**When is mindreading accurate? A commentary on Shannon Spaulding’s *How We Understand Others: Philosophy and Social Cognition***

**Abstract:** In *How We Understand Others: Philosophy and Social Cognition*, Shannon Spaulding develops a novel account of mindreading with pessimistic implications for mindreading accuracy: according to Spaulding, mistakes in mentalizing are much more common than traditional theories of mindreading commonly assume. In this commentary, I push against Spaulding’s pessimism from two directions. First, I argue that a number of the heuristic mindreading strategies that Spaulding views as especially error prone might be quite reliable in practice. Second, I argue that current methods for measuring mindreading performance are not well-suited for the task of determining whether our mental-state attributions are generally accurate. I conclude that any claims about the accuracy or inaccuracy of mindreading are currently unjustified.

**Introduction**

*Mindreading* or *theory of mind* is the ability to predict and explain the behavior of other agents in terms of their mental states. A very common view among philosophers and cognitive scientists that mindreading is an essential part of ordinary human social cognition. According to this idea, our basic ability to understand why other people act as they do, to predict what they will do next, and to plan our social interactions, all centrally involve reasoning about people’s beliefs, desires, intentions, and other mental states. Without this ability, the behaviors of others would appear to us as bizarre and unpredictable, and even basic forms of social coordination that we take for granted would become very difficult for us. Indeed, our advanced mindreading capacities are thought to be a key part of the explanation for why human beings have greatly surpassed our great ape cousins in the domains of culture, language, and large-scale cooperation (Baron-Cohen, 1997; Scott-Phillips, 2014; Tomasello, 2014). Mindreading, in other words, is our primary means for gaining the knowledge we need to succeed in the social world.

In *How We Understand Others: Philosophy and Social Cognition*, Shannon Spaulding is ambivalent about this view of mindreading (Spaulding, 2018). On the one hand, she endorses the idea that mindreading is a pervasive and important feature of social cognition – what she dubs the “broad scope of mindreading claim” (Spaulding, 2018, p. 9) – and forcefully rejects attempts by other philosophers to undermine it. In this sense, her book can be read as consistent with the mainstream view in the mindreading literature. But on the other hand, she argues that her fellow mainstream mindreading theorists have too often taken it for granted that mental-state attribution is generally accurate – a habit that she attributes to excessive emphasis on measures like the false-belief task, and the tacit assumption that accurate belief-attribution ought to be the primary *explanadum* of mindreading theories (Spaulding, 2018, p. 43). In contrast, Spaulding maintains, our mental-state attributions regularly misrepresent the real mental states of others; indeed, Spaulding thinks that accuracy is often not even our primary goal when we reason about the minds of others. This pessimism about the accuracy of mindreading is a dominant theme throughout the book, as well as in Spaulding’s earlier work (Spaulding, 2016, 2017). Drawing on a wide range of empirical findings from social psychology (which have received surprisingly little attention from mindreading theorists until now), Spaulding paints a picture of mindreading as a complex, messy process that takes on different forms in different social contexts and is regularly distorted by a variety of unreliable heuristics and biases. This culminates in an argument for a version of the Model Theory of mindreading (Godfrey-Smith, 2005; Maibom, 2003), which takes the heterogeneity and context-sensitivity of mindreading as its primary explanandum, rather than its putative accuracy. Thus, while Spaulding agrees with the received view that mental-state attribution is a central and pervasive part of everyday social cognition, she pushes back against the idea that mindreading is a reliable source of social knowledge.

Spaulding also argues that this pessimism about theory of mind has consequences for debates in mainstream philosophy. For example, she argues that our unreliable mindreading abilities should affect how we deliberate in contexts of apparent peer disagreement. The problem of peer disagreement refers to contexts in which two agents who are epistemic peers find themselves in possession of the same evidence and yet nevertheless come to different conclusions regarding a particular question; each agent is then faced with the question of whether she should remain steadfast in her position (Kelly, 2011), or if she should treat her epistemic peer’s position as evidence that she should conciliate and decrease confidence in her own position (Christensen, 2007). To the extent that the problem of peer disagreement has any bearing upon how we should deliberate in real-world scenarios (and is not a mere idealization), it presupposes that we are usually able to recognize when we are in the company of an epistemic peer. But Spaulding suggests that this assumption is not tenable in light of her arguments about the unreliability of mindreading: not only do we regularly overestimate our own epistemic status, we also display a wide range of biases when it comes to judging the knowledge and competence of others. If this is right, then most perceived cases of peer disagreement are probably no such thing, because we are not normally able to recognize whether another person is an epistemic peer. Because most of the biases relevant to evaluations of potential epistemic peers tend to inflate how we view our own knowledge and competence relative to others, Spaulding thinks that our default policy in contexts of apparent peer disagreement should be to decrease our confidence in our own judgments (Spaulding, 2018, pp. 81–88).

Like Spaulding, I think that the question of accuracy has largely been taken for granted in the mindreading literature, and I agree that she has identified an extremely important issue in the study of theory of mind. However, I am not convinced that the empirical considerations she raises support pessimistic conclusions about overall reliability of mindreading. In the first part of this commentary, I will argue that several of the efficient, heuristic mental-state attribution strategies that Spaulding views as unreliable might actually produce fairly accurate mental-state attributions most of the time. I will go on to suggest that a better way to describe the trade-off that we make when we alternate between deliberate and efficient modes of mindreading is in terms of *precision* rather than accuracy. In the second half of the paper, I will turn to a question that arises from these debates about the reliability of mindreading: how exactly could we *measure* the accuracy of mental-state attribution anyway? After surveying a several different methodologies, I conclude that we don’t currently possess *any* reliable means for determining whether a particular act of mental-state attribution is accurate. A consequence of this conclusion is that any empirical claim about the accuracy of mindreading is probably unjustified. This creates problems for Spaulding’s pessimistic take on mindreading, but also for the optimists about mindreading whom she criticizes.

**Mindreading for accuracy versus mindreading for efficiency**

A central part of Spaulding’s critique of traditional theories of mindreading like the theory-theory and the simulation theory is that they all tacitly assume that the process of mental-state attribution is uniform and consistent across mindreaders, targets, and social contexts. When explaining how mindreading works in the controlled experimental conditions of the false belief task (for example), this is excusable. But in real-world scenarios, Spaulding argues, the mindreading process is influenced by a myriad of factors that affect what information it takes in and how that information is processed. As a result, mental-state attribution can end up functioning very differently in different kinds of social contexts, and different individuals in the same context may well arrive at divergent mental-state attributions. This variability is a key component of Spaulding’s case for pessimism about the accuracy of mindreading: different people in the same situations with the same evidence can nevertheless arrive at vastly divergent psychological interpretations of a particular individual’s actions (Spaulding, 2017, 2018, p. 36).

One kind of factor that can influence the cognitive strategies and resources that we employ while mindreading is our social goals. Sometimes, we are very motivated to make correct mental-state attributions and behavioral predictions about others (say, when interviewing a potential nanny for one’s child). In these cases where mindreading “aims at accuracy,” Spaulding suggests that we tend to engage in conscious, deliberative forms of reasoning that are cognitively taxing and liable to break down under cognitive load (Spaulding, 2018, p. 45). At other times, a person’s actual mental states are not particularly relevant to our plans (e.g. we do not need to know what a cashier at a grocery store is *really* thinking in order to buy eggs and milk). In these cases, we do not engage in effortful forms of reasoning about mental states. Instead, we employ fast and efficient mindreading strategies that require less cognitive effort but are more prone to error. Thus, whether or not we care about getting our mental-state attributions right can dramatically affect how mindreading functions in a given situation.

The particular type of mindreading strategy we employ when we are motivated by efficiency may depend upon how we have socially categorized the mindreading target (Spaulding, 2018, pp. 45–46). If we see the target as similar to us in some respect (e.g. as part of our in-group), we are more likely to employ a mental-state attribution heuristic called *egocentric projection*. Egocentric projection occurs when a mindreader assumes that her target shares her mental states, or otherwise uses her own mental states as an initial anchor from which she can subsequently adjust (Ames, 2004a; Epley, Keysar, Van Boven, & Gilovich, 2004). The downsides of this heuristic are revealed whenever the target is not actually all that similar to the mindreader. In these cases, the mindreader may infer that certain propositions are common knowledge when they are not, or mistakenly come to believe that certain goals are shared. Even when a mindreader does recognize that a target is not entirely similar to her, using her own mental states as an anchor may nevertheless lead her to underestimate the gap between her target’s beliefs and her own.

On the other hand, if a mindreader sees her target as different from herself, she may instead draw on stereotypes about the target’s perceived social category (Ames, 2004b). As Spaulding notes, we tend to automatically categorize people upon first encountering them. These categorizations activate stereotypes associated with a particular group, which often include beliefs about mental states and traits that members of that group are likely to share. For example, categorizing a person as a “mother” might lead to the expectation that she cares more about family than work, while categorizing a someone as a “millennial” might lead to the expectation that she is entitled, self-absorbed, and overly fond of avocado toast. Because these stereotypical beliefs are stored in an efficient format, they become readily accessible whenever we are faced with the task of predicting and interpreting a target’s behavior. This can lead to biased interpretations and predictions of a target’s actions, such that the same behavior might be interpreted one way when coming from a mother, another way when coming from a millennial, and so on.

Spaulding’s insight that a mindreader’s goals and social categorizations can affect how mentalizing functions is a good one. It is not clear, however, that the modes of mindreading that aim for efficiency rather than accuracy are in fact unreliable. Consider first egocentric projection: as noted, this strategy will fail when it turns out that we have overestimated the similarity between ourselves and our target. However, when it comes to the vast majority of our everyday beliefs, we *are* oftenquite similar to the people around us (Heal, 1996). Consider, for instance, the belief that the Allies won the second World War, the belief that dark clouds mean rain, or the belief that today is Tuesday. Because these beliefs are so mundane, I do not need to engage in deliberative reasoning in order to attribute them to the people around me, nor do I need any particular reasons for this attribution beyond my evidence for holding those beliefs in the first place. In projecting these beliefs of mine onto others, there is vanishingly little chance of error. The chance of error further decreases when one takes into account the fact that people naturally sort themselves into groups of like-minded individuals, like when we go to philosophy conferences, spend time with our friends, or cluster together online in political echo chambers. In these familiar social contexts, we can reliably expect a wide range of facts to be common knowledge, and tailor our mental-state inferences accordingly. For these reasons, the default assumption that our beliefs tend to be shared (effectively, a true-belief default) is built into a variety of different accounts of mindreading not as a bug, but as a feature (Heal, 1996; A. M. Leslie, Friedman, & German, 2004). Thus, while egocentric projection may occasionally lead to striking errors (Epley et al., 2004), there is good reason to think of it as an ecologically rational heuristic rather than as a bias (c.f. Goldstein & Gigerenzer, 2002).

It is also not obvious that stereotype-based mindreading is particularly unreliable. Certainly, one can find many examples of inaccurate stereotypes that lead to mindreading errors and perpetuate harmful social inequalities and structural injustices. To cite just one example that Spaulding provides, many falsely believe that people with disabilities tend to have a lower than average quality of life, which could in turn lead to a number of misleading mental-state inferences (Spaulding, 2018, p. 47). However, stereotypes are not false by definition. As a cognitive structure, stereotypes are just an efficient way of storing and retrieving semantic information associated with different social categories (Amodio, 2014). The truth or falsity of a given stereotype thus depends upon the content that it encodes, not on the mere fact that it is a stereotype.[[1]](#endnote-1)

For example, consider the stereotype that most American academics are politically liberal, which happens to be true (Duarte et al., 2014). If one were to rely upon this stereotype when anticipating such a U.S. academic’s beliefs about a divisive political issue such as gun control or abortion, one would probably arrive at the correct answer most of the time. Nor is this example unique: there is a large (though controversial – see for example Bian & Cimpian, 2017) empirical literature showing that many of our demographic stereotypes are surprisingly accurate when compared to real-world statistics (Jussim, 2012; Jussim, Crawford, & Rubinstein, 2015). Relying upon any of these stereotypes during mental-state attribution would therefore seem like a fairly safe bet, provided one does not have any additional individuating information about the mindreading target (Crawford, Jussim, Madon, Cain, & Stevens, 2011).

This is not to deny that the stereotype-based errors that Spaulding cites occur, or that they are not morally worrisome: they are certainly both. And even when true, relying upon stereotypes can be morally problematic, both because doing so perpetuates the underlying social injustices that make them true, and because treating a person as a mere instance of a kind seems to reflect a failure to respect that person as a unique individual. But when we are trying to understand the relationship between stereotyping and mindreading, and the role of stereotypes in social cognition more generally, we should nevertheless be careful not to treat cases of error as representative. The ubiquity of stereotyping in social cognition does not imply the ubiquity of erroneous stereotyping. Focusing on these cases obscures the fact that stereotype-based mindreading often works quite well.

The fact that mindreading processes that “aim for efficiency” may actually be quite reliable problematizes Spaulding’s analysis of modes of mindreading that “aim for accuracy.” Certainly it is true that sometimes we are motivated to engage in conscious, deliberative forms of mindreading that are more *effortful* than stereotyping and projection, but it is not clear that what this effort buys is increased accuracy. To better characterize the function of this mode of mindreading, we must distinguish between accuracy and *precision*. To illustrate, consider the contrast between the following two predictions:

1. It will rain in July.
2. It will rain on the morning on July 3.

If it rains on the morning of July 3, then it will be the case that both predictions were accurate. However, the two predictions nevertheless differ in terms of their precision. The first prediction is relatively easy to make, but also quite imprecise: it doesn’t tell us when in July it will rain. This places limits on its utility, because it does not help us if we are trying to make specific plans (e.g. if we are trying to decide which day we should have a barbecue, or an outdoor wedding). In contrast, the second prediction is very precise and would support highly specific plans. But making it with any confidence would also require more cognitive effort, and it has a higher chance of being wrong.

I suspect that the difference between efficient and deliberative forms of mindreading involves a similar trade-off between cognitive effort and precision. While heuristics like egocentric projection and stereotyping may lead to reliably accurate mental-state attributions, they are also relatively imprecise. This might be fine as long as our plans do not depend upon having a very precise representation of a person’s mental states – for instance, when we are interacting with a cashier at the checkout counter or chit-chatting with a stranger while waiting for a bus. Relying upon egocentric projection or stereotyping in these contexts may lead us to construct informationally sparse models of other agents that fail to capture many details about their actual attitudes. But in many cases, a person’s specific mental states do matter to us, and it is worth it for us to engage in explicit perspective taking to understand that they are thinking – say, during a serious conversation with a romantic partner, or when a therapist attempts to understand her patient’s suffering. By consciously reasoning about a person’s mental states and imagining what their experiences are like, we buy ourselves an added degree of precision that can be crucial when negotiating significant interpersonal relationships or coordinating on complex plans (Westra, 2018, 2019b).

In short, while our goals and social categorizations might influence the kind of mindreading strategy we adopt, this does not give us reason to be worried about the overall accuracy of mindreading. Efficient mindreading strategies may trade off on precision, but they can nevertheless function as ecologically rational and reliable heuristics, and need not be viewed as “biases” in an epistemically pejorative sense (though they do result in morally problematic judgments some of the time).

**How do we measure accurate mental-state attribution?**

Spaulding’s case for inaccuracy in mindreading is not limited to stereotyping and egocentric projection, however. In her book, she cites a long list of domain-general biases discovered by social psychologists such as naïve realism (Pronin, Lin, & Ross, 2002), self-serving and group serving attribution biases (Miller & Ross, 1975; Pettigrew, 1979), and the Dunning-Krueger effect (Kruger & Dunning, 1999), and describes how these might negatively impact the accuracy of mental-state attribution – for example, by distorting our appraisals of others’ knowledge and competence relative to our own (Spaulding, 2018, pp. 81–85). These, she might argue, are genuine biases, and seem much less amenable to an explanation in terms of ecological rationality. Should not the existence of such a wide range of biases should cause us to be pessimistic about the accuracy of mindreading and other social judgments?[[2]](#endnote-2)

Perhaps. But the mere fact that these biases are likely to lead to mindreading errors does not tell us much about the overall frequency of those errors relative to accurate mental-state attributions. Even the fact that a large number of these biases exist does not tell us about their overall effects on mindreading. Consider an analogous argument where I try to convince you that human vision is unreliable by listing off the dozens of optical illusions to which we are prone. This kind of argument would be unconvincing, because we know that these kinds of illusions comprise only a minority of our perceptual experiences, and that our visual capacities are generally reliable despite them. Likewise, the fact that we are prone to systematic patterns of error in certain contexts is consistent with the claim that mindreading is overall reliable and trustworthy (Westra, 2019a). Thus, in order to make a claim about the overall accuracy of mindreading, it is not sufficient to provide a list of all the biases that might cause mindreading errors. What is really needed is a way of measuring the accuracy of mental-state attribution in its own right.

Problematically, many of the most commonly used mindreading paradigms are not actually concerned with measuring accuracy at all. For example, developmental theory-of-mind paradigms such as the false-belief task (Wimmer & Perner, 1983) and the theory-of-mind scale (Wellman & Liu, 2004) only purport to measure are whether a participant possesses certain mental-state concepts. In tasks like these, young children are asked to make basic action predictions based on what fictional characters want, think, or know in a particular scenario; succeeding in these tasks requires children to understand that other agents can have mental states that differ from their own (e.g. that a character with a false belief might search for an object in a location the child knows to be incorrect). In order to isolate conceptual competence from extraneous performance factors, these paradigms are intentionally designed to be as simple as possible, and hardly reflect the complexities of real-world social interactions.[[3]](#endnote-3) Of course, possessing mental-state concepts is a necessary condition for accurate mental-state attribution: you can’t correctly attribute beliefs unless you know what beliefs *are*. But beyond establishing a basic prerequisite for accuracy, these tasks are silent on the matter.

Tasks that are used to measure more advanced mindreading abilities in adults are likewise unsuited for measuring accuracy. The Reading-the-Mind-in-the-Eyes Test (RMET) (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001), for example, measures how well people are able to apply various emotion words to specific facial expressions. Participants in this task view a series of images depicting the eyes and part of the face of a model, and must choose which among four possible emotion words best describe the model’s current psychological state. To succeed in the task, they must be able to make fine-grained discriminations between different facial expressions, and also express those discriminations using sophisticated psychological vocabulary.[[4]](#endnote-4) Again, this ability might be relevant to the accuracy of mindreading, since it taps into participants’ ability to individuate different types of psychological states according to their observable correlates. However, the being able to classify a face according to the *type* of emotion it is displaying sets the bar for accuracy fairly low: recognizing that a person appears to be *pensive* is not nearly as hard as judging what they are being pensive *about* – that is, the particular *content* of the emotional state in question. And even in terms of emotional categorization, the decontextualized nature of the RMET leaves much to be desired, since facts about the context in which a face occurs (e.g. body posture) can radically change whether we interpret it as expressing positive or negative emotions (Aviezer, Ensenberg, & Hassin, 2017; Aviezer, Trope, & Todorov, 2012). It is thus not clear that performance on the RMET says much at all about the accuracy of mental-state attribution.

In contrast, recursive mindreading tasks (e.g. Stiller & Dunbar, 2007) do involve the attributions of particular contents, but not as the result of any sort of inference. Rather, these tasks test participants’ ability to keep track of the mental states of characters from a story when this information is presented in a complex, multiply embedded format (e.g. “Bill knows that John thinks that Sally doesn’t know that Jenny thinks she’s nice”). Here, the correct mental-state information is simply *given* to the participants in a narrative format; the only way to be inaccurate in this task is to *forget* the correct answer (hence why Stiller and Dunbar (2007) also found a significant correlation between recursive mentalizing and short-term memory). Of course, maintaining relevant information in memory is sure to be relevant to any kind of complex problem-solving, including problems that involve mentalizing. Thus, this task also taps into something that is relevant to accurate mindreading. But it is certainly not a test of mindreading accuracy as such.

The closest the we get to a way of measuring the accuracy of mental-state attribution comes from the literature on ‘empathic accuracy’ (Hodges, Lewis, & Ickes, 2014; Ickes, Stinson, Bissonnette, & Garcia, 1990). In these studies, experimenters videotape a target individual as she is being interviewed or during a conversation. Afterwards, they show the tape with to the target and ask her to indicate what she was thinking at different time points. The same video is then shown to another participant. At different points during the video, this participant is asked to guess what the target is thinking. Afterwards, independent coders rate how similar the judge’s guesses are to the target’s self-reports.[[5]](#endnote-5)

This use of self-reported mental states as an accuracy criterion is problematic for several reasons. First, it assumes that we have reliable introspective access to our own thoughts. The problem with this is that there is a large body of evidence from both neurotypical and clinical cases showing that we are prone to confabulate about both our prior decision-making and our occurrent conscious experiences (Johansson, Hall, Sikström, & Olsson, 2005; Nisbett & Wilson, 1977; Roser & Gazzaniga, 2004; Schwitzgebel, 2002). This research suggests that when targets review their recorded interviews and indicate what they were thinking at the time, they may simply be generating plausible mental-state attributions about themselves after the fact. This ties in a further problem: according to theories of introspection like Peter Carruthers’ Interpretive-Sensory Access account, introspection just is a form of mindreading directed at the self (Carruthers, 2009, 2011). If something like this theory is correct, then empathic accuracy researchers’ criterion for accurate third-party mindreading *presupposes* the accuracy of a different kind of mindreading. And finally, the mental-state attributions that these studies evaluate are explicit, verbal, and conscious. Yet most mindreading theories hold that this kind of reasoning about other minds constitutes just a small part of our overall theory of mind, and that most mentalizing is implicit, unconscious, and non-verbal. If this is right, then it may be that we make many accurate mental-state attributions that we are unable to put into words.

Despite these objections to the empathic accuracy approach, it is admittedly difficult to come up with a criterion for accurate mental-state attribution that does not rely upon introspection. One possibility might be to employ advanced neuroimaging methods such as multivariate pattern analysis to decode the content of a person’s occurrent mental states, and then use this as an accuracy criterion for third-party mental-state attribution (but see Ritchie, Kaplan, & Klein (2017) for some of the limitations of this technique). But barring the significant technological and methodological advances that this would require, introspective accuracy criteria are probably the best that we can expect to achieve. If the above-mentioned concerns about introspection are warranted, then this bodes ill for the measurement of accuracy in mindreading.

This is shocking, considering how often mindreading theorists appeal to the adaptive significance of mindreading and its importance in social learning. But contrary to Spaulding’s pessimistic view, the problem with this assumption is not that we are *less* accurate than these theorists would have us believe. It is rather that we are not in an epistemic position to know whether mindreading is accurate *one way or the other*. In other words, claims about the accuracy and inaccuracy of mindreading are both equally unjustified. If this is right, then mindreading theorists cannot claim to offer advice to philosophers worried about peer disagreement, or about any other philosophical issue that presupposes the accuracy of our mental-state attributions. It could be that Spaulding is in fact correct that our judgments about who are our epistemic peers are not to be trusted. Alternatively, it could be that strategies like egocentric projection and stereotyping actually make mindreading more reliable, as I argued above. But without any way to empirically test such claims, neither position enjoys any direct evidential support.

**Conclusion:**

In this commentary, I have focused on one theme that emerges in Spaulding (2018): the accuracy of mindreading. I have offered some pushback against Spaulding’s arguments about the inaccuracy of mindreading, and then stepped back to point to what I think is a bigger problem about accuracy than the one that Spaulding has focused on. In spite of these points of disagreement, I am largely sympathetic to Spaulding’s overall project. Theories of mindreading ought to take into account the role that different motivational and social factors have on the way that mindreading functions; Spaulding’s Model Theory provides a promising framework for talking about these effects. Moreover, by taking up the issue of accuracy, Spaulding has advanced the conversation around mindreading beyond questions about its internal and conceptual underpinnings to questions about how our mental-state attributions actually reflect the way the world is. With this commentary, I hope to have built upon Spaulding’s advances by highlighting the complexity of this issue, and shown that it is worthy of further philosophical attention.

**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), 340–353. https://doi.org/10.1037/0022-3514.87.3.340

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), 573–585. https://doi.org/10.1037/0022-3514.87.5.573

Amodio, D. M. (2014). The neuroscience of prejudice and stereotyping. *Nature Reviews: Neuroscience*, *15*(10), 670–682. https://doi.org/10.1038/nrn3800

Aviezer, H., Ensenberg, N., & Hassin, R. R. (2017). The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*. https://doi.org/10.1016/j.copsyc.2017.06.006

Aviezer, H., Trope, Y., & Todorov, A. (2012). Body cues, not facial expressions, discriminate between intense positive and negative emotions. *Science*, *338*(6111), 1225–1229.

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind*. MIT Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=MDbcNu9zYZAC&pgis=1

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The ’ ’ Reading the Mind in the Eyes ’ ’ Test Revised Version : A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *Journal of Child Psychology and Psychiatry*, *42*(2), 241–251. Retrieved from https://docs.autismresearchcentre.com/papers/2001\_BCetal\_adulteyes.pdf

Bian, L., & Cimpian, A. (2017). Are stereotypes accurate? A perspective from the cognitive science of concepts. *Behavioral and Brain Sciences*, *40*, e3. https://doi.org/10.1017/s0140525x15002307

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences*, *32*(02), 121–138.

Carruthers, P. (2011). *The opacity of mind: an integrative theory of self-knowledge*. Oxford University Press.

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *Philosophical Review*, *116*(2), 187–217. https://doi.org/10.1215/00318108-2006-035

Crawford, J. T., Jussim, L., Madon, S., Cain, T. R., & Stevens, S. T. (2011). The use of stereotypes and individuating information in political person perception. *Personality and Social Psychology Bulletin*, *37*(4), 529–542. https://doi.org/10.1177/0146167211399473

Dodell-Feder, D., Ressler, K. J., & Germine, L. T. (2018). Social cognition or social class and culture? on the interpretation of differences in social cognitive performance. *Psychological Medicine*. https://doi.org/10.1017/S003329171800404X

Duarte, J. L., Crawford, J. T., Stern, C., Haidt, J., Jussim, L., & Tetlock, P. E. (2014). Political Diversity Will Improve Social Psychological Science. *Behavioral and Brain Sciences*, *38*, 1–54. https://doi.org/10.1017/S0140525X14000430

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology*, *87*(3), 327.

Godfrey-Smith, P. (2005). Folk Psychology as a Model. *Philosophers’ Imprint*, *5*(6), 1–16. https://doi.org/10.3366/jqs.2011.0012

Goldstein, D. G., & Gigerenzer, G. (2005). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*(1), 75–90. https://doi.org/10.1037/0033-295x.109.1.75

Heal, J. (1996). Simulation, theory, and content. In P. Carruthers & P. K. Smith (Eds.), *Theories of Theories of Mind* (pp. 75--89). Cambridge, UK: Cambridge University Press.

Helming, K. A., Strickland, B., & Jacob, P. (2016). Solving the Puzzle about Early Belief-Ascription. *Mind & Language*, *31*(4), 438–469. https://doi.org/10.1111/mila.12114

Hodges, S. D., Lewis, K. L., & Ickes, W. (2014). The matter of other minds: Empathic accuracy and the factors that influence it. *APA Handbook of Personality and Social Psychology, Volume 3: Interpersonal Relations.*, *3*, 319–348. https://doi.org/10.1037/14344-012

Ickes, W., Stinson, L., Bissonnette, V., & Garcia, S. (1990). *Naturalistic Social Cognition: Empathic Accuracy in Mixed-Sex Dyads*. *Journal of Personality and Social Psychology* (Vol. 59). Retrieved from https://pdfs.semanticscholar.org/fd94/e2ae60224abe60759a790083a922fc58b442.pdf

Johansson, P., Hall, L., Sikström, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119. https://doi.org/10.1126/science.1111709

Jussim, L. (2012). Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy. *Social Perception and Social Reality: Why Accuracy Dominates Bias and Self-Fulfilling Prophecy*, 1–486. https://doi.org/10.1093/acprof:oso/9780195366600.001.0001

Jussim, L., Crawford, J. T., & Rubinstein, R. S. (2015). Stereotype (In)Accuracy in Perceptions of Groups and Individuals. *Current Directions in Psychological Science*, *24*(6), 490–497. https://doi.org/10.1177/0963721415605257

Kelly, T. (2011). Peer disagreement and higher-order evidence. In A. I. Goldman & D. Whitcomb (Eds.), *Social Epistemology: Essential Readings* (pp. 183–217). Oxford: Oxford University Press.

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.

Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in “theory of mind”. *Trends in Cognitive Sciences*, *8*(12), 528–533. https://doi.org/10.1016/j.tics.2004.10.001

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology*, *50*(1), 45–85. https://doi.org/10.1016/j.cogpsych.2004.06.002

Leslie, S.-J. (2014). Carving up the Social World with Generics. In J. Knobe, T. Lombrozo, & S. Nichols (Eds.), *Oxford Studies in Experimental Philosophy: Volume One* (pp. 208–232). Oxford: Oxford University Press.

Lewis, K. L., Hodges, S. D., Laurent, S. M., Srivastava, S., & Biancarosa, G. (2012). Reading between the minds: the use of stereotypes in empathic accuracy. *Psychological Science*, *23*(9), 1040–1046. https://doi.org/10.1177/0956797612439719

Maibom, H. (2003). The Mindreader and the Scientist. *Mind & Language*, *18*(3), 296–315. https://doi.org/10.1111/1468-0017.00229

McGeer, V. (2007). The regulative dimension of folk psychology. In D. D. Hutto & M. Ratcliffe (Eds.), *Folk psychology re-assessed* (pp. 137–156). Springer.

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

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.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2017). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *The British Journal for the Philosophy of Science*, *00*, 1–27. https://doi.org/10.1093/bjps/axx023

Roser, M., & Gazzaniga, M. S. (2004). Automatic Brains, Interpretive Minds. *Current Directions in Psychological Science*, *13*(2), 56–59. https://doi.org/10.1111/j.0963-7214.2004.00274.x

Schwitzgebel, E. (2002). How well do we know our own conscious experience? The case of visual imagery. *Journal of Consciousness Studies*, *9*(5–6), 35–53. Retrieved from http://faculty.ucr.edu/~eschwitz.

Scott-Phillips, T. (2014). *Speaking Our Minds: Why Human Communication is Different, and how Language Evolved to Make it Special* (Vol. 3). Palgrave Macmillan. Retrieved from https://books.google.com/books?hl=en&lr=&id=B0yjBQAAQBAJ&pgis=1

Spaulding, S. (2016). Mind Misreading. *Philosophical Issues*, *26*.

Spaulding, S. (2017). Do you see what I see? How social differences influence mindreading. *Synthese*, 1–22. https://doi.org/10.1007/s11229-017-1404-1

Spaulding, S. (2018). *How We Understand Others: Philosophy and Social Cognition*. London: Routledge. https://doi.org/10.4324/9781315396064

Stiller, J., & Dunbar, R. I. M. (2007). Perspective-taking and memory capacity predict social network size. *Social Networks*, *29*(1), 93–104. https://doi.org/10.1016/j.socnet.2006.04.001

Tomasello, M. (2014). *A Natural History of Human Thinking*. Harvard University Press. Retrieved from https://books.google.com/books?hl=en&lr=&id=ksYXAgAAQBAJ&pgis=1

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development*, *75*(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x

Westra, E. (2018). Character and theory of mind: an integrative approach. *Philosophical Studies*, *175*(5), 1217–1241. https://doi.org/10.1007/s11098-017-0908-3

Westra, E. (2019a). Getting to know you: accuracy and error in judgments of character. *Mind & Language*.

Westra, E. (2019b). Stereotypes, theory of mind, and the action–prediction hierarchy. *Synthese*, *196*(7), 1–26. https://doi.org/10.1007/s11229-017-1575-9

Westra, E., & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind Scale. *Cognition*, *158*, 165–176. https://doi.org/10.1016/j.cognition.2016.10.021

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children’s understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Zawidzki, T. W. (2013). *Mindshaping: A New Framework for Understanding Human Social Cognition*. Cambridge, MA: MIT Press.

**Notes:**

1. Spaulding is sensitive to this point. In particular, she is careful to distinguish between stereotypes that encode false statistical generalizations and stereotypes that encode *generics* (S.-J. Leslie, 2014). For example, “most mosquitos carry West Nile” expresses a false statistical generalization, but the generic “mosquitos carry West Nile” expresses a true proposition. Spaulding allows that the stereotypes that influence mindreading may be either statistical or generic in their form, and that generic stereotypes sometimes encode true propositions. [↑](#endnote-ref-1)
2. Spaulding (2018) also argues that sometimes our goals in mindreading are not to attribute accurate mental states at all, but rather to *shape* or *regulate* the mental states of our mindreading targets (McGeer, 2007; Zawidzki, 2013). For example, one might try to encourage a person to sample a new kind of cuisine by convincing them that they *like* to try new things. Ostensibly, this mental-state attribution need not be true in order for it to achieve its goal. However, this goal is entirely consistent with – and indeed might actually depend upon –accurate mindreading. But in order for subtle acts of regulative mental-state attribution like this to be effective, they will most likely require other accurate mental-state attributions. For example, in order for the aforementioned strategy to work, I must first be able to predict that the false mental-state attribution would have a particular effect on the target’s *actual* mental states (say, by activating the person’s desire to like up other people’s expectations of her). If this part of my prediction were wrong, then the whole regulative strategy would fail. Thus, the fact that we sometimes mindread for regulative purposes does not undercut the basic claim that mindreading typically aims at accuracy. [↑](#endnote-ref-2)
3. Indeed, a common critique of these kinds of tasks is that they involve extraneous performance demands that obscure conceptual competence (Helming, Strickland, & Jacob, 2016; A. M. Leslie, German, & Polizzi, 2005; Westra & Carruthers, 2017) [↑](#endnote-ref-3)
4. It is also worth noting that performance on the Reading-the-Mind-in-the-Eyes task is strongly affected by participant race, ethnicity, and education (Dodell-Feder, Ressler, & Germine, 2018), which raises further questions about its reliability as a tool to measure mindreading competence. [↑](#endnote-ref-4)
5. Using this method, Lewis and colleagues have found that having stronger stereotypes can, at least in certain circumstances, lead to more accurate mindreading (Lewis, Hodges, Laurent, Srivastava, & Biancarosa, 2012) – a point that is consistent with the argument about stereotyping from the previous section. However, as I go on to argue, we should be very cautious about how much warrant we give to this finding. [↑](#endnote-ref-5)