Moral Evaluations Depend Upon Mindreading Moral Occurrent Beliefs

Clayton R. Critcher[1], Erik G. Helzer[2], David Tannenbaum[3], and David A. Pizarro[4]

[1]University of California, Berkeley, [2] Johns Hopkins University,

[3]University of Chicago, [4]Cornell University

Abstract

People evaluate the moral character of others not merely based on what they do, but why they do it. Because an agent's state of mind is not directly observable, people typically engage in mindreading—attempts at inferring mental states—when forming moral evaluations. The present paper identifies a heretofore unstudied focus of mindreading, *moral occurrent beliefs*—the cognitions (e.g., thoughts, beliefs, principles, concerns, rules) accessible in an agent's mind while confronting a morally-relevant decision that could provide a moral justification for a particular course of action. Whereas previous mindreading research has examined how people "reason back" to make sense of why agents behaved as they did, we instead ask how mindread occurrent beliefs (MOBs) constrain moral evaluations for an agent's *subsequent* actions. Our studies distinguish three accounts of how MOBs influence moral evaluations, show that people rely on MOBs spontaneously (instead of merely when experimental measures draw attention to them), and identify non-moral cues (e.g., whether the situation demands a quick decision) that guide MOBs. Implications for theory of mind, moral psychology, and social cognition are discussed.

KEYWORDS: moral evaluation, person perception, mindreading, occurrent beliefs, theory of mind

Moral Evaluations Depend Upon Mindreading Moral Occurrent Beliefs

1 Introduction

How do people determine whether an agent is morally praiseworthy? Such judgments extend beyond a concern with whether another's actions are good or bad (e.g., "Is donating to charity a moral activity?") to an understanding of an agent's motives or reasons for acting (Critcher, Inbar, & Pizarro, 2013; Reeder, 2009; Reeder et al., 2004; see also Monroe & Reeder, 2011). Specifically, perceivers attempt to understand whether seemingly "good" acts are done for moral reasons (Fedotova, Fincher, Goodwin, & Rozin, 2011; Gray, Young, & Waytz, 2012).

However, understanding others' reasons for acting is difficult. Unlike behavior, the contents of another's mind are not directly observable (e.g., Pronin, 2008). This means people must engage in *mindreading* (Reeder, 2009) in an effort to infer the mental states or emotions that were likely precursors to action. As Reeder (2009) put it, "Intentional acts open a window to theory of mind... [in which] the perceiver is looking for a coherent narrative that explains the known facts" (p. 3-4). People engage in mindreading to "fill in the blanks"—that is, to infer the underlying reasons and motives for a particular action.

This paper articulates a newly-identified way in which mindreading unfolds and ultimately influences moral evaluations. Our research differentiates itself in three ways. First, we examine how people engage in mindreading early in an agent's decision-making process—inferring what is going on in another's head while confronting a morally-relevant decision, and then ultimately forming a moral evaluation once the agent's behavior is observed. That is, we do not examine how people "reason back" to an explanation of a previously-observed behavior ("Now that I see everything that happened, how can I develop a story to explain it?"). Instead, we recognize that mindreading can (and does) begin even before an agent behaves. As will

become clear, it is not obvious how this type of mindreading will ultimately influence moral evaluation. Second, given our interest in how people form inferences about others based on their mental contents while deliberating (instead of how people construct narratives, *post hoc*, to explain why a person behaved as he or she did), we introduce the concept of *occurrent beliefs*, a philosophical concept that has been neglected in moral psychology (Audi, 1994). In the present context, occurrent beliefs are the thoughts, beliefs, and concerns (essentially, a summary of an agent's mental content) active in an agent's mind when confronting a morally-relevant decision. Third, and as will be more fully developed later, we differentiate ourselves from past moral judgment research by examining how people infer and rely on others' *moral* (as opposed to immoral) mental states in the service of moral evaluation.

We set out to answer three research questions. First, we ask whether people mindread occurrent beliefs in the service of forming moral evaluations. Second, we test three accounts of how mindread occurrent beliefs (MOBs) guide moral evaluation. Third, we identify features of the agent and the agent's decision context—including features that are not in themselves morally relevant—that guide inferences about an agent's moral occurrent beliefs, and thus, how that agent is morally evaluated.

1.1 Occurrent Beliefs in Relation to Other Mental States

Our purpose is not to empirically contrast the use of occurrent beliefs against other forms of mental state inference (i.e., assess their relative contribution), but instead to determine whether and how mindread occurrent beliefs may factor into moral evaluation. But to better understand what occurrent beliefs are, it is useful to contrast them against related mental states. We take an intentionally broad perspective on occurrent beliefs, using the phrase to capture the mental contents active in an agent's mind at a given point in time. Occurrent beliefs can be

distinguished from dispositional beliefs—those beliefs that an agent has, but that are not activated (Audi, 1994). In social cognitive terms, occurrent beliefs are accessible, whereas dispositional beliefs are merely available (e.g., Markus & Kunda, 1986). For example, our readers no doubt possess the dispositional belief that two plus two equals four, but such a belief likely did not rise to the level of an occurrent belief until encountering this sentence. In the domain of moral decision making, an occurrent belief may play a role in guiding a decision (e.g., a person may experience concern for the welfare of a sick child…and then donate to a medical charity) or may be experienced and then discarded (e.g., …and then decide to send her money elsewhere). In this way, occurrent beliefs can come online as people formulate reasons for acting, but people do not necessarily act in accordance with them.

Two properties of occurrent beliefs make them particularly interesting for the study of moral psychology. First, occurrent beliefs are often "visited upon" a person involuntarily due to features of the decision context. This means perceivers can mindread occurrent beliefs merely from knowing the decision an agent confronts, or particular features of the context in which that decision unfolds. For example, as people enter the ballot box to vote on a new education tax, those voting within a school may be more likely to experience the occurrent belief "Schools really need the money" than those voting in other civic buildings (see Berger, Meredith, & Wheeler, 2008). Given the premium placed on perceived intentionality in many aspects of moral judgment (Baird & Astongton, 2004; Cushman, 2008; Karniol, 1978; Knobe, 2004; Miller et al., 2010; Piaget, 1932; Young et al., 2007; Young & Saxe, 2008; Yuill, 1984; Yuill & Perner, 1988), it is not immediately clear whether such unbidden cognitions—that may reflect more about the situation a person finds oneself in as opposed to a person's moral character—would factor into moral judgment. Second, people can (and do) draw inferences about an agent's moral

occurrent beliefs even before the agent decides what to do. When *Sophie* is confronted with her tragic choice, moviegoers begin to guess what is going through her mind long before they know her decision. It is atypical to consider, say, the intentionality of a behavior before it occurs, but inferences about occurrent beliefs—as beliefs about beliefs instead of beliefs about actions—are more natural to consider in this way.

1.2 How Mindread Occurrent Beliefs May Inform Moral Evaluation

When confronting a morally-relevant decision, an agent may have a variety of occurrent beliefs. In the present work, we consider the *moral* occurrent beliefs that agents are assumed to experience. We take an intentionally broad perspective on moral occurrent beliefs in order to capture the full range of mental content—e.g., convictions, principles, thoughts, rules—that might be active in an agent's mind, and as such, have the potential to serve as the moral basis of a certain course of action. Before turning to the question of how people lean on moral occurrent beliefs, it is useful to note the relative novelty of focusing on how people infer (and rely upon) moral, as opposed to immoral, mental states.

Previous research has examined how mindreaders are sensitive to potential *immoral* motives that suggest praise should be withheld for seemingly good behavior. For example, people receive less praise when they stand to gain from their "good" actions—both when actors make explicit their ulterior motive (Knobe, 2003; Mikhail, 2002), or when perceivers merely notice the possibility for the agents' self-gain (Critcher & Dunning, 2011; Fein, 1996). And although other work has identified circumstances when mindreading *amplifies* praise, that work has also focus on how people reason about selfish temptations—in this case, those that were foregone (Reeder & Spores, 1983). Instead of focusing on inferences about alternative motives, we examine how mindreading about beneficent mental states contribute to moral judgment.

What makes the exploration of mindread moral occurrent beliefs especially intriguing is that it is unclear exactly *how* they might guide moral evaluation. We identified three possibilities. They differ on the basis of a key distinction: whether occurrent beliefs are seen to be direct reflections of moral character (Possibility 1), or whether mindreading occurrent beliefs changes how much praise or blame agents later receive for their subsequent actions (Possibilities 2 and 3, which are not mutually exclusive). In clarifying each possibility, we will draw on the familiar example of the trolley problem, in which an agent "John" must choose whether to alter the course of a runaway trolley (and kill one person) or elect not to switch the trolley's track (allowing five people to die).

1.2.1 Possibility #1: People are praised for mindread moral occurrent beliefs, regardless of whether they act on them.

One possibility, the *direct-information hypothesis*, holds that agents are judged more positively when they are assumed to have moral occurrent beliefs while considering what they should do. In other words, moral occurrent beliefs are assumed to offer direct information about a person's moral character. By this account, social perceivers consider John, his context, and the choice he is confronting and ask themselves, "What is going through John's head?" To the extent he is assumed to have moral cognitions—whether those are occurrent beliefs that would justify a utilitarian decision ("By killing one person, I could save more lives") or a deontology-backed decision ("It is just wrong to actively cause the death of an innocent person"), John would receive moral praise. This account does not argue that John's behavior is irrelevant to moral evaluations of him. Instead, the direct-information account predicts that John receives some praise for how much moral thinking he was assumed to be doing, and then some praise for the course of action he ultimately took.

Two lines of reasoning support the direct-information hypothesis. First, people make *spontaneous trait inferences* merely following co-consideration of a person and a trait-relevant behavior (Crawford, Skowronski, Stiff, & Scherer, 2007; Uleman, 1999). Thus, the mere assumption that John was entertaining a moral occurrent belief might boost moral evaluations of John, even if his subsequent actions cause perceivers to tweak those assessments. Second, the assumption that an agent held a moral occurrent belief may lead to a more charitable inference about why the person did not act on it. For example, if a perceiver is sure that Jeanie is (vs. is not) experiencing the occurrent belief "It is important to donate to children's charities because they will help to alleviate suffering," but then observes her walking past a donation jar, the perceiver may give her something of a pass, assuming that she is going to use her money to do something even more morally worthwhile.

On the other hand, we see two reasons to doubt this hypothesis. First, in many cases, people experience occurrent beliefs not because of any feature of their character, but because some feature of their environment visited the occurrent beliefs upon them. In other cases, contextual features reduce the likelihood a belief will occur to a person. Our studies will detail effects of both types, which will highlight that the occurrence of moral beliefs may not offer a direct glimpse of an agent's moral character. Second, previous research has shown that even though people tend to give themselves credit for their unrealized good intentions, people withhold crediting others when their actions do not live up to their praiseworthy ambitions (Kruger & Gilovich, 2004). This suggests that merely assuming someone else has moral occurrent beliefs may not be sufficient to dole out praise.

1.2.2 Possibility #2: Mindread moral occurrent beliefs constrain how much praise will be offered for each action

An alternative account, the *matching-praise hypothesis,* rests on the premise that people think that properly-motivated, praiseworthy behavior unfolds in a specific temporal sequence: Agents have a moral occurrent belief (e.g., John thinks, "I can save the most lives by diverting the trolley") that precedes the *matching* behavior (pulling the switch). By this account, the presence of a particular occurrent belief constrains how much praise the agent will receive for subsequently acting in that matching way. In other words, agents are praised to the extent that they were assumed to have had the matching occurrent belief—i.e., the one that could provide a moral justification for the behavior.

We see two reasons to endorse the matching-praise hypothesis. First, the matching-praise account is rooted in in the idea that if a belief did not occur to a person, then it could not have been his or her basis for acting (Malle, Knobe, O'Laughlin, Pearce, & Nelson, 2000). But if a moral occurrent belief was mindread, then it is at least a possible (moral) basis for action. Second, the matching-praise account accepts that even if the mere occurrence of a belief is not directly informative (e.g., because it is prompted by the situation instead of by a person's moral character), that a decision to act on that (involuntary) occurrent belief can reveal moral character. Thought of differently, even if we don't select the situations we find ourselves in, those situations can define how our moral character will be revealed.

1.2.3 Possibility #3: Mindread moral occurrent beliefs determine how much blame will be offered for foregoing each action

A complementary possibility, the *competing-blame hypothesis*, is that agents deserve blame when they fail to act on moral occurrent beliefs they were assumed to be experiencing. On this account, how much praise John receives for diverting the trolley (thereby killing one to save five) depends on whether he was assumed to have had the competing (i.e., behavior-

mismatching) moral occurrent belief, "It is wrong to actively kill an innocent person." More specifically, John should receive *less* praise to the extent he was assumed to ignore a morally-relevant occurrent belief. This account may be seen as a plausible extension of Reeder and Spores's (1983) finding that moral agents are praised for not succumbing to selfishness. Instead, the competing-blame account suggests moral agents may be seen as possessing blameworthy character for failing to follow the moral guidance of an occurrent belief.

1.2.4 Summary of empirical predictions

The three hypotheses make different (but overlapping) predictions of how mindreading occurrent beliefs influence moral evaluations. In summarizing these predictions, it is helpful to differentiate *matching* from *competing* occurrent beliefs—the occurrent beliefs that do or do not match the ultimate behavior, respectively. The direct-information hypothesis (Possibility 1) predicts that both occurrent beliefs will be positive predictors of moral evaluations. The matching-praise account (Possibility 2) predicts that the matching occurrent belief will be a positive predictor of moral evaluation. The competing-blame hypothesis (Possibility 3) predicts that the competing occurrent belief will be a negative predictor moral evaluation. Note that either Possibility #1, Possibility #2, Possibility #3, both Possibilities #2 and #3, or none could be true.

1.3 Overview of the Present Studies

We conducted six studies, using three different moral dilemmas, to test whether and how people mindread occurrent beliefs in determining moral praise. In each dilemma, a moral agent decided between a utilitarian action (maximizing total welfare that also causes harm in the process) against a deontological one (refusing to violate a moral rule). We employed dilemmas of this sort for several reasons. First, utilitarian and deontological decisions cleanly correspond to matching occurrent beliefs—deontology-backed aversions to direct harm and utilitarian

justifications for promoting the greater good (Nichols & Mallon, 2006). This facilitates a crisp test of whether and how people rely on mindread occurrent beliefs in forming moral evaluations. Second, in order to differentiate between our three accounts, which make different predictions depending on whether participants acted or failed to act on different MOBs, we wanted participants to make moral evaluations of a person deciding between different courses of action. Our approach, applied to this context, is depicted in Figure 1.[1]

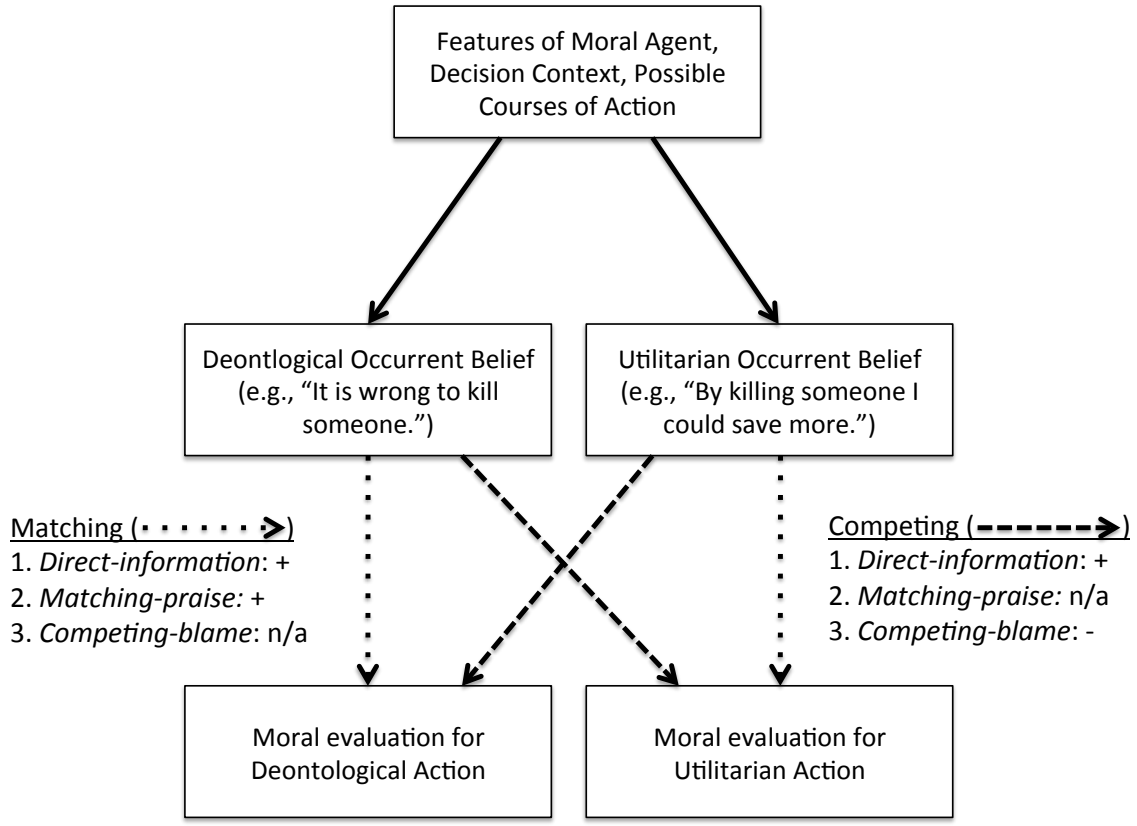

Figure 1 How mindreading occurrent beliefs influence moral evaluations, applied to dilemmas pitting deontological against utilitarian courses of action. Features of the moral agent, the decision context, and the possible actions themselves influence what moral occurrent beliefs are assumed to transpire in an agent's mind. The moral agent can choose between competing actions. The three accounts—direct information, matching-praise, and competing-blame—differ in whether and how they predict that the occurrent beliefs will influence moral evaluations. The direct information and matching-praise accounts predict a positive influence of the matching occurrent belief on moral evaluation for an action (i.e., a positive effect along the dotted lines). The direct information account also predicts a positive effect of the competing occurrent belief on moral evaluation, whereas the competing-blame account predicts a negative effect (i.e., positive or negative effects along the dashed lines).

The goal of Studies 1a-1c was to distinguish among our three accounts of how mindread occurrent beliefs influence moral evaluation. These studies take the form of typical moral judgment studies, offering information about an agent who is confronting a moral dilemma, with no additional information as to what moral beliefs the particular agent is likely experiencing. With one of our mechanistic accounts supported in Studies 1a-1c, Studies 2-4 offer a causal test of this model. Furthermore, these demonstrate how our present approach can be extended to make novel predictions about how extradecisional features should influence moral judgment. In particular, these studies manipulated features of the agent or the decision context that were assumed to change agents' moral occurrent beliefs: whether the agent lacked basic emotional or cognitive capacities (Study 2), whether the decision was made under time constraints (Study 3), and who was focal in the agent's visual field (Study 4).

2 Studies 1a, 1b, and 1c

In Studies 1a–1c, we investigated whether and how people rely on mindread occurrent beliefs in forming moral evaluations. In each study, participants considered a moral agent who was confronted with a different moral dilemma. Before learning the agent's decision, participants indicated the likelihood that the agent was experiencing each relevant occurrent moral belief. Thus, consistent with our aims, we measured what agents were presumed to be thinking *before* they had actually acted. Next, participants were randomly assigned to learn that the agent had actually chosen the utilitarian or the deontological action. Finally, participants offered their moral evaluations of the target.

Our three accounts—direct-information, matching-praise, and competing-blame—make different predictions concerning whether and how the matching MOB (the one that matches the chosen behavior) and the competing MOB (the one that mismatches the chosen behavior) should

influence moral evaluation. According to the *direct-information* account, any mindread moral occurrent belief should lead to positive moral evaluations; thus, both the matching and competing occurrent beliefs should be positive predictors of moral evaluation. By the *matching-praise* account, an agent should be praised more to the extent he is assumed to have the matching occurrent belief. By the *competing-blame* account, the agent should be blamed more (i.e., receive less praise) to the extent he is assumed to have the competing occurrent belief (i.e., one he did not act on).

We also tested (and hoped to rule out) an uninteresting account of mindread occurrent beliefs—that they may merely measure participants' expectations of what a moral person (or what the participants themselves) would and would not do. By this artifactual *MOBs-as-expectations hypothesis,* participants decide that it would be better to do X and not Y, and thus infer that the agent is likely experiencing the occurrent belief that matches X but not Y. This artifactual account makes two predictions. First, it predicts that the two MOB measures will be negatively correlated and likely strongly so (consistent with the idea that a strong expectation that a moral person would do X entails a weak expectation that the agent would do Y). Second, it predicts that we should find support for *both* the matching-praise and competing-blame accounts. That is, if moral occurrent beliefs merely reflect the pathways that perceivers think the agent clearly should versus should not follow, then people should be praised or blamed for taking or failing to take, respectively, the expected course of action.

2.1 Method

2.1.1 Participants and design

Participants in Studies 1a-1c were recruited from Amazon.com's Mechanical Turk labor market for a small cash payment. The three studies had 95, 97, and 108 participants, respectively.

In each study, participants were randomly assigned to one of two decision conditions: *utilitarian* or *deontological*.

2.1.2 Procedure and materials

In each study, participants considered a different moral dilemma whereby a moral agent was confronted with two options, one that cold be justified by utilitarian calculus and one that could be supported by deontological reasoning (i.e., maximizing lives saved vs. adhering to moral rules against harming). Before learning how the agent acted, participants were asked to indicate whether the agent would experience each of two occurrent beliefs— one suggesting an appreciation of the utilitarian consequence, and one that suggested a deontology-backed aversion to direct harm. Finally, participants learned the agent's decision and offered a moral evaluation.

In Study 1a, participants read a modified version of Tetlock et al.'s (2000) "sick Johnny" moral dilemma. A hospital director, Robert, has to decide whether to spend $3 million of the hospital's limited resources to save the life of a sick five-year-old named Johnny. Spending the money to save Johnny would prohibit the hospital from updating hospital infrastructure— updates that could be used to save many future lives. Thus, the hospital director has to choose between letting Johnny die in order to save more lives in the future (utilitarian decision) or spend the money and thereby save the life of Johnny (deontological decision: avoiding violation of what Tetlock et al. [2000] called a "taboo tradeoff").[2]

In Study 1b, participants read a moral dilemma about Jewish townspeople hiding in a secret basement while Nazi soldiers searched the town (e.g., Greene, Nystrom, Engell, Darley, & Cohen, 2004). The townspeople were maintaining careful quiet, for the Nazis would kill anyone they discovered. Suddenly, a small baby in the arms of a townsperson, Jack, began to bawl. Left unabated, the crying would attract the attention of the Nazis, which would result in the certain

death of all of the townspeople. Jack can choose to smother the child, which would kill the baby but save everyone else (utilitarian decision); otherwise, Jack could let the child continue to cry, though this means the Nazis will find the hidden townspeople (deontological decision).

Study 1c introduced a new dilemma not used in previous research. Participants read about a high-level military commander working to root out Al Qaeda terrorist cells in Afghanistan. Intelligence had led the military commander, Michael, to a rural inn on the Ukraine-Poland border. There, a meeting of top Al Qaeda leaders planning a 9/11-style attack was scheduled to take place. Several of these men were among the FBI's "Most Wanted Terrorists." The night of the meeting, Michael looked down at the inn from the surrounding mountains and could clearly see the Al Qaeda leaders enter the inn, just as was expected. He also saw their translator, an innocent man kidnapped by the terrorists and forced to work for them against his will. Michael had to decide whether to recommend an airstrike, which would kill all of those present in the inn, both the terrorists and the innocent translator (utilitarian decision). To make sure that utilitarian motives would push for a strike, we added that "if a strike is not ordered now, it is doubtful that one will occur in time to stop the 9/11-style attack." Alternatively, Michael could decide against ordering the strike (deontological decision).

These dilemmas and choices, which are also used in Studies 2-4, are summarized in Table 1.

Table 1 Summary of dilemmas used in all studies

Dilemma	Summary	Utilitarian Action	Deontological Action	Studies
Sick Johnny	A hospital director must decide whether to save the life of a sick five-year-old Johnny by funding an expensive organ transplant. If the surgery is denied, Johnny will die, but the hospital will retain funds to improve hospital quality, thereby saving more lives in the future.	The hospital director denies the surgery.	The hospital director funds the surgery.	1a, 2
Crying Baby	A Jewish townsperson, along with other Jewish townspeople, hide in a secret basement as Nazi soldiers search the town above. A baby in the basement begins to cry. If the baby continues to cry, it will attract the attention of the Nazi soldiers. The Nazis will kill any Jews—children or adults—whom they discover.	The Jewish townsperson smothers the child to death.	The Jewish townsperson does not smother the child to death.	1b, 3
Terrorist-Inn	An American military commander must decide whether to launch an airstrike on a rural inn. Top Al Qaeda operatives are meeting inside. A strike on the inn would kill everyone (including an innocent bystander), but would stop the terrorists from launching a 9/11-style terrorist attack.	The military commander orders a strike on the inn.	The military commander does not authorize the strike.	1c, 4

2.1.2.1 Occurrent Beliefs

Participants then completed measures that asked them to indicate to what extent the a
likely "appreciated, experienced, or possessed" each of two relevant moral occurrent beliefs
Instead of using the (likely unknown) phrase "occurrent beliefs," the measure specified that
different moral concerns may be at the forefront of the agent's mind and that participants sh
indicate "to what extent you believe he is experiencing each sentiment as he is confronted w
this situation." In this way, we measured whether each idea was seen as likely occurring to t
agent. The wordings of the utilitarian and deontological occurrent beliefs were modified for
scenario, but all were anchored on nine-point scales anchored at 1 (*not at all)* and 9 (*is experiencing strongly*).

In Study 1a, participants indicated to what extent Robert experienced each occurrent
belief: "It is morally wrong or troubling to let a child die" (deontological), and "By letting th
child die, the hospital could actually save money which would allow it to ultimately save ma
more lives." (utilitarian). In Study 1b, participants indicated to what extent Jack experienced
occurrent beliefs: "It is morally wrong to actively kill a child" (deontological), and "By killi
this child, I could save everyone" (utilitarian). In Study 1c, participants rated to what extent
Michael was likely experiencing each occurrent belief: "It is morally wrong to kill innocent
civilians regardless of the circumstances" (deontological), and "It is morally right to stop the
terrorists from killing thousands of people, even if it means killing an innocent person in orc
stop the worse tragedy" (utilitarian). We followed precedent in assuming that deontological
principles would be experienced as morally-laden rules proscribing certain actions instead o
conscious appreciation of philosophical (e.g., Kantian) logic (Broeders et al., 2011; Greene,
2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008).[3]

2.1.2.2 Moral evaluation

After learning the agent's decision, participants responded to five moral evaluation items. On 8-point Likert-type scales, participants indicated to what extent the agent: is a bad versus good person, has a bad versus good conscience, is or is not "in the wrong," has blameworthy versus praiseworthy character, and is in general a moral versus immoral person. After reverse-scoring negatively-worded items, we averaged the judgments into a *praise composite* such that higher numbers reflected greater praise for the agent (Study 1a: $\alpha = .93$, Study 1b: $\alpha = .86$, Study 1c: $\alpha = .78$).

2.2 Results and Discussion

Our three accounts differ in whether and how one or both mindread occurrent beliefs (i.e., matching or competing) will predict moral evaluation. Before conducting the tests that differentiate the three accounts, we conduct initial tests that: provide a first assessment of the artifactual account (that the mindread occurrent beliefs merely measure expectations for behavior), determine how the agent's decision influenced moral evaluation of him, and assess whether there was consensus about which moral occurrent belief would be more salient to the agent in each scenario.

First, we assessed whether the measures of the two MOBs were strongly negatively correlated, as one would expect if they merely reflected expectations that the agent (or the participants themselves) would or should behave in one way versus the other. Assuaging this concern, the mindread occurrent beliefs were never negatively correlated. In fact, in one study they were (marginally) positively correlated: Study 1a, $r = .19$, $p = .07$; Study 1b: $r = .10$; Study 1c: $r = -.00$. Such a pattern is not consistent with the MOBs-as-expectations artifactual account.

That said, there were reliable patterns concerning which occurrent beliefs participants

tended to mindread in each study. Participants thought that the agents in Studies 1a and 1b would more strongly hold the deontological than utilitarian occurrent belief, whereas participants in Study 1c thought that the agent would show the reverse pattern. More specifically, participants thought that Robert in Study 1a would more strongly experience the deontological occurrent belief ("It is morally wrong or troubling to let a child die": $M = 8.04$, $SD = 1.29$) than the utilitarian occurrent belief ($M = 6.06$, $SD = 2.50$), paired $t(96) = 7.24$, $p < .001$, $d = .73$. Participants in Study 1b thought that Jack would be experiencing the deontological occurrent belief ("It is morally wrong to actively kill a child") more strongly ($M = 8.29$, $SD = 1.13$) than the utilitarian occurrent belief ($M = 7.08$, $SD = 2.32$), paired $t(94) = 4.94$, $p < .001$, $d = .51$. Study 1c participants instead thought that Michael would more strongly experience the utilitarian occurrent belief ("It is morally right to stop the terrorists from killing thousands of people…": $M = 7.47$, $SD = 1.84$) than the deontology-backed occurrent belief ($M = 6.16$, $SD = 2.36$), paired $t(107) = 4.56$, $p < .001$, $d = .44$.

Moral evaluations for each action followed a similar pattern as inferences about occurrent beliefs. In both Studies 1a and 1b, participants offered more praise to the agent who performed the deontological action. In Study 1c, participants praised the utilitarian actor more (see Table 2, Model I). More specifically, in Study 1a, the hospital director was given more praise when he made the deontological-backed decision to save the child ($M = 6.47$, $SD = 1.26$) instead of the utilitarian decision ($M = 3.97$, $SD = 1.72$), $t(93) = 8.11$, $p < .001$, $d = 1.68$. In Study 1b, the townsperson was praised more when he made the deontological-backed decision to spare the infant's life ($M = 5.87$, $SD = 1.55$) than the utilitarian decision ($M = 4.63$, $SD = 1.35$), $t(92) = 4.14$, $p < .001$, $d = .86$.[4] In Study 1c, participants judged Michael as more moral when he made the utilitarian decision to order the strike on the inn, thereby killing an innocent man in the

process (M = 5.64, SD =1.32), as opposed to when he made the deontological-backed decision (M = 5.13, SD = 1.24), $t(105) = 2.05$, $p = .04$, $d = .40$.

Although the striking consistency between mindread occurrent beliefs and moral evaluations for each action is consistent with our accounts—especially the matching-praise and competing-blame possibilities—they do not yet distinguish between them because we have yet to test how and which mindread occurrent beliefs predict moral evaluations. For each study, we regressed moral evaluation on the behavior, the matching MOB, and the competing MOB. As a reminder, when the agent made the utilitarian [deontological] decision, the matching occurrent belief was the utilitarian [deontological] one. The other occurrent belief is the competing one.

As can be seen in Table 2 (Models II – IV), regardless of whether the mindread occurrent beliefs were entered as individual (Models II-III) or simultaneous predictors (Model IV), we found consistent support only for the matching-praise account. In all three studies, we found the matching MOB was a positive predictor of moral evaluations. In other words, the amount of praise agents received for each action was determined by whether they were assumed to have the matching occurrent belief. Using Preacher and Hayes's (2008) bootstrapping technique with 10,000 resamples, we found significant support for this indirect effect. In all cases the 95% confidence interval of the indirect did not include 0: Study 1a, [-.5089, -.1321]; Study 1b, [-.3407, .0513]; and Study 1c, [.0012, .1556].

But in no case was the competing mindread occurrent belief a significant predictor of moral evaluation. In fact, it varied between studies whether competing MOBs trended toward being a positive or a negative predictor of moral evaluations. Not only was the direction inconsistent, but similar bootstrapping analyses to those reported above found no significant indirect effects through competing MOBs: Study 1a, [-.3429, .0157]; Study 1b, [-.0475, .2932];

Table 2 Regression models predicting moral praise (Studies 1a-1c). Model I is the direct effect of the IV (Decision) on the DV (Praise). Models II and III add two possible mediators separately: mindread matching occurrent belief (Model II) and mindread competing occurrent belief (Model III). Mode IV tests the robustness of the conclusions of Models II and III by including the candidate mediators as simultaneous predictors. *$p < .05$, **$p < .01$, ***$p < .001$

	Study 1a				**Study 1b**				**Study 1c**			
	Model I	Model II	Model III	Model IV	Model I	Model II	Model III	Model IV	Model I	Model II	Model III	Model IV
Deontological decision	.64***	.51***	.59***	.44***	.40***	.29**	.45***	.34**	-.20*	-.16	-.14	-.10
Matching MOB		.36***		.37***		.32**		.30**		.20*		.20*
Competing MOB			-.11	-.13			.19	.16			-.14	-.14

and Study 1c, [-.0340, .2134]. This means that there were no consistent or significant tendencies for agents to be praised or blamed more for being assumed to have occurrent beliefs that they did not ultimately act on.

2.3 Summary

Studies 1a-1c provide suggestive evidence that mindread occurrent beliefs influence how much agents are praised for subsequent actions. We found consistent support across all three studies for the matching-praise account, but not for the competing-blame, direct-information, or (artifactual) MOB-as-expectations accounts. People mindread moral occurrent beliefs to determine which actions could (or could not) have unfolded morally. It was not the case that mindread moral thinking was praiseworthy in itself (direct-information hypothesis) or that people were praised less when they appeared to pass up an opportunity to act on their occurrent beliefs (competing-blame hypothesis). Finally, several aspects of the data suggested that MOBs are not merely expectations of what a moral person would or would not do (e.g., the two mindread occurrent beliefs were uncorrelated).

Participants in Studies 1a-1c had little information—other than the moral choice that the participant confronted—to determine what was going through the agent's mind. As such, we suspect that many participants merely placed themselves in the context and indicated what occurrent beliefs they thought they would have. Such perspective taking is merely one way by which mindreading occurs. What would have been less interesting is if participants merely tried to guess what action they themselves would and would not take in a context and then differentially endorse the occurrent belief items to the extent they were consonant or dissonant, respectively, with participants' own forecasted behavior. Note that this is a variant of the MOBs-as-expectations artifactual account, and the same empirical arguments made earlier speak against

this account as well.

The remaining studies build on these findings in two ways. First, Studies 2-4 use experimental manipulations that permit causal tests of the matching-praise hypothesis. Second, these studies test an implication of our model. In particular, if mindread moral occurrent beliefs constrain the space of praiseworthy behavior, then features of the decision context—even those not directly related to moral character—should change moral evaluations if they seem to hint at what moral occurrent beliefs an agent has. The remaining three studies identify and test the influence of three such cues.

3 Study 2: Time to Deliberate

In Study 2, we returned to Tetlock et al.'s (2000) dilemma about a hospital director who must decide whether to spend a large sum of money to save a sick child (i.e., the dilemma used in Study 1a). But this time we varied an extradecisional factor that we suspected might shift inferences about the agent's occurrent beliefs. This feature—whether the agent was pressured to decide quickly or was able to engage in extensive deliberation—is one that varies across real-world contexts and (because it was a random occurrence of the situation and not chosen by the moral agent) is not itself a signal of an agent's moral character. Many deontological decisions are driven by quickly-appreciated, affect-backed principles. In contrast, utilitarian logic may be more easily appreciated only after additional deliberation and reflection (Greene et al., 2004, 2008). If people have some intuition of these properties, they should assume (and we empirically confirm they do) that a rushed agent would more be more likely to experience deontological than utilitarian occurrent beliefs. But given more time to deliberate, the agent should be assumed to have both deontological and utilitarian occurrent beliefs.

If our reasoning is correct, perceivers should offer relatively more praise for the

deontological (vs. the utilitarian) decision when the agent is rushed (when the deontological occurrent belief is presumed to be more accessible than the utilitarian occurrent belief). However, this gap should diminish when the agent has extensive time to deliberate (at which point both occurrent beliefs may be assumed to be present). We measured mindread occurrent beliefs and moral evaluation using different samples. The advantage is that this gives us a sense of whether people lean on MOBs spontaneously. That is, if we observe our predicted pattern of results, it cannot be that MOB measures walked participants through a reasoning process they would not have traversed spontaneously. Although the disadvantage is that we cannot test the mediation model implied by the matching-praise hypothesis, we return to such tests in Studies 3 and 4.

3.1 Method

3.1.1 Participants and design

Two hundred fifty-six undergraduates Cornell University were randomly assigned to one of four conditions in a 2 (speed: rushed or lengthy) X 2 (decision: utilitarian or deontological) between-subjects design. Participants received course credit for their participation.

3.1.2 Procedure

Participants read a modified version of Tetlock et al.'s (2000) "sick Johnny" moral dilemma. In our version, two hospital directors each must decide whether to spend $3 million of the hospital's limited resources to save the life of a sick five-year-old named Johnny. Spending the money to save Johnny would prohibit the hospital from updating hospital infrastructure, updates that could be used to save many future lives. Participants were told that the hospital's co-directors—Robert and Alan—must independently choose whether to let Johnny die (utilitarian decision) or save the life of Johnny (deontological decision). By chance, Alan was at the hospital

when this situation arose, whereas Robert was initially unreachable. By the time hospital officials could reach and explain the situation to Robert, he did not have time to engage in careful deliberation and was required to make a decision based on his immediate gut instinct. In contrast, Alan had many hours to engage in careful, thorough reflection before arriving at a decision.

In high-conflict personal moral dilemmas of this variety, people tend to quickly appreciate or experience a negative affect-backed deontological occurrent belief (e.g., "Killing a child is wrong!..."; Koenigs et al., 2007; Nichols & Mallon, 2006; Valdesolo & DeSteno, 2006), that is supplemented or replaced by a utilitarian occurrent belief (e.g., "…but by killing now, I could save the lives of many people"; Greene, 2009; Greene et al., 2001; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Kahane, Wiech, Shackel, Farias, Savulescu, & Tracey, 2012) that comes with more time and reasoned thought. Our perspective remains agnostic as to whether deontological and utilitarian occurrent beliefs actually or always map onto these properties (see Baron, 2011; Kahane et al., 2012); for our purposes it only matters that in many moral dilemmas (including the one we used) people *intuit* these properties.

In a pretest, we confirmed our assumption that utilitarian beliefs (but not deontological beliefs) are more likely to come to an agent after deliberation. We presented 129 participants from the same population with the sick Johnny dilemma and asked them to what extent Robert (rushed deliberation) and Alan (lengthy deliberation) would experience the deontological occurrent belief ("find it troubling to kill a person") and utilitarian occurrent belief ("realize that by letting the person die, the hospital would actually save money which would allow it to save many more lives"). Confirming our assumption, a 2 (speed) X 2 (occurrent belief) repeated-measures ANOVA returned a significant interaction, $F(1, 128) = 38.73$, $p < .001$, $\eta_p^2 = .23$ (see

Table 3). People assumed that Robert, who had no time to deliberate, would be more likely to have the deontological than the utilitarian belief occur to him, paired $t(128) = 7.20$, $p < .001$, $d = .63$. In contrast, people assumed that Alan, who had time to engage in more lengthy deliberation, would experience both the deontological and the utilitarian occurrent beliefs equally, $t < 1$.

In our main sample, participants learned of the existence and constraints of both directors, but only learned the decision of (and morally evaluated) one director. Participants made five *moral evaluation* judgments about the target, each on 7-point, Likert-type scales. Participants indicated whether he should be praised (versus blamed), had a good (versus bad) moral conscience, was a good (versus bad) person, was the type of person who would be a good (versus bad) friend, and was a moral (versus immoral) person ($\alpha = .86$).

3.2 Results and Discussion

We submitted the praise composite to a 2 (speed) X 2 (decision) ANOVA. Although there was no main effect of speed, $F(1, 252) = 1.34$, $p > .24$, $\eta_p^2 = .01$, we did observe a main effect of decision, $F(1, 252) = 24.79$, $p < .001$, $\eta_p^2 = .09$. But consistent with our central hypothesis, there was a significant Speed X Decision interaction, $F(1, 252) = 4.80$, $p = .03$, $\eta_p^2 = .02$ (see Table 3). Robert, who had to make a decision immediately, was praised much more for saving Johnny's life than for letting Johnny die, $d = .63$: $t(252) = 5.07$, $p < .0001$. In contrast, Alan, who had considerable time to think about his decision received only somewhat more praise for the deontological vs. the utilitarian decision, $d = .25$: $t(252) = 1.97$, $p = .05$. Thus, although participants in both cases had a preference for the agent who made the deontological decision (replicating Tetlock et al., 2000), the effect was (as hypothesized) reliably attenuated when the agent had sufficient time to consider the utilitarian course of action. In more general terms, the significant interaction is consistent with the mindreading occurrent beliefs approach, whereas the

fact that there was still a sight preference for the deontological action (saving the sick child's life) shows that mindread occurrent beliefs are not the *only* influence on moral judgment. Given we measured mindread occurrent beliefs and praise with different samples, the significant Speed X Decision interaction suggest that people rely on MOBs *spontaneously* in crediting targets. That is, the predicted interaction on moral evaluation emerged even though the measures never drew participants' attention to occurrent beliefs.

Note that this pattern of results is inconsistent with an alternative prediction that when under situational duress, a decision may be seen as less intentional and thus less useful in assessing blame or praise (see Monroe & Reeder, 2011). To the contrary, we found that the agent's decision was viewed as offering a more diagnostic, differentiated moral signal under rushed conditions. Our pretest indicated that only under time duress did participants intuit a difference in the occurrence of deontological and utilitarian beliefs. As the MOB account predicts, it is under these rushed circumstances that the hospital director's behavior offers the most diagnostic, differentiating signal of his moral character.

Relatedly, it is worth noting that decision speed was a useful cue even though the agent himself did not have control over the amount of time he had to deliberate. The present findings can be contrasted against recent research that has examined what is signaled when moral agents arrive at moral decisions quickly or slowly of their own accord (Critcher et al., 2013; Tetlock et al., 2000). In the present research, the length of time participants had to deliberate was no chosen by the moral agent, but was instead governed by the situation. As a result, deliberation time in the present study was not an endogenous variable that provided direct information about the specific agent and his motives (Critcher et al., 2013), but was an exogenous cue that reflected the

Table 3 Mindread occurrent beliefs and moral evaluations for utilitarian and deontological occurrent beliefs and behaviors, respectively. Each mean is followed parenthetically by the corresponding standard deviation. Note: Within each study and measure (belief or moral evaluation), means with a different subscript are significantly different, $p < .05$

	Moral occurrent belief		Moral evaluation following behavior	
	Utilitarian	Deontological	Utilitarian	Deontological
Study 2: Speed				
Rushed	5.32 (1.69)$_c$	6.80 (1.30)$_a$	5.14 (1.11)$_a$	6.07 (1.12)$_c$
Lengthy	6.27 (1.41)$_b$	6.30 (1.56)$_b$	5.22 (1.11)$_a$	5.61 (1.00)$_b$
Study 3: Skill intact				
Emotion	4.01 (2.05)$_b$	5.89 (1.87)$_a$	4.35 (1.28)$_a$	4.92 (1.04)$_c$
Reason	5.76 (2.21)$_a$	3.12 (2.10)$_c$	4.64 (0.91)$_{bc}$	4.59 (1.31)$_{ab}$
Study 4: Visual salience				
Innocent bystander	4.34 (2.07)$_d$	7.53 (1.76)$_a$	5.16 (1.19)$_{ab}$	5.35 (1.10)$_{ab}$
Terrorist	6.76 (1.69)$_b$	6.17 (2.00)$_c$	5.49 (0.98)$_a$	5.03 (1.31)$_b$

presence or absence of a situational constraint. The MOB account explains how this serves as an indirect signal of what moral occurrent beliefs the agent could have been experiencing, and then how those mindread moral occurrent beliefs altered how much praise each action elicited.

4 Study 3: Emotional or Rational Deficits

Study 3 built on the previous study in two ways. First, we manipulated a different extradecisional clue, a feature of the moral agent himself. Whereas all participants learned the agent had a neurodefect, we varied the supposed nature of the deficit. Some participants were told the agent had a rational deficit, meaning the agent was able to rely on emotional impulses only to guide his sense of right and wrong. Other participants were told the agent had an emotional deficit, meaning the agent could rely only on rational deliberation and calculation to differentiate right from wrong. Due to the earlier-reviewed connection between utilitarianism and reason, and deontology and emotion, we thought it likely (if participants intuit these properties) that the emotion-intact and reason-intact agents would be seen to more strongly possess the deontological and utilitarian occurrent beliefs, respectively. Note that we use brain deficit manipulation merely to test how assumptions about an agent's emotionality or rationality affects mindread occurrent beliefs and moral evaluation, not because of a specific interest in generalizing the results to those with neural deficits. Second, we measured MOBs and moral evaluations in the same sample. This permitted a test of the mediation model suggested by the matching-praise account, as well as the opportunity to rule out competing and artifactual accounts.

Participants in Study 3 considered the Nazi-baby dilemma used in Study 1b, in which a Jewish townsperson must decide whether to actively kill a baby whose crying will alert Nazi soldiers to the hidden location of Jewish townspeople. If our participants have the intuition that

the deontological beliefs would be more likely to occur to the emotion-intact agent, and utilitarian beliefs would be more likely to occur to the reason-intact agent, then our MOB account predicts that the two agents should receive different moral evaluations for deciding to kill (utilitarian) or not kill (deontological). Furthermore, and consistent with the support for the matching-praise account (in Studies 1a-1c), we expected that moral evaluation would be mediated by the assumed presence of the matching MOB (but not the assumed absence of the competing MOB).

4.1 Method

4.1.1 Participants and design

Four hundred sixty-four undergraduates from Cornell University were randomly assigned to one of four conditions in a 2 (intact faculty: emotion or reason) X 2 (decision: utilitarian or deontological) full-factorial, between subjects design. Participants received course credit for their participation.

4.1.2 Procedure

As in Study 1b, participants read the moral dilemma about Jewish townspeople hiding from Nazi soldiers in a basement. But this time, we included a manipulation that was designed not to be directly informative about the decision-maker's moral character, but that we expected would influence inferences about the agent's occurrent beliefs.

Those in the *reason intact* condition were told that Jack was "missing the part of his brain that allows him to have strong emotional impulses that signal what is morally right or wrong. Instead, all he can do is use rational calculation to calculate what is the right thing to do." In this way, it was noted Jack was "like a computer." Those in the *emotion intact* condition were told that Jack's deficit kept him from "engaging in rational calculations to arrive at his decision.

Instead, all he can do is use his strong emotional impulses that signal what is morally right or wrong." In both conditions it was noted that Jack was simply "born this way."[5]

Before learning Jack's course of action, participants indicated the occurrent beliefs they mindread in Jack. In particular, they estimated whether he was having the thoughts "Killing [the child] is wrong" and "By killing the child I could save more people." Both responses were made on 8-point scales anchored at 1 (*not at all*) and 8 (*completely*).

Participants then learned that Jack let the baby continue to cry (deontological decision) or that Jack smothered the baby (utilitarian decision). Finally, participants made judgments on 8-point scales that offered a moral evaluation (α = .82), indicating whether Jack: was a good person, should be praised, had a good moral conscience, had blameworthy moral character (reverse-scored), was an immoral person (reverse-scored), and was "in the wrong" (reverse-scored).

4.2 Results

Participants' inferences about Jack's occurrent beliefs depended on the nature of his brain deficit. A 2 (intact faculty: emotion-intact or reason-intact) X 2 (occurrent belief: utilitarian or deontological) mixed-model ANOVA, with only the second factor measured within-subjects, showed that MOBs depended on the type of neurological deficit, $F(1, 462) = 244.57$, $p < .001$, $\eta_p^2 = .35$ (see Table 3). Participants assumed that reason-intact Jack was more likely to have the utilitarian occurrent belief than was emotion-intact Jack, paired $t(231) = 9.60$, $p < .001$, $d = .63$. Emotion-intact Jack was instead assumed to have the deontological occurrent belief more so than reason-intact Jack, paired $t(231) = 12.42$, $p < .001$, $d = .82$.

As expected, moral evaluations followed a similar pattern: The Intact Faculty X Decision interaction was also significant, $F(1, 454) = 8.52$, $p = .004$, $\eta_p^2 = .02$. Jack was praised more for

smothering the child when he possessed reason compared to when he possessed emotion, $t(454) = 1.97$, $p = .05$, $d = .27$. In contrast, Jack was praised more for not killing the child when he possessed emotion compared to when he possessed reason, $t(454) = 2.16$, $p = .03$, $d = .28$.

To distinguish between our three accounts of how MOBs influence moral evaluation (see Figure 2), we again created a *matching occurrent belief* variable that reflected the extent to which Jack was assumed to have the occurrent belief that matched his ultimate behavior. We submitted the praise composite to a two-way 2 (intact faculty) X 2 (decision) ANCOVA, with appreciation of the matching occurrent belief and competing occurrent belief as covariates. Consistent only with the matching-praise account of MOBs, Jack was praised more to the extent he was assumed to have the matching occurrent belief, $F(1, 452) = 14.36$, $p < .001$, $\eta_p^2 = .03$, but was praised no differently for being assumed to have the competing occurrent belief, $F < 1$. Consistent with full mediation, the Intact Faculty X Decision interaction dropped to non-significance, $F < 1$. More formally, we tested the indirect effect of our manipulations (specifically, the Intact faculty X Decision interaction) on moral evaluation through the assumed presence of the matching occurrent belief. Using bootstrapped standard errors (10,000 resamples; Preacher & Hayes, 2008), we find a reliable indirect effect through inferences of matching occurrent beliefs, 95% CI [.0402, .1610], but not through the mindread competing occurrent belief, 95% CI [-.0627, .0437].

Note that like in Study 2, there also remained a main effect of Decision, $F(1, 452) = 7.05$, $p = .01$, $\eta_p^2 = .02$. As in Study 1b, there was a general tendency to think that it reflects better on a person to avoid actively killing a child. Note that the full mediation and the lingering main effect of Decision permit two distinct conclusions. The pattern of full mediation shows that the deficit manipulation's influence on the moral evaluation elicited by each behavior is *entirely*

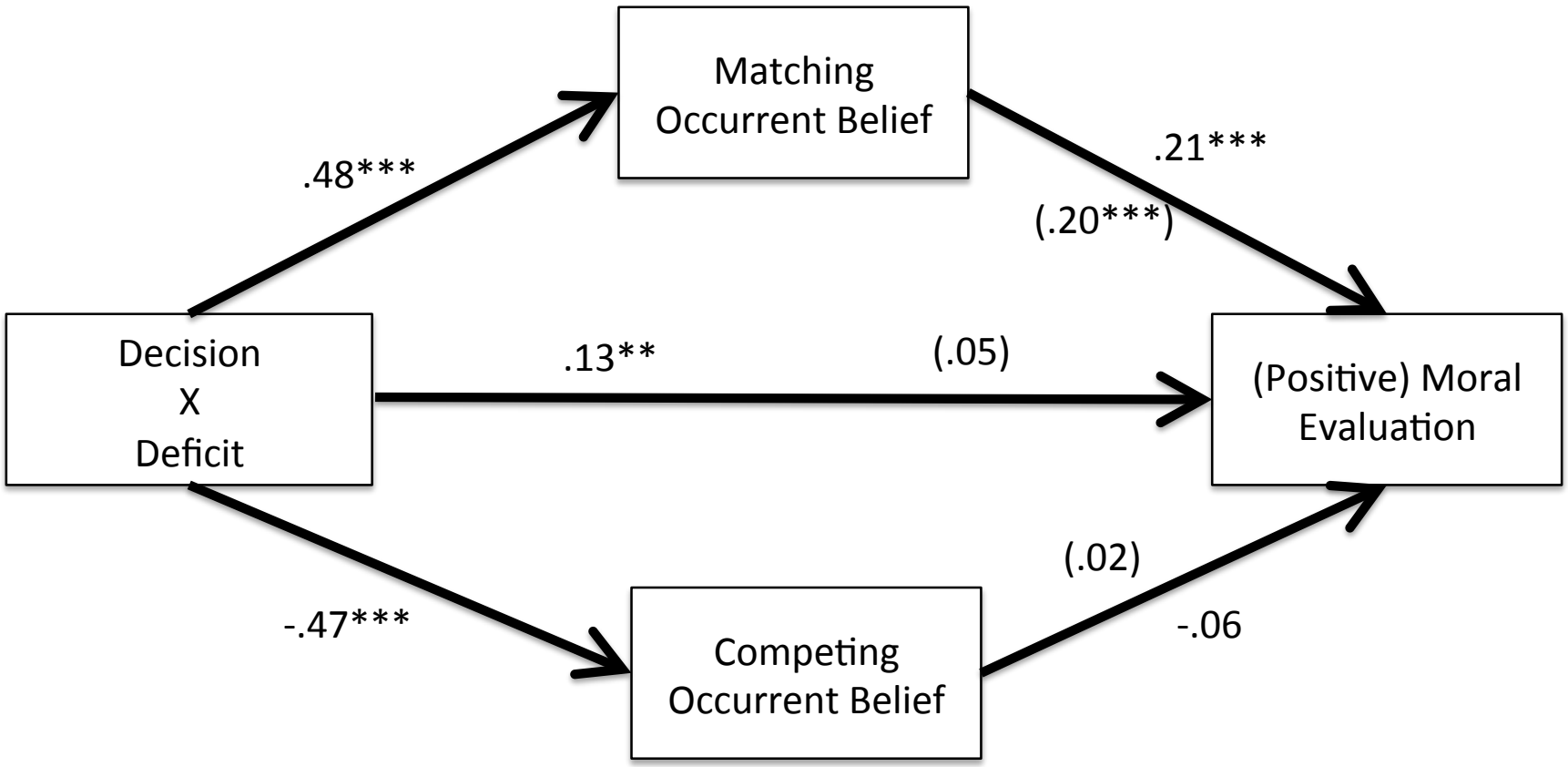

Figure 2 Mindread matching occurrent beliefs fully mediate the interactive influence of the manipulations (decision and deficit) on positive moral evaluations. There is no similar indirect effect through assumed appreciating of the competing occurrent belief. All numbers are standardized betas. Standardized betas in parentheses are estimated simultaneously in a single model. (Study 3).

explained by mindread occurrent beliefs. The lingering main effect of Decision instead reflects that mindreading occurrent beliefs are not the *only* influence on moral evaluations.

4.3 Discussion

Recall that in Study 1b, participants assumed that an agent would experience the deontological occurrent belief that killing a child is wrong, which explained elevated praise for making that choice. But in Study 3, when we introduced an extradecisional factor (i.e., emotion or reasoning deficits) that shifted participants' inferences about the agent's occurrent beliefs, moral evaluations for the agents' actions then shifted accordingly. Consistent with the matching-praise account, the moral agent was praised if he was assumed to have an occurrent belief justifying his action (i.e., the matching occurrent belief). In other words, participants offered praise to the extent it was plausible that an action followed from a relevant moral belief. It was not the case that agents were blamed more for failing to act on a competing occurrent belief (i.e., the competing-blame account), or that the assumed presence of any moral belief was a positive predictor of praise (i.e., the direct influence account). Furthermore, moral occurrent beliefs did not merely reflect participants' expectations about what the agent should or should not do. Had this been the case, both matching and competing occurrent beliefs would have each mediated (in opposite ways) the Decision X Deficit interaction on moral evaluation.

One strength of Study 3 is that the design allowed us to directly test for the influence of matching and mismatching MOBs. But one concern is that our measures may have walked participants through a reasoning process that they would not have spontaneously employed. To address this concern, we conducted a follow-up study using participants from the Berkeley Decision Research Group's on-line subject pool. Participants ($N = 115$) considered the same scenario, but only completed the moral evaluation measures. Would the brain-deficit

manipulation still influence moral evaluations, even when participants' attention had not been called to occurrent beliefs? A main effect of decision showed that Jack was judged more positively when he did not kill the child, $F(1, 111) = 15.43$, $p < .001$, $\eta_p^2 = .12$. (As noted before, this main effect hints that mindread occurrent beliefs are not the only influence on moral judgment.) More critically, there was a significant Intact Faculty X Decision interaction, $F(1, 111) = 4.43$, $p = .04$, $\eta_p^2 = .04$. Looking to the means by condition confirmed that the same pattern on moral evaluation was replicated. That is, Jack was praised relatively more for refusing to kill the child when he had his emotion vs. his reason intact (*M*s = 5.75 and 5.09, respectively). In contrast, Jack was praised relatively more for killing the child when he had reason vs. his emotion in tact (*M*s = 4.63 and 4.24, respectively).

To appreciate the usefulness of the MOB perspective, consider the present findings in light of recent developmental psychology research. Danovitch and Keil (2008) found that even young children report an emotionally-deficient computer to be a worse moral advisor than a rationally-deficient one. This suggests that people may prize emotional sentiments over rational calculation as a source of moral knowledge. Consistent with this possibility, participants in Study 1b thought that the action driven by the emotion-laden occurrent belief was superior (as reflected by the main effect of Decision). But participants in the present study showed no tendency to see the emotion-intact person as more morally praiseworthy than the reason-intact person. Instead, the intact faculty manipulation changed the praiseworthiness of each action. The moral evaluators seemed to care little that moral agents experienced one type of occurrent belief or the other, but instead were sensitive to the match between the agent's mindread occurrent beliefs and subsequent behavior.

5 Study 4: Visual Salience

Study 4 moved beyond situational (Study 2) or person (Study 3) factors that *limited* the agent's (perceived) ability to have a moral occurrent belief. Study 4 examined a heretofore unstudied factor that might be seen to *enhance* the salience of one of two competing moral beliefs: the agent's visual perspective. Study 4 used the terrorist-inn dilemma introduced in Study 1c, in which an agent must decide whether to bomb an inn containing both terrorists and innocent civilians. We varied the agent's visual perspective, such that either a terrorist or an innocent bystander loomed large in the agent's visual field while deliberating on what to do. We speculated that when the innocent bystander was said to be visually salient, that participants would assume the deontological occurrent belief (proscribing taking innocent human life) would become accessible. We hypothesized that when the terrorist was visually salient, that participants would assume the utilitarian occurrent belief (that through killing an innocent more lives could be saved) would occur to the agent. As in our previous studies, we predicted that the extent to which people had each occurrent belief would guide how much praise they received for the matching behavior.

5.1 Method

5.1.1 Participants and design

Two hundred nineteen undergraduates at the University of California, Berkeley, were randomly assigned to one of four conditions in a 2 (visible target: terrorist or innocent bystander) X 2 (decision: utilitarian or deontological) full-factorial, between-subjects design.

5.1.2 Procedure

We modified the terrorist-inn scenario used in Study 1c to facilitate a manipulation of visual salience. Participants read about two high-level military commanders, Michael and Matt, working to root out Al Qaeda terrorist cells in Afghanistan. The same information about

terrorists and an innocent bystander in a rural inn was again provided. The night of the meeting, the two military commanders look down at the inn from separate vantage points in the surrounding mountains. From Michael's lookout, the only person he can see through a window is the nervous-looking innocent translator. From Matt's lookout, the only person he can see through a window was a terrorist "who is #3 on the FBI's 'Most Wanted Terrorist' list". We reminded participants that despite their different vantage points "both Michael and Matt know who all is in the room." Michael and Matt each have to decide independently whether to recommend an airstrike, which would kill all of those present.

At this point, participants rated the likelihood that Michael and Matt would each have the relevant deontological occurrent belief ("One should not kill innocent people regardless of the circumstances") and utilitarian occurrent belief ("One must stop people from killing thousands of people, even if one must kill an innocent person to do this").

Next, participants learned about the behavior of only one of the commanders, either Michael (innocent bystander salient) or Matt (terrorist salient). The agent was said to have ordered the attack (utilitarian decision) or not ordered the attack (deontological decision). Participants rated the agent on the same five items used in our previous studies. All but one item were reverse-scored (praiseworthy vs. blameworthy character) so that higher numbers would reflect greater praise ($\alpha = .79$).

5.2 Results and Discussion

Participants believed that visual salience would influence the occurrence of the two moral beliefs: A 2(visible target: terrorist or bystander) X 2(principle: deontological or utilitarian) interaction emerged, $F(1, 218) = 224.09$, $p < .001$, $\eta_p^2 = .51$. Participants mindread that Michael, who was looking at the innocent bystander, would be more likely to have the deontological

occurrent belief than the utilitarian one, paired $t(218) = 16.14$, $p < .001$, $d = 1.09$. Matt who was looking at a terrorist, was instead assumed to be experiencing the utilitarian occurrent belief more than the deontological one, paired $t(218) = 3.02$, $p = .003$, $d = .20$. (See Table 3 for all means).

We then tested whether moral praise for each decision depended on who was salient in the agent's visual field. As expected, a Decision X Visible Target interaction emerged, $F(1, 215) = 4.20$, $p = .04$, $\eta_p^2 = .02$. When the terrorist was visually salient, Matt was praised more for ordering the strike than for failing to do so, $t(215) = 2.12$, $p = .04$, $d = .40$. In contrast, when the innocent bystander was visible, there was a non-significant reversal by which the agent was praised directionally more for failing to order the strike, $t < 1$, $d = .16$.

Using a similar analytic strategy to our earlier studies, we tested whether the matching and competing occurrent beliefs mediated the effects on moral evaluation (see Figure 3). The more the target was thought to have the matching moral occurrent belief, the more he was praised, $F(1, 213) = 7.60$, $p = .01$, $\eta_p^2 = .03$. The mindread competing occurrent belief had no influence on moral evaluations, $F(1, 213) = 1.98$, $p > .16$, $\eta_p^2 = .01$. With the two potential mediators included as covariates, the Decision x Visual Salience interaction was no longer statistically significant, $F(1, 213) = 1.06$, $p > .30$, $\eta_p^2 < .01$. Much as in Study 3, the influence of the manipulation (in this case, visual salience) on judgments of one action versus the other was *fully* accounted for by the mindread matching occurrent belief. We formally tested for mediation (10,000 resamples), and found that matching occurrent beliefs reliably mediated praise judgments, 95% CI [.0141, .2033], whereas the competing occurrent beliefs did not, 95% CI [-.1101, .0183]. Now for the fifth time, only the matching-praise account of MOBs was supported.

It is worth nothing that unlike Studies 2 and 3, the results of Study 4 did not yield a

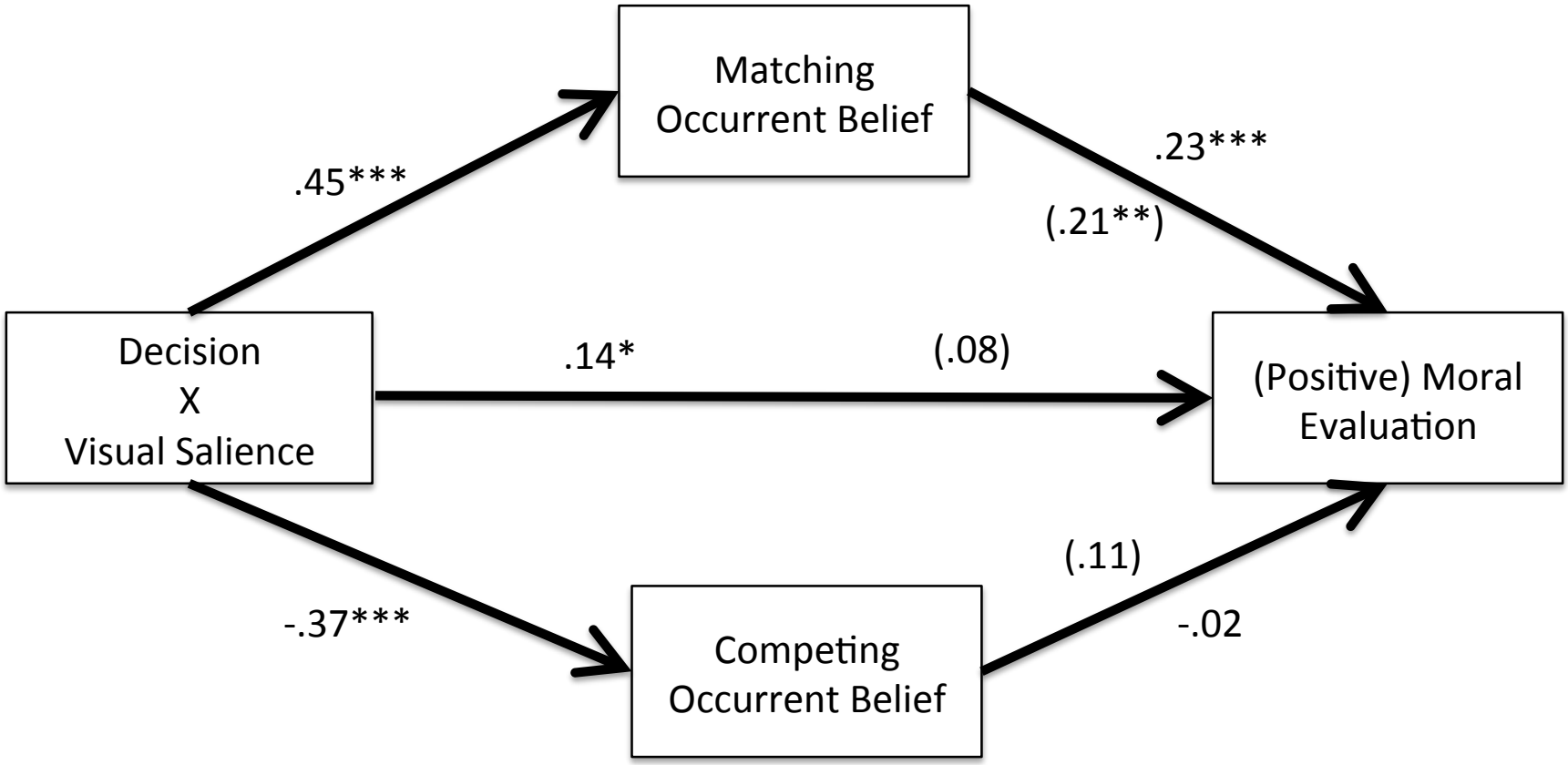

Figure 3 Mindreading the matching occurrent belief fully mediates the interactive influence of the manipulations (decision and visual salience) on praise. There is no similar indirect effect through assumed appreciating of the competing occurrent belief. All numbers are standardized betas. Standardized betas in parentheses or brackets come from the same model. (Study 4).

significant main effect of decision, $F < 1$. That said, there still remained a non-significant tendency to see the utilitarian decision as more praiseworthy than the deontological decision (Study 1c). This can be seen in looking at the asymmetric strength of the simple effects reported earlier. Nonetheless, the results do show that mindreading occurrent moral beliefs were the only significant predictor of moral judgment: Participants relied on the visibility of the two victims to infer what beliefs were salient to the moral agent, and this inference guided judgments of praise. This finding was foreshadowed in Study 1, in which this terrorist-inn dilemma (Study 1c) was the only scenario for which moral principles *fully* mediated the effect of decision on moral evaluation.

As in Study 3, the disadvantage of measuring both mindread occurrent beliefs and moral praise is that we may be documenting a meditational pathway that participants would not have proceeded through spontaneously. To address this limitation, we conducted a follow-up study in which we omitted the occurrent belief measures. Undergraduates ($N = 312$) at the University of California, Berkeley, saw one of the 4 versions of the terrorist-inn dilemma used in Study 4 before completing the moral evaluation measures. Bolstering confidence in the robustness of our effect, a reliable Decision X Visual Field interaction emerged, $F(1, 308) = 8.55$, $p = .004$, $\eta_p^2 = .03$. The commander was given more praise for ordering the strike when the terrorist (as opposed to the innocent translator) was visible (Ms = 5.01 and 4.77, respectively). By contrast, the commander was given more praise for deciding not to order the strike when the innocent bystander (as opposed to the terrorist) was visible (Ms = 4.98 and 4.37, respectively). Thus, the findings of Study 4 do not appear to be driven by explicitly asking participants to infer the agent's occurrent beliefs before formulating their moral evaluations.

Whereas the actual influence of the features manipulated in Studies 2 and 3 (decision

speed and emotion vs. reason) has been the subject of previous research, Study 4 introduced a novel feature, visual perspective. Recent research, though, has examined the role of mental imagery in moral judgment. Amit and Greene (2012) found that one reason people find it more acceptable to kill one person in order to save five people (a utilitarian action) when that involves flipping a switch (switch dilemma) as opposed to pushing the single victim to his death (footbridge dilemma) is that people are more likely to create a vivid mental image of the victim in the footbridge versus the switch dilemma. If one treats the visibility manipulation as analogous to more vivid mental imagery, then Amit and Greene's (2012) study could be cited as support for the reasonableness of our participants' intuitions.[6] What is important for the present purposes is that perceivers assume that visual perspective affects occurrent beliefs, and perceivers use this information in assigning moral praise.

6 General Discussion

Our studies document a novel means by which mindreading unfolds, distinguish between three accounts of how such mindread content influences moral evaluations, and leans on this account to predict how and why various non-moral, extradecisional cues (e.g., who is visible to an agent) affect judgments of agents' moral character. In particular, we establish the importance of mindread moral occurrent beliefs to moral evaluation for later-observed actions. We distinguished between three accounts of how MOBs might influence moral evaluation and found consistent support for the matching-praise account: Initial mindreading of agents' moral occurrent beliefs determines how much praise agents will receive for the subsequent, matching actions. Because moral beliefs may be assumed to occur (or not occur) to people due to extra-decisional features of the situation (e.g., the degree to which an agent must make a rushed decision), our account suggests a wide range of heretofore unappreciated influences on moral

evaluation.

Studies 1a-1c showed that mindread occurrent beliefs help explain which actions do or do not receive praise. The studies employed moral dilemmas similar to those used in much previous research on moral reasoning. These scenarios focused squarely on the details of a choice confronting an agent instead of on the moral cognitions or beliefs of the agent. Consistent with the matching-praise MOB account, the extent to which an agent was praised for each course of action was mediated by the extent to which the agent was assumed to experience the matching occurrent belief. There was no support for the direct-information or the competing-blame accounts: Competing mindread occurrent beliefs neither led to more praise, nor more blame, as these two accounts would have predicted, respectively. This evidence, combined with the finding that there was no significant negative correlation between the extent to which agents were assumed to experience one occurrent belief vs. the other, ruled out the MOBs-as-expectations artifactual account. In other words, MOBs did not merely identify the perceived wisdom of choosing to act or not to act in each way.

Studies 2-4 offered experimental tests of our model by varying features that were assumed to shift the assumed occurrence of different moral beliefs. Study 2 varied whether an agent was rushed in his decision; Study 3, whether an agent suffered deficits in emotion or reason; and Study 4, who was visually salient to the agent. These manipulations affected inferences about the agent's occurrent beliefs, and in turn, how much praise the agent received for each course of action. We found consistent evidence that people *spontaneously* relied on moral occurrent beliefs to inform moral evaluations: Although mediation models found consistent support for the matching-praise account alone (Studies 1a-1c, 3-4), we continued to observe effects consistent with this account even when we did not measure moral occurrent

beliefs (and thus did not call participants' attention to a construct they might not have spontaneously considered).

Our studies highlight how people rely on contextual information not only to determine whether actions are caused by the person or the situation—the historical focus of attribution theory (e.g., Kelley, 1967)—but instead to help them identify the underlying meaning of a behavior. Trope (1986) noted that many behaviors are inherently ambiguous (e.g., an emotional facial expression), and people rely on information about the situation (e.g., the fact that the emoter just won a bet) to resolve that ambiguity. Our MOB account similarly emphasizes that people may look to contextual factors to help resolve ambiguity about a behavior's underlying meaning. The present work details one general way in which this disambiguation unfolds: The context provides cues about what occurrent beliefs are likely active in an agent's mind, which changes the meaning of the subsequent behavior.

In light of recent findings that moral judgments can be pushed around by influences as trivial and incidental as hypnotically-induced disgust (Wheatley & Haidt, 2005), humorous film clips (Valdesolo & DeSteno, 2006), a bitter beverage (Eskine, Kacinik, & Prinz, 2011), and odious "fart spray" (Inbar, Pizarro, & Bloom, 2012; Schnall, Haidt, Clore, & Jordan, 2008), our depiction of moral perceivers as engaging in a sophisticated mindreading process may seem inconsistent. A similar apparent contradiction was considered by Simonson (2008), who asked how it is that people's preferences show signs of being constructed in the moment they are asked to report those preferences, even as people's underlying preferences show signs of stability. Simonson's resolution applies to both his question about preferences and ours about moral judgment: The error is in thinking that the psychological process must be characterized by one or the other, for in actuality both can apply. Moral evaluation may be shaped by fairly sophisticated

processes like mindreading occurrent beliefs even as such judgments are also (and perhaps simultaneously) influenced by incidental, biasing factors.

On this point, we should note that we did not predict (nor did our findings suggest) that mindreading occurrent beliefs is the *only* influence on moral evaluations. In Studies 1a and 1b, mindreading of moral occurrent beliefs partially mediated effects on praise, and in Studies 2 and 3 the deontological action was still praised more than the utilitarian action even after controlling for mindread occurrent beliefs. The fact that MOBs fully mediated the interactive effects of our manipulations on judgments indicates MOBs fully account for these manipulations' influence on praise. However, the places where main effects of decision lingered (even after controlling for MOBs) are the circumstances in which features other than MOBs affected moral evaluations as well. For example, although in Study 2 the relative praiseworthiness of funding sick Johnny's surgery compared to letting the child die was weaker when the hospital director had more time to consider his decision (and thus more time to come to appreciate the utilitarian occurrent belief), participants still thought it was relatively worse to trade off a child's life for money. That mindread occurrent beliefs are not the only influence on moral evaluation can also be seen in the fact that our manipulations' effects on mindread occurrent beliefs tended to be stronger than their effects on moral evaluations (given they are multiply determined—e.g., by the decision itself). Notwithstanding, in some circumstances mindreading occurrent beliefs fully accounted for moral evaluations following one action vs. another (Studies 1c and Study 4). One reading of this variability is that mindreading occurrent beliefs may always underlie moral evaluations, but sometimes other influences may matter as well.

One implication of the present findings is that the research question "What features of an action make it permissible or impermissible?" should be supplemented with "What features of a

decision-making context will change an agent's occurrent beliefs?" In our studies, participants' intuitions about occurrent beliefs conformed to certain patterns that need not (and likely do not) apply in all situations. For example, although participants in Study 2 intuited that deontological beliefs quickly occur to an agent, in other moral dilemmas it is actually the utilitarian beliefs that are quick and intuitive (Kahane et al., 2012). Whereas participants in Study 3 intuited a relationship between deontology and emotion, in other contexts it may be the utilitarian beliefs that are emotion-rich (Baron, 2011). Of course, there need not be a one-to-one correspondence between the social-cognitive reality of what moral beliefs spring to mind and perceivers' assumptions about this reality. Stated differently, the validity of our findings does not hinge on people having accurate predictions about what contextual features drive morally occurrent beliefs. That said, the usefulness of our model will depend on the ability to identify contextual factors that have a systematic effect on what MOBs are assumed to be made accessible.

Future research may also explore whether the MOB approach can explain why certain features of actions turn an otherwise permissible action into an impermissible one (Baron & Spranca, 1997; Mikhail, 2007). For example, people typically find it permissible to kill one person in order to save five if doing so requires flipping a switch (switch dilemma), but not when doing so requires pushing a man to his death (footbridge dilemma). In explaining this divergence, researchers have identified how the kill and no-kill actions take different forms in each scenario (e.g., Greene et al., 2009; Waldmann & Dieterich, 2007). But instead of explaining people's judgments by referencing descriptive rules governing each action (e.g., "Directly pushing someone to their death is morally outrageous!"), it may be helpful to consider how these same contextual variations may shift inferences about the agent's occurrent beliefs. For example, intentionally applying personal force to a victim likely requires that the agent hold the victim in

his visual field. If this visual perspective is assumed to make the occurrent belief condemning harm salient (as in Study 4), this could explain why perceivers believe agents should take the deontological no-harm action. This suggests that a greater of understanding of what influences mindread moral beliefs may help us to preemptively predict which actions will vs. will not earn an agent moral praise.

One may ask whether the current research can be directly extended to understanding *im*moral occurrent beliefs. That is, if people receive praise to the extent that they are assumed to have accessible an occurrent belief that would provide a moral justification for an action, would it also be the case that people are blamed more to the extent that they are assumed to have immoral occurrent beliefs prior to their actions? We suspect immoral occurrent beliefs may be treated differently. Such beliefs, because they are counternormative, may instead be consistent with the direct-information model of MOBs—one we considered but ultimately rejected in considering moral occurrent beliefs. If a person encounters a charity donation box and has the occurrent belief, "I could reach through the slot and grab $20 without being caught" perceivers may see this as direct information about the person's immoral character, even if the person ultimately does not act on the thought. Future research should extend the MOB approach to understand what model best characterizes the influence of mindread immoral beliefs on moral evaluation.

6.1 Conclusion

Although we have focused on understanding what guides evaluations of moral character, we suspect that the logic underlying our model can be extended to other types of person perception. In the moral domain, mindreading is central because perceivers are interested not merely in observed behavior, but in understanding whether such behavior was undertaken for the

right reason. This interest in mental precursors likely applies to non-moral evaluations as well. For example, a calculus teacher interested in judging her student's ability would want to know not just whether the student answered a multiple-choice problem correctly, but whether the student solved it in the correct manner. If the student answers correctly after a single second, it may be assumed that there was no time to actually work through the complex derivative that the problem required. As a result, praise for the student's calculus ability may be withheld. We look forward to future efforts to apply our model to additional domains, as well as attempts to better understand what cues people do (and also should) use to understand what beliefs occur to others.

References

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 23*, 861-868.

Audi, R. (1994). Dispositional beliefs and dispositions to believe. *Nous, 28*, 419-432.

Baird, J. A., & Astington, J. W. (2004). The role of mental state understanding in the development of moral cognition and moral action. *New Directions for Child and Adolescent Development, 103*, 37-49.

Baron, J. (2011). Utilitarian emotions: Suggestions from introspection. *Emotion Review, 3*, 286-287.

Baron, J., & Spranca, M. (1997). Protected values. *Organizational Behavior and Human Decision Processes, 70*, 1-16.

Bartels, D. M. (2008), Principled moral sentiment and the flexibility of moral judgment and decision making," *Cognition, 108,* 381-417.

Berger, Jonah, Meredith, M., & Wheeler, S. C. (2008). Contextual priming: Where people vote affects how they vote. *Proceedings of the National Academy of Sciences, 105,* 8846-8849.

Broeders, R., van den Bos, K., Müller, P. A., & Ham, J. (2011). Should I save or should I not kill? How people solve moral dilemmas depends on which rule is most accessible. *Journal of Experimental Social Psychology, 47*, 923-934.

Crawford, M. T., Skowronski, J. J., Stiff, C., & Scherer, C. R. (2007). Interfering with inferential, but not associative, processes underlying spontaneous trait inference. *Personality and Social Psychology Bulletin, 33*, 677-690.

Critcher, C. R., & Dunning, D. (2011). No good deed goes unquestioned: Cynical reconstruals

maintain belief in the power of self-interest. *Journal of Experimental Social Psychology, 47*, 1207-1213.

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2013). How quick decisions illuminate moral character. *Social Psychological and Personality Science, 4*, 308-315.

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

Cushman, F.A. and Greene, J.D. (2012) Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience, 7*, 269-279.

Danovitch, J. H., & Keil, F. C. (2008). Young Humeans: The role of emotions in children's evaluation of moral reasoning abilities. *Developmental Science, 11*, 33-39.

Eskine, K. J., Kacinik, N. A., & Prinz, J. J. (2011). A bad taste in the mouth: Gustatory disgust influences moral judgment. *Psychological Science, 22*, 295-299.

Fedotova, N. O., Fincher, K. M., Goodwin, G. P., & Rozin, P. (2011). How much do thoughts count? Preference for emotion versus principle in judgments of antisocial and prosocial behavior. *Emotion Review, 3*, 316-317.

Fein, S. (1996). Effects of suspicion on attributional thinking and the correspondence bias. *Journal of Personality and Social Psychology, 70*, 1164–1184.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, *23*, 101-124.

Greene, J.D. (2007). Why are VMPFC patients more utilitarian?: A dual-process theory of moral judgment explains. Trends in Cognitive Sciences, Vol. 11, No. 8, 322-323.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social*

Psychology, 45, 581-584.

Greene, J.D., Cushman, F.A., Stewart. L.E., Lowenberg, K., Nystrom, L.E., and Cohen, J.D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364-371.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144-1154.

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.

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*, 2105-2108.

Inbar, Y., Pizarro, D. A., & Bloom, P. (2012). Disgusting smells cause decreased liking of gay men. *Emotion, 12*, 23-27.

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience, 7*, 393-402.

Karniol, R. (1978). Children's use of intention cues in evaluating behavior. *Psychological Bulletin, 85*, 76-85.

Kelley, H. H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), Nebraska symposium of motivation (Vol. 15). Lincoln: University of Nebraska Press.

Knobe, J. (2003). Intentional action and side effects in ordinary language. Analysis, *63*, 190–193.

Knobe, J. (2004). Intention, intentional action and moral considerations. *Analysis, 64*, 181-187.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007).

Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature*, *446*, 908–911.

Kruger, J., & Gilovich, T. (2004). Actions, intentions, and trait assessment: The road to self-enhancement is paved with good intentions. *Personality and Social Psychology Bulletin, 30*, 328-339.

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.

Markus, H., & Kunda, Z. (1986). Stability and malleability of the self-concept. *Journal of Personality and Social Psychology, 51*, 858-866.

Mikhail, J. (2002). Aspects of the Theory of Moral Cognition: Investigating Intuitive Knowledge of the Prohibition of Intentional Battery and the Principle of Double Effect. Georgetown University Law Center Public Law & Legal Theory Working Paper No. 762385. Available at http://ssrn.com/abstract1/4762385

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Science, 11*, 143-152.

Miller, M. B., Sinnott-Armstrong, W., Young, L., King, D., Paggi, A., Fabri, M., Polonara, G., & Gazzaniga, M. S. (2010). Abnormal moral reasoning in complete and partial colostomy patients. *Neuropsychologia, 48*, 2215-2220.

Monroe, A. E., & Reeder, G. D. (2011). Motive-matching: Perceptions of intentionality for coerced action. *Journal of Experimental Social Psychology, 47*, 1255-1261.

Nichols, S., & Mallon, R. (2006). Moral dilemmas and moral rules. *Cognition, 100*, 530-542.

Piaget, J. (1932). *The moral judgment of the child.* London: Kegan Paul, Trench, Trubner and

Co.

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40,* 879–891.

Pronin, E. (2008). How we see ourselves and how we see others. *Science*, *320*, 1177–1180.

Reeder, G. D. (2009) Mindreading and dispositional inference: MIM revised and extended. *Psychological Inquiry, 20*, 73-83.

Reeder, G.D. & Spores, J.M. (1983). The attribution of morality. *Journal of Personality and Social Psychology, 44*, 736-745.

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology, 86*, 530-544.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin, 34*, 1096-1109.

Schopenhauer, A. (2009). *The two fundamental problems of ethics* (C. Janaway, Trans.). Cambridge, UK: Cambridge University Press. (Original work published 1841)

Simonson, I. (2008).Will I like a “medium” pillow? Another look at constructed and inherent preferences. *Journal of Consumer Psychology, 18*, 157−171.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology, 78*, 853-870.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Journal of Personality and Social Psychology, 93*, 239-257.

Uleman, J. S. (1999). Spontaneous versus intentional inferences in impression formation. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 141-160). New York: Guilford.

Valdesolo, P. & DeSteno, D.A. (2006). Manipulations of Emotional Context Shape Moral Decision Making. *Psychological Science, 17 (6),* 476-477.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: intervention myopia in moral intuitions. *Psychological Science, 18*, 247-253.

Wheatley, T., & Haidt, J. (2005). Hypnotically induced disgust makes moral judgments more severe. *Psychological Science, 16,* 780-784.

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*, 8235–8240.

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

Yuill, N. (1984). Young children's coordination of motive and outcome in judgementsC of satisfaction and morality. *British Journal of Developmental Psychology, 2*, 73-81.

Yuill, N., Perner, J., (1988). Intentionality and knowledge in children's judgments of actors responsibility and recipients emotional reaction. *Developmental Psychology, 24*, 358–365.

FOOTNOTES

1. We test our hypotheses using these dilemmas for two additional reasons. First, there has been extensive research in moral psychology on sacrificial moral dilemmas of this type, largely in an effort to develop a descriptive account of moral judgment (Bartels, 2008; Cushman & Greene, 2012; Mikhail, 2007). Relying on similar methodologies permits comparisons between our investigations. Second, and relatedly, this previous research has typically focused on what features of *actions* change moral judgments. This offers a particularly conservative context in which to test our mindreading occurrent beliefs accounts, given our interest in how inferred, but unobservable, occurrent beliefs may mediate moral judgments.

2. In Studies 1b and 1c, the deontological decision avoids a violation of Kant's categorical imperative.

3. It has even been suggested that this more psychologically-realistic route to deontological behavior is actually more praiseworthy than a dispassionate deduction from Kantian principles (Schopenhauer, 1841/2009).

4. It is worth noting that Bartels (2008) found, in an almost-identical dilemma, that people indicated that they would smother the child in this context (utilitarian behavior). We find that participants praise the agent more for not smothering the child (deontological behavior). This highlights that studies that examine how people would resolve dilemmas are not a substitute for studies of moral evaluation.

5. We used two questions to check whether participants in fact believed that appreciation of the deontological and utilitarian principles stemmed from emotionality and reason, respectively. Participants indicated on 8-point scales whether a decent person whose morals told him he should [not] kill the baby would be influenced more by his emotional impulses (1) or

dispassionate "mathematical" calculation (8). Participants indicated that a decent person's decision to kill the child would be driven more by mathematical calculation than by emotional impulses ($M = 6.05$, $SD = 1.69$), $t(457) = 19.58$, $p < .001$, but that a decision to let the child cry would be driven more by emotional impulse than mathematical calculation ($M = 2.54$, $SD = 1.72$), $t(457) = 24.41$, $p < .001$. These two tests against the midpoint (4.50) confirm our assumption that in this dilemma the utilitarian principle is assumed to be appreciated through reason, and the deontological principle, through emotion.

6. Bartels (2008) found that more vividly-written moral dilemmas—those that include affectively-rich details that more fully capture the emotions and tragedy of potential victims—elicit less utilitarian personal endorsements. Our scenarios hold explicitly-presented vividness constant, for they merely vary who is said to be visible through a window. That said, participants may have assumed that visual salience would make different occurrent beliefs salient because the salience of the innocent bystander or the terrorist may have made different outcomes more salient—i.e., the innocent taking of a life or a terrorist attack that would kill many people, respectively.