Social and Personality Psychology Compass 7/8 (2013): 585–604, 10.1111/spc3.12044

When Mental States Matter, When They Don't, and What That Means for Morality

Liane Young* and Lily Tsoi
Boston College

Abstract

Research has shown that moral judgments depend on the capacity to engage in mental state reasoning. In this article, we will first review behavioral and neural evidence for the role of mental states (e.g., people's beliefs, desires, intentions) in judgments of right and wrong. Second, we will consider cases where mental states appear at first to matter less (i.e., when people assign moral blame for accidents and when explicit information about mental states is missing). Third, we will consider cases where mental states, in fact, matter less, specifically, in cases of "purity" violations (e.g., committing incest, consuming taboo foods). We will discuss how and why mental states do not matter equivalently across the multi-dimensional space of morality. In the fourth section of this article, we will elaborate on the possibility that norms against harmful actions and norms against "impure" actions serve distinct functions – for regulating *interpersonal* interactions (i.e., harm) versus for protecting the *self* (i.e., purity). In the fifth and final section, we will speculate on possible differences in how we represent and reason about *other people*'s mental states versus our *own* beliefs and intentions. In addressing these issues, we aim to provide insight into the complex structure and distinct functions of mental state reasoning and moral cognition. We conclude that mental state reasoning allows us to make sense of other moral agents in order to understand their past actions, to predict their future behavior, and to evaluate them as potential friends or foes.

Many of us find morality interesting because reasonable people – ordinary folks and professional philosophers alike – disagree about what's right and wrong. Yet it is precisely this disagreement that presents a challenge for students of moral psychology. If we wish to understand the moral mind, whose mind should we study? One solution is to tackle as many minds as we can find and not simply undergraduate minds sitting in introductory psychology classes (a current source of many research participants; Henrich, Heine, & Norenzayan, 2010). A complementary approach, though, is to identify universal moral rules (e.g., DeScioli, Asao, & Kurzban, 2012; Hauser, 2006; Mikhail, 2007) that apply to all moral judgments, independent of the culture-specific content of culture-specific moral codes. One candidate rule (i.e. "rule of intent") is this: *intentional* wrongdoings are morally worse than *accidental* wrongdoings. For example, if we were to discover that aliens on another planet believe certain seemingly arbitrary actions to be morally wrong – touching elbows, making eye contact, and mixing salt with water – we might infer that, within the alien culture, these "crimes" would be perceived as even more wrong when committed on purpose versus by accident. In our own culture, intent makes the difference between murder and manslaughter (Hart & Honore, 1959). As such, mental states, including people's intentions, beliefs, and desires, may represent a cognitive constant in the messiness of morality.

In this article, we will first present behavioral and neural evidence for the role of mental states in moral judgments of right and wrong. Second, we will consider cases where mental states appear at first to matter less (i.e., when people assign moral blame for accidents and

when explicit information about mental states is missing). Third, we will consider cases where mental states, in fact, matter less, specifically, in cases of "purity" violations (e.g., committing incest, consuming taboo foods). We will discuss how and why mental states do not matter equivalently across the space of morality. In the fourth section of this article, we will elaborate on the possibility that norms against harmful actions and norms against "impure" actions serve distinct functions – for regulating *interpersonal* interactions (i.e., harm) versus for protecting the *self* (i.e., purity). In the fifth and final section, we will speculate on possible differences in how we represent and reason about *other people*'s mental states versus our *own* beliefs and intentions. In addressing these issues, we aim to provide insight into the complex structure and distinct functions of mental state reasoning and moral cognition.

For Moral Judgment, It's the Thought that Counts

Folk intuition and the law converge on the notion that murder is worse than manslaughter, as mentioned above. In this first section, we will review behavioral and neural evidence for the more general hypothesis that mental states matter for moral judgments – when judging innocent and malicious intentions. We will also review how reduced or impaired mental state reasoning influences moral judgment.

Importantly, empirical work supports a role for intent not only when lives are at stake but also across diverse contexts (Cushman, 2008; Decety, Michalska, & Kinzler, 2012; Killen, Mulvey, Richardson, Jampol, & Woodward, 2011; Young, Cushman, Hauser, & Saxe, 2007). For example, allocating resources selfishly or unfairly is seen as wrong, but doing so *on purpose* is seen as worse than doing so *by accident* (Cushman, Dreber, Wang, & Costa, 2009). In addition, information about mental states informs not simply moral judgments of individuals but also moral judgments of groups (e.g., corporations, unions, political parties; Waytz & Young, 2012). Put plainly, innocent intentions in the case of accidents serve to mitigate blame.

Meanwhile, *malicious* intentions can lead to assignments of blame even in the absence of actual harm, as in failed murder attempts. Indeed, unsavory desires on their own are often enough to evoke blame even when those desires are causally disconnected from the harmful event (e.g., a man who is forced at gunpoint by attackers to kill his wife's lover, whom he wants dead anyway; Woolfolk, Doris, & Darley, 2006). We even assign moral blame for harmful desires when no attempt at harm occurs (Inbar, Pizarro, & Cushman, 2012). For example, we blame agents who seek to benefit from others' misfortune even when the unfortunate events are out of those agents' control (e.g., traders who profited from but did not cause or even attempt to cause the subprime mortgage crisis; Inbar, Pizarro & Cushman, 2012; see also Pizarro, Uhlmann, & Bloom, 2003). These findings reveal that information about intentions and desires often dominate our moral judgments. It is the failure to process emotionally salient intentions (e.g., murderous desires) that results in abnormally *lenient* judgments of failed attempts to harm (including failed murder attempts), as in the case of patients with focal lesions to the ventral sub-region of the medial pre-frontal cortex (VMPFC), an area implicated in social–emotional processing (Young et al., 2010).

Recent work has zeroed in on the neural mechanisms supporting mental state reasoning for moral judgments (for reviews, see Dungan & Young, 2011; Young, 2013). This work builds directly on prior (and ongoing) work on the neural basis of social cognition. Indeed, a number of studies using functional magnetic resonance imaging (fMRI) have revealed a consistent neural network for social cognition, including sub-regions of the medial pre-frontal cortex (MPFC), right and left temporo-parietal junction (RTPJ and LTPJ), and precuneus (Fletcher et al., 1995; Gallagher et al., 2000; Gobbini et al., 2007; Saxe &

Kanwisher, 2003). A recent activation likelihood estimation (ALE) meta-analysis of 135 studies showed overlapping neural networks for social cognition (e.g., mental state reasoning) and moral cognition (Bzdok et al., 2012). Much of our own work has targeted the role of one key region, the RTPJ (Perner et al., 2006; Saxe & Kanwisher, 2003), for moral cognition (Young et al., 2007). We note that other work indicates additional roles for the RTPJ, for example, in prosocial behavior, but our focus here will be on moral judgment (Morishima, Schunk, Bruhin, Ruff, & Fehr, 2012; see also Carter, Bowling, Reeck, & Huettel, 2012).

Recent evidence reveals a role for the RTPJ in supporting mental state reasoning for moral judgment. The RTPJ supports the initial encoding of the mental state and its eventual integration with other task-relevant information (e.g., outcome information) for moral judgment (see Figure 1; Young & Saxe, 2008). At the time of integration, the magnitude of response in the RTPJ is correlated with moral judgment; participants with a higher RTPJ response to accidental harms, on average, deliver more forgiveness and less blame for accidents, compared to participants with a lower RTPJ response (Young & Saxe, 2009). More recently, the use of multi-voxel pattern analysis (MVPA) in fMRI research has allowed us to determine whether the spatial pattern of neural activity across voxels within the RTPJ (as well as other brain regions) differentiates between intentional harms and accidental harms. In other words, does the RTPJ support this computation, and, if so, how? We found that even though the magnitude of neural response (averaged across voxels) in the RTPJ is high for intentional and accidental harms, MVPA reveals that the voxel-wise pattern in the RTPJ distinguishes between intentional and accidental harms (Koster-Hale, Saxe, Dungan, & Young, 2013); also, individual differences in neural discriminability correlate with individual differences in behavioral discriminability – the extent to which participants distinguish between intentional and accidental harms in their moral judgments. In addition, research using high-density event-related potentials (ERPs) reveals differentiation in the RTPJ between intentional and accidental harms as fast as 62 ms post-stimulus while participants

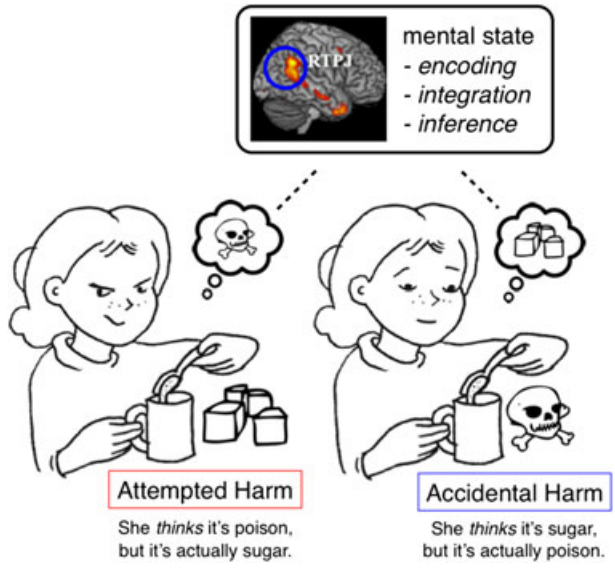


Figure 1 The right temporo-parietal junction (RTPJ) is a key brain region for processing mental states such as people's beliefs and intentions (e.g., "she wanted to poison him", left; "she thought it sugar", right) during moral cognition (e.g., judgments of attempted harms, left; accidental harms, right).

view morally relevant visual stimuli (Decety & Cacioppo, 2012). Convergent behavioral evidence suggests that intent represents an early input to moral judgment (Malle & Holbrook, 2012). Finally, temporarily disrupting activity in the RTPJ using transcranial magnetic stimulation (TMS) leads to more outcome-based (i.e., fewer intent-based) moral judgments (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010).

Other work has targeted the role of mental state reasoning in moral cognition across the life span. As children develop the capacity to reason fully and flexibly about mental states, including false beliefs, they are more likely to forgive an accidental agent, e.g., someone who throws away a classmate's cupcake, which is hidden in a brown paper bag that looks like trash (Killen et al., 2011). However, younger children between the ages of three and four years focus more on the bad outcome (e.g., cupcake in the trash) and less on the lack of negative intent (e.g., she didn't *mean* any harm, she *didn*'t *know* it was a cupcake, she *thought* it was trash) and therefore assign more blame for an accident (see also Cushman, Sheketoff, Wharton, & Carey, 2013; Hebble, 1971; Piaget, 1965; Shultz, Wright, & Schleifer, 1986; Yuill & Perner, 1988; Yuill, 1984; Zelazo, Helwig, & Lau, 1996). At the other end of the age spectrum, older adults (mean of 71.8 years) also show increased reliance on bad outcomes (Moran, Jolly, & Mitchell, 2012), associated with reduced activity in the dorsal sub-region of the MPFC, another node in the neural network for social cognition and mental state reasoning. Outcome-based judgments may therefore be described by a U-shaped curve, appearing most robustly at either end of the age spectrum; on one end are young children who have not yet developed mature mental state reasoning capacities, and on the other end are older adults who show a decline in mental state reasoning and associated neural activity.

A similar focus on outcomes versus intentions emerges in the case of autism, a neurodevelopmental disorder associated with impairments in social cognition, including specific deficits in mental state reasoning (Moran et al., 2011). Like young children and older adults, high-functioning individuals with autism also judge accidents more harshly on the basis of the bad outcome rather than the neutral intent. This behavioral pattern in high-functioning autism is also consistent with a recent fMRI study using the approach of MVPA (described above): in participants with autism, the spatial pattern of neural activity in the RTPJ does not reliably distinguish between intentional and accidental harms (Koster-Hale et al., 2013). We suggest that accidents pose a particular challenge because they pit salient information about a bad outcome against relatively neutral information about a false belief or lack of knowledge or intent. Forgiving an accident may require overriding a pre-potent response to an emotionally salient outcome from a lost cupcake to a lost life (cf. Miller et al., 2010).

As we discuss further in the next section, forgiving an accident may pose a challenge not simply for younger and older populations and individuals with autism but for everyone else as well. According to a recent study, the individuals who are able to *ignore* accidental outcomes (e.g., salient harms) in favor of "hyper-rational", intent-based judgments are those with a clinical diagnosis of psychopathy (Young et al., 2012). Other studies reveal individual differences in the tendency to use mental state information for exculpation (Cohen & Rozin, 2001; Young & Saxe, 2009). In one study, participants were instructed to weigh "mitigating circumstances" (relevant for the sentencing of guilty defendants); regions for mental state reasoning (including the TPJ, dorsal MPFC, and precuneus) and for emotional empathy (including the right middle insula) were recruited, and individual differences in the inclination to mitigate were correlated with activity in the latter set of regions (Yamada et al., 2012). The ability to integrate cognitive inputs for moral judgment appears to depend also on functional connectivity between areas for social cognition and emotional processing, i.e., ventral MPFC and the amygdala (Decety et al., 2012). These results highlight the multiple mechanisms that

may interact and compete during moral cognition (i.e., outcome processing versus mental state processing; Cushman, 2008; Greene et al., 2001; Greene et al., 2004; Young et al., 2007).

In sum, mental states matter for moral judgments of harmful actions – from poisoning other people to tossing other people's cupcakes in the trash. Considering innocent intentions can lead people to assign less blame for accidents, and considering guilty intentions can lead to moral condemnation even in the absence of a harmful outcome.

The Thought Counts Even More than You Think

Given the evidence above, the simple point that intentions matter for moral judgment may seem exceedingly uncontroversial. In this section, we will therefore consider two cases in which intentions may appear, upon initial consideration, to count less – first, when we assign moral blame for accidents in spite of the actor's innocent intentions and, second, when intent information isn't available. In brief, it would seem (at first) that (1) when observers blame people for causing accidents, they do so on the basis on the bad outcomes that are brought about *in spite of* benign intentions or false beliefs, and (2) in the absence of mental state information, observers would be forced to rely on other information, such as readily observable, external actions and outcomes. Do these two cases count as "exceptions" to the rule of intent? We show, on the contrary, that moral judgments in both of these cases nevertheless depend on inferences about unobservable, internal mental states.

As mentioned at the end of the first section above, many of us are familiar with the struggle of forgiveness. Like young children, older adults, and individuals with autism, we may find it difficult, on occasion, to forgive even in the face of obviously unintentional harm. Imagine a friend who spills ink on your white carpet (e.g., she thought the container held pencils), a colleague who emails you a debilitating virus (e.g., he believed his own computer to be fixed), or a neighbor who feeds your seriously peanut-allergic child a PB&J sandwich (e.g., she heard you say "almonds"). Whatever blame we cannot help but assign seems due in large part (if not in full) to the bad outcome ranging from annoying to deadly (e.g., the ruined carpet, the corrupted hard drive, the dead child). Indeed, outcomes play an important role especially in judgments of blame and punishment, as opposed to evaluations of an action as wrong or permissible (Cushman, 2008; Cushman et al., 2009) or assessments of personal moral character (e.g., Critcher, Inbar, & Pizarro, 2012; Pizarro & Tannenbaum, 2011; Tannenbaum, Uhlmann, & Diermeier, 2011). We suggest, though, that while outcomes contribute to some of the blame we assign for accidents, mental state factors contribute as well and perhaps even more than outcomes (Young, Nichols, & Saxe, 2010). The case of accidents is the first of the two "exceptions" introduced at the start of this section.

Consider the last example – death by PB&J. Suppose that your neighbor, Mrs. Smith, holds a false belief about your child's allergy – she thinks, falsely, that your child is allergic to almonds, not peanuts – and she accidentally *kills* your child. Imagine a second, happier version of this story in which Mrs. Smith holds a true belief about your child's allergy and therefore causes *no harm* at all. Surely the most salient difference between these two versions is the presence of harm in one case and the absence of harm in the other. However, a more subtle but, as we will argue, a more important difference is the status of the *belief* – whether it is true or false. Does this difference matter? Does it matter more? Imagine now a third "hybrid" version of the story in which Mrs. Smith holds a false belief – she thinks, *falsely*, that your child is allergic to almonds, not peanuts; she prepares a PB&J sandwich for your child, but luckily it gets eaten by her husband, so – hungry child aside – no harm is done. Participants in a recent study judge Mrs. Smith in this third "hybrid" version (false belief, *no harm*) to be nearly as blameworthy as she is in the first version (false belief, *extreme harm*); and both cases

are judged much worse than the happy version (true belief, *no harm*; Young et al., 2010). Why would this difference matter? It turns out that participants assess *false* beliefs as *unjustified* or *unreasonable*; these judgments of negligence drive judgments of moral blame. In other words, we may be harsh on accidents not simply because of the bad outcome but also because of mental state assessments (e.g., Mrs. Smith *should have known* better!).

Now we turn to the second of the two "exceptions". What happens when information about mental states isn't accessible (i.e., when we don't know whether beliefs are true or false, justified or unjustified; whether intentions are innocent or malicious)? Are we forced to rely on alternate routes to moral evaluations? In other words, if we lack direct access to mental state information, we may base our judgments instead on other facts, namely, facts about the actions and the outcomes. Notably, actions and outcomes are often observable, unlike mental states, which are invisible and "hidden" inside people's heads. On the contrary, recent evidence reveals spontaneous mental state reasoning for moral judgment even when explicit information about mental states is not provided. For example, the RTPJ, a key brain area for mental state reasoning, as discussed above, is selectively recruited for *morally relevant* facts versus *morally irrelevant* facts about an action (Young & Saxe, 2009). After hearing about a person who puts a powdery substance into someone's coffee, participants are told either that the powder is poisonous or safe (morally relevant) or that the powder fills the container it is in (morally irrelevant). This differential neural activation (i.e., more RTPJ activity for morally relevant information) suggests that moral judgments depend on spontaneous mental state inference even in the absence of explicit mental state information. Participants may be motivated to consider what moral agents may be thinking (e.g., did she *think* it was sugar?) and whether they know what they are doing (e.g., did she *know* she was poisoning her friend?). Of course, whether RTPJ activation in this instance reflects specific answers to these questions or the mere effort to answer them is worth exploring, which we do next.

How might participants infer mental states in the absence of explicit mental states? A growing body of literature suggests that information about an actor's moral character, or prior record, may inform assessments of his or her harmful (or helpful) actions as intentional or not (Alicke, 2000, 2008; Knobe, 2005, 2010). Recent fMRI studies have combined behavioral and neural approaches to examine mental state inferences in moral contexts. In one fMRI study, participants were set up to interact with "other players" in an economic game, where these "players" behaved fairly or unfairly (Kliemann et al., 2008). After these "interactions", participants read, in the scanner, a series of stories presented as "written by the players" about the players' past actions of ambiguous intent (e.g., broke roommate's lamp, spilled sister's nail polish, shrunk friend's sweater). Harmful actions performed by previously unfair players were judged as more blameworthy and more intentional than the same actions performed by previously fair players, across participants. Notably, these judgments were also associated with increased RTPJ activity, reflecting participants' inference of blameworthy intent based on negative prior record (e.g., whether they had been unfair to the participants in the economic game). Additional fMRI studies have uncovered broadly similar patterns in which increased RTPJ activity reflects inferences of negligence (Young et al., 2010b), negative intent, or even the lack of positive intent (Young, Scholz, & Saxe, 2011). In other words, mental state inferences and assessments, supported by the RTPJ, may depend in part on "background" information about agents' past behavior and moral character.

Together, these findings suggest that (1) considering false beliefs as unreasonable or unjustified can lead people to assign more blame for accidents, and (2) in the absence of explicit mental state information, people make spontaneous inferences about beliefs and intentions for moral evaluation. Thus, mental states matter even in cases where they might

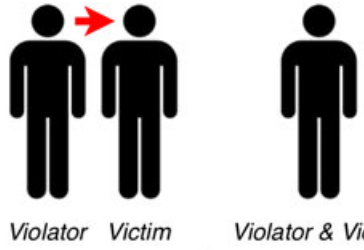
have appeared at first to matter less. By contrast, the next sections focus on cases where mental states, in fact, matter less.

Mental States Matter More for Harmful Actions, Less for Purity Violations

Mental state factors (e.g., whether intentions are good or bad, whether beliefs are true or false, justified or unjustified) represent robust inputs to moral judgments of harmful actions, as revealed in the first two sections above. The aim of this next section is to investigate whether mental states matter equivalently across distinct types of moral actions or distinct *moral domains* (Graham et al., 2011; Graham, Haidt, & Nosek, 2009; Haidt, 2007). Recent neuroimaging evidence using voxel-based morphometry (VBM) suggests that values across distinct moral domains (e.g., harm, purity) are associated with distinct neurological bases; in particular, individual differences in regional brain volume reflect variation in sensitivity to individualizing values (e.g., harm) versus binding values (e.g., purity; Lewis et al., 2012). Thus, we will consider not simply harmful actions (as in the previous sections) but also actions that violate purity norms (e.g., consensual incest, eating taboo foods) – actions that don't necessarily cause harm but that nevertheless appear to defile our moral selves or reflect poorly on our moral character (Inbar, Pizarro, & Cushman, 2012; Tannenbaum et al., 2011).

Extensive work linking distinct moral emotions to distinct moral domains reveals that harmful actions elicit *anger*, whereas purity violations (including "taboo" behaviors related to food and sex) elicit *disgust* (e.g., Horberg, Oveis, Keltner, & Cohen, 2009; Rozin, Lowery, Imada, & Haidt, 1999; Russell & Giner-Sorolla, 2011a, 2013; Russell, Piazza, & Giner-Sorolla, 2013; but see Salerno & Peter-Hagene, in press, for evidence on the interactive effect of anger and disgust on moral outrage). For example, in one emotion induction study, anger-eliciting sounds ("noise music") led uniquely to harsher moral judgments of harm violations (e.g., "crimes against persons"), whereas disgust-eliciting sounds (the sound of an emetic event, vomiting) led uniquely to harsher moral judgments of purity violations (e.g., "crimes against nature"; Seidel & Prinz, 2013). In addition, anger reactions are flexibly influenced by contextual cues and social justifications, whereas disgust reactions are largely immune to these factors (Russell & Giner-Sorolla, 2011b, 2011c; Russell & Giner-Sorolla, 2013). For example, anger is modulated by information about intent (i.e., whether the harm was intentional or accidental), whereas disgust is modulated solely by the presence or absence of an impure action (Russell & Giner-Sorolla, 2011a). Specifically, intentional harms elicit more anger (but not more disgust) than accidents, whereas taboo actions elicit more disgust than non-taboo actions (see Figure 2). An outstanding question is this: do moral *judgments* of harmful actions depend more on *intent* than moral judgments of purity violations?

To test the specific prediction that the "rule of intent" applies differently to distinct moral domains (harm versus purity), we presented participants with a series of intentional and accidental harms, similar to the ones described in the first section, as well as intentional and accidental purity violations (Young & Saxe, 2011). In one example of an accidental harm, participants were asked to imagine preparing a person's dish with peanuts without knowing about the person's peanut allergy. In one example of an accidental purity violation (e.g., incest), participants were asked to imagine sleeping with someone who turned out, the next day, to be a long-lost sibling. As predicted, we found that for a series of violations that were judged equally harshly across domains (harm and purity), participants perceived a large moral difference between intentional and accidental harms and a relatively small (but still statistically significant) moral difference between intentional and accidental purity violations (represented by bodily violations related to food and sex; Chapman & Anderson, 2013; Russell & Giner-Sorolla, 2013; Tybur, Lieberman, Kurzban, & DeScioli, 2013). In a separate



Harm	Purity
Anger / Guilt	Disgust / Shame
Other-Directed	Self-Directed
Intent matters more	Intent matters less

Figure 2 Distinct moral norms serve distinct functions for regulating interpersonal interactions (left) versus for protecting the self (right): it is wrong to harm others, and it is wrong to defile the self. Differences concern content (harm versus purity), emotional signatures (anger/guilt versus disgust/shame), function (other-focus versus self-focus), and reliance on mental state information such as intent.

experiment, not only were accidental harms judged *less harshly* (based on false beliefs and *innocent* intentions) than accidental purity violations such as accidental incest, but failed attempts to harm others were judged *more harshly* (based on false beliefs and *guilty* intentions) than failed attempts to commit incest (e.g., imagine sleeping with someone you believe to be your long-lost sibling only to discover, after the fact, that you're not actually related).

In a final twist, we presented participants with different kinds of failed attempts – some based on false beliefs (as in the scenarios above) and others based on true beliefs but that involved otherwise thwarted actions (Young & Saxe, 2011). Imagine the following attempted harm: you know your nemesis is allergic to peanuts, and you decide to sneak a few peanuts into her lunch; however, you have misplaced your peanuts. Now imagine the following attempt at incest: you know your one-night stand is your long-lost sibling, and you decide to sleep together; however, the fire alarm sounds, and the opportunity passes. If harmful intent is what matters most for judging attempted harms, then participants should not distinguish between different types of attempts (e.g., whether they involve false beliefs or true beliefs), and indeed they did not. By contrast, participants did distinguish between the two types of attempted incest. Nearly sleeping with one's *true* sibling was judged to be much worse than actually sleeping with a person *falsely* believed to be a sibling. In other words, moral judgments of (attempted) incest were based more on features of the (attempted) act, determined by the facts of the world (e.g., whether they were *actually* related), and less on the contents of the actor's mind (e.g., whether they *thought* they were related).

Consistent with these behavioral findings, fMRI evidence suggests that the RTPJ (discussed above as a brain area for mental state reasoning) is recruited more robustly for evaluating harms versus purity violations (Young et al., in prep). In addition, although the magnitude of response (averaged across voxels) in the RTPJ is high for intentional and accidental harms, multi-voxel pattern analyses (MVPA) reveal that the spatial pattern of neural activity across voxels within the RTPJ holds information about intent – but only for harmful actions and not for purity violations (see also Koster-Hale et al., 2013). In other words, the voxel-wise pattern in the RTPJ distinguishes between intentional and accidental harms but not between intentional and accidental purity violations.

Together, the neural and behavioral findings suggest that harmful actions elicit greater attention to agents' mental states compared to purity violations, and judgments of harmful (versus impure) actions rely more on specific computations for discriminating between

intentional and accidental acts, supported by neural substrates for mental state reasoning. In the next section, we will consider a theoretical account for the different role of mental states in the harm domain versus the purity domain.

Distinct Moral Norms for You and Me

The previous section suggests that mental states matter more for judgments of harmful actions versus purity violations. In this section, we propose an adaptive account for this cognitive difference. We propose that distinct moral norms serve distinct functions – for regulating *interpersonal* relationships versus for protecting the *self*. In other words, (1) it is wrong to *harm others*, and (2) it is wrong to *defile the self*. We also present preliminary evidence for the link between *Harm* and *Other*-focus and *Purity* and *Self*-focus ("HOPS" model; see Figure 2). As such, the HOPS account falls in line with other accounts that motivate drawing distinctions between moral norms by appealing to distinct evolved functions (e.g., Haidt & Joseph, 2004; see also Russell & Giner-Sorolla, 2013; Tybur et al., 2013).

Why might intent matter more for judgments of harms? One possibility is that moral norms against causing *harm* function as moral norms against causing *harm to others*. More specifically, the proposal is that harm norms (e.g., don't hurt others) serve to limit our negative impact on each other. Indeed, paradigmatic cases of harm (physical or psychological) feature at least one agent (the violator) who harms at least one patient (the victim; Gray & Wegner, 2009, 2012; Gray, Young, & Waytz, 2012; Young & Phillips, 2011). The victim may demand an explanation from the violator, and the violator might appeal to her innocent intent (e.g., "I didn't mean to do it"). As far as typical harms are concerned, it takes (at least) two (Gray & Wegner, 2012). What does *interpersonal* morality have to do with *intent*? Information about intent supports not only explanations and evaluations of other people's past actions but also reliable predictions of their future behavior. Especially in the case of accidents, only knowing a person's true intentions can afford an accurate identification of "friend or foe". Mental state reasoning is thus crucial for social interaction and moral judgment (for a review, see Young & Waytz, in press). In particular, moral judgments of harms (which tend to be interpersonal) rely on mental state reasoning.

Why might intent matter less for judgments of purity violations? Moral norms against sleeping with blood relatives, eating taboo foods, or touching taboo substances may have evolved as a means for us to protect *ourselves*, for our own good, from possible contamination. In particular, researchers have proposed that our reactions of disgust, elicited by purity violations, evolved initially for pathogen avoidance and food rejection (e.g., Chapman & Anderson, 2013; Chapman, Kim, Susskind, & Anderson, 2009; Rozin, Haidt, & Fincher, 2009; Rozin, Haidt, & McCauley, 2008; Russell & Giner-Sorolla, 2013; Tybur et al., 2013; but see Rottman & Blake, submitted; Royzman & Kurzban, 2011). When we worry about negatively impacting ourselves, we may care less about whether the impact is accidental or intentional; the key for ourselves is to avoid the contamination, the unsavory outcome. Indeed, purity violations like consensual incest or eating taboo foods (e.g., horse meat, dog meat, rat meat, depending on the culture) are often deemed morally offensive even in the absence of any victims (e.g., see Haidt, 2001; Haidt, Koller, & Dias, 1993). Often, such "impure" acts directly affect only the participating parties and not third parties. When we think about engaging in these acts ourselves even by accident, they may nevertheless feel wrong.

Recent work on moral emotions, discussed briefly in the previous section, is consistent with the HOPS (harm-other, purity-self) proposal. Anger (associated with harm) is elicited more not only for intentional versus accidental harms but also for harms that affect *another person* versus *one's own self* (Gutierrez & Giner-Sorolla, 2007; Russell & Giner-Sorolla,

2011a). By contrast, disgust (associated with impurity) is modulated not by intent (intentional versus accidental) or target (self versus other) but by the presence or absence of moral impurity (e.g., taboo behavior). Other work reveals, more generally, the inflexible nature of disgust compared to anger (Hutcherson & Gross, 2011; Russell & Giner-Sorolla, 2011b, Russell & Giner-Sorolla, 2011c). People are able to mitigate their own reactions of anger but not disgust by imagining "different circumstances" (Russell & Giner-Sorolla, 2011b); people are able to provide good reasons in response to an elicitor (e.g., pedophilia) for their anger ("Pedophiles violate other people's human rights") but not disgust ("Pedophiles are gross"; Russell & Giner-Sorolla, 2011b); and people would rather evoke anger in others than disgust (Hutcherson & Gross, 2011). Anger, elicited by interpersonal harms, motivates people to take (interpersonal) action – to punish the offenders or to repair relationships or both; by contrast, disgust, elicited by purity violations, typically motivates simple avoidance – increasing distance between the self and the stimulus (Gutierrez & Giner-Sorolla, 2007; but see Widen & Russell, 2013).

Similar patterns have been observed for the self-conscious emotions of guilt and shame, which may arise in the agents who violate harm and purity norms in response to their own actions (Giner-Sorolla & Espinosa, 2011; for a recent review of guilt and shame, see Cohen, Panter, & Turan, 2012). Recent work indicates that individuals may also experience self-conscious emotions (guilt, shame) in response to other-directed emotions (anger, disgust) that typically arise in observers who are judging the actions: guilt emerges in response to others' anger and shame in response to others' disgust (Giner-Sorolla & Espinosa, 2011). Guilt, an "approach" emotion, motivates relationship repair, whereas shame (like disgust) motivates avoidance (Sheikh & Janoff-Bulman, 2010). Research on guilt and shame also supports a possible dissociation between self-directed versus other-directed violations. For example, failures in *self*-restraint or *self*-control (e.g., body-related transgressions including excessive eating, spending, gambling) typically elicit shame, whereas failures to be prosocial (e.g., failures to help or to care for others) often lead to guilt (Sheikh & Janoff-Bulman, 2010). Thus, research on guilt and shame (as well as anger and disgust) may reveal links between the HOPS model and models of moral cognition that highlight both an approach/avoid (i.e., prescriptive/proscriptive) dimension as well as a self/other dimension (Janoff-Bulman, Sheikh, & Hepp, 2009; Janoff-Bulman, 2011). Although here we focus on the *proscriptive* aspects of our model – avoiding harming others and avoiding defiling the self – future exploration should target other areas of this multi-dimensional space. For example, it would be worthwhile to know whether, in moral contexts, the approach dimension applies more easily to other-oriented actions, linked by guilt and anger, while the avoid dimension applies primarily to self-directed actions, linked by shame and disgust (but see discussion in the final section below on helping the future self).

Our own preliminary evidence supports the link between *harmful* actions (associated with anger, guilt) and *other-focus*, on the one hand, and *impure* or defiling actions (associated with disgust, shame) and *self-focus*, on the other hand. Suppose you are forced by an experimenter to choose between two options: (1) harm: deliver a mild electric shock to another person, and (2) defile: expose another person to a bad smell using fart spray, a popular technology among moral psychologists studying disgust (e.g., Inbar, Pizarro, & Bloom, 2012; Rottman & Kelemen, 2012; Schnall, Haidt, Clore, & Jordan, 2008). Now suppose you are faced with the same choice for yourself: to spray (defile) or to shock (harm)? In a series of experiments, we found participants to be more resistant to harming another person, but more resistant to defiling themselves (Dungan, Chakroff, Wu, & Young, in prep). For example, one group of participants sorted a series of harmful and impure outcomes from most preferred to least preferred, while a different group of participants sorted the same outcomes for a friend.

Again, participants did not want to cause harm to others (and also judged *harming* others to be more wrong than *defiling* others). By contrast, participants did not want to defile themselves (and also judged defiling versus harming themselves to be more wrong). In addition, when asked to recount instances of their own past violations, participants reported their own harmful actions as impacting *others* more and purity violations as impacting *themselves* more.

If one key difference between harm versus purity violations is that harms are typically directed at *others*, and purity violations typically impact the *self* (see also Gray & Keeney, in prep, for how harm and purity violations may differ along the dimension of atypicality), then two further predictions follow. First, intent should matter less for self-directed versus other-directed actions whether the action is harmful or impure (e.g., cutting own/another's arm with knife; smearing feces on own/another's arm). Second, other-directed actions should elicit more anger, and self-directed actions should elicit more disgust, again, whether the action is harmful or impure. We tested both of these predictions in a 2 (intentional vs. accidental) × 2 (harmful vs. impure) × 2 (other-directed vs. self-directed) design (Chakroff, Dungan, & Young, submitted). In support of the first prediction, intent mattered more for judgments of harmful versus impure actions and also, importantly, more for other-directed versus self-directed actions. In line with the second prediction, harmful actions and other-directed actions elicited more anger, whereas impure actions and self-directed actions elicited more disgust. Participants were also more likely to deliver harsh "moral *character*" judgments in the case of self-directed actions (e.g., defiling oneself reflects poorly on one's moral character) and harsh "moral *action*" judgments in the case of other-directed actions (e.g., hurting someone else is an immoral act; cf. Pizarro & Tannenbaum, 2011; Tannenbaum et al., 2011).

Consistent with this overall pattern, other work reveals that participants' moral judgments of suicide (the ultimate self-directed harm) are uniquely correlated with their (1) endorsement of purity morals as measured by the Moral Foundations Questionnaire (MFQ; Graham et al., 2009, 2011), (2) self-reported ratings of disgust in response to fabricated obituaries of people who committed suicide, and (3) judgments that the people who committed suicide had tainted the purity of their souls (Rottman, Kelemen, & Young, submitted). By contrast, moral judgments of homicide (the ultimate other-directed harm) are uniquely correlated with participants' endorsement of harm morals on the MFQ. A follow-up experiment also revealed moral judgments of suicide to be uncorrelated with any judgment of harm – including judgments that people in the obituaries had harmed themselves, harmed other people, or even harmed God. Moral judgments of suicide were correlated only with purity judgments. Notably, although conservative, religious participants were more likely to judge suicide as immoral overall, compared to liberal, secular participants, the observed patterns emerged robustly even in participants who reported being both liberal and secular. In brief, participants perceived suicide, an extreme instance of self-harm, to be morally wrong insofar as they perceived suicide to be a purity violation.

We do note that, broadly speaking, self-directed violations may be understood within the framework of dyadic morality (Gray, Schein, & Ward, submitted; Gray, Young, & Waytz, 2012) insofar as the self itself may be dyadic; for example, our present self can punish our past self or reward our future self (see the section "The Dyad Within Us" in Gray, Waytz, & Young, 2012). We also note though that the dyadic self may nevertheless be qualitatively distinct from the standard interpersonal dyad. We explore evidence for these ideas in greater detail in the final section below.

Finally, before we conclude this section, we suggest that while we have focused on connecting Harm to Other-related cognition and Purity to Self-related cognition in the HOPS model, ongoing and future work aims to extend HOPS more broadly to accommodate other targets (e.g., ingroups and outgroups) and also other domains (e.g., loyalty, authority, justice,

fairness). For now, we make two suggestions. First, concerns about one's self may track with concerns about one's group (ingroup). In the same set of experiments described above (Dungan, Chakroff, Wu, & Young, in prep), participants judged purity violations within one's group more harshly and harm violations outside one's group more harshly (cf. Schmidt, Rakoczy, & Tomasello, 2012). The guiding hypothesis here is that keeping yourself pure will do you good only insofar as the others around you are also pure; in other words, concerns about contagion or contamination apply more to ingroup than outgroup members. Second, other concerns beyond purity may also apply more to the ingroup. In particular, we may assign greater moral weight to "binding" values such as loyalty within our own group. That is, we may prefer loyal friends and family, whereas we may value justice and fairness within but also, importantly, beyond our group. Recent work investigating whistle-blowing decisions directly reveals the tension between norms concerning loyalty (to friends and family who support the self) and norms concerning justice and fairness (Waytz, Dungan, & Young, submitted).

To summarize the key points of this section, we propose distinct functions (other versus self) for distinct moral domains (harm versus purity): it is wrong *to harm others*, and it is wrong to *defile the self*. We consider mental states more when evaluating *interpersonal harms*. By contrast, we are not as likely to consider (our own) mental states when evaluating *self-directed purity violations* (see Figure 2). Indeed, when we aim to interpret or evaluate the actions of *other people*, including harmful actions, we benefit from understanding their beliefs and intentions. Rarely, though, as we will propose in the next section, do we need to interpret or evaluate our own actions in the same way – and, perhaps as a direct result, we are less often in the position of monitoring and updating our model of our own mental states across moral and non-moral contexts alike. In brief, key differences (at the levels of function and mechanism) between harm and purity norms may arise because of differences between self-related and other-related cognition outside the moral domain, the focus of the next section.

The Challenge of Thinking about One's Own Thoughts

The previous sections presented evidence showing a reduced role for mental states in the case of purity or *self-directed* violations, compared to harmful or *other-directed* violations. As suggested above, we have many reasons to figure out what others are thinking and intending in determining "friend or foe", whereas we are not often in the position of needing to represent our own mental states. Are we less likely to base moral judgments of our own actions on information about our own mental states compared to when we evaluate others? Are we less likely to spontaneously construct representations of our own mental states? Are we worse at retrieving such representations? In this final section, we propose preliminary answers to these questions. This final section focuses on mental state reasoning in moral and non-moral contexts, to establish deeper cognitive roots for the links proposed in the previous sections: a reduced role for intent for self-related cognition and an enhanced role for intent for other-related cognition.

We think about other minds in order to make predictions about people's future behavior and judgments about their past behavior. From carrying on simple conversations to evaluating accidents, social interaction and moral evaluation require us to process, in an ongoing fashion, other people's mental states. Does she *understand* what I'm telling her? Is he *interested* in hearing more? Did she *realize* that my feelings got hurt? Will he *misunderstand* this message? As such, mental state reasoning allows us to interact with other moral agents and to identify the agents with whom we even want to interact in the first place; indeed, mental state reasoning may even be in primary service to moral cognition (Young & Waytz, in press). Yet while we think many thoughts, including thoughts about what others may be thinking, do we need to reason about

our *own* minds to the same extent? We suggest not. We suggest that we may hold many beliefs without representing those beliefs *qua* beliefs (Malle et al., 2000). For instance, when you spoon sugar into your coffee, you hold the belief "this is sugar", although you are unlikely to attribute to yourself that belief (e.g., "I believe that this is sugar"). An exceptional case may be one where you start feeling ill and, consequently, you may reflect on the (false) beliefs of your own past self, "While I *thought* it was sugar, perhaps it was poison". We suggest that such reasoning about one's own mental states occurs infrequently and typically when we discover our beliefs to be false and/or to lead to bad consequences. By contrast, social interaction with and moral evaluation of *other people* depend crucially on our ability to update our mental models of their mental states.

Is there any evidence for reduced or even impaired reasoning about one's own mental states versus others' mental states? For example, when people judge their own actions, versus others' actions, are they more or less likely to use mental state information? One possibility is that mental states will actually exert a *greater* influence on judgments of the self. After all, we have access to our own minds in ways that other people (luckily) do not. Conversely, we often rely on indirect approaches (i.e., inferences) when thinking about other minds. A second possibility is that we do not have special access to our own mental states; we must infer our own thoughts just as we infer the thoughts of others based on external, observable behavior (Bem, 1972). Indeed, developmental evidence shows that children are unable to reason about their *own* (past) false beliefs until they learn to reason about *others*' (current) false beliefs (Atance & O'Neill, 2004). Thus, we may be equally bad (or good) at thinking about our own thoughts and others' thoughts. A third possibility, as suggested above, is that judgments of the *self* reflect a *reduced* (though not absent) role for mental states – in other words, we are *worse* at thinking about our own thoughts.

To disambiguate between these possibilities, we presented hypothetical moral scenarios describing intentional versus accidental harms written in the second and third person (Cushman & Young, unpub.). Some participants read scenarios written in the second person, featuring the participant as the actor (e.g., "*You* accidentally/intentionally hit your neighbor's dog"), while other participants read scenarios in the third person (e.g., "*Susan* accidentally/intentionally hit her neighbor's dog"). Participants judged their own and others' intentional harms similarly harshly, but they judged their own accidents more harshly than others' accidents. Although more work is needed, this pattern suggests preliminarily that participants may have focused more on what they themselves did (the bad outcome) than on what they meant to do (their innocent intent). Recent fMRI research also suggests that taking different perspectives on moral actions (e.g., harming someone versus being harmed) rely on distinct but interacting neural networks (Decety & Porges, 2011). At first glance, these results seem to contradict research on Fundamental Attribution Error (FAE; Heider, 1958; Jones & Harris, 1967): we should be more lenient on ourselves as long as we attribute our behavior to external circumstances (e.g., "the roads were slick", "the dog came out of nowhere") rather than internal character (e.g., "I am a reckless driver"). And indeed if and when we are forced to defend or justify ourselves to others we may begin to attribute our behavior to external, attenuating circumstances. However, while FAE research distinguishes between environmental (external) factors and stable (internal) traits, we focus on neither. We focus instead on transient mental states, such as beliefs (which are internal) about the situation (which is external; Malle, 1999, 2004; Malle et al., 2000; Young & Saxe, 2010).

The preliminary evidence indicates that mental state information may exert less of an influence on moral judgments of the self. But are we actually worse at representing our own mental states? To test this prediction, we designed an fMRI study to measure people's reasoning about their own versus others' mental states in a non-moral context (Gweon, Young, & Saxe,

2011). In the "Self" version of the task, participants viewed a series of pictures; some of these pictures were designed to be misleading: participants would be misled about the object in the picture when seeing only a part of the picture. After completing a word association task for the partially occluded pictures, full pictures were revealed one after the other. Participants then saw all previous pictures along with a set of new, previously unseen pictures; participants were instructed to think back to their belief about the object during the word association task and to mark whether they were right or wrong in the word association task, or whether the picture was new. A different group of subjects participated in the "Other" version of the task; these participants first saw the pictures revealed one after the other. Afterwards, they were told that another person who had not seen the full pictures would perform the word association task, and that they (the participants) would see the other person's response on their screen. Subsequently, as participants viewed the previous pictures along with new pictures, participants were told to think back to the *other* person's belief and to indicate whether the other person was right or wrong, or whether the picture was new. If people are worse at thinking about their own thoughts, "Self" participants should perform worse than "Other" participants, which is precisely what we found. Participants were less accurate at retrieving their own beliefs versus other people's beliefs. This difference in accuracy emerged regardless of whether the beliefs were initially false or true, suggesting that impression management (e.g., "I was right") could not account for the full pattern. Importantly, "Self" participants were not less accurate than "Other" participants in general (i.e., for identifying the new pictures).

In addition, explicit assessments of one's own past beliefs (e.g., "was I right or wrong?") elicited higher activation in the RTPJ than assessments of pictures as new, suggesting that simply thinking thoughts (in response to new stimuli) does not rely on the same neural mechanisms as thinking *about* our own past thoughts (when we are instructed to do so by an experimenter) or thinking about other people's thoughts (which we may do spontaneously; Gweon, Young, & Saxe, 2011). Meanwhile, similar neural patterns (enhanced activity in the RTPJ) are associated with thinking about the thoughts of *our own past selves* and thinking about the thoughts of *other people*, hinting at the possibility that we consider our past selves much like we consider other people.

Here we return to the notion introduced earlier that self-directed violations including, importantly, purity violations, may be understood within dyadic morality (Gray, Young and Waytz, 2012) insofar as the self itself is also dyadic (see the section "The Dyad Within Us" in Gray, Waytz, & Young, 2012). Extensive work reveals that the self can be both an agent and a patient (as we have argued to be the case for many purity violations). Indeed, prior research shows that we perceive our past and future selves from an observer-like perspective, but we perceive our present selves from an actor-like perspective (Pronin & Ross, 2006; but see Quoidbach, Gilbert, & Wilson, 2013 for asymmetric perceptions of past and future selves). Moreover, the tendency to think about our future selves like other people is also associated with an unwillingness to make short-term sacrifices (in the present moment) for long-term (future) well-being (Bartels & Rips, 2010; Ersner-Hershfield, Wimmer, & Knutson, 2009). Neuroimaging evidence provides further support for this notion: participants who thought about how much they would enjoy an event in the *future* versus in the present showed less activity in brain regions implicated in self-referential processing, including the ventral MPFC (Mitchell et al., 2011). On the flip side, when people are made to feel psychologically connected to their future selves (Bartels & Urminsky, 2011; Ersner-Hershfield et al., 2009), or when their future selves are otherwise made salient (Bryan & Hershfield, 2011), people show greater self-control, reflecting greater regard for their future selves. Finally, priming people to feel responsible to their future selves just as they might feel responsible for close others (e.g., family members) also leads to the same effect – more

"moral" treatment of the future self (Bryan & Hershfield, 2011). To link these findings back to the work discussed above: when we do engage in mental state reasoning for ourselves, we may do so mostly in cases where we see ourselves as dyadic – we may think about our past self (e.g., we should have known better) as we think about another person. Mental state reasoning for self and others may therefore rely on the same neural machinery. Nevertheless, because we don't typically need to engage in mental state reasoning for ourselves, we may be worse at it – leading to the observed differences between self-related and other-related cognition as well as between harm and purity norms.

In sum, even though we may have introspective access to the contents of our own minds, in the moment, we may not spontaneously construct explicit representations of our current mental states, and therefore, we may be worse at retrieving these states after the fact. This may also relate to how we make think about (and treat) our past and future selves. Representing mental states may serve a special function – to help us understand the external actions of *other* moral agents in terms of their unobservable mental states, so that we can explain, predict, and evaluate their actions.

Conclusion

We have shown that reasoning about mental states – beliefs and intentions – for moral judgment relies on specific neural substrates (e.g., RTPJ, MPFC). Mental state reasoning is especially important for judging violations of norms against harmful actions versus norms against purity violations. Critically, moral norms may serve distinct functions for regulating our impact on each other versus for protecting ourselves from contamination: *it is wrong to harm others, and it is wrong to defile the self.* Additional evidence reveals differences in how we think about our own versus others' mental states. We may be less motivated to represent our own mental states, and thus retrieving our own mental states presents an unexpected challenge. Together, these findings reveal a special role for mental state reasoning in moral cognition: mental state reasoning allows us to make sense of other moral agents – to understand people's past actions, to predict their future actions, and, most importantly, to evaluate the people around us as potential social partners.

Acknowledgments

We thank Jesse Graham for thoughtful comments on a previous draft of this manuscript as well as for giving the "HOPS" model its name. We thank James Dungan, Alek Chakroff, and Josh Rottman for helpful discussions. We acknowledge support from the John Templeton Foundation and the Dana Foundation.

Short Biographies

Liane Young is an assistant professor in the Department of Psychology at Boston College, where she directs the Morality Lab. Her research focuses on the cognitive and neural bases of human moral judgment, including the roles of mental state reasoning and emotional processing. Her research relies on the tools of social psychology and cognitive neuroscience, including functional magnetic resonance imaging, transcranial magnetic stimulation, and the study of patients with cognitive and neural deficits. Her research in this area has been published in *Proceedings of the National Academy of Sciences*, *Neuron*, *Nature*, *Cognition*, and *Psychological Science*. Young received her BA in philosophy (2004) and her PhD in cognitive psychology (2008) from Harvard University, after which she did post-doctoral work in Cognitive Neuroscience at MIT's Brain and Cognitive Sciences Department.

Lily Tsoi is currently a research assistant in the Department of Psychology at Boston College and will start her graduate studies there in the fall of 2013. Her research interests lie at the intersection of social psychology and cognitive neuroscience. She is currently studying the role of brain regions for theory of mind in moral judgment. She received her BA in neuroscience from Wellesley College.

Endnotes

* Correspondence: Department of Psychology, Boston College, 140 Commonwealth Avenue, McGuinn 347, Chestnut Hill, MA 02467, USA. Email: liane.young@bc.edu

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, **126**, 556–74.

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

Atance, C. M., & O'Neill, D. K. (2004). Acting and planning on the basis of a false belief: Its effects on 3-year-old children's reasoning about their own false beliefs. *Developmental Psychology*, **40**, 953–64. doi:10.1037/0012-1649.40.6.953

Bartels, D. M., & Rips, L. J. (2010). Psychological connectedness and intertemporal choice. *Journal of Experimental Psychology. General*, **139**, 49–69. doi:10.1037/a0018062

Bartels, D. M., & Urminsky, O. (2011). On intertemporal selfishness: How the perceived instability of identity underlies impatient consumption. *Journal of Consumer Research*, **38**, 182–198. doi:10.1086/658339

Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 1–62). New York: Academic Press.

Bryan, C. J., & Hershfield, H. E. (2011). You owe it to yourself: Boosting retirement saving with a responsibility-based appeal. *Journal of Experimental Psychology. General*. doi:10.1037/a0026173

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 & Function*, **783–796**. doi:10.1007/s00429-012-0380-y

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**, 109–11. doi:10.1126/science.1219681

Chakroff, A.. Dungan, J., & Young, L. (submitted). Harming ourselves and defiling others: What defines a moral domain?

Chapman, H. A., & Anderson, A. K.. (2013). Things rank and gross in nature: A review and synthesis of moral disgust. *Psychological Bulletin*, **139**, 300–27. doi:10.1037/a0030964

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**, 1222–6. doi:10.1126/science.1165565

Cohen, A. B., & Rozin, P. (2001). Religion and the morality of mentality *Journal of Personality and Social Psychology*, **81**, 697–710. doi:10.1037//O022-3514.81.4.697

Cohen, T. R., Panter, A. T., & Turan, N. (2012). Guilt proneness and moral character. *Current Directions in Psychological Science*, **21**, 355–359. doi:10.1177/0963721412454874

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2012). How quick decisions illuminate moral character. *Social Psychological and Personality Science*. doi:10.1177/1948550612457688

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, **108**, 353–80. doi:10.1016/j.cognition.2008.03.006

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, **4**, e6699. doi:10.1371/journal.pone.0006699

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

Cushman, F., & Young, L. (2011). The role of intent for moral judgments of self versus other. Unpublished raw data.

Decety, J., & Cacioppo, S. (2012). The speed of morality: A high-density electrical neuroimaging study. *Journal of Neurophysiology*. doi:10.1152/jn.00473.2012

Decety, J., & Porges, E. C. (2011). Imagining being the agent of actions that carry different moral consequences: An fMRI study. *Neuropsychologia*, **49**, 2994–3001. doi:10.1016/j.neuropsychologia.2011.06.024

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, **22**, 209–20. doi:10.1093/cercor/bhr111

DeScioli, P., Asao, K., & Kurzban, R. (2012). Omissions and byproducts across moral domains. *PLoS One*, **7**, e46963. doi:10.1371/journal.pone.0046963

Dungan, J., & Young, L. (2011). Multiple moralities: Tensions and tradeoffs in moral psychology and the law. *Thurgood Marshall Law Review*, **36**.

Dungan, J., Chakroff, A., Wu, H., & Young, L. (in prep). Purity versus pain: Distinct moral concerns for self versus other.

Ersner-Hershfield, H., Wimmer, G. E., & Knutson, B. (2009). Saving for the future self: Neural measures of future self-continuity predict temporal discounting. *Social Cognitive and Affective Neuroscience*, **4**, 85–92. doi:10.1093/scan/nsn042

Fletcher, P., Happe, F., Frith, U., Baker, S., Dolan, R., Frackowiak, R., & Frith, C. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, **57**, 109–128. doi:10.1016/0010-0277(95)00692-R

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**, 11–21. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10617288

Giner-Sorolla, R., & Espinosa, P. (2011). Social cuing of guilt by anger and of shame by disgust. *Psychological Science*, **22**, 49–53. doi:10.1177/0956797610392925

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**, 1803–1814.

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**, 1029–1046. doi: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**, 366–385. doi:10.1037/a0021847

Gray, K., & Keeney, J. (in prep). Moral dimensions, not domains: Severity and typicality explain the role of intent across moral content.

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

Gray, K., & Wegner, D. M. (2012). Morality takes two: Dyadic morality and mind perception. In M. Mikulincer, & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil.* (pp. 109–127). Washington, DC: American Psychological Association. doi:10.1037/13091-006

Gray, K., Schein, C., & Ward, A. (submitted). The harm hypothesis: Moral judgment is unified by a dyadic template of perceived harm, 1–125.

Gray, K., Waytz, A., & Young, L. (2012). The moral dyad: A fundamental template unifying moral judgment. *Psychological Inquiry*, **23**, 206–215. doi:10.1080/1047840X.2012.686247

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, **23**, 101–124. doi:10.1080/1047840X.2012.651387

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron*, **44**, 389–400. doi:10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, **293**, 2105–8. doi:10.1126/science.1062872

Gutierrez, R., & Giner-Sorolla, R. (2007). Anger, disgust, and presumption of harm as reactions to taboo-breaking behaviors. *Emotion*, **7**, 853–68. doi:10.1037/1528-3542.7.4.853

Gweon, H., Young, L., & Saxe, R. R. (2011). Theory of Mind for you, and for me: Behavioral and neural similarities and differences in thinking about beliefs of the self and other. *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society* (pp. 2492–2497).

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, **108**, 814–834. doi:10.1037//0033-295X.108.4.814

Haidt, J. (2007). The new synthesis in moral psychology. *Science*, **316**, 998–1002. doi:10.1126/science.1137651

Haidt, J., & Joseph, C. (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Dædalus*, **133**, 55–66. doi:10.1162/0011526042365555

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**, 613–628.

Hart, H. L. A., & Honore, T. (1959). *Causation in the law*. Oxford: Clarendon Press.

Hauser, M. D. (2006). *Moral minds: How nature designed our universal sense of right and wrong*. New York: Harper Collins.

Hebble, P. W. (1971). The development of elementary children's judgment of intent. *Child Development*, **42**, 1203–1215.

Heider, F. (1958). *The psychology of interpersonal relations*. New York: Wiley.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, **33**, 61–83; discussion 83–135. doi:10.1017/S0140525X0999152X

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology*, **97**, 963–76. doi:10.1037/a0017423

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social-functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology*, **100**, 719–37. doi:10.1037/a0022408

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012a). Disgusting smells cause decreased liking of gay men. *Emotion*, **12**, 23–27. doi:10.1037/a0023984

Inbar, Y., Pizarro, D. A., & Cushman, F. (2012b). Benefiting from misfortune: When harmless actions are judged to be morally blameworthy. *Personality and Social Psychology Bulletin*, **38**, 52–62. doi:10.1177/0146167211430232

Janoff-Bulman, R. (2011). Conscience: The do's and don'ts of moral regulation. In M. Mikulciner & P. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 131–148). Washington, DC: American Psychological Association.

Janoff-Bulman, R., Sheikh, S., & Hepp, S. (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology*, **96**, 521–537. doi:10.1037/a0013779

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, **3**, 1–24. doi:10.1016/0022-1031(67)90034-0

Killen, M., Mulvey, K. L., Richardson, C., Jampol, N., & Woodward, A. (2011). The accidental transgressor: Morally-relevant theory of mind. *Cognition*, **119**, 197–215. doi: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**, 2949–57. doi:10.1016/j.neuropsychologia.2008.06.010

Knobe, J. (2005). Theory of mind and moral cognition: Exploring the connections. *Trends in Cognitive Sciences*, **9**, 355–7. doi:10.1016/j.tics.2005.06.015

Knobe, J. (2010). Person as scientist, person as moralist. *The Behavioral and Brain Sciences*, **33**, 315–29; discussion 329–65. doi:10.1017/S0140525X10000907

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 of the United States of America*. doi:10.1073/pnas.1207992110

Lewis, G. J., Kanai, R., Bates, T. C., & Rees, G. (2012). Moral values are associated with individual differences in regional brain volume. *Journal of Cognitive Neuroscience*, **24**, 1657–63. doi:10.1162/jocn_a_00239

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, **3**, 23–48. doi:10.1207/s15327957pspr0301_2

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

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**, 661–84. doi:10.1037/a0026790

Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology*, **79**, 309–326.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences*, **11**, 143–52. doi:10.1016/j.tics.2006.12.007

Miller, M. B., Sinnott-Armstrong, W., Young, L., King, D., Paggi, A., Fabri, M., Polonara, G., et al. (2010). Abnormal moral reasoning in complete and partial callosotomy patients. *Neuropsychologia*, **48**, 2215–20. doi:10.1016/j.neuropsychologia.2010.02.021

Mitchell, J. P., Schirmer, J., Ames, D. L., & Gilbert, D. T. (2011). Medial prefrontal cortex predicts intertemporal choice. *Journal of Cognitive Neuroscience*, **23**, 857–66. doi:10.1162/jocn.2010.21479

Moran, J. M., Jolly, E., & Mitchell, J. P. (2012). Social-cognitive deficits in normal aging. *Journal of Neuroscience*, **32**, 5553–5561. doi:10.1523/JNEUROSCI.5511-11.2012

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 of the United States of America*, **108**, 2688–92. doi:10.1073/pnas.1011734108

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., & Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, **75**, 73–9. doi:10.1016/j.neuron.2012.05.021

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. *Social Neuroscience*, **1**, 245–58. doi:10.1080/17470910600989896

Piaget, J. (1965). *The moral judgment of the child*. New York: Free Press.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: American Psychological Association.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, **39**, 653–660. doi:10.1016/S0022-1031(03)00041-6

Pronin, E., & Ross, L. (2006). Temporal differences in trait self-ascription: When the self is seen as an other. *Journal of Personality and Social Psychology*, **90**, 197–209. doi:10.1037/0022-3514.90.2.197

Quoidbach, J., Gilbert, D. T., & Wilson, T. D. (2013). The end of history illusion. *Science*, **339**, 96–8. doi:10.1126/science.1229294

Rottman, J., & Blake, P. (submitted). " Hows " and " whats " inform the " whys ": Evolution, development, and the emergence of disgust. *Evolutionary Psychology*, **10**, 1–18.

Rottman, J., & Kelemen, D. (2012). Aliens behaving badly: Children's acquisition of novel purity-based morals. *Cognition*, **124**, 356–60.

Rottman, J., Kelemen, D., & Young, L. (submitted). Suicide taints the soul: Purity as a fundamental moral concern.

Royzman, E., & Kurzban, R. (2011). Minding the metaphor: The elusive character of moral disgust. *Emotion Review*, **3**, 269–271. doi:10.1177/1754073911402371

Rozin, P., Haidt, J., & Fincher, K. (2009). From oral to moral. *Science*, **323**, 1179–1180.

Rozin, P., Haidt, J., & McCauley, C. R. (2008). Disgust. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed, pp. 757–776). New York: Guilford Press.

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**, 574–586. doi:10.1037/0022-3514.76.4.574

Russell, P. S., & Giner-Sorolla, R. (2011a). Moral anger, but not moral disgust, responds to intentionality. *Emotion*, **11**, 233–40. doi:10.1037/a0022598

Russell, P. S., & Giner-Sorolla, R. (2011b). Social justifications for moral emotions: When reasons for disgust are less elaborated than for anger. *Emotion*, **11**, 637–46. doi:10.1037/a0022600

Russell, P. S., & Giner-Sorolla, R. (2011c). Moral anger is more flexible than moral disgust. *Social Psychological and Personality Science*, **2**, 360–364. doi:10.1177/1948550610391678

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**, 328–51. doi:10.1037/a0029319

Russell, P. S., Piazza, J., & Giner-Sorolla, R. (2013). CAD revisited: Effects of the word moral on the moral relevance of disgust (and other emotions). *Social Psychological and Personality Science*, **4**, 62–68. doi:10.1177/1948550612442913

Salerno, J., & Peter-Hagene, C. L. (in press). The interactive effect of anger and disgust in moral outrage and judgments. *Psychological Science*.

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. doi:10.1016/S1053-8119(03)00230-1

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator's group affiliation. *Cognition*, **124**, 325–33. doi:10.1016/j.cognition.2012.06.004

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, **34**, 1096–109. doi:10.1177/0146167208317771

Seidel, A., & Prinz, J. (2013). Sound morality: Irritating and icky noises amplify judgments in divergent moral domains. *Cognition*, **127**, 1–5. doi:10.1016/j.cognition.2012.11.004

Sheikh, S., & Janoff-Bulman, R. (2010). The "shoulds" and "should nots" of moral emotions: A self-regulatory perspective on shame and guilt. *Personality and Social Psychology Bulletin*, **36**, 213–24. doi:10.1177/0146167209356788

Shultz, T. R., Wright, K., & Schleifer, M. (1986). Assignment of moral responsibility and punishment. *Child Development*, **57**, 177–184.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, **47**, 1249–1254. doi:10.1016/j.jesp.2011.05.010

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

Waytz, A., & Young, L. (2012). The group-member mind trade-off: Attributing mind to groups versus group members. *Psychological Science*, **23**, 77–85. doi:10.1177/0956797611423546

Waytz, A., Dungan, J., & Young, L. (submitted). Valuation of fairness over loyalty is associated with whistleblowing.

Widen, S. C., & Russell, J. A. (2013). Children's recognition of disgust in others. *Psychological Bulletin*, **139**, 271–99. doi:10.1037/a0031640

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**, 283–301. doi:10.1016/j.cognition.2005.05.002

Yamada, M., Camerer, C. F., Fujie, S., Kato, M., Matsuda, T., Takano, H., Ito, H., et al. (2012). Neural circuits in the brain that are activated when mitigating criminal sentences. *Nature Communications*, **3**, 759. doi:10.1038/ncomms1757

Young, L. (2013). Moral thinking. In D. Reisberg (Ed.), *The Oxford handbook of cognitive psychology*. New York: Oxford University Press.

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**, 845–51. doi:10.1016/j.neuron.2010.03.003

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 of the United States of America*, **107**, 6753–8. doi:10.1073/pnas.0914826107

Young, L., & Phillips, J. (2011). The paradox of moral focus. *Cognition*, **119**, 166–78. doi: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**, 1912–20. doi:10.1016/j.neuroimage.2008.01.057

Young, L., & Saxe, R. (2009). An fMRI investigation of spontaneous mental state inference for moral judgment. *Journal of Cognitive Neuroscience*, **21**, 1396–405. doi:10.1162/jocn.2009.21137

Young, L., & Saxe, R. (2010). It's not just what you do, but what's on your mind: A review of Kwame Anthony Appiah's "Experiments in Ethics". *Neuroethics*, **3**, 201–207. doi:10.1007/s12152-010-9066-4

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, **3**, 202–214. doi:10.1016/j.cognition.2011.04.005

Young, L., Chakroff, A., Dungan, J., Koster-Hale, J., & Saxe, R. (in prep). Neural evidence for when the thought counts less.

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 of the United States of America*, **104**, 8235–40. doi:10.1073/pnas.0701408104

Young, L., Koenigs, M., Kruepke, M., & Newman, J. P. (2012). Psychopathy increases perceived moral permissibility of accidents. *Journal of Abnormal Psychology*, **7–15**. doi:10.1037/a0027489

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. doi:10.1007/s13164-010-0027-y

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**, 302–15. doi:10.1080/17470919.2010.529712

Young, L., & Waytz, A. (in press). Mind attribution is for morality. In S. Baron-Cohen, H. Tager-Flusberg, & M. Lombardo (Eds.), *Understanding other minds*. Oxford: Oxford University Press.

Yuill, N. (1984). Young children's coordination of motive and outcome in judgements of satisfaction and morality. *British Journal of Developmental Psychology*, **2**, 73–81. doi:10.1111/j.2044-835X.1984.tb00536.x

Yuill, N., & Perner, J. (1988). Intentionality and knowledge in children's judgments of actor's responsibility and recipient's emotional reaction. *Developmental Psychology*, **24**, 358–365. doi: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**, 2478–2492.

