Do You See What I See? How Social Differences Influence Mindreading

Shannon Spaulding

Forthcoming in Synthese. Please cite final version.

Abstract:
Disagreeing with others about how to interpret a social interaction is a common occurrence. We often find ourselves offering divergent interpretations of others' motives, intentions, beliefs, and emotions. Remarkably, philosophical accounts of how we understand others do not explain, or even attempt to explain such disagreements. I argue these disparities in social interpretation stem, in large part, from the effect of social categorization and our goals in social interactions, phenomena long studied by social psychologists. I argue we ought to expand our accounts of how we understand others in order to accommodate these data and explain how such profound disagreements arise amongst informed, rational, well-meaning individuals.

1. Introduction[1]

On July 17, 2014 in Staten Island, NY, Eric Garner was stopped by police officers allegedly for selling loose cigarettes. After conversing with five police officers for several minutes, one police officer grabbed Garner's wrist and Garner swatted away his hand. That police officer then put his arm around Garner's neck and pulled him backwards onto the ground, putting him in a chokehold position. Once Garner was

[1] I have talked about the ideas in this paper with many people. I am particularly grateful for my conversations with the following people: Lauren Ashwell, Mikkel Gerken, Brie Gertler, Suilin Lavelle, Karen Neander, Carlotta Pavese, Guillermo del Pinal, Sarah Robins, Armin Schulz, Robert Thompson, Evan Westra, and Tad Zawidzki. Thanks also to the audiences at Coastal Carolina University, Mississippi State University, University of Houston, George Washington University, and University of Kansas. Finally, thanks to two anonymous referees at this journal. Their feedback helped me develop my critiques and and positive arguments.

on the ground, the other four police officers helped to restrain him by kneeling on his back and cuffing his hands behind his back. Over the course of several minutes while lying face down on the sidewalk, Garner repeated 11 times "I can't breathe." The police officers did nothing to help him breathe. After Garner lost consciousness, the police officers turned him on his side. The ambulance arrived 7 minutes later. They did not perform CPR or administer oxygen. Garner was transported to the hospital and pronounced dead on arrival. The medical examiners determined that the cause of death was the police officer's chokehold and subsequent restraint, and the death was ruled a homicide. The entire event was captured on video.

Despite having a clear video recording of the interaction between Garner and the police officers, interpretations of this event varied widely. Some interpreted Garner's behavior as innocuous and viewed the police officers' behavior as unjustified police aggression. Others interpreted Garner's behavior as threatening and viewed the police officers' behavior as an appropriate, justified response to a dangerous situation. This disagreement was widely discussed online and on television in the aftermath of the event itself and after a grand jury decided not to indict the officer who put Garner in a chokehold. The radio program This American Life (episode 548) documented two people watching the video together to see if they could find common ground. One person was a producer for This American Life who had worked in a correctional facility, and the other person was her friend, a New York City police officer. The producer is Black, and the police officer is White.

After a four-hour conversation, much to their dismay, the two women could find almost no common ground in their interpretation of the event.[2]

How can there be such strong disagreement among those who saw the very same video? Many hostile encounters between police and citizens are unrecorded, and both police advocates and civil rights leaders say that things would be much clearer if we could just see exactly what happened. The disagreements over the Garner case suggest that that is not true. Disagreements persist despite clear video evidence. How can this be?

Interpreting the encounter between Garner and the police involves evaluating the behaviors, for example, as aggressive or unthreatening, inferring the motives of Garner and the police officers, judging what the police knew and what Garner was trying to do. This kind of social interpretation is a task of *theory of mind*. Theory of mind is the capacity to understand people in terms of mental states. A central part of theory of mind is *mindreading*, the process of attributing mental states to others, e.g., beliefs, desires, intentions, and emotions, in order to interpret and anticipate their behavior.

People starkly disagree about how to mindread Garner and the police officers, despite the seemingly unambiguous video evidence. For example, in the conversation reported in the This American Life episode, the police officer interpreted Garner as *refusing* to be arrested and *stalling* the police. Even when he was lying on the ground, the police officer interpreted Garner as *willfully* disobeying the police's instructions. In contrast, the producer interpreted Garner as *trying* to

[2] You can access this episode at http://www.thisamericanlife.org/radio-archives/episode/548/cops-see-it-differently-part-two?act=0#play.

reason with the police, not resisting. And when Garner was lying on the ground, the producer could see him only as thoroughly detained and *sincerely* reporting that he cannot breathe. She asked, "Don't you see that this person is in pain struggling to communicate?" The police officer and producer disagreed about Garner's mental states, e.g., what he desired and intended. Both women were frustrated by their disagreement because they were convinced that their own interpretation obviously was correct.[3]

This divergence in interpretations of Garner's mental states raises the question: How do we attribute mental states to others? Philosophers have converged on a general answer, viz., through a range of processes involving simulation and theorizing (Goldman, 2006; Nichols & Stich, 2003). Such hybrid accounts of mindreading aim to explain the various strategies we employ to understand others' behavior. Remarkably, however, these accounts of mindreading

[3] Of course, not all disagreements about this and other such cases stem solely from disagreements about the mental states of the observed agents. Disagreements have many causes, e.g., misinformation, limited knowledge, false beliefs. Three kinds of disagreement stand out as theoretically interesting, though: those that result from (1) disagreements about mental states (about what someone was thinking, feeling, or intending), (2) disagreements about norms (about what is appropriate to think and do in a given situation), and (3) disagreements about both mental states and norms. I take it that many disagreements about the Garner case and others like it are of the third sort. Disagreeing observers tend to have different opinions about how citizens and police ought to behave, but they also disagree about what Garner and the police were thinking and trying to do, e.g., whether a police officer felt mortally threatened or about what a citizen was intending to do. The commentary from This American Life illustrates this well. In these cases, disagreements about norms and mental states often are intertwined and difficult to pull apart. This is one reason that conversations about such cases are so difficult. I focus on the different inferences about mental states because, unlike disagreements about norms, this has received relatively little attention in the analyses of these deep-rooted disagreements. Thanks to a reviewer for highlighting norms as an additional source of disagreement in these cases.

do not explain, or even attempt to explain, deep disagreements about the interpretation of social interactions.[4] Reading the philosophical literature on mindreading, one easily could come away with the impression that neurotypical adults rarely disagree about what others are thinking, feeling, intending, etc. The Garner case is a tragic and dramatic reminder that this is not true. Despite having access to the same evidence, intelligent, rational, well-meaning people profoundly disagree about how to interpret this situation. In fact, such disagreements over social interpretation are common in our everyday lives, both in cases where the events are well documented and in cases where they are not. I shall argue that figuring out *how* these disagreements arise illuminates aspects of mindreading that have received little attention from philosophers.

As it stands, contemporary philosophical accounts of mindreading do not address these profound disagreements about social interpretation. We can rectify this problem by expanding accounts of mindreading to incorporate psychological phenomena long studied by social psychologists. I will argue that the kind of disparity in mental state attributions present in the Garner case stems, in large part, from the effect of social categorization and the ways in which our goals influence our approaches in a social interaction. These phenomena play an important role in shaping our social interpretations. Specifically, I will argue that they influence both the input and processing of mindreading. With such an expansion we can explain these disagreements and offer a more complete, accurate account of mindreading.

---

[4] In fact, neither of the two Stanford Encyclopedia of Philosophy entries on mindreading address such disagreements (Gordon, 2009; Ravenscroft, 2010).

There is a high theoretical payoff for examining the intersection of theory of mind and social psychology. But there is more than just a theoretical payoff. It is important for us to understand how informed, well meaning individuals can disagree so profoundly about the interpretation of a social interaction. Understanding the factors that shape our social interpretations may help us understand one another better, which is essential for flourishing in diverse, multicultural societies.

2. Theory of Mind

In this section, I shall give a brief introduction to theory of mind. This introduction serves two purposes: It shows what the theory of mind literature focuses on and by implication what it does not focus on, and it sets the stage for my critique of the theory of mind literature later in the paper.

We are intensely social creatures, and as such theory of mind is central to our everyday lives. The developmental psychologist Henry Wellman offers the following compelling description of the role of theory of mind in our lives.

> We humans live socially—raised by parents, in familial communities constantly interacting with, caring about, and working with other people. We not only live socially, we think socially—developing and depending on extensive knowledge about social life, social entities (persons, friends, rivals, clans, families), as well as social actions and interactions (loving, aggressing, advising). Conceivably, the vast array of social cognition that humans acquire could be a loosely connected, even disconnected, array of separate facts, ideas, and conventional truisms. But the claim behind the phrase theory of mind is that human social cognition is founded on an understanding of ourselves and others in terms of our inner, mental, psychological states (Wellman, 2014, p. 2)

Two competing accounts of mindreading have dominated the theory of mind literature: the Theory Theory (TT) and the Simulation Theory (ST). The TT holds that we explain and predict behavior by employing a tacit folk psychological theory about how mental states inform behavior (Carruthers & Smith, 1996; Davies & Stone, 1995a; Nichols & Stich, 2003). With our folk psychological theory, we infer from a target's behavior what his or her mental states probably are, and thereby explain the behavior. From these inferences, plus the psychological principles in the theory connecting mental states to behavior, we predict the target's behavior.

The ST, in contrast, holds that we explain and predict a target's behavior by using our own minds as a simulation of the other person's mind (Currie & Ravenscroft, 2002; Davies & Stone, 1995b; Gordon, 1992; Heal, 1998). To explain a target's behavior, we put ourselves in another's shoes, so to speak, and imagine what our mental states would be and how we would behave if we were that agent in that particular situation. To predict a target's behavior, we take the attributed mental states as input and simulate the target's decision about what to do next.

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

The TT, ST, and the various hybrid theories focus on when, how, and what mental states we attribute to others, i.e., attributional cognitive processes. Debates within theory of mind concern theorizing vs. simulational mindreading (Davies & Stone, 1995a, 1995b), the ontogenetic development of theory of mind (Apperly & Butterfill, 2009), phylogenetic development of theory of mind (Andrews, 2008; Lurz, 2009; Premack & Woodruff, 1978), the kinds of mental state concepts required for mature theory of mind (Apperly & Butterfill, 2009; Goldman, 2000), the extent to which theory of mind relies on self knowledge (Carruthers, 2011; Goldman, 2006; Gordon, 1995), the explanation of pretense (Carruthers, 2006; Nichols & Stich, 2000), and a host of other related behaviors. Each of these debates is about the cognitive processes that make successful, accurate mindreading possible.

Mindreading is an impressive and incredibly useful ability. Given how different each person's experiences are and how complex the social world is, it is remarkable that we ever generate similar and accurate mentalistic explanations and predictions of behavior. Thus, it is natural to frame the investigation of mindreading around explaining this impressive feat, as the theory of mind literature has. However, given this way of framing the investigation of mindreading, these debates rarely focus on individual differences in mindreading. Indeed, even outsiders' critiques of the theory of mind literature do not address the kinds of disagreements that are common in our everyday lives (Gallagher, 2005; Hutto, 2008). As a result, individuals' deep disagreements about the interpretation of social interactions simply do not come up in these discussions. This, I shall argue below, is a serious

problem. These disagreements reveal that philosophers have missed important facts about how mindreading works.

Social psychologists have investigated how such disagreements arise,[5] and yet philosophers have paid little attention to this. The phenomena studied by social psychologists clearly are relevant to when, how, and what mental states we attribute to others. For example, categorizing people as part of our in-group or part of an out-group affects the extent to which we attribute to them secondary emotions and complex mental states. When we judge someone to be different from us, e.g., someone who has a different race, age, and socioeconomic status, we tend to attribute simpler, baser motives and emotions to that person. However, when we judge someone to be similar to us, e.g., someone who shares our race, age, and socioeconomic status, we usually do not to caricature this person's mental states. Instead, we tend to project our own motivations, emotions, and beliefs to that person. Perceived similarity plays an important role in how we understand others.[6]

---

[5] In a foundational study on this topic, Hastorf and Cantril (1954) found that students from rival universities interpreted a video of a football game between the rivals dramatically differently. Disagreements emerged over whether the game was played fairly, which team played dirty, whether particular charges were justified, whether a non-call was justified, the proportion of infractions the other team made, etc. As any sports fan knows, one's allegiance to a team colors one's interpretation of what happens in the game. For excellent contemporary social psychology research, see Bertram Malle (2004), Daniel Ames (2004a), Susan Fiske and colleagues (2005), Nicholas Epley and Adam Waytz (2010), John Barresi (2004), and Jay Van Bavel and colleagues (2014). Not all of these are disagreements are about mental states – some are simply disagreements about the facts – but many of the disagreements are about mental states, e.g., whether the referees are biased and whether an interaction was incidental contact or targeting.

[6] Though it is presented in different terms, this is one reason why civil rights advocates urge that police departments' racial makeup be similar to the communities they police.

Thus, the phenomena studied by social psychology clearly are relevant to philosophical discussions of mindreading.

Philosophers do not systematically examine how these social psychology findings bear on theory of mind.[7] These data play almost no role in constraining our accounts of how we understand other people. The participants in the various philosophical debates described above rarely, if at all, discuss these data and how they influence mindreading. Presumably this segregation between theory of mind and social psychological phenomena began as a benign simplification. In order to build a theory of how we attribute mental states, theorists have to abstract away from some of the messy empirical details. But the field has advanced dramatically, and accounts of theory of mind aim to explain the subtle details of real-life, complex human social behavior. For example, philosophical debates concern the real life conditions under which we confabulate (Carruthers, 2009), children's successes and failures in figuring out when someone is pretending (Nichols & Stich, 2000), the idiosyncratic mindreading mistakes of people with autism (McGeer, 2001), etc. Thus, philosophical discussions of mindreading nowadays focus on the messy empirical details, but they neglect the deep mindreading disagreements common among neurotypical adults. I aim to rectify this mistake in this paper.

The next sections discuss two phenomena in social psychology: social categorization and the goals and corresponding approaches in social interaction. I

[7] Perhaps one reason for this is that social psychologists often are interested in global character traits whereas philosophers studying mindreading are interested in our ability to infer specific propositional attitudes. However, as I argue in the main text, these social psychological phenomena are highly relevant to how we infer specific propositional attitudes. Thanks to a reviewer for pointing out this difference between the two literatures.

shall argue that these phenomena affect both the input and processing of mindreading. Specifically, social categorization clarifies how we narrow down the available social information and, as a result, the sort of information that serves as input for mindreading. The goals and approaches of social interaction reveal the mindreading processes we employ under various conditions. These findings are crucial for understanding when, how, and what mental states we attribute to others, and thus any adequate account of how we understand others must take them into consideration.

3. Social Categorization

Social categorization – sorting people, behaviors, and events into social categories – is essential for successful navigation of the social world. It helps make the social world more comprehensible and predictable, and thereby allows us to manipulate the social world for our purposes. We would be hopelessly lost without the ability to detect patterns and categorize people, situations, and events.

In social interactions, we reflexively and rapidly sort people into categories. For example, within 100 milliseconds of seeing a face, we can sort people by age, gender, and race (Ito, Thompson, & Cacioppo, 2004; Liu, Harris, & Kanwisher, 2002). We tend to sort people by the most salient category, and age, gender, and race often are the most salient categories.[8] However, the categories we employ can be

[8] Race is salient in multicultural societies. However, in a racially uniform society in which one never encounters a person of another race, race would not be a salient social category.

modulated by cognitive load, task, and context (Gilbert & Hixon, 1991; Wheeler & Fiske, 2005). Context, cognitive load, and our goals influence what is salient to us, and what is salient determines which social categories we employ when we reflexively sort people into social groups. Thus, social categories are activated reflexively, but which ones are activated depends on the context.[9]

On the basis of rapid facial recognition, we spontaneously infer personality traits, as well, e.g., trustworthiness, competence, aggressiveness, dominance (Olivola & Todorov, 2010; Rule, Ambady, & Adams Jr, 2009). Of course these may be inaccurate inferences. Nevertheless, we do make these inferences, and we do so very quickly even when we are under cognitive load. (Malle & Holbrook, 2012; Todorov & Uleman, 2003).[10] Inferences about personality traits play an important role in the mental states we attribute to people and our interpretation of their behavior.

The categories we use to sort people are associated implicitly with specific characteristics. For example, we associate old and incompetent, female and warm,

---

[9] Social psychologists call social categorization conditionally automatic (Macrae & Bodenhausen, 2000). The term "automatic" often invites confusion because people use it many different ways. In this context, automatic means reflexive or spontaneous. It does not mean fast, innate, subconscious, etc. Thus, social categorization is *automatic* in the sense that it is reflexive, and it is *conditionally* automatic in the sense that the categories we employ are conditional on the context. So-called conditionally automatic social categorization is compatible with sorting by some categories faster than others – indeed, the evidence is that we sort faces by age, race, and gender faster than other categories. It is also compatible with it taking longer for us to process the entire stereotype associated with a category than it takes to sort people into basic categories like age, race, and gender. Thanks to a reviewer for pushing me to clarify these issues.

[10] The speed of spontaneous trait inferences is a disputed matter. For example, Todorov and Uleman (2003) report that spontaneous trait inferences occur as quickly as 100 milliseconds after exposure to a face. In contrast, Malle and Holbrook (2012) find that spontaneous trait inferences occur within 1400 to 1600 milliseconds, depending on the task and type of stimulus. In either case, spontaneous personality trait inferences occur very rapidly in social interactions.

baby-face and unthreatening. These associations are the sort of thing probed by the Implicit Association Task (IAT) (Greenwald, McGhee, & Schwartz, 1998; Greenwald, Poehlman, Uhlmann, & Banaji, 2009). IAT measures how quickly and accurately subjects categorize stereotypic and counter-stereotypic associations, e.g., Black :: gun, White :: cellphone vs. Black :: cellphone, White :: gun. The task measures the strength of a person's implicit associations.[11]

In addition to – and on the basis of – social categorization and personality trait inferences, we quickly identify people as part of our in-group or as part of an out-group (Tajfel, 1974). In-grouping and out-grouping appear to be a function of perceived similarity (Ames, 2004a, 2004b). That is, those who we judge to be like us are categorized as part of in our in-group, and those who we judge to be unlike us are categorized as part of an out-group. Perceived similarity is a subjective, contextually relative, and sometimes idiosyncratic judgment, not an objective measure of actual similarity.

In heterogeneous societies, age, race, and gender are salient features of people, and thus one tends to identify people who share one's age, race, or gender as part of one's in-group. However, social categorization extends beyond these classifications. People have multiple, overlapping identities, and perceived similarity

[11] There is no consensus on how to interpret what IAT and other such tasks are measuring. Some philosophers interpret IAT and other such tasks as measuring of our *implicit attitudes,* e.g., Gendler (2008) and (Mandelbaum, 2015). On that interpretation, IAT reveals that despite explicit egalitarian attitudes toward Blacks and Whites, many White people have implicit White supremacist attitudes. In contrast, Levy (2014) argues that IAT measures our *patchy endorsements*, which is something more fragmented than ordinary beliefs. Machery (2016) offers an alternative interpretation according to which IAT and other such tasks measure *traits* rather than attitudes. See Del Pinal and Spaulding (in progress) for an alternative interpretation of what IAT and other implicit bias tasks are measuring.

is relative to a context. Who counts as part of one's in-group varies depending on the social context. For example, in a context where hobbies are salient, only runners count as part of my in-group, and non-runners are the out-group. However, when political views are salient, only liberals are part of my in-group. In this context, both runners and non-runners may be categorized as part of my political in-group. Moreover, in categorizing individuals into in-groups and out-groups, we gloss over differences within the groups and exaggerate differences between the groups (Linville, Fischer, & Salovey, 1989; Mullen & Hu, 1989). Thus, my runner friends may not be as similar, to me and to each other, as I assume, and my runner and non-runner friends may not be so different as I assume.

These patterns are interesting in and of themselves, but they are important in this discussion because we treat in-group members quite differently from out-group members. We usually have more favorable attitudes toward and empathize more with in-group members, especially people who share our gender, race, age, religion, or nationality than toward people do not share these features (Rudman, Greenwald, Mellott, & Schwartz, 1999). By making idiosyncratic features artificially salient, experimenters can elicit in-group favoritism even for very minimal, arbitrary groups (Ashburn-Nardo, Voils, & Monteith, 2001). The effects of in-group favoritism especially are strong in a context of competition or threat (Cikara, Bruneau, Van Bavel, & Saxe, 2014).

More troublingly, we regard in-group members as more human, in a sense, than out-group members (Hackel et al., 2014). We regard them as more capable of experiencing secondary emotions (such as pride and guilt) and as having richer,

more complex mental experiences than out-group members (Haslam, 2006). We tend to attribute more simplistic, caricatured mental states to those we perceive to be unlike us. At the extreme, people tend to dehumanize those individuals who are perceived to be least like them, e.g., the homeless and drug addicts (Harris & Fiske, 2006).

In sum, social categorization influences our interpretation of others' behaviors, personality traits, and our expectations of what they will do next. On the basis of perceived similarity, we judge others as part of an in-group or an out-group, and in-group/out-group status significantly affects when, how, and what mental states we attribute to others. Individuals from different social backgrounds will tend to have different implicit associations, spontaneous personality trait inferences, and judgments of perceived similarity, which will result different patterns of social categorization. These patterns of social categorization can be modulated but only if one is aware of the patterns and their effect and one has sufficient cognitive resources to detect and revise these inferences and judgments. This discussion of social categorization demonstrates how our social differences influence our social interpretations.

4. Goals and Approaches in Social Interaction

We have various goals in social interactions. The following is not an exhaustive list, but it covers several of the most common goals. Sometimes we aim for accuracy, especially when something important depends on getting it right, when we will be

held responsible for our interpretation of the interaction, or when the situation is unusual and unexpected (Fiske & Neuberg, 1990; Kelley, 1973; Tetlock, 1992). However, often times we aim for efficiency primarily and accuracy only secondarily. When the social interaction seems ordinary and familiar, when not much hangs on it, or when we are otherwise cognitively taxed, we use cognitive shortcuts, e.g., stereotypes and projection (Fiske & Taylor, 2013, pp. 177-199). Another cluster of goals within social interaction includes anxiety reduction, self-esteem, and confirmation of one's worldview. In these cases, our inferences, in one way or another, serve self-interested purposes (Dunning, 1999; Kunda, 1990). These three goals are not mutually exclusive or exhaustive. We have, to varying degrees, each of these goals with respect to different aspects of a social interaction. Moreover, our goals can change within a social interaction, and there can be tradeoffs among these different goals. In what follows, I shall address these goals as distinct perspectives for simplicity and ease of explication, but this is just a theoretical simplification. In real social interactions, these goals dynamically interact with each other.

Corresponding to this array of goals are various approaches for social interpretation.[12] With the goal of accuracy, we tend to search for relevant

---

[12] An approach for social interpretation is like a strategy that need not be conscious or deliberate. Each of these goals and corresponding approaches may be conscious or non-conscious. Like the goals that they correspond to, these approaches are not mutually exclusive. Social interactions often are complex diachronic events, and we may adopt multiple mindreading approaches when interpreting a social interaction. Many mindreading episodes involve some deliberation (for salient aspects of the situation that we want to get right), heuristics (for aspects of the situation that seem familiar or unimportant to us), and self-interested biases (for aspects of the situation that may threaten our self-image or ideology). Moreover, these approaches may interact in the sense that self-interested biases influence our careful deliberation, heuristics inform and influence the judgments we make when

information in a controlled and deliberative fashion. For example, when members of a job search committee make judgments about the candidates (e.g., who will be a friendly colleague, who will be willing to serve on committees, who will want to stay for the long term), the stakes are high. Thoughtful members of the committee will want to ensure that their judgments are accurate, consider all the relevant evidence, and make sure their decision is not based on merely superficial cues. This kind of reasoning is effortful, cognitively taxing, and difficult if one is under cognitive load or not well practiced in this kind of reflective reasoning (Gilbert, Krull, & Pelham, 1988).

When aiming for efficiency, there are several approaches to social interpretation. For example, when we perceive an individual to be *similar* to us in some salient respect, we often simply project our own mental states to that individual (Ames, 2004a, 2004b). We also use our mental states as an anchor and adjust the interpretation based on how similar the individual is to us. These are egocentric heuristics. They generate the "curse of knowledge," a phenomenon wherein we falsely assume that others know what we know, and the "false consensus effect," which occurs when we falsely assume that others share our opinion on some matter (Clement & Krueger, 2002; Epley & Waytz, 2010, p. 512).

When we perceive an individual to be *different* from us, we tend to take alternative efficient approaches. Often we use stereotypes about the individual's salient in-group (Ames, 2004a; Krueger, 1998; Vorauer, Hunter, Main, & Roy, 2000). Stereotypes may be positive, negative, or neutral beliefs about some group. These

---

deliberating, and careful deliberation may correct the heuristics we use and combat the pull of self-interested biases.

are efficient approaches because they rely on long-term memory rather than working memory, which is quite limited and already taxed by other online processes (Baddeley, 2012). Reliance on stereotypes is a shortcut that reduces cognitive load. Moreover, once the stereotype is activated, processing stereotype-consistent information is less cognitively demanding than processing stereotype-inconsistent information. When subjects are under cognitive load, they will use the most efficient approaches, e.g., employing a stereotype and attending to stereotype-inconsistent information only if it is highly salient (Gilbert & Hixon, 1991).

We tend to use efficient approaches like stereotyping and projection when we are in very familiar situations. As particular situations become more familiar to us, the interpretation of those situations will become more accessible and more difficult to override (Higgins, King, & Mavin, 1982). The tendency to habitually code situations and others' behavior in a particular way can become proceduralized. Well-practiced judgments make social interpretation easier, more efficient, and more predictable, but they preempt equally reasonable but less practiced judgments (Smith, 1990). Thus, in very familiar situations it may be difficult for people *not* to employ particular stereotypes or project their own perspective.

The two previous types of approaches differ with respect to how effortful they are. The approaches we take when we have self-serving goals may be effortful or efficient. Consider first the Self-Serving Attributional Bias, which describes our tendency to take credit for success and deny responsibility for failure. (Miller & Ross,

1975).[13] For example, we often attribute our successes to some internal factor, e.g., diligence or talent, and attribute our failures to external mitigating factors, e.g., bad luck or bias. In this way, we come to feel good about our successes and brush off our failures.

This is relevant to social interpretation because this pattern is found for judging in-group and out-group behaviors, as well. This is called the Group-Serving Attributional Bias (Brewer & Brown, 1998; Pettigrew, 1979). One tends to judge the success of an out-group to be the result of situational factors and the failure of an out-group as the result of internal factors, whereas one judges the success of one's in-group to be the result of internal factors and the failure of one's in-group to be the result of situational factors. This pattern is very clear in sports fans. When my sports team wins it is because they are talented and hard working, but when my sports team loses it is because they were off their game, the other team got lucky, and the referees were biased against my team. One explains a target's behavior differently depending on whether the target is part of one's in-group or an out-group.

Another approach corresponding to self-serving goals is called Naïve Realism. It describes the tendency to regard others as more susceptible to bias and misperception than oneself (Pronin, Lin, & Ross, 2002). We think we simply see

---

[13] The Self-Serving Attributional Bias is distinct from the Actor-Observer Effect, which holds that people explain others' behavior in terms of dispositional factors and their own in terms of situational factors. In other words, behavioral explanations differ depending on whether one is the actor or the observer. The empirical evidence for this effect is mixed (Malle, 2006), but see Malle, Knobe, and Nelson (2007) for a novel interpretation of the asymmetries in patterns of explanation.

things as they are, but others suffer from bias. This tendency is prevalent in interactions in which people disagree, e.g., political debates. One tends to regard those of a different political party as likely to be misguided and biased by their personal motivations, whereas one regards oneself (and to some extent other members of one's political party) simply as correct. This self-serving approach influences how we interpret others' mental states and behavior, especially out-group members who disagree with us about some salient issue.

Finally, confirmation bias describes a general tendency to consider only information that confirms one's preconceived ideas, prior knowledge, and relevant expertise and interpret ambiguous information in light of this prior information. With respect to social cognition, we have preconceived ideas, prior knowledge, and some expertise about other individuals and groups, and we tend to interpret social interactions in terms of this prior information. For example, racists notice when individuals behave in ways that confirm their racist beliefs but they often do not attend to the many cases in which individuals act in ways that disconfirm their racist beliefs. Confirmation bias affects both deliberative, controlled processes and efficient processes like stereotyping. It occurs regardless of how the prior information originated, how likely it is to be true, and whether accuracy is incentivized (Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992; Snyder, Campbell, & Preston, 1982).

Sometimes, of course, information glaringly contradicts our preconceived ideas and we have to reconcile the information with our ideas. However, social interactions can be ambiguous, and the interpretation of these ambiguous

interactions differs greatly by individuals who have radically different preconceived ideas. Thus, when we are called to interpret an ambiguous social interaction, we tend to consider only information that confirms our ideas. The result is consistent validation of one's worldview.

In summary, the various goals we have in social interactions determine the various approaches we take to understand other people. Sometimes we have the motivation and ability to exhaustively review the available social information and attribute mental states to others in that way. Other times we take shortcuts because we lack the motivation or ability to do an exhaustive search. The shortcuts we take partly are a function of social categorization: we are more likely to project our own mental states on those we perceive to be similar to us and stereotype those we perceive to be different from us. In addition, many of our social interpretations are guided by self-interest, and as a result we interpret others' behavior and mental states in light of what we antecedently believe or want to believe. Thus, our interpretation of a social interaction is influenced by our goals and whether the interaction involves in-group or out-group members. It is important to note again that these approaches are not mutually exclusive and often co-occur in an episode of social interpretation. Deliberation can influence and be influenced by heuristics and self-interested biases. One can start see how two people who differ in their goals and judgments of perceived similarity might come away with profoundly different interpretations of the same evidence.

5. Implications for Mindreading

The literature on how we understand others started in earnest nearly 40 years ago with the publication of Premack and Woodruff's (1978) seminal article on theory of mind. Since then, research on theory of mind has flourished in developmental and comparative psychology, linguistics, cognitive neuroscience, and philosophy. Above I described just a few of the many ongoing philosophical debates about how we understand others' minds and behavior. These debates generally embrace the nuanced empirical details about how mindreading operates, but the social psychological phenomena discussed above strangely are left out of these discussions.[14]

In the philosophy literature, the two most authoritative books on theory of mind are Shaun Nichols and Stephen Stich's (2003) *Mindreading* and Alvin Goldman's (2006) *Simulating Minds*. Nichols and Stich do not discuss any of these social psychology findings in their book. Goldman discusses only egocentric biases (as evidence that we use our own minds to simulate others). In these two pillars of the literature, there is no systematic discussion of the ways in which social categorization and goals and approaches of social interactions. These books are not alone in neglecting the way in which these phenomena constrain and distort mindreading; the field as a whole overlooks these findings.[15]

---

[14] Developmental psychologists working in theory of mind face this same lacuna (Apperly, 2012; Rakoczy, 2014). For example, Ian Apperly, writes, "although it has long been recognized in principle that there should be important links between ToM and research on social psychology, reasoning, and experimental pragmatics, these literatures have seldom meshed well in practice" (2012, p. 837).
[15] See, for example, Carruthers (2011); Carruthers and Smith (1996); Davies and Stone (1995a, 1995b); Gordon (2009); Ravenscroft (2010).

This fact is puzzling because the social psychological phenomena discussed in the previous sections have significant implications for how we attribute mental states to others in order to understand their behavior. Specifically, two factors determine the output of our mindreading processes: the input and the processing.[16] Social categorization influences the input to mindreading, and our goals and approaches influence how we process this input. These findings are directly relevant to debates about how mindreading is achieved. I address input in the next section and processing in the following section.

5.1 Mindreading Input

In complex, dynamic social situations, the amount of available information is enormous. There is too much information for human beings to process, and we attend to only a tiny portion of that information. Furthermore, social situations can be ambiguous, especially when they involve people outside one's close circle of family and friends. Sometimes this ambiguity is superficial insofar as one is not in a position to know the correct interpretation but with sufficient investigation could figure it out. Other times, however, the ambiguity is deep in the sense that one simply cannot know the right interpretation.

[16] In what follows, I draw broadly on Nicholas Epley's framework (2008), which holds that the accuracy of our mindreading attributions depends on what we take as input and how we process that information. One difference between his framework and mine is that he argues that we tend to reason about others' mental states by reasoning about our own mental states first, and this serves as an anchor that we may subsequently adjust with deliberation. In contrast, I argue that when we perceive others to be different from ourselves, we tend not to use an egocentric anchor. In such cases, we rely on our stereotypes of members of that out-group.

The abundance of information and ambiguity of social situations highlight the fact that one does not simply read off the social facts from the environment. We take as input some subset of the available information and base mental state attributions on this information. If we want to given a complete account of how we attribute mental states, we have to understand which information is taken as input for mindreading.

There are some universal patterns regarding what is salient to us, e.g., biological movement and faces are especially salient. But we attend to more than just biological movement and faces. We also attend to social features. The social features that are salient to us depend on our situation, expectations, and goals (Fiske & Taylor, 2013, pp. 66-68). And it is here that the data from social psychology are relevant.

Social categorization is especially important for understanding what gets taken as input for mindreading. Studies on social categorization reveal that we spontaneously and rapidly sort people by the most salient features, usually gender, age, and race. Social categories involve implicit associations, e.g., old and incompetent, female and warm. Thus, social categories and the features we implicitly associate with these categories are salient to us and hence can serve as input to mindreading. In other words, these features are the basis for mental state attributions, which we tend to use to interpret and anticipate others' behavior.[17]

---

[17] Social categorization may shape the inputs to mindreading, as I argue in this section. However, social categorization may also run in parallel with mindreading, and in some cases social categorization and mindreading may influence each other. In the latter case, mindreading a target may cause us to re-categorize a target. In this

We do not regard all individuals as the same. Rather, on the basis of perceived similarity, which is relative to a context, we spontaneously and rapidly classify individuals as like-us or not like-us. When we sort people into in-groups or out-groups, we tend to ignore heterogeneity within groups and exaggerate differences between groups. In other words, we attend to the features that mark an individual as part of an in-group or out-group in a certain context. These features are salient and serve as input for mindreading.

In addition to social categories and group status, we also attend to others' behaviors. But we do not simply read off the meaning of others' behaviors from the world. Behavior interpretation is modulated by the situational context. Whether we evaluate behavior as normal, funny, aggressive, or rude depends on the situational context. On the basis of behavior evaluation, we infer individuals' personality traits, e.g., friendly, pretentious, or cheerful. The features of a situation that we notice are a function of the situational context, which makes certain interpretations more accessible to us. Thus, to understand the input to mindreading, one has to understand how the situational context modulates behavior evaluation, personality trait inferences, and accessible interpretations.[18]

Social categorization sheds light on some of the ways in which we narrow down the available social information and, as a result, the sort of information that serves as input for mindreading. Philosophical discussions of theory of mind tend to focus on what is common among everyone's mindreading processes, but in doing so

---

kind of case, mindreading serves as a corrective for our categorizing. Thanks to a reviewer for pointing out this possibility.

[18] For more on how the situational context modulates our social interpretations, see Spaulding (2017).

these discussions miss out on interesting, important, and dramatic differences. We do not all attend to the same features of a social interaction, nor do we interpret these features in the same way. Our subjective sense of perceived similarity to others in the social interaction, implicit associations, and situational context influence the features, concepts, and interpretations that are to salient to us. Individuals who differ in their social characteristics will tend to take different information as input to mindreading, and as a result they will tend to generate different social interpretations. This is not simply a matter of a theoretical discussion not capturing the messy, empirical details. Individuals with different social backgrounds generate different mindreading judgments in predictable ways, and the mindreading literature simply fails to detect these patterns and the disagreements they generate.

This discussion highlights the fact that even if two people are looking at the same video, if the individuals differ with respect to these social psychological phenomena they will tend to interpret the situation very differently. How one individual interprets an event will be quite different from how the other person interprets it. This is dramatically evinced by the divergent interpretations of the Garner video. Though it is difficult to say with any precision exactly what features of a situation two particular individuals are taking as input for mindreading, we can take the disagreement reported on This American Life as a schematic example of how such disagreements arise. The producer for This American Life likely identifies with Garner as part of her racial in-group, whereas the police officer likely identifies with the other police offers as part of her in-group. As explained above, this

subjective sense of perceived similarity influences how one sees and interprets a target's behavior. Furthermore, the past experiences and narratives familiar to the producer and the police officer likely differ, so they will notice different features of the situation and interpret those features in light of narratives familiar to them. Though the producer and police officer are looking at the same interaction, they are attending to different aspects of the situation and interpreting those aspects differently. This is just a sketch of an explanation of how differences in social backgrounds influence individuals' social interpretation. To fully understand when, how, and what mental states we attribute to others – and to make sense of profound disagreements about social interpretation – one must understand how these social psychological determine the inputs to mindreading.[19]

5.2 Mindreading Processing

Two factors are crucial in understanding the output of mindreading: the input and the processing (Epley, 2008). The previous section argues that social psychological data shed light on how we narrow down the available information and what gets taken as input to mindreading. This section addresses how we process that information. The data on goals and approaches of social interactions are especially relevant here.

---

[19] One could understand this discussion of the inputs to mindreading in terms of multiple mindreading systems, e.g., Apperly and Butterfill (2009). On such a view, the social categorization processes I describe may serve as input to either low-level/perceptual/system-1 processes or high-level/inferential/system-2 processes. Thanks to a reviewer for suggesting this idea.

Depending on the goals that we have for social interactions, we will process information to different degrees. To the extent that accuracy is an important goal (e.g., when we will be held responsible for our attribution), the depth of processing will be relatively deep. We will tend to deliberate for longer, in a controlled fashion, and about more aspects of the interaction. We may reflect on our personality trait inferences, behavior evaluations, our own perspectives, and our mental state attributions and adjust them according to how well they fit the situation as a whole. This mindreading strategy will be effective to the extent that we are well practiced in this type of reasoning and not under serious cognitive load (Spaulding, 2016).

However, to the extent that efficiency is a primary goal (e.g., when we lack the motivation to deliberate), the depth of processing will be more limited. In these cases, we use projection and stereotypes and engage in very little adjustment. These heuristics are reliable when we accurately estimate the degree of similarity between the target and ourselves. Efficiency goals limit the depth of processing, e.g., the adjustment of our mental state inferences. Thus, if we have misjudged the degree of similarity, these heuristics are likely to lead to error and be corrected only by deliberative processing (Spaulding, 2016).

Finally, we often are motivated by self interest in social interactions, and accordingly we will tend to limit the depth of processing of information indicating that we are failures, biased, or have false beliefs about the world. Of course, as noted above, these goals and approaches are not mutually exclusive, and many social events will involve interaction and tradeoffs among these different goals and approaches. To take just a couple examples, one could have the goal of accurate

mindreading but also have the deliberative processes warped by self- and group-enhancing biases, or one's interpretation based on stereotyping could be further cemented by confirmation bias.

Extant theories of mindreading do not predict the ways in which our goals restrict the depth of processing with respect to mental state inferences. In many cases, theories of mindreading assume that our goals in social interactions are the same in every circumstance. However, these data show that our goals differ in various situations and as a result the mindreading strategies we adopt differ accordingly. In fact, it is even more complicated than that. In each situation, we have various goals that affect how we interpret various aspects of the situation. As a result, deliberation, heuristics, and self-interested biases interact and influence each other. Understanding the output – the mental state attributions – requires understanding how our goals shape our search for information

To be sure, each of the existing theories of mindreading could integrate some of these data. For instance, Theory Theorists will find evidence for their view in accuracy-oriented approaches. Theory Theorists argue that the process of inferring an agent's mental states relies on a rich body of information. On the basis of our folk psychological knowledge, situational knowledge, and beliefs about the target (including group membership, personal history, etc.), we make an abductive inference about the agent's mental states. Although the theorizing of TT need not be conscious or explicit, when it is conscious it resembles the careful deliberation that we engage in when we have accuracy as a goal. When accuracy is a goal, we try to take into consideration various aspects of the situation and use this information to

come up with the right interpretation of the target's mental states. Thus, Theory Theorists easily can incorporate the data on accuracy-oriented approaches.

In contrast, Simulation Theorists will find evidence for their view in one kind of efficiency-oriented approach, namely projection. The ST holds that we understand others by mentally putting ourselves in a target's situation and figuring out what we would think, feel, and do in that situation. This is similar to the approach we take when we prioritize efficiency in mindreading, regard the target as relevantly similar to us, and project our own mental states to a target. In the latter kind of case, we attribute to the target the beliefs, desires, emotions, and intentions that we think we would have in that situation. Clearly Simulation Theorists easily can incorporate these data into their theory.

The information presented highlights some of the strengths of both Theory Theory and Simulation Theory. Both theories clearly are getting *something* right about mindreading. The empirical evidence suggests that we do deliberate in some circumstances, and we do project in some circumstances. However, clearly the monolithic TT and ST are getting something wrong too, because we do not *only* deliberate and we do not *only* project. Hybrid theorists who posit a conjunction of theoretical and simulational processes will find the most vindication here, as a central finding is that we use a variety of mindreading strategies. The data on deliberation and projection can be integrated into hybrid theories fairly straightforwardly.

However, as they stand, TT, ST, and the various hybrid theories do not integrate much of this evidence. In particular, none of these theories explain or

predict how stereotypes and self-interest biases influence our mental state attributions. This is important because stereotypes and self-interest biases are pervasive. One may wonder whether one of these theories would have an easier time making room for the data on stereotypes and self-interested biases. It seems to me that there is more flexibility in the TT and TT-oriented hybrid approaches. Because TT posits an information-rich mindreading process, part of this information base could be beliefs about how various groups of people behave in particular situations, i.e., stereotypes. Furthermore, there is a rich literature on how various cognitive biases influence deliberation, and we could add this kind of data to TT and TT-oriented hybrid approaches. The ST may not as easily integrate stereotypes about how various groups of people behave in particular situations. Such an addition would seem ad hoc given that ST's purported theoretical advantage over TT is that it holds that mindreaders do not rely on a rich body of information when interpreting and predicting others' behavior. The ST may be able to include background information about cognitive biases, e.g., about how we selectively ignore information that we are failures, biased, or have false beliefs about the world. These could be construed as constraints on the simulation process. Finally, a hybrid ST theory could incorporate data on how we sometimes deliberate carefully and make an inference to the best explanation of what a target's mental states are.

The previous paragraph sketches a few ways (some ad hoc, some more principled) in which existing mindreading theories could integrate some of the data I have discussed here. However, the implications of my arguments are deeper than simply pointing out unexplained data and divvying up the evidence in favor of one

mindreading theory over another. The evidence and arguments presented here provide the connective tissue to unify theories of mindreading.

Contemporary hybrid accounts, e.g., Goldman (2006) and Nichols and Stich (2003), gesture at the plurality of mindreading processes by noting that we theorize *and* simulate, but in order for an account of how we attribute mental states to be unified and informative it has to do more than simply posit a conjunction of processes. The existing models of how we understand others do not tell us much about the sort of information that gets taken as input for these approaches, how the information received as input is processed, how deeply we process this information with each of these approaches, when and why we shift between these mindreading approaches, and how accurate any of them are (Spaulding, 2016). Rather, they simply posit a motley collection of psychological processes with little explanation of how, when, and why we use these different psychological processes in social cognition. The discussion here helps fill in some of these gaps. The arguments and data I present explain how our goals in a social interaction determine our approaches, how these approaches limit or enhance the depth of processing, and the ways in which our subjective perceived similarity judgments influence information processing. One theoretical upshot of this discussion is that identifying how these social psychological phenomena influence how we attribute mental states to others strengthens hybrid theories by filling in these gaps, and it moves them closer to the goal of a unified, coherent account of how we understand others.

From the previous section, we learned that individuals with different social characteristics tend to take different information as input. Individuals' goals are a

further source of divergence. Even if individuals with different social backgrounds have the same goals in a social interaction, their approaches will differ when the interaction involves in-group or out-group members and when individuals feel threatened in some way by the interaction. If individuals have different goals – as is often the case in real world social interpretation – not only will they take different information as input, they will process that information very differently as well.

I do not claim that individuals who have different stereotypes and make different perceived similarity judgments *never* converge on the same interpretation. They can converge on the same interpretation, especially when they both have accuracy as a goal and deliberate carefully about the situation and their interpretations. However, this kind of dialogue is effortful and it requires recognizing the ways in which your perspective is idiosyncratically shaped by your social background. Moreover, careful deliberation is a not a guarantee of convergence among informed, rational, well-meaning individuals. The conversation between the This American Life producer and police officer is an example of this. Despite their explicit goal of coming to some consensus about the Eric Garner video, their deliberative mindreading yielded different results. Even if two people are presented with the same evidence, what they notice and how they approach the interaction may be very different, they may have different self-interested biases shaping their interpretation, and therefore their interpretations may be very different as well. Putting all of this together, we can see precisely how rational, well-meaning individuals can come to profoundly different interpretations of a social interaction.

6. Conclusion

Philosophical accounts of mindreading tend to present a simplified picture of how we understand others that suggests that individuals take the same information as input, process this information in a similar way, and generate very similar mindreading judgments. Because of this simplified picture, these accounts do not explain or predict the sort of divergent social interpretation common in our everyday lives and dramatically evinced in the Eric Garner case. The arguments advanced here show that this simplified picture is inaccurate. Social categorization influences what we attend to and how we interpret social interactions. Categorizing individuals, behaviors, and events depends on the situational context, our experiences, and our expectations, and these may differ predictably and dramatically among individuals with different social backgrounds. Augmenting existing accounts of mindreading with these data will allow us to explain how social disagreements arise and provide us with a more realist, accurate account of how we understand others. The arguments highlight a further important theoretical benefit of examining these data: we now have the resources to construct a unified account of mindreading. Instead of simply presenting a motley conjunction of mindreading processes, now we are in a better position to explain more of the mindreading processes we employ, the conditions under which we use them, and how information is processed with these various approaches. This is the goal of

philosophical accounts of mindreading, and these arguments bring us a step closer to realizing this goal.

I began this paper with a particularly striking example of diverging mental state attributions: disagreements about the lethal encounter between Eric Garner and Staten Island police officers. Despite the seemingly unambiguous video evidence, rational, informed, well-meaning individuals differ in their social interpretation of this encounter. Accounts of how we understand others do not explain or predict these divergent interpretations. A theoretical payoff of the discussion in this paper is that we now have the resources to explain how this and other such disagreements can arise among informed, well-intentioned people. Furthermore, supplementing existing theory of mind accounts in the way I suggested yields a more realistic (and perhaps more depressing) picture of the process of social interpretation.

There is more than just theoretical payoff, though. We often find ourselves disagreeing with family, friends, neighbors, and strangers on social media about the social interpretation of some behavior. Especially in contentious disagreements, like the social interpretation of Eric Garner's death, our response to people who disagree with us is, bluntly put, to regard them as intellectually or morally deficient. Certainly some people who disagree with us *are* intellectually and/or morally deficient, but not in every grave disagreement. Examining the ways in which backgrounds and experiences shape our interactions with others can help us pinpoint the sources of disagreement among informed, well meaning individuals. Moreover, recognizing that *everyone's* perspectives are shaped by these social psychological factors should

make one less dogmatic about one's social interpretations. Reducing dogmatism paves the way for productive conversations about contentious social interpretations. Such conversations may not resolve all disagreements or directly prevent the kind of interaction that led to Eric Garner's death, but they are a good start.

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.

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.

Andrews, K. (2008). It's in Your Nature: A Pluralistic Folk Psychology. *Synthese, 165*(1), 13-29.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *The Quarterly Journal of Experimental Psychology, 65*(5), 825-839.

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states. *Psychological Review, 116*(4), 953-970.

Ashburn-Nardo, L., Voils, C. I., & Monteith, M. J. (2001). Implicit associations as the seeds of intergroup bias: How easily do they take root? *Journal of Personality and Social Psychology, 81*(5), 789-799.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29.

Barresi, J. (2004). Intentional relations and divergent perspectives in social understanding. In S. Gallagher & S. Watson (Eds.), *Ipseity and Alterity: Interdisciplinary Approaches to Intersubjectivity* (pp. 74-99). Rouen: Presses Universitaires de Rouen.

Brewer, M. B., & Brown, R. J. (1998). Intergroup relations. In D. T. Gilbert, S. T. Fiske, & G. Lindzey (Eds.), *The Handbook of Social Psychology* (4th ed., Vol. 1-2, pp. 554-594). New York: McGraw-Hill.

Carruthers, P. (2006). Why pretend? In S. Nichols (Ed.), *The Architecture of the Imagination* (pp. 89-109). Oxford: Oxford University Press.

Carruthers, P. (2009). How we know our own minds: The relationship between mindreading and metacognition. *Behavioral and Brain Sciences, 32*(2), 1-18.

Carruthers, P. (2011). The Opacity of Mind: An Integrative Theory of Self-Knowledge. New York: Oxford University Press.

Carruthers, P., & Smith, P. K. (1996). *Theories of Theories of Mind*. Cambridge: Cambridge University Press.

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

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

Currie, G., & Ravenscroft, I. (2002). *Recreative Minds: Imagination in Philosophy and Psychology*. Oxford: Oxford University Press.

Davies, M., & Stone, T. (1995a). *Folk Psychology: The Theory of Mind Debate*. Oxford: Blackwell.

Davies, M., & Stone, T. (1995b). *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell.

Del Pinal, G., & Spaulding, S. (in progress). Conceptual Centrality and Implicit Bias.

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

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

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

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

Fiske, S. T., & Taylor, S. E. (2013). *Social cognition: From brains to culture* (2nd ed.). London: Sage.

Gallagher, S. (2005). *How the Body Shapes the Mind*: Oxford University Press, USA.

Gendler, T. S. (2008). Alief and belief. *Journal of Philosophy, 105*(10), 634-663.

Gilbert, D. T., & Hixon, J. G. (1991). The trouble of thinking: activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology, 60*(4), 509-517.

Gilbert, D. T., Krull, D. S., & Pelham, B. W. (1988). Of thoughts unspoken: Social inference and the self-regulation of behavior. *Journal of Personality and Social Psychology, 55*(5), 685-694.

Goldman, A. I. (2000). Folk psychology and mental concepts. *Proto Sociology, 14*, 4-25.

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

Gordon, R. M. (1992). The Simulation Theory: Objections and Misconceptions. *Mind & Language, 7*(1-2), 11-34.

Gordon, R. M. (1995). Simulation Without Introspection or Inference From Me to You,. In M. Davies & T. Stone (Eds.), *Mental Simulation: Evaluations and Applications* (pp. 53-67). Oxford: Blackwell.

Gordon, R. M. (2009). Folk Psychology as Mental Simulation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2009 ed.).

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464-1480.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*(1), 17-41.

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology, 52*, 15-23.

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low neuroimaging responses to extreme out-groups. *Psychological Science, 17*(10), 847-853.

Harris, L. T., Todorov, A., & Fiske, S. T. (2005). Attributions on the brain: Neuro-imaging dispositional inferences, beyond theory of mind. *NeuroImage, 28*(4), 763-769.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review, 10*(3), 252-264.

Hastorf, A. H., & Cantril, H. (1954). They saw a game; a case study. *The Journal of Abnormal and Social Psychology, 49*(1), 129-134.

Heal, J. (1998). Co-cognition and off-line simulation: Two ways of understanding the simulation approach. *Mind & Language, 13*(4), 477-498.

Higgins, E. T., King, G. A., & Mavin, G. H. (1982). Individual construct accessibility and subjective impressions and recall. *Journal of Personality and Social Psychology, 43*(1), 35-47.

Hutto, D. D. (2008). *Folk Psychological Narratives: The Sociocultural Basis of Understanding Reasons*. Cambridge, MA: MIT Press.

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

Kelley, H. H. (1973). The processes of causal attribution. *American psychologist, 28*(2), 107-128.

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

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480-498.

Levy, N. (2014). Neither Fish nor Fowl: Implicit Attitudes as Patchy Endorsements. *Nous*, 800-823.

Linville, P. W., Fischer, G. W., & Salovey, P. (1989). Perceived distributions of the characteristics of in-group and out-group members: empirical evidence and a computer simulation. *Journal of Personality and Social Psychology, 57*(2), 165-188.

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

Lurz, R. W. (2009). *The philosophy of animal minds*. Cambridge: Cambridge University Press.

Machery, E. (2016). De-Freuding Implicit Attitudes. In M. Brownstein & J. Saul (Eds.), *Implicit Bias & Philosophy* (Vol. 1, pp. 104-129). Oxford: Oxford University Press.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51*(1), 93-120.

Malle, B. F. (2004). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. Cambridge, MA: MIT Press.

Malle, B. F. (2006). The actor-observer asymmetry in attribution: a (surprising) meta-analysis. *Psychological Bulletin, 132*(6), 895-919.

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology, 102*(4), 661-684.

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: New answers to an old question. *Journal of Personality and Social Psychology, 93*(4), 491-514.

Mandelbaum, E. (2015). Attitude, Inference, Association: On the Propositional Structure of Implicit Bias. *Nous, 50*(3), 629-658.

McGeer, V. (2001). Psycho-practice, psycho-theory and the contrastive case of autism: How practices of mind become second-nature. *Journal of Consciousness Studies, 8*(5-7), 109-132.

Miller, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin, 82*(2), 213-225.

Mullen, B., & Hu, L.-t. (1989). Perceptions of ingroup and outgroup variability: A meta-analytic integration. *Basic and Applied Social Psychology, 10*(3), 233-252.

Nichols, S., & Stich, S. (2000). A cognitive theory of pretense. *Cognition, 74*(2), 115-147.

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

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

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

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences, 1*(04), 515-526.

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.

Rakoczy, H. (2014). What are the relations of thinking about groups and theory of mind? *British Journal of Developmental Psychology, 32*(3), 255-256.

Ravenscroft, I. (2010). Folk Psychology as Theory. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2010 ed.).

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

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

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

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

Smith, E. R. (1990). Content and Process Specificity in the Effects of Prior Experiences. In T. K. Srull & R. S. Wyer (Eds.), *Advances in Social Cognition* (Vol. 3, pp. 1-59). Hillsdale, NJ: Erlbaum.

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

Spaulding, S. (2016). Mind Misreading. *Philosophical Issues, 26*(1), 422-440.

Spaulding, S. (2017). How we think and act together. *Philosophical Psychology*, 1-17. doi:10.1080/09515089.2017.1295640

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

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

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

Vorauer, J. D., Hunter, A., Main, K. J., & Roy, S. A. (2000). Meta-stereotype activation: evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *Journal of Personality and Social Psychology, 78*(4), 690-707.

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*: Oxford University Press.

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