*** *Manuscript Draft* ***

In S. Baron-Cohen, H. Tager-Flusberg, M. Lombardo (eds.),

Understanding Other Minds. Oxford University Press.

Mind attribution is for morality

Liane Young[1*] & Adam Waytz[2*]

[1] Department of Psychology, Boston College

[2] Management and Organizations Department, Kellogg School of Management,

Northwestern University

* These authors contributed equally.

Morality – judging others' behavior to be right or wrong, as well as behaving in a right or wrong manner towards others – is a critical component of social life. Morality depends critically on our ability to attribute minds to entities that engage in moral actions (towards ourselves and others) and the entities that experience these actions (our own actions and others').

The cognitive capacities for attributing minds to others and considering the specific contents of those minds (i.e. mental state reasoning or theory of mind) allow us to understand and interact with individuals and even entire groups of individuals. More specifically, mental state reasoning represents a critical cognitive input for behavior explanation, action prediction, and moral evaluation. We deploy our mental state reasoning abilities in order to explain people's past actions (e.g., Lisa looked for her shoes in the garage because she *forgot* her mother had moved them to closet); to predict people's future behavior (e.g., Mike will tell Barbara his favorite dog joke *not knowing* that Barbara's dog was just hit by a car); and to make moral judgments (e.g., Grace must be a bad person for putting what she *thinks* is poison into someone else's coffee). Our capacity to consider other people's mental states, including their thoughts, their true or false beliefs, and their helpful or harmful intentions, helps us to navigate our social environment. Indeed, as much research has shown, mental state reasoning functions flexibly across domains, one of which is morality, the focus of this chapter.

The novel claim we make in this chapter is that the *primary* service of mental state reasoning may be for moral cognition and behavior, broadly construed. In particular, the cognitive capacities for mental state reasoning become less relevant when morality is not at stake. We are motivated to understand the actions of relevant moral agents, to

predict people's actions when those actions affect us, directly or indirectly, and to evaluate moral agents as current or future allies or enemies. Computations like these crucially elicit mental state reasoning.

In this chapter, we will therefore review the literature on mental state reasoning for moral cognition – both for judging other moral actors, from the position of "judge" on high, and also for figuring out, as "actors" on the ground, so to speak, who might help us or hurt us, to whom we have moral obligations (for helping or, minimally, not hurting), and whom we ought to trust or avoid (see Figure 1).

Morality on high

In this first section, we discuss the critical role of mental states for third-party moral judgments, including how people judge moral agents who harm others. Mental state reasoning is a key cognitive process for evaluating the guilty and innocent intentions of moral agents (Hart, 1968; Kamm, 2001; Mikhail, 2007). Indeed, recent research on the interaction of mental state reasoning and moral cognition has focused on the dominant role of agents' mental states versus the outcomes of agents' actions for our moral judgments (Cushman, 2008; Young, Cushman, Hauser, & Saxe, 2007).

To target the distinct roles of mental states and outcomes, many of these studies present scenarios in which agents produced either a negative outcome (harm to another person) or a neutral outcome (no harm), based on the belief that they would cause the negative outcome ("negative" belief / harmful intention) or the neutral outcome ("neutral" belief / innocent intention). Participants deliver a moral judgment – evaluating

the agent's action as permissible or forbidden, or deciding how much moral blame the agent deserves for his or her behavior.

An example illustrates the possible tension between mental states and outcomes:

> Grace and her co-worker are taking a tour of a chemical factory. Grace stops to pour herself and her co-worker some coffee. Nearby is a container of sugar. The container, however, has been mislabeled "toxic", so Grace thinks that the powder inside is toxic. She spoons some into her co-worker's coffee and takes none for herself. Her co-worker drinks the coffee, and nothing bad happens.

This scenario pits harmful intentions against neutral outcomes in representing a failed attempt to harm. In an alternative scenario:

> A container of poison sits near the coffee. The container, however, has been mislabeled "sugar", so Grace thinks the powder inside is sugar. She spoons some into her co-worker's coffee. Her co-worker drinks her coffee and ends up dead.

In this key scenario, an accident occurs – a bad outcome due to a false belief (but not malicious intent). Across studies relying on similar stimuli, participants assigned more moral weight to the agent's belief and intent, compared to the outcomes (Young, et al., 2007). A simple metric of this effect is that participants almost universally judge an attempted harm (e.g., trying but failing to poison someone) as morally worse than an accidental harm (e.g., accidentally poisoning someone).

Other research has investigated not only the simple contrast between intentions and outcomes but also the relative contributions of distinct internal and external factors (e.g., outcome, causation, belief, and desire) for different kinds of moral judgments (e.g., character, permissibility, blame, and punishment) (Cushman, 2008; Cushman, Dreber, Wang, & Costa, 2009). Importantly, the agent's *belief* about whether his or her action would cause harm dominated moral judgments across the board, followed by the agent's

desire to cause harm. The relative contribution of beliefs versus outcomes was greatest for judgments about the moral character of the agent or the moral permissibility of the action. Punishment judgments depended relatively more on outcomes. Nevertheless, these findings indicate the key role of mental state factors for moral judgments.

Notably, mental state factors may underlie moral judgments even in cases where outcomes appear, on the surface, to determine moral judgments. Consider the case of accidents. Many people assign some blame to agents who cause harmful outcomes, even when they didn't intend to cause the harmful outcomes. (An interesting exception is the psychopath – in the absence of an emotional response to the harmful outcome, psychopaths rely primarily on the stated innocent intent and deliver abnormally lenient judgments of accidents; Young, Koenigs, Kruepke, & Newman, 2012). Recall the scenario in which Grace accidentally poisons her co-worker because she mistakes the poison for sugar. Again, participants mostly excuse Grace on the grounds of her false belief and innocent intention, but they nevertheless assign some moral blame to Grace for the harm done. Behavioral and neural evidence suggests that this moral blame is determined not simply by the harmful outcome of Grace's action; instead, participants' assessment of Grace's mental state drives this judgment (Young, Nichols, & Saxe, 2010). Participants judge Grace's false belief as more unjustified or unreasonable when it leads to a bad (versus neutral) outcome, and therefore they judge Grace to be more morally blameworthy. Consistent with this behavioral pattern, activity in brain regions for mental state reasoning, including the right temporo-parietal junction (RTPJ) (Jenkins & Mitchell, 2009; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; R. Saxe & Kanwisher, 2003; Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010), is selectively

enhanced when participants make moral judgments in response to bad outcomes. In other words, participants revise their evaluations of agents' mental states (e.g., whether beliefs were justified or reasonable) in light of the outcome. In sum, even when we judge accidents harshly, we may do so primarily by considering important mental state factors (e.g., belief justification, negligence, recklessness) and not simply the outcome of the action.

Most of the time, then, internal, unobservable mental states (e.g., beliefs, intentions, desires) carry more moral weight than external outcomes. Extraordinarily, recent research suggests that mental states overwhelm even other external factors, including external, situational constraint, e.g., whether an agent could have done otherwise (Woolfolk, Doris, & Darley, 2006). In one study, participants read variations of the following story:

> Bill discovers that his wife Susan and his best friend Frank have been involved in a love affair. All three are flying home from a group vacation on the same airplane. In one variation of the story, their plane is hijacked by a gang of ruthless kidnappers who surround the passengers with machine guns, and order Bill to shoot Frank in the head; otherwise, they will shoot Bill, Frank, and the other passengers. Bill recognizes the opportunity to kill his wife's lover and get away with it. He wants to kill Frank and does so.

In another variation: "Bill forgives Frank and Susan and is horrified when the situation arises but complies with the kidnappers' demand to kill Frank." When Bill *wanted* to kill Frank, participants actually judged Bill to be more responsible for Frank's death, and the killing to be more morally wrong, even though Bill's desire played no causal role in Frank's death in either case. Mental state factors are clearly at the forefront of our minds when we're making moral judgments.

Blaming immoral agents for their harmful desires and intentions, as in the case of vengeful Bill above, may be easy and automatic for most people (though a key exception, patients with focal lesions to the ventromedial prefrontal cortex, is discussed further below). Forgiving accidents, however, presents a greater challenge. Prior research indicates substantial individual differences among healthy adults in the moral judgments of accidents (Cohen & Rozin, 2001; Sargent, 2004; Young & Saxe, 2009). In one study, participants who showed greater recruitment of brain regions for mental state reasoning, i.e. the RTPJ, were more likely to forgive accidents, showing greater consideration of the agent's innocent intention (versus the action's harmful outcome), compared to participants with lower RTPJ responses during moral judgment (Young & Saxe, 2009a).

In development, full forgiveness or exculpation for accidents does not emerge until approximately seven years of age, surprisingly late in childhood. Interestingly, five-year-old children appear to be capable of reasoning about false beliefs: in the paradigmatic "false belief task", children predict that observers will look for a hidden object where they last saw the object and not in its true current location (Flavell, 1999; Wellman, Cross, & Watson, 2001). However, these same children will largely fail to forgive accidents to the same extent as healthy adults: if a false belief leads an agent to unknowingly cause harm to another (e.g., as a result of mistaking poison for sugar), the agent is judged just as bad as though the harm had been caused on purpose (Piaget, 1965/1932). Thus, the ability to integrate mental states (like beliefs and intentions) into moral judgments, versus the ability to simply encode mental states, may reflect distinct developmental achievements, with distinct functional profiles in the RTPJ (Young & Saxe, 2008). Consistent with this hypothesis, adults diagnosed with Asperger's

Syndrome, who pass standard false belief tasks, also deliver especially harsh moral judgments of accidents (Moran et al., 2011).

Whereas neurotypical adults have particular difficulty exculpating accidents, another population shows a specific deficit in delivering moral judgments of failed attempts to harm, including failed murder attempts – harmful intentions in the absence of harmful outcomes (Young et al., 2010). Patients with focal lesions to the ventromedial prefrontal cortex (vMPFC) judged attempted harms as more morally permissible compared to neurotypical control participants. Strikingly, vMPFC patients even judged attempted harms as more morally permissible than accidents – a reversal of the normal pattern of moral judgments (Cushman, 2008). Consistent with these behavioral data, a recent fMRI study indicates a positive correlation between vMPFC activity and moral judgments of failed attempts to harm; neurotypical participants with high vMPFC responses judged failed attempts more harshly than individuals with low vMPFC responses (Young & Saxe, 2009). Together, these results suggest that vMPFC patients may be unable to trigger an appropriate emotional response to abstract mental state information, i.e. harmful intentions. The vMPFC may not play a role in encoding mental states per se; rather, the vMPFC supports emotional responses to mental state content. This account is consistent with prior work revealing a role for the vMPFC in generating emotional responses to any abstract information (Bechara & Damasio, 2005). Thus, vMPFC patients deliver moral judgments based primarily on the neutral (permissible) outcome, reflecting a "no harm, no foul" mentality.

What then are the neural mechanisms that directly support the encoding and integration of mental states in moral judgments? Recent evidence suggests that specific

brain regions support multiple distinct cognitive components of mental state reasoning for moral judgment: the initial encoding of the agent's mental state (Young & Saxe, 2008), the use and integration of mental states (e.g., with outcomes) for moral judgment (Young, et al., 2007), spontaneous mental state inference when mental states are not explicitly provided in the scenario (Young & Saxe, 2009b), and even post-hoc reasoning about beliefs and intentions to rationalize or justify moral judgments (Kliemann, Young, Scholz, & Saxe, 2008; Young, Nichols, et al., 2010; Young, Scholz, & Saxe, 2011).

Building on prior research on the neural substrates for mental state reasoning in the service of action prediction and explanation (Perner, et al., 2006; R. Saxe & Kanwisher, 2003), recent research suggests that a key brain region for moral judgment is the right temporo-parietal junction (RTPJ). In one study, mentioned above, individual differences in moral judgments were significantly correlated with individual differences in the RTPJ response (Young & Saxe, 2009a). Participants with a high RTPJ response during moral judgment, and a putatively more robust mental state representation (e.g., of the false belief and innocent intention), assigned less blame to agents causing accidental harm. Participants with a low RTPJ response (and weaker mental state representation) assigned more blame, similar to young children and individuals with Asperger's Syndrome (Moran, et al., 2011). One source of developmental change in moral judgments (from a reliance on outcomes to a reliance on mental states) may therefore be the maturation of specific brain regions for representing mental states such as beliefs – consistent with recent research suggesting the RTPJ may be late maturing (Gweon, Dodell-Feder, Bedny, & Saxe, in press; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009).

Finally, disrupting RTPJ activity also disrupts the use of mental state information for moral judgment. A recent study probing moral judgments used transcranial magnetic stimulation (TMS) to produce a temporary "virtual lesion" in the RTPJ (Young, Camprodon, et al., 2010). After using fMRI to functionally localize the RTPJ in each participant, offline TMS and online TMS were used to modulate neural activity in two experiments. In both experiments, TMS to the RTPJ versus the control region reduced participants' reliance on mental states in their moral judgments, and consequently increased the role of outcomes. For example, disrupting RTPJ activity led to more lenient judgments of failed attempts to harm; participants based their moral judgments more on the neutral outcome (versus the harmful intent). Thus, compromised mental state reasoning in the case of neurodevelopmental disorders (e.g., high functioning autism) or via TMS leads to abnormal moral cognition.

The findings reviewed in this section provide behavioral and neural evidence for mental state reasoning as a key cognitive process for moral judgment. In sum, evaluating moral agents and their actions requires observers to represent and assess the underlying mental states.

Morality on the ground

In this second section, we argue that the key relationship between mind attribution and morality extends beyond the domain of judgment. As social animals, we are not merely passive observers or judges of other people's moral and immoral actions; instead, we are active participants in the social world. We engage in good and bad behaviors toward others, and we must decide how to act toward whom and, in turn, determine who

is capable of helping or hurting us. In other words, as moral actors, we must determine who is friend and who is foe. Indeed, the motivation for affiliation with others (e.g., potential allies) and the motivation for action prediction (e.g., potential enemies) are major determinants of mind attribution (Epley, Waytz, & Cacioppo, 2007; Waytz, Gray, Epley, & Wegner, 2010). It is the *moral* salience of these social contexts that requires and engages mind attribution both for understanding others and for anticipating their actions.

Whether reasoning about allies or enemies, people must engage in mind attribution. Determining who's with us and who's against us (and, at a more basic level, who counts as "us" versus "them") through intergroup categorization, is typically an automatic and spontaneous process (Brewer, 1979). Minimal cues to ingroup and outgroup status lead people to encode alliances and coalitions (Kurzban, Tooby, & Cosmides, 2001; Turner, Brown, & Tajfel, 1979). Furthermore, the same neural architecture responds to ingroup and outgroup members after minimal exposure to these individuals. The amygdala, a region involved in processing motivationally relevant information, is responsive to faces of both ingroup members and outgroup members depending on the processing goals of the perceiver (Lieberman, Hariri, Jarcho, Eisenberger, & Bookheimer, 2005; Van Bavel, Packer, & Cunningham, 2008). Intergroup categorization thus allows us to determine who in our social environment is capable of helping and harming us and whom we ourselves might be able to help or harm. Thus, allies and enemies alike require social reasoning but elicit distinct motivational strategies. As we argue below, the motivation for affiliation underlies our reasoning about allies, whereas the motivation for action prediction, for anticipating future actions or even attacks, underlies our reasoning about enemies.

The motivation to affiliate with others, and to do good for others, triggers the desire to know others' minds. Understanding the minds of other people is critical for coordination, cooperation, and communication (Epley, & Waytz, 2010). Indeed, a number of research programs have suggested that the capacity for understanding other minds is precisely the capacity that has allowed humans to operate effectively in large social groups (Baron-Cohen, 1995; Humphrey, 1976; Tomasello, Carpenter, Call, Behne, & Moll, 2005). Furthermore, interpersonal liking is often correlated with mind attribution (Kozak, Marsh, & Wegner, 2006), and people will attribute particular mental states, such as secondary emotions, preferentially to ingroup members versus outgroup members (Harris & Fiske, 2006; Leyens et al., 2000). Thus, the motivation for social connection, especially with those within our own moral circle, is a major determinant of mind attribution.

In particular, motivation for social connection leads people to more accurately infer people's emotions from facial or vocal cues (Pickett, Gardner, & Knowles, 2004). This motivation can also increase people's tendency to perceive mental states in nonhuman entities such as supernatural agents, technology, and pets, thereby anthropomorphizing them (Aydin, Fischer, & Frey, 2010; Epley, Akalis, Waytz, & Cacioppo, 2008; Epley, Waytz, Akalis, & Cacioppo, 2008). Furthermore, neuroimaging studies have shown that cooperation and generous behavior toward others elicit activity in brain regions that support social cognition including the MPFC (McCabe, Houser, Ryan, Smith, & Trouard, 2001; Waytz, Zaki, & Mitchell, 2012), demonstrating the deployment of mind attribution for good moral behavior. These findings show that when people seek positive social interactions with other moral agents, they engage in mental

state reasoning and may even become hyperattentive to specific features (e.g., emotions) of their social partners' mental states.

Likewise, the motivation to harm others, including our enemies, and to defend ourselves against others' harmful actions, also requires a robust understanding of other minds, especially for predicting future actions or attacks. Thus, negative moral interactions are also accompanied by the desire to know others' mental states. As we describe below, the motivation to understand and predict others' actions is therefore another major determinant of mind attribution (Dennett, 1987; Epley, et al., 2007).

A number of studies have demonstrated that motivation to attain mastery over others leads to mind attribution. In one instance, this effect obtains for non-human agents; entities that operate unpredictably and that require explanation elicit more attribution of humanlike mental states (i.e., anthropomorphism) (Waytz et al., 2010; Morewedge, 2009). When people are motivated to gain control or to explain events in the environment, they will often do so by looking to anthropomorphic Gods or other mentalistic agents (K. Gray & Wegner, 2010a; Kay, Gaucher, Napier, Callan, & Laurin, 2008; Kay, Moscovitch, & Laurin, K., 2010; Kelemen, 2009). Together, these studies support the idea that the motivation to explain, predict, and understand—the motivation to attain mastery over others—increases mental state reasoning.

Functional neuroimaging evidence suggests that when people are placed in competitive situations with others, in which they must predict and understand others' behavior, brain regions for mental state reasoning including the MPFC (Decety, Jackson, Sommerville, Chaminade, & Meltzoff, 2004) and TPJ (Halko, Hlushchuk, Hari, & Schurmann, 2009) are robustly recruited. One study using positron emission topography

(PET) demonstrated that during a competitive game, the MPFC was preferentially engaged when participants believed they were playing an entity capable of strategic moral or immoral behavior (a human being) versus an entity incapable of such behavior (Gallagher, Jack, Roepstorff, & Frith, 2001). Together, these studies suggest that mind attribution supports not only good moral behavior, such as cooperation with allies, but also strategic interaction with unpredictable others, including enemies.

To demonstrate the relationship between mind attribution and distinct moral motivations towards enemies and allies, we conducted a series of studies targeting both the motivation for social connection and the motivation for action prediction in a single paradigm (Waytz & Young, 2012). In a first study, American participants answered questions about the United States Army and the Taliban, obvious ally and enemy groups, respectively. Participants rated how much they desired social connection with each group and how much they were motivated to predict the actions of each group. Motivation for social connection predicted attribution of mind to the U.S. Army (ingroup / ally), whereas motivation for action prediction did not. By contrast, motivation for action prediction predicted attribution of mind to the Taliban (outgroup / enemy), whereas motivation for social connection did not. A second study asked American Democrats and Republicans (during the contentious 2010 mid-term elections) to evaluate both the Democratic and Republican party on similar measures, and the same pattern of results emerged. Motivation for social connection uniquely predicted mind attribution toward participants' own political party, whereas motivation for action prediction uniquely predicted mind attribution toward the opposing political party. Taken together, these findings

demonstrate that anticipating both positive and negative social interactions (with other moral agents) provokes mind attribution.

Although these dual motivations for effective social interaction engage mind attribution, they may engage different forms of mind attribution. In fact, fMRI research demonstrates that different nodes of the neural network for theory of mind are preferentially engaged by cooperation versus competition. In one study, in which participants were instructed to play a strategic game, the posterior cingulate was more involved in cooperation, whereas the MPFC was more involved in competition (Decety et al., 2004). Another study demonstrated that reasoning about others' cooperative mental states versus deceptive mental states recruited distinct brain regions for theory of mind. Whereas both cooperation and deception elicited activation in the TPJ and precuneus, deception selectively increased activation in the MPFC (Lissek et al., 2008). Based on this pattern, the authors suggest that different systems are involved in processing mental states that match an observer's expectations (in this case, cooperative intentions) versus mental states intended to undermine the observer's expectations. More broadly, these neural findings suggest distinct cognitive processes for mental state reasoning in cooperative versus competitive contexts.

One hypothesis regarding the differential types of minds attribution for cooperation versus competition suggests two distinct dimensions of mind. People think about mind in terms of *agency* (i.e., the capacity to plan, to think, and to intend) as well as *experience* (i.e., the capacity to feel pain and pleasure) (H. M. Gray, Gray, & Wegner, 2007). The attribution of experience grants a person status as a moral *patient* (i.e., someone who is capable of *experiencing* the moral acts of others), whereas the attribution

of agency grants a person status as a moral *agent* (i.e., someone who is capable of *doing* moral acts to others) (K. Gray & Wegner, 2009; K. Gray & Wegner, 2010b). Therefore, people express more moral concern toward moral patients, whereas they view moral agents as morally responsible and therefore blameworthy or praiseworthy for their actions (H. M. Gray, et al., 2007).

The tendency to associate experience and agency with distinct moral characters suggests the motivation for social connection and the motivation for action prediction might differentially trigger attributions of experience and agency, respectively. The motivation for social connection involves the desire to give and receive moral care from another person through prosocial behavior, including cooperation. Therefore, the motivation for social connection should preferentially increase the attribution of *experience* to others. By contrast, the motivation for action prediction entails identifying entities that are capable of planning and acting intentionally and, furthermore, determining the content of those plans and intentions. This motivation should be uniquely linked to the preferential attribution of *agency* to others (Kozak & Czipri, 2011). Although no existing evidence speaks to this distinction, future behavioral and neural approaches should uncover whether differential motivations for positive and negative moral interactions map onto the attributions of distinct dimensions of mind.

Most important, considerable research suggests that moral action, and the motivation to engage in moral action—whether positive or negative—depends crucially on mind attribution. People consider the minds of other moral actors not only when judging third-party behavior, but also when attempting themselves to engage with others, either allies or enemies. Behaving well and behaving badly may reside on opposite ends

of the moral spectrum, but both depend crucially on mental state reasoning – reasoning about the mind of friends and foes.

From the mind on the ground to the mind on high

In this chapter, we have described how mind attribution is critical for judging moral actions as well as for engaging in good and bad actions towards others. Yet another link between mind attribution and morality, to be explored in future research, is the moral actor's consideration of an evaluative mind or an ultimate judge (see Figure 1). A number of studies suggest that when people decide whether to engage in righteous or reproachable actions, they consider whether others are watching, a tendency commonly known as impression management (Leary & Kowalski, 1995). For instance, in monetary exchange games that allow people to behave selfishly or generously, people behave more cooperatively when merely primed with reminders of a judgmental God (Shariff & Norenzayan, 2008) or cues that others are watching (Haley & Fessler, 2005). Perceiving the presence of a mindful, nonhuman agent also increases honesty and hesitance to cheat in a game (Bering, McLeod, & Shackleford, 2005; Waytz, Cacioppo, & Epley, 2010).

Future research should investigate whether personal decisions about acting morally or immorally in fact engage the tendency to search for or perceive a mind on high – either the mind of peer observers or an ultimate moral judge. For now, though, it is clear that mind attribution plays a primary role in both moral judgment and social interactions between moral actors.

References

Aydin, N., Fischer, P., & Frey, D. Turning to God in the face of ostracism: Effects of social exclusion on religiousness. *Personality and Social Psychology Bulletin*, *36*, 742-753

Baron-Cohen, S. (1995). *Mindblindness : an essay on autism and theory of mind*. Cambridge, Mass.: MIT Press.

Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. *Games and Economic Behavior, 52*, 336-372.

Bering, J. M., McLeod, K. A. & Shackelford, T. K. (2005). Reasoning about dead agents reveals possible adaptive trends. *Human Nature*, *16*, 60–81.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin, 86*, 307-324.

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality. *Journal of Personality and Social Psychology, 81*(4), 697-710.

Cushman, F. (2008). Crime and Punishment: Distinguishing the roles of causal and intentional analysis in moral judgment. *Cognition, 108*(2), 353-380.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PLoS One, 4*(8), e6699.

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.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, Mass.: MIT Press.

Epley, N., & Waytz A. (2010). Mind Perception. In D. T. G. S.T. Fiske, & G. Lindzey (Ed.), *The Handbook of Social Psychology* (5th ed., pp. 498-541). New York: Wiley.

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*, 114–120.

Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition, 26*(2), 143-155.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864-886.

Flavell, J. H. (1999). Cognitive Development: children's knowledge about the mind. *Annual Review of Psychology 50*, 21-45.

Gallagher, H. L., Jack, A.I., Roepstorff, A., Frith, C.D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, *16*, 814-821.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619-619.

Gray, K., & Wegner, D. A. (2009). Moral Typecasting: Divergent Perceptions of Moral Agents and Moral Patients. *Journal of Personality and Social Psychology, 96*(3), 505-520.

Gray, K., & Wegner, D. M. (2010a). Blaming God for Our Pain: Human Suffering and the Divine Mind. *Personality and Social Psychology Review, 14*(1), 7-16.

Gray, K., & Wegner, D. M. (2010b). Torture and judgments of guilt. *Journal of Experimental Social Psychology, 46*(1), 233-235.

Gweon, H., Dodell-Feder, D., Bedny, M., & Saxe, R. (in press). Theory of Mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. . *Child Development*.

Haley, K. J., & Fessler, D. M.T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, *26*, 245– 256.

Halko, M.-L., Hlushchuk, Y., Hari, R., & Schurmann, M. (2009). Competing with peers: Mentalizing-related brain activity reflects what is at stake. *Neuroimage, 46*(2), 542-548.

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

Hart, H. L. A. (1968). *Punishment and Responsibility*. Oxford: Oxford University Press.

Humphrey, N. K. (1976). The social function of intellect. In P. P. G. Bateson & R. A. Hinde (Eds.), *Growing points in ethology*. Oxford, UK: Cambridge University Press.

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under uncertainty: dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cereb Cortex, 20*(2), 404-410.

Kamm, F. M. (2001). *Morality, Mortality: Rights, duties, and status*. New York: Oxford University Press.

Kay, A. C., Gaucher, D., Napier, J. L., Callan, M. J., & Laurin, K. (2008). God and the Government: Testing a compensatory control mechanism for the support of external systems. *Journal of Personality and Social Psychology, 95*, 18-35.

Kay, A. C., Moscovitch, D. M., & Laurin, K. (2010). Randomness, attributions of arousal, and belief in god. *Psychological Science, 21*, 216-218.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. . *Cognition, 111*, 138-143.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949-2957.

Kozak, M.J., & Czipri, A. (2011). *Behind enemy minds: Mind attribution and perceived threat*. Manuscript in preparation, Pace University, New York, NY.

Kozak, M. J., 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*, 543–555.

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences of the United States of America, 98*(26), 15387-15392.

Leary, M.R., & Kowalski, R.M. (1995). *Social anxiety*. London: Guildford Press.

Leyens, J., 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 in-groups and out-groups. *Personality and Social Psychology Review, 4,* 186-197.

Lieberman, M. D., Hariri, A., Jarcho, J. M., Eisenberger, N. I., & Bookheimer, S. Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience, 8*(6), 720-722.

Lissek S, Peters S, Fuchs N, Witthaus H, Nicolas V, Tegenthoff M, et al. (2008). Cooperation and deception recruit different subsets of the theory of mind network. *PLoS ONE*, *3*, e2023.

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 of the United States of America, 98*(20), 11832-11835.

Mikhail, J. M. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143-152.

Moran, J. M., Young, L. L., Saxe, R., Lee, S. M., O'Young, D., Mavros, P. L., et al. (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proc Natl Acad Sci U S A*.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology: General, 138*, 535-545.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: the roles of left and right temporo-parietal junction. *Soc Neurosci, 1*(3-4), 245-258.

Piaget, J. (1965/1932). *The Moral Judgment of the Child*. New York: Free Press.

Pickett, C. L., Gardner, W. L., & Knowles, M. (2004). Getting a cue: The need to belong and enhanced sensitivity to social cues. *Personality and Social Psychology Bulletin, 30*, 1095-1107.

Sargent, M. J. (2004). Less thought, more punishment: need for cognition predicts support for punitive responses to crime. *Pers Soc Psychol Bull, 30*(11), 1485-1493.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind". *Neuroimage, 19*(4), 1835-1842.

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., & Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child Dev, 80*(4), 1197-1209.

Shariff, A.F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science, 18*, 803–809.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691.

Turner, J. C., Brown, R. J., & Tajfel, H. (1979). Social comparison and group interest in ingroup favouritism. *European Journal of Social Psychology, 9*(2), 187-204.

Van Bavel, J. J., Packer, D. J., & Cunningham, W. A. (2008). The neural substrates of in-group bias: a functional magnetic resonance imaging investigation. *Psychological Science, 19*(11), 1131-1139.

Waytz, A., Cacioppo, J. T., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*, 219-232.

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in cognitive sciences, 14*(8), 383-388.

Waytz, A., & Young, L. (submitted). *Attributing mind across enemy lines: When we treat outgroups as mental agents*.

Waytz, A., Zaki, J., Mitchell, J.P. (2012). Response of the dorsal medial prefrontal cortex predicts altruistic behavior. *Journal of Neuroscience*, *32*, 7646-7650.

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

Woolfolk, R. L., Doris, J. M., & Darley, J. M. (2006). Identification, situational constraint, and social cognition: Studies in the attribution of moral responsibility. *Cognition, 100*(2), 283-301.

Young, L., Bechara, A., Tranel, D., Damasio, H., Hauser, M., & Damasio, A. (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron, 65*, 845-851.

Young, L., Camprodon, J., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporo-parietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgment. *Proc Natl Acad Sci U S A, 107*, 6753-6758.

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.

Young, L., Nichols, S., & Saxe, R. (2010). Investigating the Neural and Cognitive Basis of Moral Luck: It's Not What You Do but What You Know. *Review of Philosophy and Psychology, 1*, 333-349.

Young, L., & Saxe, R. (2008). The neural basis of belief encoding and integration in moral judgment. *NeuroImage, 40*, 1912-1920.

Young, L., & Saxe, R. (2009a). Innocent intentions: a correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47*(10), 2065-2072.

Young, L., & Saxe, R. (2009b). An fMRI Investigation of Spontaneous Mental State Inference for Moral Judgment. *Journal of Cognitive Neuroscience, 21*, 1396-1405.

Young, L., Scholz, J., & Saxe, R. (2011). Neural evidence for "intuitive prosecution": The use of mental state information for negative moral verdicts. *Social Neuroscience, 6,* 302-315.

Young, L., Koenigs, M., Kruepke, M., & Newman, J. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology.*

Figure 1. Mental state reasoning for moral cognition occurs at multiple levels. Arrows indicate direction of mind attribution. Observers who make third party judgments ("Morality on high") attribute mind to moral actors. Moral actors who interact with allies and enemies engage in mental state reasoning for affiliation, action understanding and prediction ("Morality on the ground"). Actors may also infer the mind of an evalutive judge ("From the mind on the ground to the mind on high").