Full citation:
Tsoi, L. & Young, L. (2018). Moral reasoning. In S. Thompson-Schill (Ed.), Language & Thought. Volume 3 of The Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience (4th ed.). Editor-in-chief: J. Wixted. Wiley.

Moral reasoning

Lily Tsoi and Liane Young

BOSTON COLLEGE

Abstract

Understanding people's minds is essential for effectively navigating our social world. This chapter focuses on the capacity to attribute and reason about minds (theory of mind; ToM) and its role in moral cognition. The section on moral judgment focuses on the circumstances in which people rely on mental states for moral judgments and how ToM may differ depending on the moral domain. We provide a functional explanation for these patterns of mental state reasoning that contrasts the need to regulate interpersonal relations with the need to protect the self. The section on moral behavior focuses on interactions with moral agents (e.g., friends, foes, ingroups, outgroups). We examine how ToM is deployed during two fundamental social contexts (i.e., cooperation and competition) and elaborate on the circumstances in which people fail to consider the minds of others. We end by providing some evidence that ToM can improve interpersonal and intergroup relations.

Key Terms: morality, theory of mind, mentalizing, cooperation, competition, dehumanization, social cognition

Introduction

Understanding other minds—that other people have minds as well as the specific contents of those minds—is essential for effectively navigating our social world. People deploy their capacity for theory of mind (ToM) across many contexts, including for communication and coordination, for forming and maintaining relationships, and for explaining people's past actions and predicting people's future actions. This chapter focuses on the role of ToM in moral cognition. Indeed, a key function of ToM is for moral judgment and behavior. We are especially motivated to understand morally relevant actions, to predict people's actions when those actions affect us directly or indirectly, and to evaluate moral agents as future allies or enemies. Here we discuss the ways in which ToM is crucial for moral cognition in two parts: judgment and behavior.

In the first part of the chapter, we focus on moral judgment—judgments of right or wrong, innocent or guilty. We discuss the circumstances in which people rely on mental states for moral judgments and how ToM may differ depending on the moral domain. Ultimately, we provide a functional explanation for these patterns of mental state reasoning that contrasts the need to regulate interpersonal relations with the need to protect the self. In the second part of the chapter, we focus on moral behavior—interactions with moral agents, for example, friend or foe, acquaintance or stranger, ingroups or outgroups. We focus on how ToM is deployed during two fundamental social contexts: cooperation and competition. We describe circumstances in which people fail to consider the minds of others, honing in on outgroup dehumanization. Finally, we end by providing some evidence that ToM can improve interpersonal and intergroup relations.

How people make moral judgments

In this section, we review (1) research uncovering the role of mental state reasoning, i.e., theory of mind (ToM), in moral judgment, (2) evidence that people rely on information about mental states differently across moral domains, (3) and our proposed explanation for this difference: that mental states matter more for judgments impacting others versus the self.

Mental states matter for moral judgment

Adults make frequent, rapid inferences about mental states such as intent and desire when evaluating others' actions, especially when judging whether those actions are right or wrong (Malle & Holbrook, 2012). Generally, harms brought about intentionally are seen as worse than harms brought about accidentally. Indeed, intent is what separates murder from manslaughter and represents a main factor in determining culpability (Hart & Honoré, 1985). This heuristic applies not just to extreme acts like murder or manslaughter but also to acts that don't involve physical harm to people, such as breaking objects (Decety, Michalska, & Kinzler, 2012) or allocating money in an unfair fashion (Cushman, Dreber, Wang, & Costa, 2009). Furthermore, information about a person's mental states appears to matter even when that person isn't the one causing harm—for example, someone who bets that a natural disaster will occur is not even capable of causing the natural disaster but is nevertheless blamed more for it when it does happen (Inbar, Pizarro, & Cushman, 2012).

The capacity to explicitly represent mental states such as beliefs emerges between three and four years of age (for reviews, see Saxe, Carey, & Kanwisher, 2004; Wellman, Cross, & Watson, 2001; but see Onishi & Baillargeon, 2005). At

this age, children are able to distinguish belief from reality in location-change tasks: if Maxi puts chocolate in the kitchen cupboard, and his mother moves the chocolate to a drawer, children will correctly judge that Maxi will search in the cupboard; younger children typically fail the test and judge that Maxi will search the drawer (Wimmer & Perner, 1983). Even though children around 4 years of age are able to pass tests of false belief understanding like the one above, these children's ability to flexibly use ToM for moral judgment develops later. That is, if a person puts his cupcake inside a bag and walks away, and his friend throws away the bag into the garbage thinking it's trash, young children between three and four years old still tend to focus more on the negative outcome (e.g., wasted cupcake in the trash) and less on the lack of negative intent (e.g., she didn't know it was a cupcake; she thought it was trash); these young children therefore assign more blame for accidental acts (Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; see also Cushman, Sheketoff, Wharton, & Carey, 2013; Hebble, 1971; Shultz, Wright, & Schleifer, 1986; Yuill & Perner, 1988; Zelazo, Helwig, & Lau, 1996). Thus, young children—whose ToM capacities are just starting to develop—focus more on outcomes rather than intentions during moral judgments. Only later when the capacity to integrate belief information for moral judgment emerges do older children start to judge accidental harms less harshly (Baird & Astington, 2004; Killen et al., 2011).

While much work focuses on moral development in children (for a review, see Rottman & Young, 2015), most of the work we present in this chapter is on adult moral cognition. One effective approach to moral cognition in adults has been to systematically vary information about both intent and outcome in a 2 x 2 design (Young, Cushman, Hauser, & Saxe, 2007). In studies using this design, participants are presented with four types of vignettes: someone with a negative intention causes a negative outcome (intentional harm), someone with a negative intention causes a neutral outcome (attempted harm), someone with a neutral intention causes a negative outcome (accidental harm), and someone with a neutral intention causes a neutral outcome (neutral act). For example, if a character named Grace puts powder into her friend's coffee thinking the powder is sugar when it is in fact poison, Grace would be committing accidental harm. By contrast, if Grace thinks the powder is poison when it is in fact sugar, Grace would be committing attempted harm. Participants then judge the moral permissibility of the action.

This 2 x 2 design reveals a key behavioral pattern in adults: people weigh information about the agent's belief more heavily than information about the action's outcome when making moral judgments (Young et al., 2007). Notably, attempted harm (when the agent intended to cause harm but failed to cause harm) is judged as less morally permissible than accidental harm (when the agent caused harm but did not mean to). In fact, in many instances, failed attempts to harm are judged just as morally wrong as successful attempts. Other work has also revealed that an agent's belief about whether his or her action would cause harm is the

single most important factor in judgments of moral permissibility (Cushman, 2008).

One notable point is that mental states matter beyond judgments of moral permissibility. Evidence suggests that judgments of whether a person is good or bad (moral character), whether an act is right or wrong (moral wrongness), or whether a person deserves to be punished for his or her actions (moral punishment) all depend (to differing degrees) on mental state information (Cushman, 2015). By contrast, other factors such as the action itself or the outcome of the action are not always taken into consideration. For example, people care about the action when making judgments of wrongness but not punishment, whereas people care about the outcome of the action when making judgments of punishment but not wrongness.

Neuroimaging research has provided convergent evidence supporting the use of mental state information for moral judgment. This research builds on extensive work in social cognitive neuroscience more generally revealing brain regions that are consistently recruited for ToM. This network of brain regions, known as the ToM network, includes bilateral temporoparietal junction (TPJ), precuneus, and dorsomedial prefrontal cortex (dmPFC), and ventromedial prefrontal cortex (vmPFC) (Fletcher et al., 1995; Gallagher et al., 2000; Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007; Saxe & Kanwisher, 2003). Evidence suggests that these brain regions play a key role in moral cognition: for example, according to a recent meta-analysis of over 240 experiments, brain regions consistently involved in making moral decisions converge with regions involved in ToM (Bzdok et al., 2012).

To investigate the role of ToM regions in moral judgment, some researchers have relied on systematically varying information about mental states (using the same basic 2 x 2 design described above) and examining how neural activity within the ToM network differs when people receive different types of information (e.g., neutral vs. negative beliefs; neutral vs. negative outcomes). For the most part, this work has targeted two neural measures: response magnitude in a given region (mean level of activity averaged across voxels in a region) and spatial patterns of voxel-wise activity within a region. While activation-based univariate analyses indicate the involvement of a region in a given task (Mur, Bandettini, & Kriegeskorte, 2008), multivariate pattern analyses can reveal whether a particular feature or dimension (e.g., whether an act is intentional or accidental) within the domain of the target task (e.g., theory of mind) is encoded in a region (e.g., Koster-Hale, Saxe, Dungan, & Young, 2013).

Recent work using both measures reveals a selective role for the right TPJ (rTPJ) in moral cognition. Thus far, multivariate pattern analyses reveal that the spatial patterns of activity in the rTPJ, but no other ToM regions, distinguish between intentional and accidental harm, although the magnitudes of rTPJ

response are no different for these two types of harm (Koster-Hale et al., 2013). This result suggests that, while intentional and accidental harms elicit similar levels of activity in the rTPJ, the rTPJ nevertheless differentiates between intentional and accidental harm in its spatial patterns of activity, indicating that the rTPJ encodes information about intent. The rTPJ represents harmful acts, in particular, as either intentional or accidental. Meanwhile, univariate analyses reveal that the rTPJ contributes to the initial encoding of an agent's beliefs for moral judgment (Young & Saxe, 2008). Specifically, this prior work has found that initially reading about Grace's belief that the powder is poison elicits a higher rTPJ response compared to reading that the powder is poison. Moreover, the rTPJ supports the integration of belief information with other morally relevant features of an action such as the outcome, as indicated by significantly above-baseline responses during the presentation of the agent's action following the presentation of belief information (Young & Saxe, 2008). Finally, the magnitude of response in the rTPJ during the time of integration correlates with moral judgment; in particular, people with a higher rTPJ response assign less blame for accidental harm (Young & Saxe, 2009). This finding suggests that people with more robust mental state representations (e.g., of false beliefs and/or innocent intentions) are more forgiving of accidents.

While it may be unsurprising that ToM regions are recruited for processing explicit mental state information for moral judgment, other work reveals that even when people are not provided with explicit mental state information, people still reason about mental states when making moral judgments (Young & Saxe, 2009a). This work shows that the rTPJ is recruited for processing morally relevant facts (e.g., powder is toxic or sugar) versus morally irrelevant facts about an action (e.g., powder fills the container). When people read morally relevant facts, they may spontaneously wonder what moral agents know (e.g., did she know she was poisoning her friend?) or believe (e.g., did she think it was sugar?). This finding—greater rTPJ activity for morally relevant over morally irrelevant information—suggests that moral judgments depend on spontaneous mental state inference in the absence of explicit mental state information.

When no explicit mental state information is provided, people may use information about someone's moral character or prior record to make mental state inferences in a particular instance (Alicke, 2000, 2008; Knobe, 2005, 2010). In one study, participants first ostensibly interacted with other players in an economic investment game (Kliemann, Young, Scholz, & Saxe, 2008). In the game, an Investor invested between one and four Money Units with a Trustee; this investment was tripled and given to the Trustee who decided how much of that total amount would be repaid to the Investor. Each participant played as the Investor half the time and the Trustee the other half of the time. After the game, participants made judgments of harmful actions (e.g., shrinking a roommate's sweater) that these "players" had performed in the past. Importantly, descriptions

of these events contained no mention of mental states, leaving participants to infer whether harms were intentional or accidental. Participants judged actions performed by unfair players (i.e., those who behaved unfairly in the investment game) as both more blameworthy and more intentional than the same actions performed by fair players (i.e., those who behaved fairly in the investment game). Moreover, the rTPJ was recruited more for harmful outcomes caused by unfair players versus fair players. These findings suggest a link between participants' background knowledge (e.g., fair or unfair play) and mental state reasoning in a subsequent moral judgment task.

While fMRI studies reveal a correlation between rTPJ activity and moral judgment, they cannot provide support for a causal role for the rTPJ. One approach to this causal question is to temporarily disrupt activity in the rTPJ and examine the subsequent effect on the use of mental state information for moral judgment. In one study, transiently disrupting rTPJ activity using transcranial magnetic stimulation (TMS) both immediately before and also during moral judgment led to reduced reliance on intentions less and—by default—greater reliance on outcomes for moral judgment (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010). In particular, disrupting rTPJ activity led to more lenient judgments of failed attempts to harm—characterized by neutral outcome (e.g., the powder was sugar) and negative intent (e.g., she thought it was poison).

The study of patient populations, too, can provide evidence for the causal role of rTPJ in moral judgment. Recent work has examined moral cognition in individuals with autism spectrum disorders (ASD), a neurodevelopmental disorder characterized by difficulties with social interaction. For example, high-functioning adults with ASD focus less on intentions and relatively more on outcomes when evaluating intentional and accidental harms (Moran et al., 2011). In the case of accidental harms, adults with ASD base their judgments more on the negative outcome and less on the innocent intention, thereby making harsher moral judgments, compared to neurotypical adults. Convergent neural evidence shows that, in adults with ASD, the rTPJ does not distinguish between intentional and accidental harms in its spatial patterns of activity, in contrast to neurotypical adults (Koster-Hale et al., 2013).

To summarize, moral judgments rely on mental state information. Most of the work we have reviewed so far focuses on moral judgments of harmful actions. However, morality is complex, and moral violations can occur in many different forms. Do people rely on mental state information similarly regardless of the nature of the violation?

Mental states matter differently across moral

domains

While the prototypical moral violation does involve harm, and while harm is universally—across all cultures—considered morally relevant (Haidt, 2007), moral transgressions do not consist of only harmful acts. In fact, researchers have revealed distinct types of moral actions or moral domains (Graham et al., 2011; Graham, Haidt, & Nosek, 2009; Haidt, 2007). These moral domains include moral concerns associated with harm/care, fairness/reciprocity, ingroup/loyalty, authority/respect, purity/sanctity, and more recently proposed, liberty/oppression (Iyer et al., 2012). In other words, the morally relevant actions are not just those that cause direct physical harm to others (e.g., poisoning someone's coffee) but those involving others' emotional suffering, differential treatment of others, betrayal among members of a group, lack of respect for authority, elicitations of disgust or sense of unnaturalness, and restrictions on one's liberties. Do moral judgments of these different types of acts differ systematically from one another?

Much of the focus within moral psychology has been on the contrast between two moral domains: harm versus purity. While harms typically involve a harmful agent and a suffering patient (Gray, Schein, & Ward, 2014), purity violations typically involve bodily violations related to food, pathogens, and sex (Chapman, Kim, Susskind, & Anderson, 2009; Russell & Giner-Sorolla, 2013; Tybur, Lieberman, Kurzban, & DeScioli, 2013). Judgments in the harm and purity domains have been associated with different affective responses (Gutierrez & Giner-Sorolla, 2007; Haidt, 2001; Haidt & Hersh, 2001; Horberg, Oveis, Keltner, & Cohen, 2009; Rozin, Lowery, Imada, & Haidt, 1999; Seidel & Prinz, 2013; Wheatley & Haidt, 2005), neural activity (Borg, Hynes, Van Horn, Grafton, & Sinnott-Armstrong, 2006; Chakroff et al., 2016; Moll et al., 2005), endorsement of person-based versus situation-based attributions (Chakroff & Young, 2015), and behavioral judgment (Haidt, 2007; Haidt, Koller, & Dias, 1993; Wright & Baril, 2011).

Recent research reveals that people rely more on intent information when responding to harm versus purity violations. For instance, harmful actions typically elicit anger, whereas impure actions elicit disgust, and these emotional reactions are differently affected by information about intent (Russell & Giner-Sorolla, 2011). Anger is modulated by information about the act as intentional or accidental, whereas disgust is generally inflexible and unreasoned. Moral judgments of harm versus purity violations, too, differently depend on intent information. In one study, participants read about intentional and accidental harms (e.g., putting peanuts in your cousin's dish knowing / not knowing that he has a peanut allergy) as well as intentional and accidental purity violations (e.g., sleeping with someone knowing / not knowing that the person is a long-lost sibling) (Young & Saxe, 2011). Innocent intentions reduced blame for accidental harm, and guilty intentions increased blame for failed attempts to harm, but these

effects were muted in the case of purity violations. For example, participants perceived a large difference between intentional and accidental harms but a significantly smaller difference between intentional and accidental purity violations. Another study investigated two different kinds of failed attempts: putting peanuts in a cousin's dish falsely believing her cousin to have a peanut allergy versus intending to put peanuts in a cousin's dish but running out of peanuts and putting walnuts in instead. There was no difference between the two types of attempted harm—presumably participants focus on the presence of malicious intent in both cases. By contrast, nearly sleeping with an actual sibling (but failed to because of a fire alarm) was seen as morally worse than actually sleeping with someone falsely believed to be a sibling. These results, which have been replicated in other studies (Chakroff, Dungan, & Young, 2013), suggest that compared to judgments of harms, judgments of purity violations such as incest depend less on the agent's mental states (e.g., whether he thought they were related) and more on other features of the act (e.g., whether they were actually related) even in the case of failed attempts.

The behavioral data suggest that mental states matter more for moral judgment of harms than impure actions. However, these data leave open the question of whether people similarly engage in mental state reasoning when processing both kinds of violations but simply decide to assign less weight to mental states for moral judgments of impure actions. This question can be addressed using neuroimaging. In one study, participants read stories about harmful and impure acts and made judgments of moral wrongness (Chakroff et al., 2016). The rTPJ was recruited more robustly when evaluating harms versus purity violations, even before explicit mental state information was presented. This neural pattern reveals a difference in spontaneous mental state reasoning across moral domains. Moreover, the spatial patterns of activity within the rTPJ were distinct for intentional and accidental harms but not for intentional and accidental purity violations. This result suggests that the rTPJ—a key region for ToM—encodes information about intent, but only for harmful actions and not for purity violations.

Together, behavioral and neural evidence suggests that mental states matter more for moral judgments of harm versus purity. In the next section, we provide one possible account of this effect, honing in on the distinct functional roles of harm and purity norms.

Mental states matter more for judgments of actions impacting other people

We propose that distinct moral norms about harm versus purity serve distinct functions—for regulating interpersonal relationships versus for protecting

the self from possible contamination (Young & Tsoi, 2013). In other words, people may uphold (either explicitly or implicitly) the idea that it is wrong to harm others and that it is wrong to defile the self.

Typical cases of harm (physical or psychological) feature a harmful perpetrator or agent and a suffering victim or patient (Gray & Wegner, 2009; Gray, Young, & Waytz, 2012; Young & Phillips, 2011); in most cases, two or more individuals are involved. Norms about harm (e.g., "don't harm others") may serve to minimize suffering or, more generally, any type of negative impact on one another. Harmful actions may therefore elicit questions concerning mental states (e.g., why did he hurt her? will he try to hurt her again?). Information about intent may help explain past harmful actions and, importantly, predict future harms. To take the earlier example: if Grace attempts to poison her friend's coffee (versus if Grace accidentally poisons her friend's coffee), she is more likely to act on that ill will again in the future.

By contrast, intent information matters less for purity. People often find purity violations morally offensive even when there is no victim or even harm as in the case of consensual incest (Haidt et al., 1993; Haidt, 2001). People experience disgust in response to pathogenic substances like feces and rotting flesh, leading them to avoid these disgust triggers. Some researchers theorize that people's experience of disgust in response to purity violations evolved for avoiding potential bodily contamination (Chapman et al., 2009; Rozin, Haidt, & Fincher, 2009; Russell & Giner-Sorolla, 2013; Tybur et al., 2013; but see Rottman & Young, 2014; Royzman & Kurzban, 2011). In the case of purity, people may be mostly worried about potential contamination and focus primarily on avoiding bad outcomes for themselves rather than on underlying intentions.

Recent work provides preliminary evidence for the link between harm norms and other-directed acts as well as the link between purity norms and self-directed acts. One study examined an extreme other-directed harm—homicide—alongside an extreme self-directed harm—suicide (Rottman, Kelemen, & Young, 2014). As expected, moral judgments of homicide were correlated with concerns about harm (as measured by the Moral Foundations Questionnaire or MFQ; Graham et al., 2011; Graham, Haidt, & Nosek, 2009). By contrast, moral judgments of suicide were correlated with (1) concerns about purity, (2) ratings of disgust in response to fabricated obituaries of people who committed suicide, and (3) judgments of suicide as tainting the purity of the suicide victim's soul. Judgments of suicide were not predicted by judgments of harm—whether suicide was perceived as harmful to the self, other people, or even God. These patterns emerged even among non-religious, liberal participants, suggesting that suicide, an extreme self-directed harmful act, is associated not with harm primarily but with purity.

Another set of studies suggests that moral judgments of harm and purity violations differ for other-directed acts and self-directed acts (Dungan, Chakroff, & Young, under review). Specifically, harms (e.g., administering an electric shock) are judged as morally worse than purity violations (e.g., spraying someone with a foul odor) when the target is another person, but purity violations are judged as morally worse than harms when the target is one's own self. Additional work reveals that the target of an action (i.e., oneself versus another person) partially determines the moral domain of the action (Chakroff et al., 2013). In one study, self-directed acts were judged primarily as impure, while other-directed acts were judged primarily as harmful. Critically, although intentional acts were judged more harshly than accidental acts, this difference was significantly greater for other-directed acts than for self-directed acts. These findings suggest that prior findings of less ToM for purity versus harmful transgressions may be rooted in differences between self- versus other-directed acts.

To summarize, these studies provide initial support for a functional account of distinct moral norms. Norms against harm may serve to limit people's negative impact on each other. By contrast, purity norms against eating taboo foods or sleeping with blood relatives may have evolved as a means for us to protect ourselves from possible contamination. For purity violations, whether the act was intentional or accidental may matter less—the key is to avoid the contamination altogether. In the next section, instead of focusing on how people differentially reason about mental states for different moral norms, we focus on how people differentially reason about mental states for different moral agents.

How people interact with others

So far, we have described the role of mental state reasoning for moral judgments primarily of hypothetical third-party actions. In this section, we focus on mental state reasoning for social interaction and discuss evidence showing that the way people consider others' minds depends on the motivational context. When people are motivated to consider the mental states of others, they are able to do so readily. Some evidence suggests that people—when presented with any stimuli that can be construed as social—spontaneously use mentalistic terms to describe the situation. In a classic study, people viewed a film of geometric shapes (circle and triangle) moving in various directions and at various speeds; participants later had to describe what happened in the film (Heider & Simmel, 1944). People interpreted the movements as actions of animate beings, and some even created very complex narratives. Common descriptions included terms like "fighting", "is shut up in the house and tries to get out", "chases", "move the door". Neuroimaging work reveals that moving shapes, when perceived as animate, elicit activity in the ToM network (Wheatley, Milleville, & Martin, 2007). In addition,

merely observing social scenes—single-frame pictures containing humans—appears sufficient to elicit activity in these regions (Wagner, Kelley, & Heatherton, 2011).

The focus of this section, however, is on how people deploy ToM when they themselves are engaged in social interactions across two fundamental motivational contexts: cooperation and competition. We discuss how ToM deployment during social interactions may differ depending on people's desire for social affiliation, desire for mastery over their environment, group membership, and interpersonal and intergroup relations. In short, we address the general question of when and how people engage in ToM when interacting with other moral agents.

Affiliation and cooperation

Current theories suggest that a key function of morality is for cooperation (Greene, 2013). Evolutionary accounts of cooperation have focused on processes such as kin selection and direct and indirect reciprocity, but these processes cannot fully explain the emergence of large-scale cooperation among complete strangers and non-kin that is so common in human societies (Axelrod & Hamilton, 1981). Morality is suggested to unite people with a set of shared norms about how group members should behave, providing benefits to cooperators and imposing costs on violators and free-riders (Haidt, 2007). In other words, morality helps regulate social interactions in the direction of cooperation (Tomasello & Vaish, 2013). Groups with norms, practices, and institutions that elicit more cooperative and group-benefiting behavior can grow, outperform, and eventually replace less cooperative groups.

One major motivator of cooperative behavior is the need for affiliation or sense of involvement and belonging within a social group (Baker, 1979). The motivation to affiliate with someone increases the tendency to consider the mind of that person, and consideration of human minds in turn facilitates coordination, cooperation, and communication (Epley & Waytz, 2010). Behavioral work shows that people tend to judge those they like as more capable of acting with intention (e.g., "this person has goals"), engaging in higher order thought (e.g., "this person has a good memory"), and experiencing emotions (e.g., "this person has complex feelings") than people they don't like (Kozak, Marsh, & Wegner, 2006; McPherson Frantz & Janoff-Bulman, 2000). Moreover, people tend to attribute more secondary emotions to ingroup versus outgroup members, an effect that persists even when controlling for familiarity with ingroup and outgroup members (Cortes, Demoulin, Rodriguez, Rodriguez, & Leyens, 2005). In fact, extreme outgroup members like homeless people fail to elicit activity in the medial prefrontal cortex, a region in the ToM network (Harris & Fiske, 2006a, 2006b). Notably, the relationship between liking and mind attribution appears to be

unidirectional: instructions to take a particular person's perspective did not increase liking for that person (Kozak et al., 2006). Together, this research suggests that liking drives mind attribution but not necessarily the other way around.

Motivation for social connection can sometimes even drive people to perceive minds in non-human agents. Prior work reveals that people with fewer reported friends pay greater attention to social cues in faces and voices and are more accurate in identifying emotional facial expressions (Gardner, Pickett, Jefferis, & Knowles, 2005). In order to alleviate the pain of chronic loneliness or social disconnection, some people even attempt to find sources of connection in nonhuman entities (Epley, Waytz, & Cacioppo, 2007). Specifically, people with higher scores of loneliness deliver higher anthropomorphic mental-state ratings for gadgets (e.g., a wheeled alarm clock that requires you to get up in order to turn it off, an air purifier for people with allergies), religious deities, and pets (Epley, Akalis, Waytz, & Cacioppo, 2008). Moreover, people induced to feel lonely and isolated tend to think of their pets as having traits related to social connection (e.g., thoughtful, considerate, and sympathetic), whereas this effect was absent in people induced to feel fear—another negative emotional state.

In sum, the motivation to create and maintain social connections with others elicits consideration of other minds. In turn, understanding other minds facilitates coordination, cooperation, and communication (Epley & Waytz, 2010), evidenced by neuroimaging work revealing recruitment of ToM regions during cooperative situations (Elliott et al., 2006) as well as social games assessing cooperative intent, such as the Trust Game, Prisoner's Dilemma, and Ultimatum Game (Krueger et al., 2007; McCabe, Houser, Ryan, Smith, & Trouard, 2001; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004).

Competition and deception

Research on the evolutionary origins of ToM, however, predicts greater ToM for competition versus cooperation. Nonhuman primates are able to deploy mechanisms for reasoning about others' behaviors, but only in the context of competition (e.g., fighting over scarce resources such as food; Hare, Call, Agnetta, & Tomasello, 2000). In one study, two chimpanzees, one dominant and one subordinate, were placed in different rooms; two pieces of food were placed at various locations within a third room in between the two rooms. When doors to the rooms were opened, chimpanzees were able to enter to take the food. Unsurprisingly, if dominant individuals had visual and physical access to the food, they usually took both pieces of food. Notably, subordinate chimpanzees went to get the food that only they could see more often than the food that both they and the dominant other could see; this behavioral pattern could not be explained by merely tracking the movements of the dominant individual. In addition to being

able to reason about what competing conspecifics can and cannot see, chimpanzees can also reason about what competing others know (i.e., about what others have or have not seen; Hare et al., 2001). A follow-up study used a similar paradigm, but this time visual access of the dominant individual was manipulated in three different ways: (1) the dominant was allowed to witness where food was hidden, (2) the dominant was not allowed to see where the food was hidden, and (3) the dominant was misinformed about the location of the hidden food (the food was moved after the dominant saw the original location). In all trials, the subordinate was able to see the hiding procedures as well as monitor the dominant individual's visual access to these procedures. Subordinates chose to go for the food more often when the dominant individual did not see the food being hidden or moved—that is, when the dominant was not informed about the location of the food or was misinformed. Subordinate chimpanzees may therefore be sensitive to the false beliefs of their competitors (but see Martin & Santos, 2016).

While the above work targets interactions with conspecifics, evidence suggests that nonhuman primates are also able to deploy ToM during competition with human experimenters. Chimpanzees appear to be able to manipulate what others see (Hare, Call, & Tomasello, 2006). That is, when chimpanzees competed with a human experimenter for food, chimpanzees sometimes tried to actively conceal their approach toward the food from human competitors. This account of rudimentary ToM deployment during competition is consistent with other work showing that chimpanzees perform better in different cognitive tasks when competing than when cooperating with a human experimenter (Hare & Tomasello, 2004). Research on rhesus monkeys, too, shows similar patterns in the context of competition (e.g., Phillips, Barnes, Mahajan, Yamaguchi, & Santos, 2009; Santos, Nissen, & Ferrugia, 2006). Together, these results suggest that the capacity for rudimentary ToM in nonhuman primates is suited for the ecologically salient domain of competition.

Indeed, initial primate work showed a surprising absence of ToM in non-competitive or cooperative tasks. Several decades ago, researchers questioned whether nonhuman primates can deploy ToM (Premack & Woodruff, 1978), and early work supported the notion that nonhuman primates have no understanding of others' mental states (Heyes, 1998; Povinelli & Vonk, 2003). Specifically, chimpanzees appeared to lack understanding of visual perception: chimpanzees used gestures to beg for food from a human with a blindfold or a bucket over his head (Povinelli & Eddy, 1996). Chimpanzees also appeared insensitive to the difference between knowledge and ignorance: chimpanzees would indiscriminately follow different pointing gestures of two different human individuals even though one individual knew the location of the hidden food while the other individual didn't (Povinelli, Rulf, & Bierschwale, 1994). Moreover, chimpanzees failed the same nonverbal tests of false belief understanding that five-year old human children were able to pass (Call & Tomasello, 1999).

Together, the data led researchers to wonder whether anecdotal evidence of ToM in nonhuman primates was driven by human reinterpretations of primate behaviors (Povinelli & Vonk, 2003).

It took some time to recognize that almost all of the negative findings emerged in experiments requiring chimpanzees to understand cooperative behavior (Hare et al., 2000; Hare, Call, & Tomasello, 2001). However, cooperative contexts are not ecologically valid for nonhuman primates—instead, nonhuman primates typically interact with conspecifics in a competitive manner (Lyons & Santos, 2006). Even though different nonhuman primate species differ in their general levels of aggression, nonhuman primates across the board tend to rely on dominance when resolving disputes among conspecifics. That is, alpha males take whatever scarce food they want, and subordinates take whatever is left over. The social lives of nonhuman primates—orangutans, gorillas, chimpanzees, and bonobos—mainly revolve around competition.

Perhaps like their nonhuman primate counterparts, humans are also motivated to engage in ToM more for competitive contexts. One study examined this question in children with autism. As previously mentioned, autism is characterized by persistent difficulty with social interactions. Overall, children with ASD tend to perform poorly on ToM tasks (Baron-Cohen, Leslie, & Frith, 1985; Happé, 1995; Yirmiya, Erel, Shaked, & Solomonica-Levi, 1998). One suggested explanation for this poor performance is that these tasks are "motivationally barren" (Peterson, Slaughter, Peterson, & Premack, 2013). That is, children with autism may not be motivated to keep up with a conversation or getting an arbitrary question correct in a laboratory setting. Even receiving material reward such as candy for correct responses may not be sufficient to elicit ToM in children with autism (Begeer, Rieffe, Terwogt, & Stockmann, 2003). Instead, children with autism may be more motivated to consider others' mental states in more naturalistic and relevant situations such as competition (e.g., fighting with a sibling over a specific toy). Indeed, prior work has found that children with autism are able to engage in ToM and attribute mental states when they are playing a competitive game with another even if they are unable to perform well on standard ToM tests such as the location-change task (e.g., Maxi's chocolate was moved from the cupboard to the drawers without Maxi knowing; where will Maxi look for the chocolate?) (Peterson et al., 2013).

Related to competition is deception. Sometimes effectively competing with another person requires deceiving that person. A number of studies reveal a link between ToM and deceptive behavior in neurotypical children: the greater the ToM ability, the earlier and better children lie (Chandler, Fritz, & Hala, 1989; Polak & Harris, 1999; Talwar & Lee, 2008). Young children are more likely to confess their transgressions (e.g., playing with a toy when explicitly told not to play) when they sense that an experimenter already knows about their transgressions than when they sense that the experimenter does not know (Fu,

Evans, Xu, & Lee, 2012). Children who lie about their transgressions, compared to children who confess, tend to better understand false beliefs as measured by a task like the location-change task (Evans, Xu, & Lee, 2011; Talwar & Lee, 2008). Strikingly, a recent study provides a causal link between ToM and lying: three-year olds who were originally unable to lie were able to lie consistently after ToM training (learned mental-state concepts) (Ding, Wellman, Wang, Fu, & Lee, 2015). Three-year olds who learned about physical concepts in a control condition were less inclined to lie compared to children who learned about mental-state concepts. Generally, ToM development appears to play a role in the development of deception (Lee, 2013).

Deception and other competitive behaviors may be driven by people's motivation to gain mastery over their environment and to predict others' actions (White, 1959). This type of motivation, also known as effectance motivation, is driven by the need to make sense of and gain control of an ambiguous situation or an uncertain world—why did they attack me; what will my enemy do next; what can I do to gain or regain control of this situation? Studies with human adults and children have found that an agent's negative behavior, as compared to neutral or positive behavior, elicits ToM in the service of understanding the agent's present and future behavior (Morewedge, 2009; Vaish, Grossmann, & Woodward, 2008; Waytz et al., 2010). This type of motivation is well suited for competitive contexts—after all, in order to successfully compete against a challenging opponent, one must be able to infer the opponent's mental states, predict how the opponent will act in the future, and coordinate one's own actions accordingly. In some cases, the motivation to attain mastery even leads people to anthropomorphize God or see agents as especially mentalistic (Gray & Wegner, 2010; Kay, Moscovitch & Laurin, 2010; Morewedge, 2009; Waytz et al., 2010).

Neuroimaging research provides further evidence for ToM during competitive interactions with others (Hampton et al., 2008) and possibly differential ToM processing for competitive versus cooperative contexts (Decety, Jackson, Sommerville, Chaminade, & Meltzoff, 2004; Lissek et al., 2008). For instance, competing with different people in a pattern completion game similar to Connect Four (in which players have to block each other from building the winning pattern) versus cooperating with different people in the same game (in which players help each other build the winning pattern) preferentially recruited regions in the ToM network (Decety et al., 2004). Merely observing cooperative and deceptive social interactions also recruited bilateral TPJ and precuneus, though viewing cartoons of one person deceiving the other versus two people cooperating with each other preferentially recruited dmPFC (Lissek et al., 2008). In short, people may be particularly motivated to engage in ToM in competitive contexts, as is the case with our evolutionary ancestors.

ToM for cooperation and competition

Thus far, there are two conflicting lines of evidence for how people engage in ToM for cooperation versus competition. One line of evidence shows that people engage in more ToM for people they like and wish to affiliate with versus those they do not affiliate with. Another line of evidence shows that humans and other apes engage in more ToM during competitive contexts than cooperative contexts. Our own recent and ongoing work investigates whether cooperative and competitive interactions differentially elicit ToM (Tsoi, Dungan, Waytz, & Young, 2016). In this study, cooperation and competition were operationalized in terms of goals (shared vs. opposing goals, respectively) and payoffs (shared win/loss vs. one sole winner, respectively). We examined activity in ToM regions as participants engaged in a series of cooperative and competitive interactions with ostensibly the same individual in a game variant of "Rock, Paper, Scissors". In most studies on competition or cooperation or both, control trials involve participants' playing against the computer or playing individually—in our study, participants played with the same individual across experimental and control trials, though trial outcomes could be determined either by both players' responses or by the computer (i.e., the computer randomly chooses whether both players win or lose in cooperative trials and which single player wins in competitive trials). Overall, regions in the ToM network were recruited similarly for cooperation and competition, suggesting that ToM isn't deployed more for cooperation than competition or vice versa. Notably, though, all regions in the ToM network could discriminate between cooperative and competitive trials in their spatial patterns of activity when participants believed the outcome was determined by their and their partner's choices (i.e., experimental trials) but not when the computer determined the outcome (i.e., control trials). These results suggest that ToM regions encode information separating cooperative interactions from competitive interactions.

The results of this study help narrow the possibilities of what type of information could be encoded in ToM regions in the context of cooperation and competition. We showed that ToM regions do not simply encode goal-oriented differences (shared goal vs. opposing goals) or payoff-oriented differences (shared win/loss vs. one single winner) between cooperative and competitive interactions given that these features were present for experimental and control trials, i.e., trials in which player responses determined the outcome and trials in which the computer determined the outcome. We propose that the ToM network as a whole encodes differences in how a person processes the mental states of the other player depending on whether they are cooperating or competing—but only when the person is motivated to consider the other player's mental states (e.g., when their behavior determines the outcome). Similar accounts suggest that TPJ activity is modulated by the extent to which one perceives others' actions as affecting one's

own behavior (Bhatt, Lohrenz, Camerer, & Montague, 2010; Carter, Bowling, Reeck, & Huettel, 2012).

If people process mental states differently depending on whether they are cooperating versus competing with an individual, we speculate that this difference may reflect focus on different dimensions of mind perception. Prior social psychological research has pointed to two dimensions of mind perception: agency and experience (Gray, Gray, & Wegner, 2007; Gray & Wegner, 2009). Agency includes the capacity for planning, thinking, and intending, while experience includes the capacity for emotion and feeling. In one study, participants made judgments of the mental capacities of different entities: seven living human forms (i.e., a human fetus, a 5-month-old infant, 5-year old girl, an adult woman, an adult male, the participant, a man in a persistent vegetative state), three nonhuman animals (i.e., frog, family dog, and wild chimpanzee), a dead woman, God, and a sociable robot (Gray et al., 2007). Unsurprisingly, participants perceived a dead person as having little agency or experience. On the other end, participants perceived alive adults, including themselves, as high in both agency and experience. God was perceived as high in agency but low in experience, while infants and nonhuman animals were perceived as high in experience but low in agency.

More recent work has shown that different motivations lead to preferential focus on different dimensions of mind perception (Waytz & Young, 2014). Previously, we described two motivational factors that drive behavior: affiliation motivation and effectance motivation (Epley et al., 2007; Waytz, Gray, Epley, & Wegner, 2010; White, 1959). Effectance motivation elicits preferential focus on agency, whereas affiliation motivation elicits preferential focus on experience (Waytz & Young, 2014). This effect was found when American participants were tasked with evaluating hypothetical outgroups as well as specific outgroups (e.g., China, Iran). Other research has revealed that focus on different mental states even leads to differential success during a negotiation (Galinsky, Maddux, Gilin, & White, 2008). Specifically, focusing on the other person's thoughts, interests, and purposes (i.e., agency) helps people reach a deal, whereas focusing on the other person's feelings and emotions (i.e., experience) does not provide any unique advantage. It is possible that cooperative and competitive interactions may primarily rely on different motivational states (though the mapping need not be one-to-one). Given social psychological evidence for different dimensions of mind perception alongside our own finding that ToM regions encode information separating cooperation and competition, we propose that people deploy ToM for both cooperation and competition but focus on different aspects of mental states (e.g., experience versus agency). To summarize, whether people are motivated to compete with opponents or cooperate with allies, they robustly represent the minds of their interaction partners.

When social interactions lead to less consideration of mental states

Above we described research indicating robust ToM for social interaction in the case of cooperation and competition alike. In a previous section, we described research revealing certain limitations on mental state reasoning—for moral judgments of hypothetical purity violations. In this section, we describe research investigating the circumstances in which people attribute less mind to others in interpersonal and intergroup contexts.

Especially relevant are situations of violence and conflict, in which people may be motivated to think of their enemies as less than fully human—as savages or barbarians without culture, self-restraint, or other more sophisticated cognitive capacities. Members of different ethnic and racial groups, especially victims of massacres and genocides like the Jews during the Holocaust and the Tutsis in Rwanda, have been compared to rats, cockroaches, and vermin (Haslam, 2006). To see victims as sub-human or mere objects may facilitate aggression (Bandura, 2002). This phenomenon—dehumanization—has been observed explicitly and implicitly (Haslam & Loughnan, 2014) and can even occur outside the extreme contexts of violence and conflict, in the face of gender and disability disparities, just to name a few (Haslam, 2006).

Extensive social psychological research reveals the relevance of group membership on mind attribution. Research suggests that people are less likely to spontaneously mentally simulate the actions of outgroup members (Gutsell & Inzlicht, 2010). This may be unsurprising given that people tend to attribute uniquely human characteristics less comprehensively to outgroup members than ingroup members. These characteristics include the capacity for compassion and morality (Kelman, 1973; Opotow, 1990; Struch & Schwartz, 1989) and secondary emotions like admiration or remorse (Leyens et al., 2000). In fact, evaluating others who are extremely dissimilar to the self (e.g., homeless people) fails to elicit activity in the medial prefrontal cortex, a key region within the "social brain" (Harris & Fiske, 2006). Moreover, people are even less likely to perceive animacy in faces when they are told that the faces belong to outgroup versus ingroup members (Hackel, Looser, & Van Bavel, 2014). Researchers found that when participants were presented face morphs of ingroup and outgroup members that varied along a spectrum from human (animate) to non-human (inanimate), an outgroup face, compared to an ingroup face, needed to be more animate to be equally likely to be perceived as having or lacking a mind.

Importantly, though, people do not simply dehumanize outgroup members across the board. For example, people who perceive an outgroup as more threatening also rely on more lenient thresholds for detecting animacy in outgroup faces (Hackel et al., 2014). Other work has shown that in the context of intergroup

conflict or threat, people do consider mental states of outgroup members, though they do so in a biased fashion (Waytz, Young, & Ginges, 2014). Specifically, American Democrats and Republicans, as well as Israelis and Palestinians in the Middle East, attribute their own group's support for aggression during conflict to ingroup love (i.e., compassion and empathy toward their own group) more than outgroup hate (i.e., dislike and animosity toward opposing group), but they attribute the opposing group's support for aggression to outgroup hate more than ingroup love. These results raise the possibility that threat specifically may lead to greater effectance motivation—again, the motivation to master one's environment—which may serve to reverse people's default tendency to disregard the minds of outgroup members.

Consideration of others' minds may improve interpersonal and intergroup relations

In 2006, President Barack Obama, then a senator, called graduating seniors of Northwestern University to cultivate their empathy in a commencement speech:

> "There's a lot of talk in this country about the federal deficit. But I think we should talk more about our empathy deficit – the ability to put ourselves in someone else's shoes; to see the world through those who are different from us – the child who's hungry, the laid-off steelworker, the immigrant woman cleaning your dorm room.
>
> As you go on in life, cultivating this quality of empathy will become harder, not easier. There's no community service requirement in the real world; no one forcing you to care. You'll be free to live in neighborhoods with people who are exactly like yourself, and send your kids to the same schools, and narrow your concerns to what's going in your own little circle." (Obama, 2006).

Obama and others have honed in on the powerful impact of empathy—the capacity to understand and feel what another person is experiencing—on interpersonal and intergroup relations. Research has revealed an effect of empathy training on different behavioral outcomes: for instance, training physicians to be more empathic improves physician empathy as rated by patients (Riess, Kelley, Bailey, Dunn, & Phillips, 2012). Likewise, training children to be more empathic decreases bullying behaviors (Şahin, 2012). Typically, these types of training programs teach people how to be aware of people's feelings, how to decode the emotional facial expressions of others, and how to respond appropriately to others. While prior work tended to treat empathy as a unitary construct, more recent work suggests that empathy consists of at least two components (Shamay-Tsoory, 2011; Singer, 2006): affective empathy, the capacity to respond with an appropriate emotion to, or resonate with, others' emotional states, and cognitive empathy, the

capacity to understand the mental states of others (akin to ToM). We focus on this latter component of empathy.

Inducing empathy for members of stigmatized groups (e.g., people with AIDS, homeless people) improves explicit evaluations of those members (Batson et al., 1997). In one study, participants listened to an interview of a woman with AIDS and were either instructed to take an objective perspective about what happened (control condition) or instructed to imagine how that woman felt about what happened and how it had affected her life (perspective taking condition). Participants in the perspective taking condition, compared to those in the control condition, expressed greater disagreement for items like "for most people with AIDS, it is their own fault that they have AIDS" and greater agreement for items like "our society does not do enough to help people with AIDS". This improvement in explicit evaluations after perspective taking is not limited to stigmatized groups. In groups involved in ideological conflict (e.g., regarding immigration laws in the US state of Arizona; regarding the conflict in the Middle East between Israelis and Palestinians), members in empowered groups (e.g., Israelis) showed more positive attitudes toward outgroup members after perspective taking; interestingly, members in disempowered groups (e.g., Palestinians) showed more positive attitudes toward outgroup members after perspective giving or sharing (Bruneau & Saxe, 2012). This finding suggests that attitudinal changes resulting from enhanced perspective taking may depend on the power dynamics of the group members taking the perspective. Generally, though, when people are asked to take the perspective of others, evidence suggests that people tend to report improved explicit and implicit evaluations that appear to endure over time (Todd & Galinsky, 2014).

Taking the perspective of an outgroup member also increases support for policies that attenuate intergroup inequality (Todd, Bodenhausen, & Galinsky, 2012). In one study, White participants who took the perspective of a Black or Latino person reported greater support of affirmative action; this effect was mediated by increased perceptions of intergroup discrimination. Furthermore, the effect of perspective taking on perceptions of intergroup discrimination is mediated by associations between the self and outgroup member—that is, when people take the perspective of an outgroup member, they associate the self more with the outgroup member, which in turn increases their sensitivity to discrimination.

Theories of how people form impressions of others may point to a possible prerequisite of empathy. One such theory is the continuous model of impression formation, which describes several component processes of impression formation, from initial categorization or stereotype to individuation (Fiske & Neuberg, 1990). According to this model, people typically rely on quick heuristics such as stereotypes when forming impressions of a person unless those stereotypes cannot satisfactorily allow people to draw conclusions and final impressions of that

person. When people are unable to form a satisfactory impression of a person based on stereotypes, they begin to assess and integrate individual attributes of the person and individuate the person. Perhaps in order to take the perspective of an outgroup member, people need to think of that person as an individual and not simply as a member of a larger group. Indeed, research shows that the empathy gap between ingroup and outgroup members may be attenuated by reducing impressions of outgroup entitativity, or the extent to which outgroups are perceived to have the nature of an entity (e.g., unity, coherence, organization; Cikara, Bruneau, Van Bavel, & Saxe, 2014).

We do note possible limits on the effects of perspective-taking. For example, prior work has revealed that instructions to take a particular person's perspective do not increase liking for that person (Kozak et al., 2006). Perhaps treating outgroup members as individuals and taking their perspectives does not lead to increased liking of outgroup members. Nonetheless recognizing interpersonal or intergroup discrimination may lead to greater support for equality and improved interpersonal and intergroup relations.

Conclusion

Moral judgment and moral behavior, especially cooperative and competitive social interactions, rely primarily on the capacity to attribute and reason about the minds of people. The first half of this chapter focused on characterizing the role of ToM in moral judgment. Information about intent—whether an act was performed intentionally or accidentally—is an important factor for assessing the moral character of a person, the moral permissibility or wrongness of an action, and the appropriate punishment for moral misdeeds. We provided neuroimaging evidence supporting the role of ToM regions in moral judgment, focusing on the rTPJ. In addition, we argued that reliance on people's mental states differed across different moral domains. Specifically, mental states matter more for judgments of harm versus purity violations. We provided a functional explanation for this differential reliance on mental states: harm norms serve to regulate interpersonal interactions, whereas purity norms serve to protect the self.

In the second half of this chapter, we characterized the role of ToM in social interactions with moral agents, mostly focusing on two fundamental contexts: cooperation and competition. We provided neuroimaging evidence that ToM regions are recruited robustly for both interaction contexts. Furthermore, we proposed that ToM regions encode different types of mental states—specifically, agentive and experiential mental states. Similar to the first half of the chapter when we discussed how people are less motivated to care about the mental states

of purity violators, we discussed how people may be less motivated to think of outgroup members as fully human in times of violence and conflict. Lastly, we presented work revealing that greater empathy or perspective taking can improve interpersonal and intergroup relations. If a primary function of morality is to encourage people to behave cooperatively, as some researchers propose, people must overcome several factors that make cooperation so challenging. The capacity to understand other minds—theory of mind—and greater insight into this complex process may be one key solution.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556.

Alicke, M. D. (2008). Blaming badly. *Journal of Cognition and Culture*, *8*(1), 179–186.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. *Science (New York, N.Y.)*, *211*(4489), 1390–1396.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development*, *2004*(103), 37–49. http://doi.org/10.1002/cd.96

Baker, C. R. (1979). Defining and measuring affiliation motivation. *European Journal of Social Psychology*, *9*(1), 97–99. http://doi.org/10.1002/ejsp.2420090108

Bandura, A. (2002). Selective Moral Disengagement in the Exercise of Moral Agency. *Journal of Moral Education*, *31*(2), 101–119. http://doi.org/10.1080/0305724022014322

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*(1), 37–46.

Batson, C. D., Polycarpou, M. P., Harmon-Jones, E., Imhoff, H. J., Mitchener, E. C., Bednar, L. L., … Highberger, L. (1997). Empathy and attitudes: can feeling for a member of a stigmatized group improve feelings toward the group? *Journal of Personality and Social Psychology*, *72*(1), 105–118.

Begeer, S., Rieffe, C., Terwogt, M. M., & Stockmann, L. (2003). Theory of Mind--based action in children from the autism spectrum. *Journal of Autism and Developmental Disorders*, *33*(5), 479–487.

Bhatt, M. A., Lohrenz, T., Camerer, C. F., & Montague, P. R. (2010). Neural signatures of strategic types in a two-person bargaining game. *Proceedings of the National Academy of Sciences*, *107*(46), 19720–19725. http://doi.org/10.1073/pnas.1009625107

Bourgeois, P., & Hess, U. (2008). The impact of social context on mimicry. *Biological Psychology*, *77*(3), 343–352. http://doi.org/10.1016/j.biopsycho.2007.11.008

Bruneau, E. G., & Saxe, R. (2012). The power of being heard: The benefits of "perspective-giving" in the context of intergroup conflict. *Journal of Experimental Social Psychology*, *48*(4), 855–866. http://doi.org/10.1016/j.jesp.2012.02.017

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure and Function*, *217*(4), 783–796. http://doi.org/10.1007/s00429-012-0380-y

Call, J., & Tomasello, M. (1999). A Nonverbal False Belief Task: The Performance of Children and Great Apes. *Child Development*, *70*(2), 381–395. http://doi.org/10.1111/1467-8624.00028

Carter, R. M., Bowling, D. L., Reeck, C., & Huettel, S. A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, *337*(6090), 109–111. http://doi.org/10.1126/science.1219681

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, *11*(3), 476–484. https://doi.org/10.1093/scan/nsv131

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PloS One*, *8*(9), e74434. http://doi.org/10.1371/journal.pone.0074434

Chakroff, A., & Young, L. (2015). Harmful situations, impure people: An attribution asymmetry across moral domains. *Cognition*, *136*, 30–37. http://doi.org/10.1016/j.cognition.2014.11.034

Chandler, M., Fritz, A. S., & Hala, S. (1989). Small-Scale Deceit: Deception as a Marker of Two-, Three-, and Four-Year-Olds' Early Theories of Mind. *Child Development*, *60*(6), 1263. http://doi.org/10.2307/1130919

Chapman, H. A., Kim, D. A., Susskind, J. M., & Anderson, A. K. (2009). In Bad Taste: Evidence for the Oral Origins of Moral Disgust. *Science*, *323*(5918), 1222–1226. http://doi.org/10.1126/science.1165565

Cikara, M., Bruneau, E., Van Bavel, J. 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. http://doi.org/10.1016/j.jesp.2014.06.007

Cortes, B. P., Demoulin, S., Rodriguez, R. T., Rodriguez, A. P., & Leyens, J.-P. (2005). Infrahumanization or Familiarity? Attribution of Uniquely Human Emotions to the Self, the Ingroup, and the Outgroup. *Personality and Social Psychology Bulletin*, *31*(2), 243–253. http://doi.org/10.1177/0146167204271421

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380. http://doi.org/10.1016/j.cognition.2008.03.006

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97–103. http://doi.org/10.1016/j.copsyc.2015.06.003

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "Trembling Hand" game. *PLoS ONE*, *4*(8), e6699. http://doi.org/10.1371/journal.pone.0006699

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21. http://doi.org/10.1016/j.cognition.2012.11.008

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: an fMRI investigation. *NeuroImage*, *23*(2), 744–751. http://doi.org/10.1016/j.neuroimage.2004.05.025

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The Contribution of Emotion and Cognition to Moral Sensitivity: A Neurodevelopmental Study. *Cerebral Cortex*, *22*(1), 209–220. http://doi.org/10.1093/cercor/bhr111

Ding, X. P., Wellman, H. M., Wang, Y., Fu, G., & Lee, K. (2015). Theory-of-Mind Training Causes Honest Young Children to Lie. *Psychological Science*. http://doi.org/10.1177/0956797615604628

Dungan, J., Chakroff, A., & Young, L. (under review). Contextual influences on the relevance of moral norms: Harm and purity concerns for self versus other.

Elliott, R., Völlm, B., Drury, A., McKie, S., Richardson, P., & William Deakin, J. F. (2006). Co-operation with another player in a financially rewarded guessing game activates regions implicated in theory of mind. *Social Neuroscience*, *1*(3-4), 385–395. http://doi.org/10.1080/17470910601041358

Epley, N., Akalis, S., Waytz, A., & Cacioppo, J. T. (2008). Creating Social Connection Through Inferential Reproduction: Loneliness and Perceived Agency in Gadgets, Gods, and Greyhounds. *Psychological Science*, *19*(2), 114–120. http://doi.org/10.1111/j.1467-9280.2008.02056.x

Epley, N., & Waytz, A. (2010). Mind Perception. In *Handbook of Social Psychology*. John Wiley & Sons, Inc. Retrieved from http://dx.doi.org/10.1002/9780470561119.socpsy001014

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. http://doi.org/10.1037/0033-295X.114.4.864

Evans, A. D., Xu, F., & Lee, K. (2011). When all signs point to you: lies told in the face of evidence. *Developmental Psychology*, *47*(1), 39–49. http://doi.org/10.1037/a0020787

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.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*(2), 109–128.

Fu, G., Evans, A. D., Xu, F., & Lee, K. (2012). Young children can tell strategic lies after committing a transgression. *Journal of Experimental Child Psychology*, *113*(1), 147–158. http://doi.org/10.1016/j.jecp.2012.04.003

Galinsky, A. D., Maddux, W. W., Gilin, D., & White, J. B. (2008). Why it pays to get inside the head of your opponent: The differential effects of perspective taking and empathy in negotiations. *Psychological Science*, *19*(4), 378–384. http://doi.org/10.1111/j.1467-9280.2008.02096.x

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of "theory of mind" in verbal and nonverbal tasks. *Neuropsychologia*, *38*(1), 11–21.

Gardner, W. L., Pickett, C., L., Jefferis, V., & Knowles, M. (2005). On the Outside Looking In: Loneliness and Social Monitoring. *Personality and Social Psychology Bulletin*, *31*(11), 1549–1560. http://doi.org/10.1177/0146167205277208

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*(11), 1803–1814. http://doi.org/10.1162/jocn.2007.19.11.1803

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. http://doi.org/10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., & Ditto, P. H. (2011). Mapping the moral domain. *Journal of Personality and Social Psychology*, *101*(2), 366–385. http://doi.org/10.1037/a0021847

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of Mind Perception. *Science*, *315*(5812), 619–619. http://doi.org/10.1126/science.1134475

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General*, *143*(4), 1600–1615. http://doi.org/10.1037/a0036149

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505–520. http://doi.org/10.1037/a0013748

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. http://doi.org/10.1080/1047840X.2012.651387

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: The Penguin Press.

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. *Emotion*, *7*(4), 853–868. http://doi.org/10.1037/1528-3542.7.4.853

Gutsell, J. N., & Inzlicht, M. (2010). Empathy constrained: Prejudice predicts reduced mental simulation of actions during observation of outgroups. *Journal of*

Experimental Social Psychology, *46*(5), 841–845. http://doi.org/10.1016/j.jesp.2010.03.011
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. http://doi.org/10.1016/j.jesp.2013.12.001
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834. http://doi.org/10.1037//0033-295X.108.4.814
Haidt, J. (2007). The New Synthesis in Moral Psychology. *Science*, *316*(5827), 998–1002. http://doi.org/10.1126/science.1137651
Haidt, J., & Hersh, M. A. (2001). Sexual Morality: The Cultures and Emotions of Conservatives and Liberals. *Journal of Applied Social Psychology*, *31*(1), 191–221. http://doi.org/10.1111/j.1559-1816.2001.tb02489.x
Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628.
Happé, F. G. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development*, *66*(3), 843–855.
Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785. http://doi.org/10.1006/anbe.1999.1377
Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151. http://doi.org/10.1006/anbe.2000.1518
Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, *101*(3), 495–514. http://doi.org/10.1016/j.cognition.2005.01.011
Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*(3), 571–581. http://doi.org/10.1016/j.anbehav.2003.11.011
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. http://doi.org/10.1111/j.1467-9280.2006.01793.x
Harris, L. T., & Fiske, S. T. (2006). Social groups that elicit disgust are differentially processed in mPFC. *Social Cognitive and Affective Neuroscience*, *2*(1), 45–51. http://doi.org/10.1093/scan/nsl037
Hart, H. L. A., & Honoré, T. (1985). *Causation in the Law*. Oxford University Press.
Haslam, N. (2006). Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, *10*(3), 252–264. http://doi.org/10.1207/s15327957pspr1003_4
Haslam, N., & Loughnan, S. (2014). Dehumanization and Infrahumanization. *Annual Review of Psychology*, *65*(1), 399–423. http://doi.org/10.1146/annurev-psych-010213-115045
Hebble, P. W. (1971). The Development of Elementary School Children's Judgment of Intent. *Child Development*, *42*(4), 1203. http://doi.org/10.2307/1127804

Heider, F., & Simmel, M. (1944). An Experimental Study of Apparent Behavior. *The American Journal of Psychology*, *57*(2), 243. http://doi.org/10.2307/1416950
Heyes, C. M. (1998). Theory of mind in nonhuman primates. *The Behavioral and Brain Sciences*, *21*(1), 101–114; discussion 115–148.
Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, *97*(6), 963–976. http://doi.org/10.1037/a0017423
Inbar, Y., Pizarro, D. A., & Cushman, F. (2012). Benefiting From Misfortune: When Harmless Actions Are Judged to Be Morally Blameworthy. *Personality and Social Psychology Bulletin*, *38*(1), 52–62. http://doi.org/10.1177/0146167211430232
Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The Psychological dispositions of self-identified libertarians. *PLoS ONE*, *7*(8), e42366. https://doi.org/10.1371/journal.pone.0042366
Kelman, H. G. (1973). Violence without moral restraint: Reflections on the dehumanization of victims and victimizers. *Journal of Social Issues*, *29*(4), 25–61. http://doi.org/10.1111/j.1540-4560.1973.tb00102.x
Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, *119*(2), 197–215. http://doi.org/10.1016/j.cognition.2011.01.006
Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957. http://doi.org/10.1016/j.neuropsychologia.2008.06.010
Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, *9*(8), 357–359.
Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences*, *33*(04), 315–329.
Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences*, *110*(14), 5648–5653. http://doi.org/10.1073/pnas.1207992110
Kozak, M. N., Marsh, A. A., & Wegner, D. M. (2006). What do i think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, *90*(4), 543–555. http://doi.org/10.1037/0022-3514.90.4.543
Krueger, F., McCabe, K., Moll, J., Kriegeskorte, N., Zahn, R., Strenziok, M., … Grafman, J. (2007). Neural correlates of trust. *Proceedings of the National Academy of Sciences*, *104*(50), 20084–20089. http://doi.org/10.1073/pnas.0710103104
Lakin, J. L., & Chartrand, T. L. (2003). Using nonconscious behavioral mimicry to create affiliation and rapport. *Psychological Science*, *14*(4), 334–339.
Lee, K. (2013). Little Liars: Development of Verbal Deception in Children. *Child Development Perspectives*, *7*(2), 91–96. http://doi.org/10.1111/cdep.12023
Leyens, J.-P., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of

secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review*, *4*(2), 186–197. http://doi.org/10.1207/S15327957PSPR0402_06

Lissek, S., Peters, S., Fuchs, N., Witthaus, H., Nicolas, V., Tegenthoff, M., … Brüne, M. (2008). Cooperation and Deception Recruit Different Subsets of the Theory-of-Mind Network. *PLoS ONE*, *3*(4), e2023. http://doi.org/10.1371/journal.pone.0002023

Lyons, D. E., & Santos, L. R. (2006). Ecology, Domain Specificity, and the Origins of Theory of Mind: Is Competition the Catalyst? *Philosophy Compass*, *1*(5), 481–492. http://doi.org/10.1111/j.1747-9991.2006.00032.x

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. http://doi.org/10.1037/a0026790

Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835. http://doi.org/10.1073/pnas.211415698

McPherson Frantz, C., & Janoff-Bulman, R. (2000). Considering Both Sides: The Limits of Perspective Taking. *Basic and Applied Social Psychology*, *22*(1), 31–42. http://doi.org/10.1207/S15324834BASP2201_4

Moll, J., de Oliveira-Souza, R., Moll, F. T., Ignácio, F. A., Bramati, I. E., Caparelli-Dáquer, E. M., & Eslinger, P. J. (2005). The moral affiliations of disgust: a functional MRI study. *Cognitive and Behavioral Neurology: Official Journal of the Society for Behavioral and Cognitive Neurology*, *18*(1), 68–78.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., & Gabrieli, J. D. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences*, *108*(7), 2688–2692. http://doi.org/10.1073/pnas.1011734108

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General*, *138*(4), 535–545. http://doi.org/10.1037/a0016796

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2008). Revealing representational content with pattern-information fMRI--an introductory guide. *Social Cognitive and Affective Neuroscience*, *4*(1), 101–109. http://doi.org/10.1093/scan/nsn044

Obama, B. (2006, June). Commencement speech presented at Northwestern University, Evanston, IL. Speech retrieved from http://www.northwestern.edu/newscenter/stories/2006/06/barack.html

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.)*, *308*(5719), 255–258. http://doi.org/10.1126/science.1107621

Opotow, S. (1990). Moral exclusion and injustice: An Introduction. *Journal of Social Issues*, *46*(1), 1–20. http://doi.org/10.1111/j.1540-4560.1990.tb00268.x

Peterson, C. C., Slaughter, V., Peterson, J., & Premack, D. (2013). Children with autism can track others' beliefs in a competitive game. *Developmental Science*, *16*(3), 443–450. http://doi.org/10.1111/desc.12040

Phillips, W., Barnes, J. L., Mahajan, N., Yamaguchi, M., & Santos, L. R. (2009). "Unwilling" versus "unable": capuchin monkeys' (*Cebus apella*) understanding of human intentional action: Unwilling vs. unable in capuchin monkeys. *Developmental Science*, *12*(6), 938–945. http://doi.org/10.1111/j.1467-7687.2009.00840.x

Polak, A., & Harris, P. L. (1999). Deception by young children following noncompliance. *Developmental Psychology*, *35*(2), 561–568.

Povinelli, D. J., & Eddy, T. J. (1996). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, *61*(3), i–vi, 1–152; discussion 153–191.

Povinelli, D. J., Rulf, A. B., & Bierschwale, D. T. (1994). Absence of knowledge attribution and self-recognition in young chimpanzees (Pan troglodytes). *Journal of Comparative Psychology*, *108*(1), 74–80. http://doi.org/10.1037/0735-7036.108.1.74

Povinelli, D. J., & Vonk, J. (2003). Chimpanzee minds: suspiciously human? *Trends in Cognitive Sciences*, *7*(4), 157–160. http://doi.org/10.1016/S1364-6613(03)00053-6

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(04), 515. http://doi.org/10.1017/S0140525X00076512

Riess, H., Kelley, J. M., Bailey, R. W., Dunn, E. J., & Phillips, M. (2012). Empathy Training for Resident Physicians: A Randomized Controlled Trial of a Neuroscience-Informed Curriculum. *Journal of General Internal Medicine*, *27*(10), 1280–1286. http://doi.org/10.1007/s11606-012-2063-z

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*(4), 1694–1703. http://doi.org/10.1016/j.neuroimage.2004.04.015

Rottman, J., Kelemen, D., & Young, L. (2014). Tainting the soul: Purity concerns predict moral judgments of suicide. *Cognition*, *130*(2), 217–226. http://doi.org/10.1016/j.cognition.2013.11.007

Rottman, J., & Young, L. (2014). Comment: Scholarly Disgust and Related Mysteries. *Emotion Review*, *6*(3), 222–223. http://doi.org/10.1177/1754073914523043

Royzman, E., & Kurzban, R. (2011). Minding the Metaphor: The Elusive Character of Moral Disgust. *Emotion Review*, *3*(3), 269–271. http://doi.org/10.1177/1754073911402371

Rozin, P., Haidt, J., & Fincher, K. (2009). PSYCHOLOGY: From Oral to Moral. *Science*, *323*(5918), 1179–1180. http://doi.org/10.1126/science.1170492

Rozin, P., Lowery, L., Imada, S., & Haidt, J. (1999). The CAD triad hypothesis: a mapping between three moral emotions (contempt, anger, disgust) and three moral

codes (community, autonomy, divinity). *Journal of Personality and Social Psychology*, *76*(4), 574–586.
Russell, P. S., & Giner-Sorolla, R. (2011). Moral anger, but not moral disgust, responds to intentionality. *Emotion*, *11*(2), 233–240. http://doi.org/10.1037/a0022598
Russell, P. S., & Giner-Sorolla, R. (2013). Bodily moral disgust: What it is, how it is different from anger, and why it is an unreasoned emotion. *Psychological Bulletin*, *139*(2), 328–351. http://doi.org/10.1037/a0029319
Şahin, M. (2012). An investigation into the efficiency of empathy training program on preventing bullying in primary schools. *Children and Youth Services Review*, *34*(7), 1325–1330. http://doi.org/10.1016/j.childyouth.2012.03.013
Santos, L. R., Nissen, A. G., & Ferrugia, J. A. (2006). Rhesus monkeys, Macaca mulatta, know what others can and cannot hear. *Animal Behaviour*, *71*(5), 1175–1181. http://doi.org/10.1016/j.anbehav.2005.10.007
Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, *55*, 87–124. http://doi.org/10.1146/annurev.psych.55.090902.142044
Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind." *NeuroImage*, *19*, 1835–1842. http://doi.org/10.1016/S1053-8119(03)00230-1
Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, Action, and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, *18*(5), 803–817. http://doi.org/10.1162/jocn.2006.18.5.803
Seidel, A., & Prinz, J. (2013). Sound morality: irritating and icky noises amplify judgments in divergent moral domains. *Cognition*, *127*(1), 1–5. http://doi.org/10.1016/j.cognition.2012.11.004
Shamay-Tsoory, S. G. (2011). The Neural Bases for Empathy. *The Neuroscientist*, *17*(1), 18–24. http://doi.org/10.1177/1073858410379268
Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of Moral Responsibility and Punishment. *Child Development*, *57*(1), 177. http://doi.org/10.2307/1130649
Singer, T. (2006). The neuronal basis and ontogeny of empathy and mind reading: Review of literature and implications for future research. *Neuroscience & Biobehavioral Reviews*, *30*(6), 855–863. http://doi.org/10.1016/j.neubiorev.2006.06.011
Struch, N., & Schwartz, S. H. (1989). Intergroup aggression: Its predictors and distinctness from in-group bias. *Journal of Personality and Social Psychology*, *56*(3), 364–373.
Talwar, V., & Lee, K. (2008). Social and Cognitive Correlates of Childrens Lying Behavior. *Child Development*, *79*(4), 866–881. http://doi.org/10.1111/j.1467-8624.2008.01164.x
Todd, A. R., Bodenhausen, G. V., & Galinsky, A. D. (2012). Perspective taking combats the denial of intergroup discrimination. *Journal of Experimental Social Psychology*, *48*(3), 738–745. http://doi.org/10.1016/j.jesp.2011.12.011

Todd, A. R., & Galinsky, A. D. (2014). Perspective-Taking as a Strategy for Improving Intergroup Relations: Evidence, Mechanisms, and Qualifications: Perspective-Taking and Intergroup Relations. *Social and Personality Psychology Compass*, *8*(7), 374–387. http://doi.org/10.1111/spc3.12116

Tomasello, M., & Vaish, A. (2013). Origins of Human Cooperation and Morality. *Annual Review of Psychology*, *64*(1), 231–255. http://doi.org/10.1146/annurev-psych-113011-143812

Tsoi, L., Dungan, J., Waytz, A., & Young, L. (2016). Distinct neural patterns of social cognition for cooperation versus competition. *NeuroImage*, *137*, 86–96. https://doi.org/10.1016/j.neuroimage.2016.04.069

Tybur, J. M., Lieberman, D., Kurzban, R., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review*, *120*(1), 65–84. http://doi.org/10.1037/a0030778

Vaish, A., Grossmann, T., & Woodward, A. (2008). Not all emotions are created equal: The negativity bias in social-emotional development. *Psychological Bulletin*, *134*(3), 383–403. http://doi.org/10.1037/0033-2909.134.3.383

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual Differences in the Spontaneous Recruitment of Brain Regions Supporting Mental State Understanding When Viewing Natural Social Scenes. *Cerebral Cortex*, *21*(12), 2788–2796. http://doi.org/10.1093/cercor/bhr074

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388. http://doi.org/10.1016/j.tics.2010.05.006

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. http://doi.org/10.1037/a0020240

Waytz, A., & Young, L. (2014). Two motivations for two dimensions of mind. *Journal of Experimental Social Psychology*, *55*, 278–283. http://doi.org/10.1016/j.jesp.2014.08.001

Waytz, A., Young, L. L., & Ginges, J. (2014). Motive attribution asymmetry for love vs. hate drives intractable conflict. *Proceedings of the National Academy of Sciences*, *111*(44), 15687–15692. http://doi.org/10.1073/pnas.1414146111

Wellman, H. M., Cross, D., & Watson, J. (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, *72*(3), 655–684.

Wheatley, T., & Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychological Science*, *16*(10), 780–784. http://doi.org/10.1111/j.1467-9280.2005.01614.x

Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: Distinct roles for the social network and mirror system. *Psychological Science*, *18*(6), 469–474. http://doi.org/10.1111/j.1467-9280.2007.01923.x

White, R. W. (1959). Motivation reconsidered: The concept of competence. *Psychological Review*, *66*(5), 297–333. http://doi.org/10.1037/h0040934

Wiltermuth, S. S., & Heath, C. (2009). Synchrony and Cooperation. *Psychological Science*, *20*(1), 1–5. http://doi.org/10.1111/j.1467-9280.2008.02253.x

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

Wright, J. C., & Baril, G. (2011). The role of cognitive resources in determining our moral intuitions: Are we all liberals at heart? *Journal of Experimental Social Psychology*, *47*(5), 1007–1012. http://doi.org/10.1016/j.jesp.2011.03.014

Yirmiya, N., Erel, O., Shaked, M., & Solomonica-Levi, D. (1998). Meta-analyses comparing theory of mind abilities of individuals with autism, individuals with mental retardation, and normally developing individuals. *Psychological Bulletin*, *124*(3), 283–307.

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, *107*(15), 6753–6758. http://doi.org/10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240. http://doi.org/10.1073/pnas.0701408104

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, *119*(2), 166–178. http://doi.org/10.1016/j.cognition.2011.01.004

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage*, *40*(4), 1912–1920. http://doi.org/10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2009a). An fMRI investigation of spontaneous mental State inference for moral judgment. *Journal of Cognitive Neuroscience*, *21*(7), 1396–1405. http://doi.org/10.1162/jocn.2009.21137

Young, L., & Saxe, R. (2009b). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072. http://doi.org/10.1016/j.neuropsychologia.2009.03.020

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214. http://doi.org/10.1016/j.cognition.2011.04.005

Young, L., & Tsoi, L. (2013). When Mental States Matter, When They Don't, and What That Means for Morality: When Mental States Matter for Morality. *Social and Personality Psychology Compass*, *7*(8), 585–604. http://doi.org/10.1111/spc3.12044

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, *24*(3), 358–365. http://doi.org/10.1037/0012-1649.24.3.358

Zelazo, P. D., Helwig, C. C., & Lau, A. (1996). Intention, Act, and Outcome in Behavioral Prediction and Moral Judgment. *Child Development*, *67*(5), 2478. http://doi.org/10.2307/1131635