How I Know What You Know

For Oxford Handbook of Social Epistemology, Jennifer Lackey and Aidan McGlynn (eds.)

Shannon Spaulding

*Penultimate draft. Please cite only published version.*

# 1. Introduction[[1]](#footnote-1)

Mentalizing is our ability to infer agents’ mental states. Attributing beliefs, knowledge, desires, and intentions are frequently discussed forms of mentalizing. Attributing mentalistically loaded stereotypes, personality traits, and evaluating others’ rationality are forms of mentalizing, as well. This broad conception of mentalizing, which I articulate and defend in a recent book (Spaulding, 2018), has interesting and important implications for social epistemology. Several topics in social epistemology involve judgments about others’ knowledge, rationality, and competence, e.g., peer disagreement, epistemic injustice, and identifying experts. Mentalizing is at the core of each of these debates.

In this chapter, I will describe the broad conception of mentalizing and show how it is central to how we judge others’ knowledge, competence, and rationality. I will apply this perspective on mentalizing to two debates: the epistemology of peer disagreement and interventions on testimonial injustice. I argue that understanding how mentalizing works can help us see these debates in a different light. Such reframing can help us make progress on these challenging debates.

# 2. Mentalizing

Mentalizing consists in attributing mental states to agents in order to interpret, anticipate, or interact with them. Historically, the literature on mentalizing (also called theory of mind, folk psychology, or mindreading) has tended to focus on how and when we attribute beliefs to others. Belief attribution, in particular demonstrating the ability to attribute *false* beliefs, is considered a core capacity of sophisticated social cognitive abilities. The false-belief task is a cornerstone of the mentalizing literature. It was conceptualized by Daniel Dennett (1978) and experimentally implemented by Hans Wimmer and Josef Perner (1983). The insight of the task is that if you want to know whether a subject has the concept of belief, a good way to figure this is out is to see whether a subject can attribute a *false* belief. False belief attributions imply that a subject understands that others can represent the world in a certain way (i.e., they believe the world is a certain way) and that this representation can be true or false.

In a classic version of the false-belief task, subjects watch Sally, a puppet, place a toy in a basket and leave the scene. They watch Ann, another puppet, come in and move the toy to the cupboard and then leave the scene. Finally, they watch Sally return to the scene. At this point, they are asked a series of questions to ensure that they remember what happened and know where the toy actually is. The key question in this version of the task is, “Where will Sally look for her toy?” In order to pass this test, they must understand that Sally thinks the toy is in one location (the basket), and will look for it there, but actually the toy is in a different location (the cupboard). The ability to represent both how the world is and how another person represents (and sometimes misrepresents) the world is widely viewed as a core competency of folk psychology. Being able to attribute beliefs (both true and false) opens the door for a wider range of mentalizing capacities, such as elaborate pretense, irony, sarcasm, rational reconstruction of others’ reasons, using and understanding conversational implicatures, etc. Thus, there is a good reason why theorists have focused on testing whether subjects truly possess the belief concept, and there are good theoretical reasons for emphasizing the importance of passing the false-belief task.

Despite the centrality of belief in theorizing about mentalizing, mentalizing extends far beyond just belief attribution. We employ a variety of mental attributions in our social interactions. I shall highlight a few growing areas of research that study the diverse forms of mentalizing. Consider, for example, the research on a theory of mind scale, which aims to establish when children master various mentalizing skills (Wellman and Liu 2004, Wellman, Fang, and Peterson 2011). Based on a meta-analysis of over 600 primary, longitudinal, and cross-cultural studies, Wellman and colleagues argue that there is a consistent developmental pattern in mentalizing skills. From roughly age 2 to 6, children progressively demonstrate understanding of *discrepant desires* (i.e., that desires can conflict), then *discrepant beliefs* (i.e., that beliefs can conflict), *knowledge access* (i.e., that something can be true but an agent may not know it), *false belief* (e.g., that an agent’s representation of the world does not match reality), and finally *hidden emotions* (i.e., that agents can experience an emotion even when there are no outward sign of it). This research emphasizes various mentalizing skills beyond belief attribution, and it indicates that for neurotypical children there is a universal developmental trajectory of these mentalizing skills.

Recently, philosophical and empirical research has explored how and when we attribute *knowledge*. For example, one area of experimental philosophy investigates the conditions under which we attribute knowledge vs. belief (Knobe and Nichols 2017, Gerken 2017, ch. 2). Research in this area *seems* to show that our epistemic attributions depend on how much is at stake in a situation: in low-stakes situations we readily attribute knowledge rather than mere belief, but in high-stakes situations we are more cautious about knowledge attributions (Pinillos 2012, Sripada and Stanley 2012).[[2]](#footnote-2) Relatedly, while it is clear that we attribute belief without attributing knowledge (e.g., unjustified true beliefs), some work in experimental philosophy indicates that sometimes we attribute knowledge without belief (e.g., in cases where someone temporarily forgets the answer to a question they studied extensively). Importantly, knowledge attributions seem to play a different functional role than belief attributions in mentalizing (Turri 2017, Westra and Nagel 2021). Knowledge attributions lead to faster and more reliable behavioral predictions than belief attributions (Phillips et al. 2020).[[3]](#footnote-3)

The attribution of character traits and stereotypes can be forms of mentalizing, as well. Character trait and stereotype attribution are based on social categorization, the reflexive, rapid sorting of people into social categories. To understand how character traits and stereotype attribution work, we need to understand how social categorization works. When we encounter other people, we habitually and rapidly sort them into various social categories, e.g., age, gender, race, nationality, religion, and class. The most salient categories tend to be age, gender, and race, and we can sort people into these categories in a fraction of a second (Ito, Thompson, and Cacioppo 2004, Liu, Harris, and Kanwisher 2002). However, the categories we employ can be modulated by various factors, such as context, cognitive load, and our goals (Wheeler and Fiske 2005, Gilbert and Hixon 1991). Thus, which social categories are most salient to us in a given situation will vary depending on the broader context.

On the basis of social categorization, we spontaneously infer character traits, such as trustworthiness, aggressiveness, dominance (Olivola and Todorov 2010, Rule, Ambady, and Adams Jr 2009). Character trait attributions are mentalistically loaded inferences about a target’s personality. When you observe a target’s facial expressions and behavior and infer that he is aggressive, for example, you are not simply making an inference about his behavioral dispositions. You are also implicitly making an inference about his attitudes and emotions, e.g., that he is insecure or angry. Inferring character traits and explaining or predicting behavior based on character trait attributions is a form of mentalizing (Westra 2017a).[[4]](#footnote-4) We can make such character trait inferences very quickly, within a second or two, even when we are under cognitive load (Todorov and Uleman 2003, Malle and Holbrook 2012). These rapid character trait inferences turn out to be fairly accurate (Palomares and Young 2018).

Social categorization also forms the basis for social stereotype attribution. The social categories we sort people into are associated implicitly and explicitly with various features, and these associations ground social stereotypes. Stereotypes are conceptually rich systems of belief about social groups (Liberman, Woodward, and Kinzler 2017). Stereotypes are beliefs about members of particular social groups. We frequently rely on stereotypes about how individuals from particular social groups think and behave. Relying on stereotypes helps us interpret an individual’s behavior quickly and efficiently, and it helps us anticipate what that individual will do next (Westra 2017b).

Not all social stereotypes are forms of mentalizing. Some stereotypes are simply about behavior, e.g., what clothes or apparel certain groups wear. The stereotype attributions that are relevant here are ones that involve mental states, for example, psychological traits, attitudes, or emotions associated with certain social groups. Consider, for example, a stereotype about anti-vaxxers, individuals who forgo vaccines for themselves and/or their children. Part of the common stereotype about anti-vaxxers is that they have false beliefs about the safety of vaccines, they are fearful of causing harm to their children, and they do not understand the science of epidemiology. These beliefs about anti-vaxxers are beliefs about their mental states – about what they believe, fear, know – and thus are a form of mentalizing.

We frequently spontaneously attribute character traits and stereotypes in interpreting and anticipating others’ behavior (Uleman, Adil Saribay, and Gonzalez 2008). Character traits and stereotypes help us rapidly form expectations for how certain individuals will behave. Even when we encounter counter-stereotypical behavior, the stereotype forms the anchor for our expectations of that individual or group. Sometimes we are surprised by others’ behaviors precisely because we were interpreting them in light of a stereotype that ended up not fitting.

A further aspect of mentalizing is important for social epistemology: how our motivations in a social interaction influence our approach to mentalizing (Spaulding 2017). In some social interactions, what we really want is an accurate and precise account of a target’s mental states. For example, when a situation is unusual or it matters to us personally, we take a deliberative approach to mentalizing. We try to carefully weigh the evidence, ignore superficial features, and come up with a precise and accurate mental state inference. Other times we do not want or need such a careful inference. In situations where the people or events seem typical, or we are not personal invested in the situation, or we are under cognitive load, we use efficient mentalizing approaches, e.g., egocentric projection and stereotyping (Ames 2004a, Ames 2004b).[[5]](#footnote-5) Finally, sometimes mentalizing is motivated by self-interest or in-group protection. In such cases, we filter our interpretation of others’ mental states in distinctive ways. We may more charitably interpret the behavior of in-group members than out-group members. We may assume that those who are different from us or those who disagree with us are biased or misperceiving, especially in a context of threat or competition. Confirmation bias – i.e., our tendency to seek out information that confirms our interpretations and interpret ambiguous information as confirming our interpretation (Nickerson 1998) – further cements these biased interpretations. Mentalizing that is motivated by self-interest and in-group protection can be deliberative or efficient. Thus, these biases may influence a subject’s careful deliberation *and* rapid, spontaneous inferences about a target’s mind.

Finally, there is new emerging research on the relationship between mentalizing and *epistemic trust* in children (Rotenberg et al. 2015), neurotypical adults (Prochazkova et al. 2018), and in pathological cases like Borderline Personality Disorder (Bo et al. 2017) and social anxiety disorder (Sripada et al. 2009). This research indicates that mentalizing abilities mediate how much we trust other people to be authoritative on a subject. Clearly, trust also mediates how we mindread. If we are antecedently predisposed to trust certain individuals, we are more likely to infer that they *know* rather than merely believe X and that they are competent and rational assessors of X. Thus, there is a reciprocal relationship between mentalizing and trust, which is relevant when it comes to assessing others’ knowledge, competence, and rationality.

I have canvassed various aspects of mentalizing: the attribution of knowledge vs. belief, the relation between social categorization and stereotypes and character trait attributions, the ways our goals influence how we mentalize, and how trust and mentalizing are related. This is not an exhaustive list, of course. The research program in pluralistic folk psychology details many of these forms and norms of mentalizing (Andrews 2008, McGeer 2007, Spaulding 2019, Westra 2017a, Zawidzki 2013). The goal here is to demonstrate that mentalizing is a heterogeneous phenomenon. This heterogeneity is important for understanding the connection between mentalizing and social epistemology.

# 3. Mentalizing and Social Epistemology

Many debates in social epistemology involve, either directly or indirectly, attributions of knowledge, competence, and rationality. For example, debates about knowledge attribution (ch. X), how to decide who is an epistemic authority (ch. X), what to do when someone you take to be an epistemic peer disagrees with you (ch. X), how a hearer’s prejudices can result in downgrading a speaker’s credibility (ch. X). These debates concern the attribution of knowledge, competence, and rationality, all of which are forms of mentalizing. In this section, I will examine two of these debates: peer disagreement and testimonial injustice. I will argue that the broad conception of mentalizing can help us see these debates in a different light, which may lead us to different conclusions about peer disagreement and testimonial injustice.

## 3.1 Peer Disagreement

Debates about the epistemology of peer disagreement concern what we rationally ought to do when someone we take to be an epistemic peer disagrees with us on a factual matter (Matheson 2015). Two people are epistemic peers with respect to X when they possess similar evidence about X, similar background knowledge, are comparably intelligent, free from bias, competent at perceiving, reasoning, etc. (Kelly 2011). The puzzle of peer disagreement arises when someone I acknowledge as an epistemic peer with respect to X disagrees with me about X. To take a now classic example, suppose I go out to dinner with a group of friends and we decide to split the bill. My friend looks at the bill and calculates that we each owe X. However, when I look at the bill, I calculate that we each owe Y. I assume that my friend is roughly as good at simple mathematical calculations as I am, she looked at the same bill, was just as attentive as me when performing her calculation, etc. What is the rational thing for me to do here?

Perhaps it is rational for me to remain steadfast in my judgment about X (Kelly 2011, Buchak forthcoming). After all, I think I am just as good at simple math as my friend, and my calculation still seems right even after I learn that she came up with a different number. So, perhaps I should remain just as confident that my calculation is correct. Or perhaps the rational thing to do is to reduce my confidence in my judgment (Christensen 2007, Rasmussen, Steglich-Petersen, and Bjerring 2018). My friend is just as good at simple mathematical calculations as I am, and I have no independent reason to think that she’s mistaken in this case. Indeed, the fact that she disagrees with me gives me evidence that I made a mistake, so the only rational thing to do reduce is my confidence in my calculation. A middle-ground option is to maintain that whether you conciliate or remain steadfast depends on the belief, in particular your prior level of justification (Lackey, 2010) or confidence (Vavova 2014) in that belief. For instance, perhaps you should remain steadfast in the face of peer disagreement when it comes to beliefs with high antecedent justification and conciliate for beliefs without high prior justification, like the restaurant bill case. Another middle-ground option is to decide what to believe on the basis of the total evidence, both the first-order evidence for the belief and the higher-order evidence about others’ take on the belief (Kelly, 2010). If all of your friends do the math and come up with same number that your friend calculated, the total evidence may push you to conciliate. Otherwise, your total evidence (including evidence about your own attentiveness and time devoted to the calculation) may support remaining steadfast in your own conclusion.

Although this restaurant case is a simple, idealized example, such disagreements are argued to be widespread. Disagreements about philosophical views, ethical debates, public policy, etc. may count as peer disagreements. Thus, the debate is meant to have broad implications for the rationality of retaining many of our controversial beliefs.

The debate between the steadfast view and the conciliatory view sometimes hinges on whether it is legitimate to use one’s own reasoning about a topic in evaluating a peer’s reasoning about that topic. Is it appropriate to give more weight to my own decision-making process or to rely on evidence of my own attentiveness and time devoted to thinking about the problem when I disagree with an epistemic peer? Can I use the fact of disagreement as evidence that my epistemic peer is less likely to be correct? The steadfast view says yes to these questions: I can, and that tips the scales in favor of my own judgment. The conciliatory view says that I should not, and without that resource I must conclude that my epistemic peer has just as much reason to think she is correct. In that case, it is rational to reduce confidence in my judgment.

These are difficult philosophical problems. However, I think we can make progress on these problems if we bring into consideration *how* we make judgments about who is an epistemic peer. Recall that the puzzle of peer disagreement arises when someone *we* *acknowledge* to be an epistemic peer with respect to X disagrees with us about X. Understanding how we make judgments about others knowledge, competence, and rationality, especially in the context of disagreements about *controversial* beliefs, can help advance this discussion of the epistemology of peer disagreement.

So, how do we decide whether someone is an epistemic peer, inferior, or superior? Some parts of this are simple and do not involve mentalizing at all, e.g., knowing someone’s educational pedigree. If I find myself disagreeing with an epidemiologist about predictions for the Covid-19 pandemic, it is clear to me, simply in virtue of our obvious educational differences, that I am the epistemic inferior in this debate. But often epistemic differences are not so obvious. For example, when I find myself disagreeing with another *philosopher* about pandemic projections, there may not be obvious, decisive facts about our relative knowledge, competence, and rationality. In these subtler cases, we rely on mentalizing to judge epistemic status.

Several aspects of mentalizing described above are relevant to judging our own and others’ epistemic status. Take for instance attributing knowledge as opposed to mere belief. Some work in experimental philosophy suggests that we are much more cautious in attributing knowledge to others and ourselves in high-stakes situations than in low-stakes situations. It is likely that disagreements, especially disagreements about controversial topics, trigger this shift in standards.

Furthermore, when we find ourselves in emotionally, morally, or politically charged disagreements, we are more inclined toward a host of biased patterns in mentalizing. We are more likely to infer that those who disagree with us are influenced by personal motivations or misperceiving (Pronin, Lin, and Ross 2002, Pronin 2007). This pattern in mentalizing is more pronounced when the people who disagree with us are part of an out-group. In contrast we more charitably interpret our own behavior. We look for exculpatory reasons for ourselves but not for those who disagree with us (Miller and Ross 1975). Thus, if we misspeak or misremember a relevant fact, we dismiss that as a performance error. But if a target who disagrees with us similarly misspeaks or misremembers a fact, we take this to be a competence error. In this way, we come to overestimate our own competence and underestimate others’ competence.[[6]](#footnote-6)

Combining the insight about how the stakes of a situation determine our willingness to attribute knowledge (both to oneself and to others) with the evidence that emotionally, morally, and politically charged disagreements lead us to negatively assess our interlocuter’s knowledge, rationality, and competence yields the following prediction: When we enter into a high-stakes, emotionally charged disagreement, we are less likely to infer that our interlocuter *knows* X than we otherwise would be. Even in low-stakes, emotionally charged disagreements, we tend to downgrade others’ knowledge and competence and upgrade our own. This suggests that when someone is in fact an epistemic peer, we may not regard her as such because of the self-enhancing biases triggered by disagreements.

The epistemology of disagreement literature does incorporate some of these phenomena into the discussion. Some argue that egocentric biases are rational insofar as they are the basis of self-trust. Thus, some argue, it is rational to weight our own beliefs more heavily – because we trust ourselves in ways we cannot trust others – even when we regard the other person as an epistemic peer (Wedgwood 2010). Others argue that if we have high level of justification for a belief, the fact that someone disagrees with that should not cause us reduce confidence. Indeed, should we encounter an apparent epistemic peer who disagrees with these antecedently highly justified beliefs, we may have grounds for thinking they are in fact an epistemic inferior. In that case, there is no puzzle of *peer* disagreement because if she disagrees with X then it turns out she is not an epistemic peer: she was a conditional epistemic inferior (Lackey 2010).

If this were the end of the story, we might simply push for the discussion of the epistemology of peer disagreement to more explicitly reflect the fact that we sometimes overestimate our own knowledge and competence and underestimate others’ knowledge and competence, especially in the context of disagreement. On a first pass, incorporating these phenomena might seem to tip the scales toward the conciliatory view or some middle-ground account like the justificationist view insofar as the empirical phenomena indicate that our intuitive epistemic assessments may inappropriately self-enhance and other-downgrade. However, this *prima facie* tipping of the scales is by no means decisive. Cleary there are several maneuvers available for the other views. However, this is not the end of the story. Even though we tend to epistemically upgrade ourselves and downgrade others, we are not consistent or fair in the way we do this. Our judgments of others’ knowledge, competence, and rationality are idiosyncratic, shaped by our motivations and context, and riddled with biases.

We learned from the discussion of mentalizing that perceiving a social interaction involves categorizing individuals into salient social groups (e.g., age, race, gender, nationality, ethnicity, religion) that are associated with various features, stereotypes, and social biases. These associations influence how we decide who is an epistemic peer, inferior, or superior before we even begin to evaluate their evidence base or reasoning abilities. Their social category anchors our expectations of their knowledge and competence in a certain domain. For instance, we tend to associate old age with incompetence. Upon encountering an elderly person, our default expectation is that they are less competent at many physical and intellectual tasks than a young person. We often think of elderly women, for example, as affectively warm and compassionate but incompetent (Cuddy, Norton, and Fiske 2005, Fiske, Cuddy, and Glick 2007, Hummert et al. 1994). Stereotypes of the elderly are that they often get lost, are forgetful, fragile, and weak. Our habitual inclination to offer help to elderly individuals reveals our assumptions about the elderly.

The character traits (e.g., compassionate) and stereotypes (e.g., forgetful) that we instinctually apply to elderly individuals set our epistemic expectations for the elderly pretty low. Across many cultures, people habitually downgrade the epistemic status of the elderly simply in virtue of their social category. Setting the bar so low anchors our expectations for an individual elderly person’s knowledge, competence, and rationality. Compared to the non-elderly, we assume they will be less aware, less able to recall key facts, less capable of sustained focus on a complicated problem, etc. Furthermore, we rarely revise our first impressions when we are given more time to deliberate (Willis and Todorov 2006). Of course, we *can* change our minds about other people. It will take a lot of obvious and unambiguous information that the elderly individual is knowledgeable and competent in some domain to overcome the default expectations. If the elderly individual is in fact an epistemic peer – with roughly the same background knowledge, evidence, and reasoning abilities as you, a younger person – it is very likely that you will conclude that this person is your epistemic inferior.[[7]](#footnote-7) Our implicit associations, rapid character trait inferences, and stereotypes deeply shape our initial expectations and the way in which we continue to interpret a target’s behavior throughout our interactions.

Our skewed default assumptions of knowledge, competence, and rationality are not limited to the elderly, of course. We as a society habitually upgrade or downgrade the epistemic status for many social groups. This is not the place for a systematic review of these patterns, but let’s consider just a few relevant instances. Individuals in Western, developed, democratic societies tend to assume that women are less good at math and science than men (though this assumption is not supported by the data). We tend to assume that individuals of east Asian descent are better at math and science than other racial and ethnic groups. We tend to assume that Black men are more gifted athletically than intellectually. The list goes on, of course. Depending on the target individual’s social group, we may upgrade or downgrade our default assessment of their epistemic status for a particular domain. Our default assumptions about the knowledgeability and competence of certain social groups on certain topics will influence the kinds of mental state and personality trait inferences we make about individuals. And these mental state and personality trait inferences will tend to confirm the default assumptions about knowledgeability and competence. We can revise our initial judgments, but mentalizing processes often end up confirming our initial judgments.

The skewing of epistemic assessments is not limited to general self-enhancement, disagreement-based downgrading, and character-trait and stereotype-driven upgrades or downgrades. In-group/out-group status also significantly affects our judgments of other people’s epistemic credentials. In-grouping and out-grouping is a function of perceived similarity in a situational context (Tajfel 1974). Often in-grouping and out-grouping follow along social and demographic lines, e.g., age, race, gender, sexual orientation, nationality, etc. However, in-groups and out-groups can be quite fluid in different situational contexts. For instance, runners are part of my in-group when the context is hobbies, recreation, or exercise. However, these same runners may be part of an out-group for me in the context of careers, jobs, and educational achievement.

We tend to 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 (Rudman et al. 1999, Cikara et al. 2014). We more charitably interpret in-group behavior than out-group behavior. We regard mistakes and misstatements of in-group members as one-off errors, but these same mistakes and misstatements may be taken as confirmation of one’s stereotypes of the out-group. How we interpret success and failure for individuals is deeply shaped by in-group/out-group dynamics. When the in-group is a subordinate group and the out-group is a dominant group, these dynamics may reverse. For example, being one of only of a few women in a philosophy program may exacerbate feelings of impostor syndrome for some. Instead of thinking more highly of one’s own relative epistemic standing, one inappropriately downgrades oneself and upgrades others, especially out-group members. However, aside from subordinate/dominant group dynamics, typically in-group favoritism skews epistemic assessments in favor of the in-group. In-group/out-group dynamics can alter our assessment of both in-group and out-group members, again all before we even deliberate on an individual’s knowledge, competence, or rationality. We are especially subject to these patterns of bias in the context of threat or competition, like when people disagree about an important issue.

The evidence reviewed indicates that, all else equal, when we enter into a significant disagreement with someone, we upgrade ourselves and downgrade the other person. This upgrading and downgrading is more dramatic when we disagree with an out-group member (presumably especially so when the disagreement is relevant to features that mark us part of different groups). To be more specific, it is likely that when someone is modestly epistemically superior and part of our relevant in-group, we will regard her as an epistemic peer. Given the effects of in-group/out-group bias, we are likely to regard a moderately epistemically superior *out-group* member as an epistemic peer or even inferior. When someone in fact is an epistemic peer, we are likely to regard her as an epistemic inferior and potentially a more distant epistemic inferior when she is part of a relevant out-group. It is possible to see patterns of other-upgrading, as well. Members of subordinate groups in certain domains may erroneously regard an individual from a dominant group as epistemically better off than they are.

Bringing this back to the discussion of the epistemology of peer disagreement, it is very clear that our initial judgments of who is an epistemic inferior/peer/superior should not be taken at face value. Debates about what to do when someone we acknowledge to be an epistemic peer disagrees with us should not start with the presumption that we are accurate judges of relative epistemic status because data about mentalizing suggest that our judgments likely are idiosyncratic, biased, and inconsistent. We could reframe the debate to be about what to do when we disagree with someone whom we acknowledge to be *and actually* *is* an epistemic peer. This move retains all the interesting issues about self-trust, the rationality of retaining/reducing confidence in your own beliefs, the appropriate weight of your own judgments, etc. However, it comes at the cost of significantly reducing the relevance of the debate. Disagreeing with someone who is, and we acknowledge to be, an epistemic peer is pretty rare by all indications. Given that, the debate would lose some of its interestingness. It would not have the purported wide-reaching implications for the rationality of retaining our controversial beliefs.

Alternatively, we could fold these phenomena into our discussion of the epistemology of purported peer disagreement. That is, we can ask what the rational thing to do is when someone you take to be an epistemic peer disagrees with you, knowing that we have these patterns of upgrading/downgrading our own and others’ epistemic status. Having the debate on these terms seems to tilt the verdict heavily toward the conciliatory view. That is, knowing what we know about how we evaluate our own and others’ knowledge, competence, and rationality, when we have a disagreement with someone we take to be an epistemic peer, we ought to reduce confidence in our judgments.[[8]](#footnote-8) This seems to me to be the right verdict.

The broad conception of mentalizing turns out to be quite relevant to this debate in social epistemology. The narrow conception – one that focuses only on propositional attitude attribution – is of some interest, of course. Attributing true or false beliefs is relevant to questions of peer disagreement. However, the broad conception of mentalizing shows us how a host of mentalistic inferences serve as input to and affect the processing of our interpretation of others. Thus, with a broader conception of mentalizing in hand, we can pick up on important patterns in how we judge others’ knowledge, competence, and rationality.

## 3.2 Combatting Testimonial Injustice

Epistemic injustice is a topic at the intersection of ethics, epistemology, and feminist philosophy (Fricker 2007). Epistemic injustice occurs when potential knowers are denied epistemic resources to express themselves or credibility. Testimonial injustice, a form of epistemic injustice that is especially relevant to our discussion, occurs when a hearer’s prejudices discount a speaker’s credibility. That is, a hearer accords a speaker less knowledge, competence, and rationality simply in virtue of the speaker’s social group(s). For instance, evidence suggests that medical professionals more frequently discount reports of pain or adverse symptoms when expressed by Black female patients than other demographic groups. That is, Black female patients’ reports of pain and discomfort are accorded less credibility, which has predictable adverse outcomes for their health (Hoffman et al. 2016). This is a characteristic example of testimonial injustice.

Many theorists writing about testimonial injustice are well aware of the effects of stereotypes and character trait attributions on how hearers assess the knowledge, competence, rationality, and trustworthiness of speakers (Fricker 2007, Medina 2013, Goguen 2016). Unlike peer disagreement, the discussion of testimonial injustice (and epistemic injustice more generally) is anchored in these messy real-world mentalistic attributions. My focus in this section will be on how to think about interventions on testimonial injustice in light of what we know about how mentalizing works.

From our discussion thus far, it is evident that there are several impediments to combatting testimonial injustice. Sometimes biases – like the habitual and reflexive attribution of stereotypes or the split-second character trait attributions – are difficult to control because they are extremely fast. Within 100 milliseconds of seeing a face, we can sort people by age, gender, and race (Ito, Thompson, and Cacioppo 2004, Liu, Harris, and Kanwisher 2002). We spontaneously infer personality traits like trustworthiness, competence, aggressiveness, and dominance within a tenth of a second to a second and a half (Olivola and Todorov 2010, Rule, Ambady, and Adams Jr 2009, Malle and Holbrook 2012, Todorov and Uleman 2003). The attribution of stereotypes, probed by Implicit Association Task and various priming tasks, happen extremely rapidly, as well.[[9]](#footnote-9) Almost immediately upon on seeing a face or a body, we attribute character traits and mentalistically loaded stereotypes which influence our instinctual judgments of epistemic credibility. The speed of these attributions makes them difficult to notice and directly control.

However, our biases influence deliberative reasoning, as well. Slowing down and effortfully trying to notice the biases you are bringing into your assessment of others is not a simple, easy way to eliminate epistemic injustice. This leads us to a further reason that these biases are difficult to notice and control. Character trait and stereotype attributions are seamlessly woven into our perceptions of social interactions, whether they are rapid and effortless or slow and deliberative. How does this work?

On my view of mentalizing, Model Theory, we understand and interact in social interactions by rapidly deploying folk psychological models of people and situations (Spaulding 2018, 62-79). These models can be more or less elaborate, depending on our relationship with the target and our interest in the interaction. The models are a framework that help us make sense of social interactions. They take as input our background knowledge about norms, people, situations, and institutions, our stereotypes, and biases. This background information narrows the possible interpretations of the social interaction. When we are mentalizing, we do not consider all possible interpretations of a social interaction. That would be impossibly slow for real-life interactions and a costly use of cognitive resources. Instead, based on that background information we simply assume a relevant framework, and this makes a few possible interpretations of the social interaction salient for us. The result is that we simply *see* some people as more authoritative, competent, and trustworthy than others (Fricker 2007, ch. 2). When we are challenged on our perceptions of authority, competence, and lack thereof, we will sincerely reference the features we have noticed that justify these perceptions. Though we may in some cases question our judgments if they have been hasty or if someone can show that they do not fit the facts, in most cases the fact that the judgments make sense of what we are perceiving will lead us to double down on that epistemic judgment. In Model Theory terms, the judgments make sense in the context of the framework we are using because the framework provides the interpretations. And unless you can show someone that the framework is mistaken or inappropriate in this context, they will continue to find their judgments quite reasonable.[[10]](#footnote-10)

Model Theory is not the only account of mentalizing out there, of course. But it does a nice job of tying together various aspects of mentalizing. And, for our purposes here, it helps to explain some of the hurdles for successfully intervening on testimonial injustice. Suppose someone goes to mandatory diversity/inclusion/implicit bias training, and at this training they are told that they are likely to be biased against members of marginalized groups. This idea that we are biased often is met with resistance by participants. In most cases, people do not see themselves as biased. Even when they care about being unbiased and fair, they think they are seeing the world as it actually is because the vast majority of the information they attend to and call on to make their judgments seems to confirm their reasoning. One can reflect on the character traits they have attributed or the familiarity of the script for this type of situation or this type of person. But such solitary introspection will likely make their judgments seem quite reasonable. Indeed, it may make them *more* confident in their judgments. This is even more likely to be the case when the question “Am I being biased?” is sparked by an accusation of bias, which introduces a host of defensive psychological responses. Thus, even when individuals care about not being biased, it is difficult for them to see the bias.

Suppose though that you convince someone that they are exhibiting bias – or at least that they could be without realizing it. That knowledge itself will not root out bias. To take an analogy, suppose you finally convince your friend that smoking cigarettes is bad for his health and he really ought to try to quit. That’s great. He’s now on board. But simply knowing that smoking is bad for his health is not that helpful on its own. It does not help him actually deal with cravings in all the various situations in which they arise, e.g., morning coffee, work breaks, having a beer with friends, being under extreme stress, grieving, etc. He needs more than just knowledge and motivation. He needs tools – a lot of tools – to actually do the continual work to quit. The same goes for combatting bias. It’s not enough to convince people that they could be biased. We need to equip them with the tools for combatting bias in the real world, in all the various situations in which it is likely to arise. To root out patterns of biased epistemic judgments, we need new scripts or frameworks for understanding situations, interactions, and people. And that does not come from introspection. That comes from education and social interactions with people who can vividly portray different scripts and help us deploy and elaborate different folk psychological models.

But even that is not enough. Failure is almost inevitable when trying to change negative psychological patterns, so you need the psychological tools to deal with the failure and not give up. This is clear in our cultural conception of quitting smoking. We understand that smoking is a habit, and that changing the habit requires constant work (especially at the beginning), that failure is common, and successfully changing the habit requires getting back on the wagon after you fall off. Biases, like those manifested in testimonial injustice, are bad habits of the mind, as well. In order to change them, we need (1) to know they are problem for us, (2) care about changing them, (3) have various tools to instill new, better psychological habits, (4) understand that it will take a long time, may be riddled with mistakes, but it will get easier for us with practice.

Now that we have a grasp on the psychological processes underlying testimonial injustice and some of the challenges in intervening on such biases, we can see that some well-meaning, plausible-sounding interventions simply will not work. For example, mandatory diversity/inclusion/implicit bias training is common but ineffective. It backfires for the individuals who do not want to be there and is usually too short and limited to effectively change psychological habits. Confrontations, or callouts, are also common but ineffective. They are too isolated, too likely to spark defensive psychological mechanism, and do little to instill better psychological habits. Simply having people take an implicit bias test (typically an Implicit Association Task) is also unlikely to help because the results may spark defensive responses – e.g., they may not believe the results or they may argue that their “bias” is just responsive to reality – and it does not provide tools for better psychological habits in the real world (Hyde 2016).

Many well-meaning interventions do not actually support long-term change in psychological habits. So, what does work? I will focus on one methodology for decreasing bias that has been shown to be effective over the long-term. The approach is Patricia Devine’s prejudice habit-breaking intervention (Devine et al. 2012, Carnes et al. 2015, Cox and Devine 2019). There are five key elements of the intervention:

1. Emphasizing sincere personal values that oppose bias
2. Making individuals aware of their vulnerability to unintentional biases
3. Emphasizing the consequences of unintentional biases
4. Teaching effective bias reduction strategies
5. Teaching self-sustained effort over time to reduce the influence of unintentional biases

The intervention starts by appealing to one’s non-prejudiced values of fairness and equality, which are essential to motivate an individual’s sincere personal desire to behave in unbiased way. Next, the intervention creates cognitive dissonance for the individuals by contrasting their sincere personal values and sense of self with evidence that they may be vulnerable to unintentionally expressing bias and thereby harming members of marginalized groups. They do this in multiple ways. The individuals will take an Implicit Association Test, and they will discuss those results in the rest of the intervention. They read simple stories that vividly portray how negative stereotypes, like those tested in the Implicit Association Test, creep in and unfairly prejudice our judgments about others. The idea is to get them to sit with the discomfort of having non-prejudicial values and specific evidence they are vulnerable to acting in prejudiced ways. Making the intervention tied to specific evidence like this makes it more effective. Further, the intervention emphasizes that unintentional bias arises from associations, learned through socialization and experiences, that become automatically activated. The participants explore various “bias constructs” and the subtle ways in which they may manifest, giving the participants bias literacy that they can use to discuss various forms of bias. The prejudice habit-breaking model depicts unintentional bias as a common, bad habit of the mind, a framework that helps to avoid defensive responses that often result from other bias interventions.

The next part of the intervention focuses on providing a toolkit of effective, evidence-based bias reduction strategies. These include stereotype replacement, counterstereotype imagery, perspective taking, individuation, increasing intergroup contact, considering situational explanations for behavior, modifying your environment, and committing to judging others by their credentials. These strategies are described in simple language and the participants discuss how to apply the strategies in different situations, both proactively and in the moment a biased thought or expression of bias occurs. In general, the idea is to teach participants how to slow down, rely less on gut instincts, and employ various strategies for combatting the bad habit of prejudicial bias. They are taught that breaking the prejudice habit, like breaking any habit, does not happen right away, requires sustained effort, and it is likely that people will make mistakes. But, again like changing any habit, when people make a mistake, they have to go back to the strategies that work, and eventually they make progress and the new habits start to take hold. Setting the expectations this way helps to encourage resilience for when individuals face setbacks.

Unlike many well-meaning interventions, the prejudice habit-breaking intervention is proven to be effective in long-term reduction of bias (Carnes et al. 2015, Devine et al. 2012). The habit model of bias and the prejudice habit-breaking intervention fit well with the Model Theory of mentalizing I described above. Model Theory helps us see how bias is activated in our perceptions of social interactions and it highlights some of the difficulties of constructing interventions. The prejudice habit-breaking intervention gives concrete guidance on how to effectively combat bias. In broad strokes, both theories conceive of the problem of bias as the habitual deployment of faulty frameworks and the challenge is to learn to question those frameworks and effectively deploy better frameworks for understanding social interactions. Thus, reflection on these two theories helps us achieve a fuller understanding of the challenges and solutions to combatting bias, including testimonial injustice.

# 4. Conclusion

Judging others’ knowledge and competence plays a key role in the discussion of some topics in social epistemology. Here I discussed how these judgments play a role in the epistemology of peer disagreement and epistemic injustice. Central to both philosophical debates is the issue of how we judge others’ knowledge, intelligence, reasoning abilities, biases, etc. Judging a target’s knowledge and competence relative to our own is rife with mentalizing. Judgments that someone is trustworthy, competent, knows what they are talking about, etc. are based on factual information of course, but also character trait inferences, stereotypes, epistemic states like knowledge and belief. The mentalizing phenomena I describe in this paper illuminate how we make these judgments about others’ knowledge and competence, when to doubt our judgments, and provides some guidance on the kinds of interventions that are likely to succeed.

# **References**

Ames, D. L. 2004a. "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. . 2004b. "Inside the mind reader's tool kit: projection and stereotyping in mental state inference." *Journal of Personality and Social Psychology* 87 (3):340-353.

Andrews, Kristin. 2008. "It’s in your nature: A pluralistic folk psychology." *Synthese* 165 (1):13-29.

Bo, Sune, Carla Sharp, Peter Fonagy, and Mickey Kongerslev. 2017. "Hypermentalizing, attachment, and epistemic trust in adolescent BPD: Clinical illustrations." *Personality Disorders: Theory, Research, and Treatment* 8 (2):172-182. doi: 10.1037/per0000161.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2019. "What do implicit measures measure?" *Wiley Interdisciplinary Reviews: Cognitive Science*:e1501.

Brownstein, Michael, Alex Madva, and Bertram Gawronski. 2020. "Understanding implicit bias: Putting the criticism into perspective." *Pacific Philosophical Quarterly*.

Buchak, Lara. forthcoming. "A Faithful Response to Disagreement." *The Philosophical Review*.

Buckwalter, Wesley, and Jonathan Schaffer. 2015. "Knowledge, stakes, and mistakes." *Noûs* 49 (2):201-234.

Carnes, Molly, Patricia G. Devine, Linda Baier Manwell, Angela Byars-Winston, Eve Fine, Cecilia E. Ford, Patrick Forscher, Carol Isaac, Anna Kaatz, Wairimu Magua, Mari Palta, and Jennifer Sheridan. 2015. "The Effect of an Intervention to Break the Gender Bias Habit for Faculty at One Institution: A Cluster Randomized, Controlled Trial." *Academic Medicine* 90 (2):221-230. doi: 10.1097/acm.0000000000000552.

Christensen, David. 2007. "Epistemology of disagreement: The good news." *The Philosophical Review*:187-217.

Cikara, M, E Bruneau, JJ Van Bavel, and R Saxe. 2014. "Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses." *Journal of Experimental Social Psychology* 55:110-125.

Cox, William T. L., and Patricia G. Devine. 2019. "12 - The prejudice habit-breaking intervention: An empowerment-based confrontation approach." In *Confronting Prejudice and Discrimination*, edited by Robyn K. Mallett and Margo J. Monteith, 249-274. Academic Press.

Cuddy, Amy JC, Michael I Norton, and Susan T Fiske. 2005. "This old stereotype: The pervasiveness and persistence of the elderly stereotype." *Journal of social issues* 61 (2):267-285.

Del Pinal, G., and S. Spaulding. 2018. "Conceptual centrality and implicit bias." *Mind & Language* 33 (1):95-111.

Dennett, Daniel C. 1978. "Beliefs about beliefs [P&W, SR&B]." *Behavioral and Brain Sciences* 1 (4):568-570.

Devine, Patricia G, Patrick S Forscher, Anthony J Austin, and William TL Cox. 2012. "Long-term reduction in implicit race bias: A prejudice habit-breaking intervention." *Journal of experimental social psychology* 48 (6):1267-1278.

Fiske, Susan T, Amy JC Cuddy, and Peter Glick. 2007. "Universal dimensions of social cognition: Warmth and competence." *Trends in cognitive sciences* 11 (2):77-83.

Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*: Oxford University Press Oxford.

Gerken, Mikkel. 2017. *On folk epistemology: How we think and talk about knowledge*: Oxford University Press.

Gigerenzer, Gerd, and Reinhard Selten. 2002. *Bounded rationality: The adaptive toolbox*: MIT press.

Gilbert, Daniel T, and J Gregory Hixon. 1991. "The trouble of thinking: activation and application of stereotypic beliefs." *Journal of Personality and Social Psychology* 60 (4):509.

Goguen, Stacey. 2016. "Stereotype Threat, Epistemic Injustice, and Rationality." In *Implicit Bias and Philosophy*, edited by Michael Brownstein and Jennifer Saul, 216-237. Oxford University Press.

Hoffman, Kelly M, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. 2016. "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites." *Proceedings of the National Academy of Sciences* 113 (16):4296-4301.

Hummert, Mary Lee, Teri A Garstka, Jaye L Shaner, and Sharon Strahm. 1994. "Stereotypes of the elderly held by young, middle-aged, and elderly adults." *Journal of Gerontology* 49 (5):P240-P249.

Hyde, Krista. 2016. "Testimonial Injustice and Mindreading." *Hypatia* 31 (4):858-873.

Ito, Tiffany A, Erin Thompson, and John T Cacioppo. 2004. "Tracking the Timecourse of Social Perception: The Effects of Racial Cues on Event-Related Brain Potentials." *Personality and Social Psychology Bulletin* 30 (10):1267-1280.

Kelly, Thomas. 2011. "Peer disagreement and higher order evidence." In *Social epistemology: Essential readings*, edited by Alvin Goldman and Dennis Whitcomb, 183-217. Oxford: Oxford University Press.

Knobe, J., and S. Nichols. 2017. Experimental Philosophy. In *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta.

Korman, Joanna, and Bertram F Malle. 2016. "Grasping for traits or reasons? How people grapple with puzzling social behaviors." *Personality and Social Psychology Bulletin* 42 (11):1451-1465.

Kruger, Justin, and David Dunning. 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.

Lackey, Jennifer. 2010. "What Should We Do When We Disagree?" In *Oxford Studies in Epistemology*, edited by T. S. Gendler and J. Hawthorne, 274-293. Oxford: Oxford University Press.

Liberman, Z., A. L. Woodward, and K. D. Kinzler. 2017. "The origins of social categorization." *Trends in Cognitive Sciences* 21 (7):556-568.

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

Malle, B. F., and J. Holbrook. 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.

Matheson, Jonathan. 2015. *The epistemic significance of disagreement*: Palgrave Macmillan.

McGeer, V. 2007. "The regulative dimension of folk psychology." In *Folk psychology re-assessed*, 137-156. Springer.

Medina, J. 2013. *The Epistemology of Resistance: Gender and Racial Oppression, Epistemic Injustice, and the Social Imagination*: Oxford University Press.

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

Nagel, Jennifer. 2017. "Factive and nonfactive mental state attribution." *Mind & Language* 32 (5):525-544.

Nickerson, Raymond S. 1998. "Confirmation bias: A ubiquitous phenomenon in many guises." *Review of general psychology* 2 (2):175-220.

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

Palomares, J. K. S., and A. W. Young. 2018. "Facial first impressions of partner preference traits: Trustworthiness, status, and attractiveness." *Social Psychological and Personality Science* 9 (8). doi: 10.1177/1948550617732388.

Phillips, Jonathan, Wesley Buckwalter, Fiery Cushman, Ori Friedman, Alia Martin, John Turri, Laurie Santos, and Joshua Knobe. 2020. "Knowledge before belief." *Behavioral and Brain Sciences*:1-37.

Pinillos, Ángel. 2012. "Knowledge Experiments and Practical Interests." In *Knowledge Ascriptions*, edited by J. Brown and M. Gerken, 192-221. Oxford University Press.

Prochazkova, Eliska, Luisa Prochazkova, Michael Rojek Giffin, H. Steven Scholte, Carsten K. W. De Dreu, and Mariska E. Kret. 2018. "Pupil mimicry promotes trust through the theory-of-mind network." *Proceedings of the National Academy of Sciences* 115 (31):E7265-E7274. doi: 10.1073/pnas.1803916115.

Pronin, Emily. 2007. "Perception and misperception of bias in human judgment." *Trends in cognitive sciences* 11 (1):37-43.

Pronin, Emily, Daniel Y Lin, and Lee Ross. 2002. "The bias blind spot: Perceptions of bias in self versus others." *Personality and Social Psychology Bulletin* 28 (3):369-381.

Rasmussen, Mattias Skipper, Asbjørn Steglich-Petersen, and Jens Christian Bjerring. 2018. "A higher-order approach to disagreement." *Episteme* 15 (1):80-100.

Rotenberg, Ken J., Serena Petrocchi, Flavia Lecciso, and Antonella Marchetti. 2015. "The Relation Between Children's Trust Beliefs and Theory of Mind Abilities." *Infant and Child Development* 24 (2):206-214. doi: <https://doi.org/10.1002/icd.1891>.

Rudman, L. A. , A. G. Greenwald, D. S. Mellott, and J. Schwartz. 1999. "Measuring the automatic components of prejudice: Flexibility and generality of the Implicit Association Test." *Social cognition* 17 (4):437-465.

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

Spaulding, S. 2016. "Mind misreading." *Philosohical Issues* 26 (1):422-440.

Spaulding, S. 2017. "Do you see what I see? How social differences influence mindreading." *Synthese*.

Spaulding, S. 2018. *How We Understand Others: Philosophy and Social Cognition*. New York, NY: Routledge.

Spaulding, S. 2019. "What is Mindreading?" *Wiley Interdisciplinary Reviews: Cognitive Science* 11 (3).

Spaulding, S. 2021. "Beliefs and Biases." *Synthese*. doi: <https://doi.org/10.1007/s11229-021-03129-0>.

Sripada, Chandra Sehkar, Mike Angstadt, Sarah Banks, Pradeep J Nathan, Israel Liberzon, and K Luan Phan. 2009. "Functional neuroimaging of mentalizing during the trust game in social anxiety disorder." *Neuroreport* 20 (11):984.

Sripada, Chandra Sekhar, and Jason Stanley. 2012. "Empirical tests of interest-relative invariantism." *Episteme* 9 (1):3-26.

Tajfel, H. 1974. "Social identity and intergroup behaviour." *Social Science Information* 13 (2):65-93.

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

Turri, John. 2017. "Knowledge Attributions and Behavioral Predictions." *Cognitive Science* 41 (8):2253-2261.

Uleman, James S, S Adil Saribay, and Celia M Gonzalez. 2008. "Spontaneous inferences, implicit impressions, and implicit theories." *Annu. Rev. Psychol.* 59:329-360.

Vavova, Katia. 2014. "Confidence, evidence, and disagreement." *Erkenntnis* 79 (1):173-183.

Viale, Riccardo. 2020. *Routledge Handbook of Bounded Rationality*: Routledge.

Wedgwood, Ralph. 2010. "The Moral Evil Demons." In *Disagreement*, edited by R. Feldman and T. Warfield, 216-246. Oxford: Oxford University Press.

Wellman, Henry M, and David Liu. 2004. "Scaling of theory‐of‐mind tasks." *Child development* 75 (2):523-541.

Wellman, Henry M., Fuxi Fang, and Candida C. Peterson. 2011. "Sequential progressions in a theory-of-mind scale: longitudinal perspectives." *Child development* 82 (3):780-792. doi: 10.1111/j.1467-8624.2011.01583.x.

Westra, E. 2017a. "Character and theory of mind: an integrative approach." *Philosophical Studies* 175 (5):1217-1241.

Westra, E. 2017b. "Stereotypes, theory of mind, and the action-prediction hierarchy." *Synthese* 196 (7):2821-2846.

Westra, Evan, and Jennifer Nagel. 2021. "Mindreading in conversation." *Cognition* 210:104618. doi: <https://doi.org/10.1016/j.cognition.2021.104618>.

Wheeler, Mary E, and Susan T Fiske. 2005. "Controlling Racial Prejudice: Social-Cognitive Goals Affect Amygdala and Stereotype Activation." *Psychological Science* 16 (1):56-63.

Willis, J., and A. Todorov. 2006. "First impressions: Making up your mind after a 100-ms exposure to a face." *Psychological Science* 17 (7):592-598. doi: 10.1111/j.1467-9280.2006.01750.x.

Wimmer, H., and J. Perner. 1983. "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception." *Cognition* 13 (1):103-128.

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

1. Parts of this chapter are based on my book (Spaulding 2018) and a recent paper in the journal *WIREs Cognitive Science* (Spaulding 2019). [↑](#footnote-ref-1)
2. Though see Buckwalter and Schaffer (2015) for a critique of this idea. On their view, it’s not the stakes so much as increased salience that explains our willingness to attribute mere belief as opposed to knowledge. [↑](#footnote-ref-2)
3. See also Jennifer Nagel’s (2017) work on the distinction between attributing factive states (like knowledge) and non-factive states (like true and false belief). She argues that reconceiving mentalizing along these lines (factive vs. non-factive) can help us make sense of some puzzling patterns of mentalizing successes and failures in children and non-human animals. [↑](#footnote-ref-3)
4. Explicitly inferring propositional attitudes and explicitly inferring character traits may serve different cognitive functions. Some evidence suggests that we offer explicitly propositional-attitude based explanations when we encounter puzzling behavior and trait-based explanations when we encounter expected or typical behavior (Korman and Malle 2016). [↑](#footnote-ref-4)
5. See Gigerenzer and Selten (2002) and, more recently, Viale (2020) for a general review of bounded rationality and the use of heuristics. [↑](#footnote-ref-5)
6. This pattern of upgrading our own knowledge and competence and downgrading others’ knowledge and competence is amplified in domains in which we are unskilled or uneducated. In these contexts, we often do not recognize our own incompetence and ignorance, and we fail to recognize others’ equal or superior knowledge and competence (Kruger and Dunning 1999). [↑](#footnote-ref-6)
7. Should you need an example to help you imagine this situation, consider a retiree who goes back to college to pursue a postsecondary degree in a cohort of young, fresh-out-of-college graduate students. [↑](#footnote-ref-7)
8. Above I articulated a scenario in which someone regards an epistemic inferior as an epistemic peer (or epistemic peer as an epistemic superior). This is an interesting case because the subject ought to be more confident in her own judgments, however because of impostor syndrome or stereotype threat she has a very difficult time doing this. She is already inclined to conciliation because of the underlying processes that result in her downgrading her own relative epistemic status. The rational thing to do in this kind of case if you *knew* that you were in this kind of case would be to remain steadfast, but as matter of psychological reality you will not. [↑](#footnote-ref-8)
9. Implicit bias tests, like the IAT, have faced a fair amount of criticism over the last decade. See Brownstein, Madva, and Gawronski (2019, 2020) for review. Some of these critiques are fair, e.g., cautioning against interpreting scores as measures of an individual’s biases. However, I think once we have a more fine-grained understanding of what mental representations these tests are tapping into, the dismissals of such methodologies seem unwarranted. See Del Pinal and Spaulding (2018) and Spaulding (2021). [↑](#footnote-ref-9)
10. I elaborate this line of reasoning more in Spaulding (2018), chapter 6 and Spaulding (2016). [↑](#footnote-ref-10)