extends beyond these basic classifications. Thus, I may consider someone as part of my in-group in one context but not in another.[5]

First consider the heuristics we use when we perceive an individual to be similar to ourselves in some salient respect. In these cases, we often simply project our own mental states to that individual (Ames, 2004a, 2004b; Ames et al., 2012). This is an efficient strategy because we do not have to deliberate about the target's situation and likely mental states. Rather, we simply infer that the target believes, desires, or feels about some event the way we do. For example, in many contexts I consider philosophers as my in-group. I have learned that philosophers tend to have similar social and political views. If I learn that Sally is a philosopher, I assume that she shares many characteristics in common with me, including political opinions. In such a case, I simply project my own political judgments on her without any deliberation.

Sometimes we also use our mental states as an anchor and adjust the interpretation based on how similar the individual is to us. For example, if I learn that Sally specializes in social and political philosophy, I may think that she probably has more nuanced views on politics than I do and adjust my attributions accordingly. Projection and anchoring-plus-adjustment are egocentric heuristics. If our perceptions of similarity are correct, and if we accurately introspect our own mental states, these egocentric heuristics are useful and accurate. Errors occur when these two conditions are not satisfied.

[5] Importantly, perceived similarity is a subjective and sometimes idiosyncratic judgment, not an objective measure of actual similarity (Ames et al., 2012).

Errors arise when we *overestimate* the similarity between ourselves and the other person(s) and thus engage in more projection than is warranted. The resulting errors are called the False Consensus Effect and the Curse of Knowledge (Clement & Krueger, 2002; Epley & Waytz, 2010, p. 512). The False Consensus Effect occurs when we falsely assume that a group of people shares our perspective on some issue. The Curse of Knowledge is a related phenomenon in which we falsely assume that another individual knows what we know. For both kinds of mind misreading, we inappropriately project our own mental states onto others because we assume that we are more similar than we in fact are. The specific details on how this happens will differ from case to case. In general, inappropriate projection occurs when we attend to superficial similarities between others and ourselves and fail to notice or appreciate dissimilarities, e.g., in terms of situational context, personal background, knowledge, attitudes, and emotions.

A second kind of error for egocentric heuristics occurs when we correctly diagnose the similarity between the mindreading target and ourselves but inaccurately introspect our own mental states. In such a case, projecting our own mental states onto a target is warranted because we are similar to the target in the relevant respect, but we fail to understand our own beliefs, desires, motivations, and feelings and thus attribute the wrong mental states. Consider, for example, a self-unaware racist who thinks of himself as "color blind" but in fact harbors many racist attitudes. In mindreading a similar person, the mindreader correctly judges that the other person is similar and thus projects his own attitudes to that person. In this case, he attributes to the other person the belief that all races are equal. However, he

makes a mindreading error because neither he nor the similar other actually have racial egalitarian attitudes. If he had introspected correctly, he would have recognized his White Supremacist attitudes and projected those to the similar other person. This kind of error is likely to occur when we are less self-reflective and thus do not understand our own mental states.

The previous kinds of efficiency-oriented mindreading are based on egocentric heuristics, which we employ when we perceive an individual to be similar to us. When we perceive an individual to be different from us, we use alternative efficient strategies, namely, stereotypes about the individual's salient in-group (Ames, 2004a; Ames et al., 2012; Krueger, 1998; Vorauer, Hunter, Main, & Roy, 2000). Stereotypes may be positive, negative, or neutral beliefs about some group.

Table 2: Types mindreading errors for efficiency-oriented mindreading.

Efficiency-oriented Mindreading	**Types of Mind Misreading**		
Projection	Overestimate similarity; inappropriately project one's mental states	Overestimate similarity; insufficiently adjust projection	Correctly judge similarity; incorrectly introspect one's mental states
Stereotyping	Underestimate similarity; baselessly apply stereotype	Employ false stereotype	Employ misleading, unrepresentative stereotype

Stereotypes are reliable heuristics for understanding others' behavior when they are applied appropriately and the stereotypes are accurate and representative. We may make mistakes when either of these two conditions is not satisfied. When

we *underestimate* the similarity between ourselves and the other person, we baselessly apply stereotypes where projection or deliberative mindreading would be more appropriate.

Contrary to what one might expect, we are likely to make this type of error when we are in familiar situations. When we are in unusual or unfamiliar situations, we tend to deliberate about the interaction more than we do in normal and familiar cases. As particular situations become familiar to us, certain interpretations of those situations will become more accessible, more routinized, and increasingly difficult to override (Higgins, King, & Mavin, 1982). Thus, in very familiar situations we may fail to notice or appreciate stereotype-inconsistent behavior and thus inappropriately apply stereotypes.

A second type of error for stereotype-based mindreading occurs when we correctly diagnose the dissimilarity between ourselves and the other person but the stereotypes we employ are false or unrepresentative of the out-group. This pattern is evident in racist individuals' mindreading practices. The White Supremacist, for example, is inclined to use racist stereotypes to infer the motivations and perspectives of members of different racial categories. False or unrepresentative stereotypes have many sources, including explicit and implicit bias, idiosyncratic experiences with a group, poor statistical reasoning, and simply false beliefs about the group. However they arise, employing false or misleading stereotypes is likely to generate mistakes in interpreting others' mental states and behavior.

In summary, sometimes we have the motivation and ability to exhaustively review the available social information and attribute mental states to others in that

way, whereas other times we take shortcuts because we lack the motivation or ability to do an exhaustive search. In the former case, mind misreading occurs when cognitive load interferes with the information search, we apply the wrong framework to the situation, or when self-interest skews our deliberation. In the latter case, mind misreading arises when we misdiagnose the similarity or dissimilarity between ourselves and the target, fail to understand our own mental states, or apply false or inappropriate stereotypes.

4. The Limits of Mindreading

The philosophical literature on mindreading primarily focuses on successful mindreading. Although explaining how we manage the complex task of accurately attributing mental states to others is interesting and important for understanding social cognition, focusing exclusively on successful mindreading obscures the limits of our mindreading abilities. With the distinction between accuracy-oriented and efficiency-oriented mindreading, we can see that conditions for success differ for each type of mindreading. Deliberative and efficient mindreading go awry in distinctive ways. See Table 3 below.

Table 3: Types of mindreading errors for accuracy-oriented and efficiency-oriented mindreading.

Mindreading Aim	**Types of Mind Misreading**		
Accuracy	Cognitive load interferes with information search	Apply the wrong model	Self-interest biases information search
Efficiency	Misdiagnose	Correctly diagnose	Employ baseless,

	similarity between oneself and other	similarity, but incorrectly introspect one's mental states	false, or unrepresentative stereotype

Errors in mindreading reveal the limits of mindreading abilities in a way that is not possible when we focus solely on successful mindreading. To illustrate, compare what the data here indicate with respect to the processes posited by TT and ST. The evidence suggests that we successfully use the deliberative processes posited by TT and ST only when we have the motivation, time, and cognitive capacity to engage in a thorough, deliberative search for information. When we attempt to engage in such searches when we lack the cognitive capacity or have self-interested biases that skew our information search, we are likely to make mistakes. These errors, which are discussed in 3.1, are not predicted by TT or ST.

The ST predicts the use of egocentric heuristics, namely, projection and anchoring and adjustment. These efficient strategies are employed successfully only when we correctly diagnose the relevant similarity between the target and ourselves and we understand our own mental states. We are likely to err when these conditions do not hold. The other efficient strategy – stereotyping – is not predicted by either TT or ST, though it is compatible with the TT if the stereotypes are part of the theory. This efficient strategy is successful only when we correctly diagnose the relevant dissimilarity between the target and ourselves and the stereotype employed is accurate and representative of the target's relevant in-group. Stereotyping is inaccurate when it fails to meet these conditions.

Typically, mind misreading is more likely to occur when the situation is ambiguous, which social interactions often are especially when they involve people outside one's close circle of family and friends. In addition, several general psychological factors may lead to mind misreading, e.g., memory failure, psychosocial disorder, or low intelligence. The errors I discuss above arise specifically when (1) we are too cognitively taxed to engage in thorough information search, (2) we pay attention to superficial cues, (3) we are biased by self-interest, (4) we fail to understand our own mental states, (5) and we inappropriately deploy stereotypes.

Investigating these limits of our mindreading abilities paves the way for different perspectives on mindreading. An interesting upshot of this discussion is that we are likely to make mindreading errors when we are not self-aware or self-reflective. Self-awareness is a psychological state in which one takes oneself as the subject, specifically, one's traits, mental states, feelings, and behavior. Being self-aware involves reflecting on mental, physical, behavioral, and relational facts about oneself. One might have thought that the limitations on mindreading would have to do with others' behavior and mental states, i.e., that we would be unable to make sense of some behaviors in some contexts. Though that certainly happens, this investigation suggests that the more immediate limitations on mindreading are internal to the mindreader.

The idea suggested by examination of the limits of mindreading is that self-awareness predicts mindreading success. *Ceteris paribus*, an individual who is less self-aware will make more mindreading mistakes than an individual who is more

self-aware.[6] In circumstances where individuals are less self-aware, they are more likely to make mindreading errors. For deliberative, accuracy-oriented mindreading, individuals who are less self-aware are less likely to notice that they are under cognitive load, that they are being overly confident, that despite trying to deliberate carefully they are paying attention to merely superficial cues, and they likely will not notice how their own motivations skew the information search. Individuals who are less self-aware are likely to make mistakes in efficient mindreading, as well. They are less likely to consider how much or little they resemble another person, appropriately adjust their projections of their own mental states, correctly introspect their own mental states, and examine their stereotypes.

The central lesson here is that examining mind misreading sheds light on the limits of our mindreading abilities and suggests new perspectives on how mindreading works. Studying the ways in which we err in mindreading will give us a better picture of how we understand – and sometimes misunderstand – other people. I proposed a hypothesis about the role of self-awareness in mindreading. This hypothesis is not end of the debate. In fact, it is just the start. Investigating mind misreading opens up a host of new debates, which promise to advance our understanding of mindreading.

---

[6] Self-awareness does not *uniquely* predict mindreading success. Executive function will play an extremely important role in self-awareness insofar as it regulates attention, inhibitory control, and working memory. Moreover, higher intelligence and healthy psychological functioning (e.g., conscientiousness, tolerance, openness) are positively related to accurate mindreading (Hall et al., 2009). And certainly one's relation to the target and motivation to understand the target's mental states play a crucial role in the accuracy of one's mindreading judgments.

5. Implications of Mind Misreading

The discussion so far clearly is relevant to the field of social cognition, but it also has implications for social epistemology. In particular, mind misreading bears on how we judge whether others are our epistemic peers. You and I are epistemic peers with respect to some topic to the extent that we are comparably knowledgeable and competent to reason about that topic. That is, we possess the same evidence about X and are equally intelligent, free from bias, competent at perceiving, reasoning, etc. (Kelly, 2010).

The notion of epistemic peer arises in the epistemology of peer disagreement debate. Proponents of the conciliation view argue that when you disagree with someone you take to be an epistemic peer you should reduce your confidence in your judgment (Christensen, 2007), whereas proponents of the steadfast view argue that in such a case you should remain steadfast in your view (Kelly, 2010). The notion of epistemic peer comes up in the discussion of epistemic injustice, as well. Epistemic injustice, in particular testimonial epistemic injustice, occurs when a hearer's prejudices result in downgrading a speaker's credibility (Fricker, 2007). That is, in virtue of epistemically irrelevant facts about the speaker the hearer downgrades the speaker's epistemic status. Central to both philosophical debates is the issue of how we judge others' knowledge, intelligence, reasoning abilities, bias, etc. Our discussion of mind misreading sheds light on this issue.

In explaining the ways mindreading fails, I described several very common self-enhancing biases: the Self-Serving and Group-Serving Attributional Bias, which

result in overestimating our own competence and underestimating the competence of others (especially out-group members); the Dunning-Krueger Effect, wherein individuals who are not knowledgeable or competent with respect to some issue egregiously overestimate their own knowledge and competence and fail to recognize others' equal or superior knowledge and competence; and Naïve Realism, which describes the tendency to regard others as more susceptible to bias and misperception than oneself especially in the context of disagreement. These three self-enhancing biases influence how we judge our own knowledge and competence in relation to others.

In addition to the self-enhancing biases, I also discussed several biases in assessing others' knowledge and competence. Social categorization and implicit associations with social categories influence how we decide who is an epistemic peer. Simply in virtue of being part of particular social category we may upgrade or downgrade a person's knowledge or competence. For example, we tend to associate spontaneously and implicitly elderly women with warmth and incompetence. We habitually downgrade the epistemic status of an elderly woman just in virtue of her social category. Of course, we can override implicit associations, but doing so requires awareness of the associations and their effect on one's behavior, attention, and cognitive effort. For this reason implicit associations are difficult to excise from one's judgments.

Furthermore, in-group/out-group status significantly affects our judgments of other people's epistemic status. We usually have more favorable attitudes toward and empathize more with in-group members, especially people who share our

gender, race, age, religion, or nationality than toward people do not share these features. The data suggest that we are less likely to regard out-group members as epistemic peers, i.e., as being equally knowledgeable and competent. We tend to simplify and caricature the mental states of those who we perceive to be unlike us. Although we sometimes have positive stereotypes about out-groups – e.g., an American stereotype about Asians is that they are hardworking and smart – mostly we upgrade the status of our in-group and downgrade the status of out-groups. This tendency is especially strong in a context of threat, e.g., when people disagree about some important issue.

Combining self-enhancing biases with data on other-downgrading biases yields a bleak picture of how we judge others' knowledge and competence. It seems that we are most likely to regard another person as an epistemic peer when in fact she is an epistemic superior and she is part of our relevant in-group. In most other conditions, other things being equal, we are likely to regard an epistemic peer as inferior, and we are likely to regard moderately epistemically superior out-group members as inferior.

These data suggest that often we are not reliable judges of our epistemic peers. In particular, we tend to overestimate our own knowledge and competence and underestimate others', especially others who are part of an out-group. This discussion of mind misreading has implications for the epistemology of peer disagreement. In light of these facts, when we take ourselves to be in a disagreement with an epistemic peer we ought to conciliate. That is, we ought to decrease confidence in our own judgments when we disagree with someone we regard as an

epistemic peer because it is likely that that person in fact is an epistemic superior.[7] Indeed, when we take ourselves to be disagreeing with an out-group member whom we regard as moderately epistemically inferior, we should conciliate then as well because our judgments about the out-group member are likely to be even more skewed in that case.

The discussion of mind misreading in judging epistemic peers is relevant to epistemic injustice, as well. The downgrading of epistemic peers and out-group epistemic superiors just described is an instance of epistemic injustice, i.e., of a hearer's prejudices discounting a speaker's credibility. Above I discussed the conditions for successful mindreading and the various ways in which we fail to understand others when these conditions are not met. These data reveal when and how epistemic injustice is likely to arise. Social categorization, implicit bias, and in-grouping/out-grouping behaviors are particularly important for understanding when are likely to be biased in assessing others' epistemic status. Understanding when and how epistemic injustice arises is an important step in mitigating its effects.

Mind misreading so far has been an under-explored topic in philosophy, which is unfortunate because it is an interesting and important topic. Mind misreading is crucial to the study of social cognition, and it has implications beyond philosophy of mind and cognitive science. In particular, it is relevant to the epistemology of peer disagreement debate and epistemic injustice. I hope this paper

[7] The case of disagreeing experts may be more nuanced than the case of disagreeing non-experts. If one is an expert in some domain, one may be better at identifying factors that distort one's own judgments in that domain, have a more realistic assessment of one's knowledge and competence in that domain, and be better able to identify others' expertise. Even experts are not immune to many of the cognitive biases discussed in Section 3, but they may be better able to mitigate their effects.

shows just how much there is to be gained in philosophy of mind and epistemology from a systematic evaluation of the ways in which we understand and often fail to understand other people.[8]

References

Ames, D. R. (2004a). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. *Journal of Personality and Social Psychology, 87*(3), 340.

Ames, D. R. (2004b). Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573.

Ames, D. R., & Kammrath, L. K. (2004). Mind-reading and metacognition: Narcissism, not actual competence, predicts self-estimated ability. *Journal of Nonverbal Behavior, 28*(3), 187-209.

Ames, D. R., Weber, E. U., & Zou, X. (2012). Mind-reading in strategic interaction: The impact of perceived similarity on projection and stereotyping. *Organizational Behavior and Human Decision Processes, 117*(1), 96-110.

Brewer, M. B., & Brown, R. J. (1998). *Intergroup relations*: McGraw-Hill.

Christensen, D. (2007). Epistemology of disagreement: The good news. *The Philosophical Review*, 187-217.

Clement, R. W., & Krueger, J. (2002). Social categorization moderates social projection. *Journal of Experimental Social Psychology, 38*(3), 219-231.

Dunning, D. (1999). A newer look: Motivated social cognition and the schematic representation of social concepts. *Psychological Inquiry, 10*(1), 1-11.

Epley, N. (2008). Solving the (real) other minds problem. *Social and personality psychology compass, 2*(3), 1455-1474.

[8] I am grateful to the many people who have commented on these ideas at various conferences and colloquia. I am especially indebted to Suilin Lavelle, Guillermo Del Pinal, Robert Thompson, Evan Westra for their insightful comments and useful discussions about these ideas.

Epley, N., & Waytz, A. (2010). Mind perception. In S. T. Fiske, D. T. Gilbert, & G. Lindzey (Eds.), *Handbook of Social Psychology* (5th ed., Vol. 1, pp. 498-451). Hoboken, NJ: Wiley.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. *Advances in experimental social psychology, 23*, 1-74.

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*: Oxford University Press Oxford.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology, 55*(5), 685.

Goldman, A. I. (2006). *Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading*: Oxford University Press, USA.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17.

Hall, J. A., Andrzejewski, S. A., & Yopchick, J. E. (2009). Psychosocial correlates of interpersonal sensitivity: A meta-analysis. *Journal of Nonverbal Behavior, 33*(3), 149-180.

Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology, 43*(1), 35.

Ito, T. A., Thompson, E., & Cacioppo, J. T. (2004). Tracking the Timecourse of Social Perception: The Effects of Racial Cues on Event-Related Brain Potentials. *Personality and Social Psychology Bulletin*.

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist, 28*(2), 107.

Kelly, T. (2010). Peer disagreement and higher order evidence. In A. I. Goldman & D. Whitcomb (Eds.), *Social Epistemology: Essential Readings* (pp. 183--217): Oxford University Press.

Krueger, J. (1998). On the perception of social consensus. *Advances in experimental social psychology, 30*, 164-240.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480.

Liu, J., Harris, A., & Kanwisher, N. (2002). Stages of processing in face perception: an MEG study. *Nature Neuroscience, 5*(9), 910-916.

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology, 102*(4), 661.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*(2), 213.

Nichols, S., & Stich, S. P. (2003). *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding Other Minds*. Oxford: Oxford University Press.

Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology, 46*(2), 315-324.

Pettigrew, T. F. (1979). The ultimate attribution error: Extending Allport's cognitive analysis of prejudice. *Personality and Social Psychology Bulletin, 5*(4), 461-476.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin, 28*(3), 369-381.

Realo, A., Allik, J., Nõlvak, A., Valk, R., Ruus, T., Schmidt, M., & Eilola, T. (2003). Mind-reading ability: Beliefs and performance. *Journal of Research in Personality, 37*(5), 420-445.

Rule, N. O., Ambady, N., & Adams Jr, R. B. (2009). Personality in perspective: Judgmental consistency across orientations of the face. *Perception, 38*, 1688-1699.

Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology, 22*(2), 93-121.

Slowiaczek, L., Klayman, J., Sherman, S., & Skov, R. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition, 20*(4), 392-405.

Snyder, M., Campbell, B. H., & Preston, E. (1982). Testing hypotheses about human nature: Assessing the accuracy of social stereotypes. *Social cognition, 1*(3), 256-272.

Tetlock, P. E. (1992). The impact of accountability on judgment and choice: Toward a social contingency model. *Advances in experimental social psychology, 25*, 331-376.

Todorov, A., & Uleman, J. S. (2003). The efficiency of binding spontaneous trait inferences to actors' faces. *Journal of Experimental Social Psychology, 39*(6), 549-562.

Vorauer, J. D., Hunter, A., Main, K. J., & Roy, S. A. (2000). Meta-stereotype activation: evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *Journal of Personality and Social Psychology, 78*(4), 690.

Wittenbrink, B., Judd, C. M., & Park, B. (2001). Spontaneous prejudice in context: variability in automatically activated attitudes. *Journal of Personality and Social Psychology, 81*(5), 815.