Journal of Experimental Social Psychology 87 (2020) 103906

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

Moral character evaluation: Testing another's moral-cognitive machinery☆

Clayton R. Critcher[a,*], Erik G. Helzer[b], David Tannenbaum[c]

[a] University of California, Berkeley, United States of America
[b] The Johns Hopkins Carey Business School, United States of America
[c] University of Utah, United States of America

ARTICLE INFO

Keywords:
Moral evaluation
Person perception
Mental state inference
Mental occurrents
Theory of mind

ABSTRACT

People evaluate the moral character of others not only based on what they do, but also on what leads them to 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 general target of such mental state inference, *mental occurrents*—a catchall term for the thoughts, beliefs, principles, feelings, concerns, and rules accessible in an agent's mind when confronting a morally relevant decision. *Moral* mental occurrents are those that can provide a moral justification for a particular course of action. Whereas previous mindreading research has examined how people *reason back* to make sense of an agent's behavior, we instead ask how inferred moral mental occurrents (MOs) constrain moral evaluations for an agent's subsequent actions. Our studies distinguish three accounts of how inferred MOs influence moral evaluations, show that people rely on inferred MOs spontaneously (instead of merely when experimental measures draw attention to them), and identify non-moral contextual cues (e.g., whether the situation demands a quick decision) that guide inferences about MOs. Implications for theory of mind, moral psychology, and social cognition are discussed.

1. Introduction

How do people form moral evaluations of others? 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, Vonk, Ronk, Ham, & Lawrence, 2004; see also Monroe & Reeder, 2011). Specifically, perceivers attempt to understand whether seemingly benevolent acts are done for moral reasons (Fedotova, Fincher, Goodwin, & Rozin, 2011; Gray, Young, & Waytz, 2012).

Understanding another person's reasons for acting, however, is difficult. Unlike behavior, the contents of another's mind are not directly observable (e.g., Pronin, 2008). This means people often engage in *mindreading* (Reeder, 2009) in an effort to infer the thoughts, feelings, plans, and emotions that were likely precursors to action (Helms, 2019). 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" (pp. 3–4). Mindreading comprises both mental state inference (drawing conclusions about another's thoughts and experience) and simulation (placing oneself in others' shoes to understand their internal life). These two processes often work in tandem (Goldman, 2001, 2006) and can proceed in a deliberative or an automatic manner (Reeder, 2009).

This paper articulates a newly identified way in which mental state inference unfolds and ultimately influences moral evaluations. We posit that judgments of another's character are fundamentally assessments of another's "moral-cognitive machinery" (Helzer & Critcher, 2018). Agents are judged to have good character when, among other things, their moral-cognitive machinery acts properly—responding with appropriate outputs in light of relevant inputs (Helzer & Critcher, 2018). By analogy, consider how one knows whether a car functions properly. It is not enough to see that it can come to a quick stop or that its brake lights flash. If either of these outputs were not preceded by the relevant input—a tap on the brakes—one would not say the car works well. In other words, the well-functioning nature of the machinery is displayed not merely by the output, but by knowledge or inference of the entire input-output chain.

The challenge with moral evaluation is that everyday observers are not able to directly observe the inner workings of this machinery. Although drivers can directly apply inputs (e.g., a turn of a wheel, a tap of the brakes) to test for proper functioning, social perceivers must instead lean on naturalistic tests of another's inner workings. That is, they can look to features of the decision context to infer what is (or

☆ This paper has been recommended for acceptance by Aarti Iyer.

* Corresponding author.
E-mail address: claytoncritcher@haas.berkeley.edu (C.R. Critcher).

https://doi.org/10.1016/j.jesp.2019.103906
Received 15 October 2018; Received in revised form 9 September 2019; Accepted 11 October 2019
Available online 10 December 2019
0022-1031/

should be) going on inside another's mind. From this perspective, the outputs of good character—that is, what actions one should take—are context-dependent. When two people are faced with the same choice between two alternative courses of action, the extent to which one action versus the other reflects good character will depend to some degree upon context cues operating at the time of the decision.

Our account of moral evaluation differentiates itself from other accounts in three ways. First, we examine how people engage in mindreading early in an agent's decision-making processing—inferring what is going on in another's mind in light of relevant inputs, and then ultimately forming a moral evaluation once the output (i.e., a morally- relevant behavior) is observed. This contrasts with previous research that has characterized mindreading as a process of backward reasoning to explain a previously observed behavior ("Now that I see everything that happened, what explanation best fits the data?"). Second, we lean on the umbrella concept of *mental occurrents*—terminology inspired by the philosophical construct of *occurrent beliefs* (Audi, 1994)—to identify the intrapsychic connection between the situational inputs and behavioral outputs that we argue are key to moral evaluation. Following Bartlett (2018), who suggested that occurrents can characterize not only beliefs but states more generally, we define mental occurrents as the thoughts, emotions, beliefs, sentiments, and concerns (essentially, a summary of an agent's mental content) active in an agent's mind when confronting a morally relevant decision.

Third, we differentiate ourselves from past moral psychology research by examining how moral evaluators infer and rely on others' moral (as opposed to immoral or non-moral) mental states to understand moral character. This contrasts with previous demonstrations that moral evaluations turn negative when people seemingly stand to gain from a superficially "good" action—both when actors make an ulterior motive explicit (Knobe, 2003; Mikhail, 2002) or when perceivers merely notice the possibility for agents' self-gain (Critcher & Dunning, 2011; Fein, 1996). Although other work has found mindreading can prompt more positive moral evaluations, that work has also focused on how people reason about selfish temptations—more specifically, temptations forgone (Reeder & Spores, 1983). We instead consider how a determination that another has good moral character stems not merely from a failure to identify bad reasons for performing an action but also by leaning on cues that suggest the proper motivators are present.

Our aim is not to empirically contrast the use of mental occurrents against other forms of mental state inference (i.e., assess their relative contribution), but instead to determine whether and how inferred MOs may factor into moral evaluation. We take an intentionally broad perspective on mental occurrents, conceptualizing them as the mental contents active in an agent's mind at a given point in time. In social cognitive terms, we focus on what is accessible, not what is merely available (e.g., Markus & Kunda, 1986). For example, many of our readers likely resonate with the idea "One should treat others as one would hope to be treated." But for most, it likely did not rise to the level of a mental occurrent until reading the previous sentence (see Goldman, 1970, for a fuller distinction between occurrent and standing beliefs). But as Bartlett (2018) noted, "A mental state's being active is not the same as its causing the subject to act" (p. 12). Occurrent states may precede, but do not always produce, an action.

Two properties of mental occurrents make them particularly interesting to study within moral psychology. First, mental occurrents are often visited upon a person involuntarily due to features of the decision context. This means perceivers can infer mental occurrents merely from knowing the decision an agent confronts, dispositional thinking styles of an agent, 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 have the mental occurrent "Schools really need the money" compared to 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 & Astington, 2004; Cushman, 2008; Karniol, 1978; Knobe, 2004; Miller et al., 2010; Piaget, 1932; Young, Cushman, Hauser, & Saxe, 2007; Young & Saxe, 2008; Yuill, 1984; Yuill & Perner, 1988), it is not immediately clear whether such unbidden cognitions would factor into moral evaluations. Our perspective identifies such mental occurrents as part of the input-output chain that reflects well-functioning moral-cognitive machinery. Second, people can (and do) draw inferences about an agent's moral mental occurrents 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 she makes her decision. It is atypical to consider, say, the intentionality of a behavior before it occurs, but inferences about mental occurrents—as beliefs about precursors to actions instead of actions themselves—are more natural to consider in this way.

1.1. How might inferred MOs inform moral evaluation?

In this paper, we set out to answer two research questions. First, we aimed to understand how inferred mental occurrents guide moral evaluation by empirically distinguishing among three possible accounts. Second, we considered what contextual cues are assumed to change moral agents' mental contents, thereby affecting evaluators' character assessments. For illustrative purposes, we explain each of our three accounts in the context of a classic sacrificial dilemma in which an agent John must decide whether to save five people by diverting a runaway trolley, thereby causing the death of one person who would not have died otherwise.

1.1.1. Account #1: inferred MOs prompt positive moral evaluations, independent of the agent's behavior

According to this *direct-information hypothesis*, agents are judged more positively when they are assumed to have moral mental occurrents. That is, the more that one infers John is experiencing *any* moral MO relevant to his choice ("By killing one person, I could save more lives" or "It is wrong to actively cause the death of an innocent person"), the more John may be judged to be a moral person, independent of his ultimate choice. Three lines of reasoning support the direct-information hypothesis. First, people make *spontaneous trait inferences* that track co-consideration of a person and a trait-relevant behavior (Crawford, Skowronski, Stiff, & Scherer, 2007; Uleman, 1999). If moral mental occurrents are thought of as (mental) behaviors, the mere assumption that John was entertaining a moral mental occurrent might boost character evaluations of John, even if his subsequent actions cause perceivers to tweak those assessments. Implicit in this characterization is that moral mental occurrents—as relevant and morally mature approaches to a problem—may themselves be tagged as moral. Second, the assumption that an agent possessed a moral mental occurrent might 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 mental occurrent "Donating to children's charities is important because doing so will help to alleviate suffering," but then observes her walk past a donation jar, the perceiver may give her something of a pass; they may assume she is going to use her money to do something even more morally worthwhile.

1.1.2. Account #2: inferred MOs constrain moral evaluations that will be offered for taking an action

A *matching-praise hypothesis* is the most natural deduction from our account that moral evaluators are trying to assess whether moral outputs are the products of a well-functioning moral-cognitive machinery. This account is premised on the idea that properly motivated behavior—that which comes from praiseworthy character—unfolds in a specific temporal sequence: Agents have a mental occurrent (e.g., John

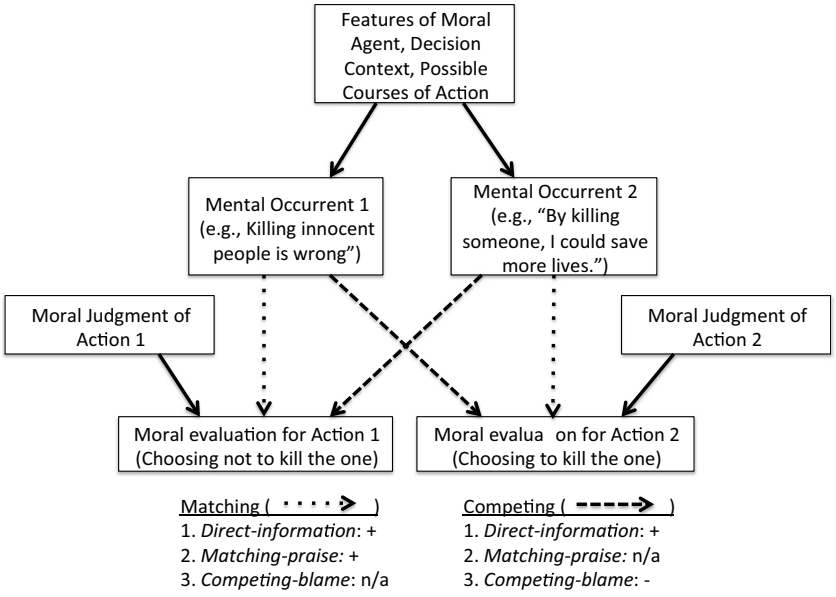

Fig. 1. How inferred moral MOs may influence moral evaluations, applied to a two-option moral dilemma. Features of the moral agent, the decision context, and the possible actions themselves influence what moral mental occurrents are assumed to transpire in an agent's mind. The moral agent can choose between competing actions. The three accounts—direct information, matching-praise, and competing-blame—differ in whether and how they predict inferred MOs will influence moral evaluations. The direct-information and matching-praise accounts predict a positive influence of the matching inferred mental occurrent on moral evaluation for an action (i.e., a positive effect along the dotted lines). The direct information account predicts a positive effect of the competing inferred mental occurrent on moral evaluation, whereas the competing-blame account predicts a negative effect (i.e., positive or negative effects along the dashed lines). An artifactual inferred-MOs-as-expectations account predicts that the matching inferred MO will have a positive effect *and* the competing inferred MO will have a negative effect. Although features of the moral agent, decision context, and possible courses of action should have strong effects on inferred mental occurrents, they should have weaker effects on moral evaluation (given moral evaluation is also influenced by assessments of the moral goodness of each possible action).

thinks, "I can save the most lives by diverting the trolley") that precedes the *matching* behavior (pulling the switch). By this account, the strength of a particular mental occurrent constrains how positively agents will be evaluated should they act in a manner that matches the preceding mental occurrent. This would reflect a sort of positive test strategy that focuses merely on actions taken (Were the precursors in place suggesting it was properly motivated?) instead of actions foregone. The matching-praise hypothesis predicts that agents will receive positive character evaluations to the extent they are assumed to have had the matching mental occurrent (i.e., the one that could provide a moral justification for the behavior) regardless of how much they were expected to possess a competing mental occurrent (i.e., one that would push for another behavior).

The matching-praise account is rooted in the idea that if a sentiment or 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 mental occurrent was inferred, then it is at least a possible (moral) basis for action. But what makes the matching-praise account particularly intriguing is that even when the mere occurrence of a belief is not directly informative (e.g., because it is prompted by the situation), a decision to act on that (involuntary) mental occurrent reveals the well-functioning moral-cognitive machinery. Thought of differently, even if people do not select the situations they find themselves in, those situations can provide the inputs that determine what outputs reflect praiseworthy moral character. This suggests that people are praised not simply because they can justify or are seen to possess the good character to think of moral justifications for their behavior. Instead, praise is offered when their actions can be justified by their mental occurrents—even when those occurrents are visited upon them by the situation.

1.1.3. Account #3: inferred MOs determine the negativity of moral evaluations that will be offered for forgoing each action

A complementary possibility, the *competing-blame hypothesis*, is that agents are seen as morally deficient to the extent they fail to act on moral mental occurrents they were assumed to be experiencing. On this account, how John is evaluated 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) mental occurrent, "It is wrong to actively kill an innocent person." More specifically, John should be viewed more negatively to the extent he was assumed to ignore a morally relevant mental occurrent. This account is a variation on Reeder and Spores's (1983) finding that moral agents are praised for not succumbing to selfish desires. The competing-blame account thus suggests moral agents may be seen as possessing blameworthy character for failing to follow the moral guidance of a mental occurrent.

1.1.4. Summary of empirical predictions

These three hypotheses make different (but overlapping) predictions for how inferred mental occurrents influence moral evaluations. In summarizing these predictions, it is helpful to differentiate *matching* from *competing* mental occurrents—those that might be seen to encourage the chosen or foregone course of action, respectively. The direct-information hypothesis (Account 1) predicts that both mental occurrents will be positive predictors of moral evaluations. The matching-praise hypothesis (Account 2) predicts that matching mental occurrents will be a positive predictor of moral evaluation. The competing-blame hypothesis (Account 3) predicts that the competing mental occurrent will be a negative predictor of moral evaluation. These accounts are not entirely mutually exclusive. For example, Accounts 2 and 3 could both be true. Instead, none of the three accounts may be correct, and people instead may evaluate moral character solely based on agents' behavior, regardless of what mental occurrents are inferred to be active.

1.2. Overview of the present studies

We conducted eight studies, using four different moral dilemmas, to test whether and how people infer mental occurrents in the service of moral evaluation. We note three considerations that relate to our reliance on dilemmas. First, agents must be deciding between competing courses of action in order to properly test and differentiate among our three accounts, which make different predictions depending on whether an agent acted or failed to act upon different mental occurrents. Second, three of our four dilemmas pit a utilitarian against a deontological course of action. These courses of action cleanly map onto two mental occurrents—deontology-backed aversions to direct harm and utilitarian justifications for promoting aggregate welfare (Nichols & Mallon, 2006). This facilitates a crisp test of whether and how people rely on inferred MOs in forming moral evaluations. Third, to extend our investigation beyond utilitarian/deontological dilemmas, we created a

fourth dilemma that did not lean on this dichotomy. Our approach and hypotheses are summarized in Fig. 1.[1]

The goal of Studies 1a–1d was to distinguish among our three accounts of how inferred MOs influence moral evaluation. These studies take the form of typical moral evaluation studies, offering information about an agent who is confronting a moral dilemma with no additional information as to what mental occurrents the particular agent is experiencing. After finding results consistent with one of our three possible accounts, Studies 2–5 test whether non-moral features have systematic effects on what moral MOs are inferred and, in turn, moral evaluation. These studies manipulated various features of the agent or decision context: whether the decision was made under time constraints (Study 2), whether the agent lacked basic emotional or cognitive capacities (Study 3), and who was focal in the agent's visual field (Studies 4–5).

We determined our sample sizes using two rules. When studies were run in the lab, research assistants recruited as many participants as they could until the end of an academic semester. When studies were run using Mechanical Turk, the funding lab's monthly budget was divided among all relevant studies to maximize the sample size of each. One exception was Study 4, which is the second version of the study we ran; we quadrupled the sample size because we added an exploratory moderator. This led us to average just over 90 participants per condition across our eight main studies, which exceeds the rough minimum threshold identified by Simmons, Nelson, and Simonsohn (2013) as requiring additional justification. All manipulations, measures, and exclusions are reported in the main text, whereas sensitivity analyses are reported in the Supplemental materials. Materials, data and analysis code can be accessed online: https://osf.io/vm6fe/?view_only=5d40d95583ee4893b79d2dc5d4162d45.

2. Studies 1a, 1b, 1c, and 1d

In Studies 1a–1d, we investigated whether and how people rely on inferred MOs 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 mental occurrent; thus, consistent with our aims, we measured what agents were presumed to be thinking *before* we provided information about how they had actually acted. Next, participants were randomly assigned to learn that the agent had actually chosen one course of action or the other. 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 inferred MO (the one that matches the chosen behavior) and the competing inferred MO (the one that mismatches the chosen behavior) should influence moral evaluation. We thus tested the significance and sign of each mental occurrent predicting moral evaluation (see Fig. 1). We also tested the possibility that inferred MOs merely track participants' expectations of what a moral person (or what the participants themselves) would and would not do. By this artifactual *inferred-MOs-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 mental occurrent that matches X but not Y. Given people tend to assume others are good people (Critcher & Dunning, 2014; De Freitas & Cikara, 2018; De Freitas, Cikara, Grossmann, & Schlegel, 2017; Helzer & Critcher, 2018), such a pattern is certainly possible, but would not suggest that inferred mental occurrents guide moral evaluation. This artifactual account makes two predictions. First, it predicts that the two inferred MO 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, this account predicts that we should find support for *both* the matching-praise and competing-blame accounts. That is, if moral mental occurrents merely reflect the pathways that perceivers think the agent clearly should versus should not follow, then people will 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–1d were Americans recruited from Amazon Mechanical Turk labor market for a small cash payment. The four studies had sample sizes of 95, 95, 108, and 146 participants, respectively. In each study, participants were randomly assigned to one of two decision conditions, which varied by the specific choice that agents confronted.

2.1.2. Procedure and materials

In each study, participants considered a different moral dilemma whereby a moral agent was confronted with two options, each of which could be supported by a different moral mental occurrent. Before learning how the agent decided, participants were asked to indicate whether the agent in the situation described was likely to experience each of two moral MOs—one supportive of each course of action. Finally, participants learned the agent's decision and offered a moral evaluation.

In Study 1a, participants read a modified version of Tetlock, Kristel, Elson, Green, and Lerner'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 spending the money to 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 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 Nazis' attention, which would result in the certain death of all of the townspeople. Jack had to choose between smothering the child, which would kill the baby but save everyone else (utilitarian decision), or letting the child continue to cry; the latter would ensure that the Nazis discover 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

[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 inferred MO accounts, given our interest in how inferred, but unobservable, MOs may mediate moral judgments.

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

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

Dilemma	Summary	Certain action	Risky action	Studies
Surgery Sequencing	A hospital director must decide whether to operate on a healthier or a more at-risk patient first. If the director operates on the healthier patient first, that patient will almost certainly live and the other will almost certainly die. If the director operates on the sicker patient first, that patient has a 20% chance of surviving and the healthier one's chance of survival drops to 55%.	The director chooses the healthier patient for the first surgery.	The director chooses the sicker patient for the first surgery.	1d, 5

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 Syrian 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 including the innocent translator (utilitarian decision). To make sure that utilitarian rationale 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).

Study 1d also introduced a new dilemma, but one that did not pit a utilitarian action against a deontological one. Participants considered a hospital director who had to decide how to sequence two surgeries. In one sequence, the director would essentially guarantee that the first patient would live, but that the second would die. In the opposite sequence, the chance of saving the sicklier patient's life rose to 20%, but the chance of saving the healthier patient's life dropped to 55%. The hospital director therefore had to decide between the first option (*certain* decision) and the second option (*risky* decision).

Table 1 provides a summary of the dilemmas and choices used in Studies 1a–1d.

2.1.2.1. Mental occurrents. Participants rated the extent the agent likely "appreciated, experienced, or possessed" each of two relevant mental occurrents. The measures specified that different moral concerns may be at the forefront of the agent's mind and that participants should indicate "to what extent you believe [the agent] is experiencing each sentiment as he is confronted with this situation." Crucially, these measures do not require participants to infer the presence of one state but not the other; both states or neither state could also be inferred. The wordings of the mental occurrents were modified for each 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 the extent to which Robert likely experienced each of the following thoughts: "It is morally wrong or troubling to let a child die" (deontological) and "By letting the child die, the hospital could actually save money which would allow it to ultimately save many more lives." (utilitarian). In Study 1b, participants indicated the extent Jack likely experienced these MOs: "It is morally wrong to let a child die" (deontological) and "By letting the child die, the hospital could actually save money which would allow it to ultimately save many more lives" (utilitarian). In Study 1c, participants estimated the extent Michael had each of two moral MOs: "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 order to stop the worse tragedy" (utilitarian). We followed precedent in assuming that deontological principles would be experienced as moral rules prohibiting certain actions rather than as a conscious application of Kantian meta-ethical beliefs (Broeders, van den Bos, Müller, & Ham, 2011; Greene, 2007; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Kahane, Everett, Earp, Farias, & Savulescu, 2015).[3] In Study 1d, participants indicated the extent the hospital director likely experienced each of the following MOs: "It is wrong to knowingly let someone die when there is a chance to save their life" (risky) and "It is wrong to risk someone's life when such a risk is not necessary" (certain).

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 the extent to which the

[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).

Table 2
Baseline moral mental occurrents and moral evaluations following each action.

	Inferred MO		Moral evaluation following behavior	
	Utilitarian	Deontological	Utilitarian	Deontological
Study 1a: Sick Johnny	6.02 (2.51)	8.02 (1.30)	3.97 (1.72)	6.47 (1.26)
Study 1b: Crying Baby	7.08 (2.03)	8.29 (1.13)	4.63 (1.35)	5.87 (1.55)
Study 1c: Terrorist-Inn	7.47 (1.84)	6.16 (2.36)	5.64 (1.32)	5.13 (1.24)
	Inferred MO		Moral evaluation following behavior	
	Risky	Certain	Risky	Certain
Study 1d: Surgery Sequencing	7.55 (1.66)	6.93 (1.94)	6.50 (1.38)	5.44 (1.54)

Note: Each mean is followed parenthetically by the corresponding standard deviation. Within each study, the two inferred MOs and the two moral evaluation measures significantly differ from each other at the $p < .05$ level.

agent: was a bad versus good person, had a bad versus good conscience, was or was not "in the wrong," had blameworthy versus praiseworthy character, and was in general a moral versus immoral person. After appropriate reverse-scoring, we averaged the judgments into *moral evaluation composites*, such that higher numbers reflect greater praise for the agent (Study 1a: $\alpha = 0.93$, Study 1b: $\alpha = 0.86$, Study 1c: $\alpha = 0.78$; Study 1d: $\alpha = 0.90$).

2.2. Results

Our three accounts differ in whether and how inferred MOs (i.e., matching or competing) are expected to predict moral evaluations. Before conducting the tests that differentiate the three accounts, we conduct initial tests that provide preliminary assessment of the artifactual account (that the inferred mental occurrents merely measure expectations for behavior), determine how the agent's decision influenced moral evaluations, and assess consensus about which moral mental occurrent would be more salient to the agent in each scenario.

First, we tested whether the measures of the two inferred MOs were strongly negatively correlated, as would be expected if inferred MOs merely reflect expectations that the agent (or the participants themselves) should behave in one way versus the other. The data fail to corroborate the artifactual account: In all four studies inferred MOs tended to be moderately positively correlated, and never significantly negatively correlated: Study 1a, $r = 0.09$, $p = .361$; Study 1b: $r = 0.19$, $p = .072$; Study 1c: $r = -0.00$, $p = .964$; Study 1d: $r = 0.36$, $p < .001$. Such a pattern is inconsistent with the artifactual, inferred-MOs-as-expectations account.

This said, there were reliable patterns concerning which MOs participants tended to mindread in each study (Table 2). Participants thought that the agents in Studies 1a ($d = 0.74$) and 1b ($d = 0.51$) would more strongly experience the deontological mental occurrent than the utilitarian mental occurent, whereas participants in Study 1c thought that the agent would show the reverse pattern ($d = -0.44$). Participants thought that the agent in Study 1d ($d = 0.30$) would experience the risk-promoting mental occurrent more strongly than the certainty-promoting one. Moral evaluations for each action followed a similar pattern to that of inferred MOs. In both Studies 1a ($d = 1.66$) and 1b ($d = 0.85$),[4] participants offered more praise to the agent who performed the deontological action than the utilitarian action. In Study 1c ($d = -0.40$), participants praised the utilitarian actor more than the deontological actor. In Study 1d ($d = 0.73$), participants praised the risk-seeking director more than the risk-averse director (see Table 3, Model I).

Although the striking consistency between inferred MOs and moral evaluations for each action is consistent with our accounts—especially the matching-praise and competing-blame possibilities—such a pattern does not necessarily distinguish between them because we have yet to test how inferred MOs predict moral evaluations. For each study, we regressed moral evaluation on behavior, the matching MO, and the competing MO. As a reminder, when the agent made the utilitarian or risky [deontological or certain] decision, the matching MO is the utilitarian or risky [deontological or certain] one. The other MO is the competing one.

As can be seen in Table 3 (Models II–IV), regardless of whether inferred MOs were entered as individual (Models II–III) or simultaneous predictors (Model IV), we found consistent support only for the matching-praise hypothesis (Account #2). In all four studies, the matching inferred MO was a significant positive predictor of moral evaluations. In other words, the amount of praise agents received for each action was determined by the degree to which they were believed to have experienced the behavior-consistent MO. Furthermore, we find support for an indirect effect of behavior on moral evaluations through inferences of matching mental occurrents (95% confidence intervals: Study 1a [−0.5089, −0.1321]; Study 1b [−0.3166, −0.0453]; Study 1c [0.0012, 0.1556]; Study 1d [0.0149, 0.1696]).

Importantly, in no case was the competing inferred MO a significant predictor of moral evaluation. Furthermore, we also failed to find any reliable indirect effects of behavior on moral evaluations through competing inferred MOs. That is, all 95% confidence intervals included 0: Study 1a [−0.3429, 0.0157]; Study 1b [−0.0475, 0.2932]; Study 1c [−0.0340, 0.2134]; Study 1d [−0.0084, 0.0955]. In short, agents were not consistently or significantly blamed or praised for the inferred mental occurrents that they failed to act on.

Of course, mediation analyses are merely consistent with, but do not definitively prove, a hypothesized causal pathway of X➔M➔Y (e.g., Fiedler, Harris, & Schott, 2018). Note that although M (the MO) was measured before X (the behavior), it was the manipulation of X that then identified which M is the matching (or competing) MO. Furthermore, M logically precedes Y (moral evaluation): Inferred MOs were measured before the moral evaluation for one action or the other could even be made, which supports the hypothesized ordering (Tate, 2015). Furthermore, in ruling out the inferred-MOs-as-expectations artifactual account, we addressed the most plausible third-variable account that could also produce data consistent with this indirect effect.

[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.

Table 3
Regression models predicting moral evaluation (Studies 1a–1d).

	Study 1a models				Study 1b models				Study 1c models				Study 1d models			
	I	II	III	IV	I	II	III	IV	I	II	III	IV	I	II	III	IV
Deontological/certain decision	0.64***	0.51***	0.59***	0.44***	0.40***	0.29**	0.45***	0.34**	−0.20*	−0.16	−0.14	−0.10	−0.28***	−0.30***	−0.34***	−0.28***
Matching MO		0.36***		0.37***		0.32**		0.30**		0.20*		0.20*		0.18*		0.23**
Competing MO			−0.11	−0.13			0.19	0.16			−0.14	−0.14			−0.04	−0.12

Notes. Model I is the direct effect of the IV (Decision) on the DV (moral evaluation). Models II and III add two possible mediators separately: matching inferred MO (Model II) and competing inferred MO (Model III). Model IV tests the robustness of the conclusions of Models II and III by including the candidate mediators as simultaneous predictors. All values are standardized betas.

* $p < .05$.
** $p < .01$.
*** $p < .001$.

2.3. Discussion

Studies 1a–1d provide suggestive evidence that inferred MOs influence the extent to which agents are praised for subsequent actions. We found consistent support across all four studies for the matching-praise account, but not for the competing-blame, direct-information, or (artifactual) inferred-MOs-as-expectations accounts. People infer moral MOs in order to determine which actions could (or could not) have been motivated by a particular mental occurrent. Our results suggest that inferred MOs in themselves did not lead to praise (direct-information hypothesis), nor were people praised less when they failed to act on a particular inferred MO (competing-blame hypothesis). Finally, the results from Studies 1a–1d suggest that inferred MOs are not merely expectations of what a moral person would or would not do (inferred-MOs-as-expectations hypothesis).

Participants in Studies 1a–1d 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 mental occurrents they thought they would have. Such mental simulation is merely one way by which mindreading occurs (Helms, 2019). 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 inferred the MOs to the extent they were consonant or dissonant, respectively, with participants' own forecasted behavior. Note that this is a variant of the inferred-MOs-as-expectations artifactual account, and the same empirical arguments made earlier speak against it.

The remaining studies build on these initial findings in two ways. First, Studies 2–5 use experimental manipulations that permit causal tests of the matching-praise hypothesis. Second, the upcoming studies test an implication of our model, namely that if inferred MOs constrain the space of praiseworthy behavior, then features of the decision context—even those not directly related to moral character—should affect moral evaluations to the extent they provide information about what moral MOs an agent is likely experiencing. The remaining four studies identify and test the influence of three such cues.

3. Study 2: time to deliberate

In Study 2, we returned to the dilemma used in Study 1a, in which a hospital director must decide whether to spend a large sum of money to save a sick child. This time we varied a feature that we suspected would shift inferences about the agent's moral MOs: whether the agent was pressured to decide quickly or was able to engage in extensive deliberation. External time constraints vary across real-world decision contexts. In Study 2, such time constraints were imposed by the situation, and thus did not offer a direct signal of the agent's 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 infer that a time-constrained agent would be more likely to experience deontological than utilitarian mental occurrents. But given more time to deliberate, the agent should be assumed to have both deontological and utilitarian mental occurrents.

If people infer MOs in this way, then the matching-praise account anticipates that perceivers should offer much more praise for the deontological decision when the agent is rushed (i.e., when the deontological MO is presumed to be more accessible than the utilitarian MO). However, this gap should diminish when the agent has sufficient time to deliberate (at which point both mental occurrents should be assumed present). For Study 2, we measured inferred MOs and moral evaluation using different samples. The advantage of this approach is that if MOs are a construct people attend to only when prompted by an experimental manipulation, then our manipulation of decision speed should have no effect. The disadvantage of measuring inferred MOs and moral evaluations in separate samples is 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 at Cornell University were randomly assigned to one of four conditions in a 2 (speed: rushed or lengthy) × 2 (decision: utilitarian or deontological) between-subjects design. Participants received course credit.

3.1.2. Procedure

Participants read a modified version of Tetlock et al.'s (2000) "sick Johnny" moral dilemma. In this version, two hospital directors must each 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 save many future lives. Participants were told 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 deciding.

In high-conflict personal moral dilemmas of this variety, people tend to quickly appreciate or experience a negative-affect-backed deontological mental occurrent (e.g., "Killing a child is wrong!...";

Table 4
Baseline inferred moral mental occurrents and moral evaluations following each decision.

	Inferred MO		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 bystanders	5.74 (2.27)$_c$	7.67 (1.56)$_a$	4.66 (1.54)$_c$	6.45 (1.16)$_a$
Terrorist	7.01 (1.99)$_b$	5.44 (2.32)$_d$	4.91 (1.41)$_b$	6.33 (1.18)$_a$

	Inferred MO		Moral evaluation following behavior	
	Risky	Certain	Risky	Certain
Study 5: Visual salience				
Emily	7.36 (1.81)$_b$	7.24 (1.97)$_b$	5.67 (1.02)$_{bc}$	5.39 (1.08)$_c$
Michael	7.73 (1.58)$_a$	6.90 (1.90)$_c$	5.86 (1.09)$_{ab}$	6.09 (0.98)$_a$

Notes. Each mean is followed parenthetically by the corresponding standard deviation. Note: Within each study and measure (inferred MO or moral evaluation), means that do not share the same letter subscript differ, $p < .05$.

Koenigs et al., 2007; Nichols & Mallon, 2006; Valdesolo & DeSteno, 2006), which with more time is supplemented or replaced by a utilitarian mental occurrent (e.g., "…but by killing now, I could save the lives of many people"; Greene, 2009; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene et al., 2008; Kahane et al., 2012). Our perspective remains agnostic as to whether deontological and utilitarian mental occurrents 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 such properties.

In a pretest on participants (N = 129) drawn from the same population, we confirmed our assumption that utilitarian occurrents (but not deontological occurrents) are seen as more likely to come to an agent after deliberation. Respondents presumed that Robert, who had no time to deliberate, would be more likely to have the deontological ("find it troubling to kill a person") than the utilitarian ("realize that by letting the person die, the hospital would actually save money which would allow it to save many more lives") mental occurrent, paired $t(128) = 7.20$, $p < .001$, $d = 0.63$. In contrast, respondents presumed that Alan, who had more time to deliberate, would experience both roughly equally, $t < 1$. The 2 (speed) × 2 (mental occurrent) interaction was significant, $F(1, 128) = 38.73$, $p < .001$, $\eta_p^2 = 0.23$ (see Table 4).

In our main sample, participants were told about the existence and constraints of both directors, but only learned the decision of (and morally evaluated) one. Participants made five moral evaluation judgments about the target, each on 8-point, Likert-type scales. They indicated whether the hospital director: should be praised (versus blamed), had a good moral conscience, was a good person, was the type of person one would want as a close friend, and was a moral person. We averaged these items to form a single index of moral evaluation ($\alpha = 0.86$).

3.2. Results and discussion

We submitted the moral evaluation composite to a 2 (speed) × 2 (decision) ANOVA. There was no significant main effect of speed, $F(1, 252) = 1.34$, $p = .248$, $\eta_p^2 = 0.005$, a significant main effect of decision, $F(1, 252) = 24.79$, $p < .001$, $\eta_p^2 = 0.090$, and consistent with our central hypothesis, a significant Speed × Decision interaction, $F(1, 252) = 4.80$, $p = .029$, $\eta_p^2 = 0.019$ (see Table 4). Robert, who had to make a decision immediately, was praised more for saving Johnny's life than for letting Johnny die, $d = 0.63$: $t(252) = 5.07$, $p < .001$. In contrast, Alan, who had sufficient time to think about his decision received only modestly more praise for the deontological than the utilitarian decision, $d = 0.25$: $t(252) = 1.97$, $p = .050$. 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 more time to consider the choice. In more general terms, the observed interaction is consistent with the inferred MO approach, whereas the fact that there was still a sight preference for the deontological decision (saving the sick child's life) shows that inferred MOs are not the only influence on moral evaluation. Given we measured inferred MOs and praise with different samples, the significant Speed × Decision interaction suggests that people rely on inferred MOs spontaneously in crediting targets (i.e., without measures that might artificially draw attention to the construct).

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 determining character (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. That is , it is only under such rushed circumstances that deontological but not utilitarian mental occurrents are assumed to be present. As such, the non-moral contextual feature permits a test of the agent's moral-cognitive machinery.

On a related point, 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; Robinson, Page-Gould, & Plaks, 2017; Tetlock et al., 2000). In the present research, the length of time participants had to deliberate was not selected by the 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 agent's dispositional motives (Critcher et al., 2013), but was

an exogenous cue that reflected the presence or absence of a situational constraint.

4. Study 3: emotional or rational deficits

Study 3 built on the previous study in two ways. First, we manipulated a contextual feature that pertained more directly more directly to the moral agent. Whereas all participants learned the agent had a neural defect, we varied the nature of the deficit. Some participants were told the agent had a "rational deficit," in that the agent was able to rely on only emotional impulses to guide his sense of right and wrong. Other participants were told the agent had an "emotional deficit," in that the agent could rely on only 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 moral occurrents, respectively. Note that we use this brain deficit manipulation merely to test how assumptions about an agent's emotionality or rationality affect inferred MOs and moral evaluation, not because of a specific interest in generalizing the results to those with neural deficits. Second, we measured inferred MOs and moral evaluations in the same sample. This permitted us to test mediation models that could distinguish among our competing 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 an infant whose crying will attract Nazi soldiers. If our participants have the intuition that the emotion-intact agent is more likely to experience deontological mental occurrents, and the reason-intact agent is more likely to experience utilitarian mental occurrents, then our favored IMO account predicts that the two agents should be evaluated differently for deciding to kill (utilitarian) or not kill (deontological) the infant. Furthermore, based on the results from Studies 1a–1d, we expected that moral evaluation would be mediated by the assumed presence of the matching IMO, but not by the assumed absence of the competing IMO.

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) × 2 (decision: utilitarian or deontological) 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. 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 rated the degree to which Jack was likely experiencing two mental occurrents: "Killing [the child] is troubling" 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 moral evaluations ($\alpha = 0.82$), indicating on 8-point scales whether Jack: was a good person, should be praised (vs. blamed), 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

Inferred MOs depended on the nature of Jack's brain deficit. A 2 (intact faculty: emotion-intact or reason-intact) × 2 (MO: utilitarian or deontological) mixed-model ANOVA, with only the second factor measured within-subjects, showed that inferred MOs depended on the type of neurological deficit, $F(1, 462) = 244.57$, $p < .001$, $\eta_p^2 = 0.346$ (Table 4). Reason-intact Jack was seen as more likely to have the utilitarian MO than was emotion-intact Jack, paired $t(231) = 9.60$, $p < .001$, $d = 0.63$. Emotion-intact Jack was instead assumed to have the deontological MO more so than reason-intact Jack, paired $t(231) = 12.42$, $p < .001$, $d = 0.82$.

As expected, moral evaluations followed a similar pattern: The Intact Faculty × Decision interaction was also significant, $F(1, 454) = 8.52$, $p = .004$, $\eta_p^2 = 0.018$ (Table 4). Reason-intact Jack was praised more for smothering the child than was emotion-intact Jack, $t(454) = 1.97$, $p = .050$, $d = 0.26$. In contrast, emotion-intact Jack was praised more for not killing the child than reason-intact Jack was, $t(454) = 2.16$, $p = .031$, $d = 0.28$.

To distinguish among our three accounts of how inferred MOs influence moral evaluation (see Fig. 2), we again created matching and competing mental occurrent variables that reflected the extent to which Jack was assumed to have the MO that matched or mismatched, respectively, his ultimate behavior. We submitted the moral evaluation composite to a two-way 2 (intact faculty) × 2 (decision) ANCOVA, with appreciation of the matching IMO and competing IMO as covariates. Consistent with only the matching-praise account, inferred matching mental occurrents were positively related to moral evaluations, $F(1, 452) = 14.36$, $p < .001$, $\eta_p^2 = 0.031$, whereas inferred competing mental occurrents were not reliably associated with moral evaluations, $F < 1$. This provides more direct support for the matching-praise account: To the extent that participants inferred that the matching MO was present, moral evaluations were more positive. Consistent with full mediation, the Intact Faculty × Decision interaction dropped to non-significance, $F < 1$. More formally, we tested the indirect effect of our manipulations (specifically, the Intact Faculty × Decision interaction) on moral evaluation through the assumed presence of the matching MO. We found a reliable indirect effect through inferences of the matching MO, 95% CI [0.0402, 0.1610], but not through the competing MO, 95% CI [−0.0627, 0.0437].

Note that like in Study 2, there remained a main effect of Decision, $F(1, 452) = 7.05$, $p = .008$, $\eta_p^2 = 0.015$. Similar to the results of Study 1b, there was a general tendency to provide more positive moral

[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 (*footnote continued*) "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.

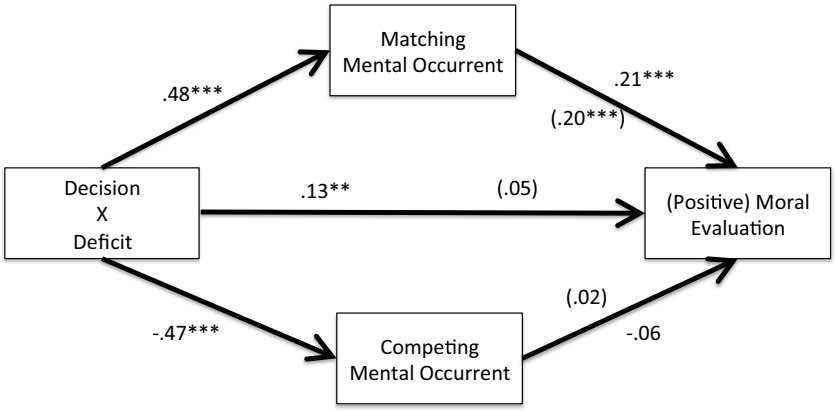

Fig. 2. Matching inferred MOs fully mediate the interactive influence of the manipulations (decision and deficit) on moral evaluations. There is no similar indirect effect through assumed appreciating of the competing MO. All numbers are standardized betas. Standardized betas in parentheses are estimated simultaneously in a single model (Study 3).

evaluations of the agent who refused to actively kill the child. Note that the full mediation and the lingering main effect of Decision permit two distinct conclusions. The full mediation indicates that the deficit manipulation's influence on the moral evaluation elicited by each behavior is entirely (statistically) explained by inferred MOs. The lingering main effect of Decision indicates that inferred MOs are not the only influence on moral evaluations.

4.3. Discussion

Recall that in Study 1b, participants assumed that an agent would have the deontological mental occurrent 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 MOs, moral evaluations for the agent's actions then shifted accordingly. Consistent with the matching-praise account, the moral agent was praised when he was assumed to have an MO justifying his action (i.e., the matching MO). In other words, participants offered praise to the extent it was plausible that the behavior had resulted because of a relevant moral mental occurrent. It was not the case that agents were blamed more for failing to act on a competing inferred MO (i.e., the competing-blame account), or that the assumed presence of any moral MO was a positive predictor of praise (i.e., the direct information account). Furthermore, inferred MOs did not merely reflect participants' expectations about what the agent should or should not do. Had this been the case, both matching and competing inferred MOs each should have mediated the direct effect (in opposite directions).

One strength of Study 3 is that the design allowed us to directly test for the influence of matching and competing inferred MOs. But one criticism is that our mental occurrent measures—by proposing two occurrents participants might not have spontaneously considered—may have suggested a cue that participants would not have spontaneously relied upon. To address this concern, we replicated the study but did not include the closed-ended inferred MO measures (N = 125, following attention check exclusions). Instead, participants were asked to report "what you think is going through Jack's head…what is he thinking and/or experiencing?" Even though this measure did not suggest potential mental occurrents that participants could consider, we continued to observe the same Intact Faculty × Decision interaction on moral evaluation, $F(1, 121) = 5.94$, $p = .016$, $\eta_p^2 = 0.05$. Emotion-intact Jack was praised more for refusing to kill the child ($M = 6.12$) than was reason-intact Jack ($M = 5.32$). By contrast, reason-intact Jack was praised relatively more for killing the child ($M = 4.76$) than emotion-intact Jack ($M = 4.39$). Furthermore, a supplemental analysis of participants' open-ended responses by a condition-blind coder revealed that Jack's alleged deficit changed the likelihood that participants spontaneously reported Jack would have the utilitarian or deontological mental occurrent, $F(1, 123) = 63.74$, $p < .001$, $\eta_p^2 = 0.05$. When participants considered reason-intact Jack, more identified the utilitarian sentiment (55%) than the deontological sentiment (1%) as a likely mental occurrent. This pattern flipped when considering emotion-intact Jack (15% vs. 44%, respectively).

To appreciate the usefulness of the inferred MO perspective, consider the present findings in light of relevant 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. 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 reason-intact vs. emotion-intact manipulation changed the praiseworthiness of each action. Moral evaluators seemed to care little that moral agents experienced one type of moral mental occurrent or the other, but instead were sensitive to whether agents optimized given the constraints of their moral cognitive machinery.

5. Study 4: visual salience

Study 4 built on our previous studies in two ways. First, we moved beyond situational (Study 2) or person (Study 3) factors that *limited* the agent's (perceived) ability to experience a particular moral MO. Study 4 examined a heretofore unstudied factor that might be seen to *enhance* the salience of one of two competing moral mental occurrents: the agent's visual perspective. Study 4 used a variant of 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 innocent bystanders loomed large in the agent's visual field while deliberating on what to do. We speculated that when the innocent bystanders were said to be visually salient, that participants would assume the deontological mental occurrent (proscribing taking innocent human life) would become accessible. We hypothesized that when the terrorist was visually salient, that participants would assume the utilitarian mental occurrent (that through killing an innocent person more lives could be saved) would occur to the agent. In other words, we suspected that the visual salience of the bystanders or the terrorists would be seen to cue thoughts related to one's moral concerns that most relate to that target. We expected to find support for the matching-praise account only, that the matching inferred MO (and not the competing inferred MO) would mediate judgments of praise.

Second, we included an exploratory measure that could yield insight into whether participants' preexisting moral intuitions might moderate our key effect: participants' political orientation. Notably, whether Americans think that foreign citizens' lives are permissible collateral damage in efforts to protect American lives is a question that divides along political lines. Liberals are more uncomfortable with the idea;

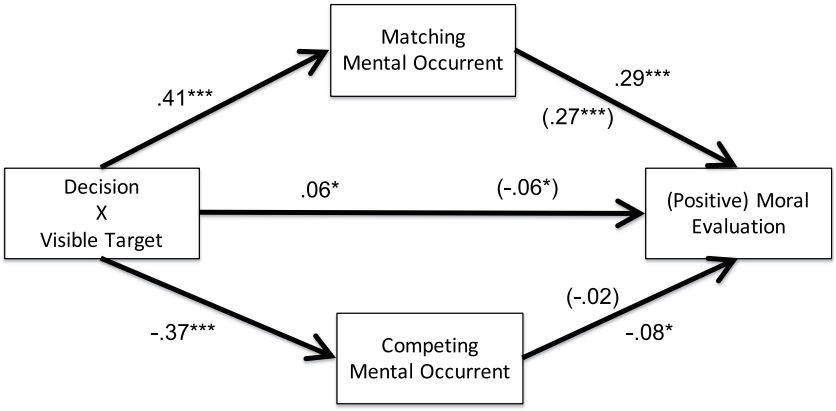

Fig. 3. Matching inferred MOs fully mediate the interactive influence of the manipulations (decision and visible target) on moral evaluations. There is no similar indirect effect through assumed appreciating of the competing MO. All numbers are standardized betas. Standardized betas in parentheses come from the same model (Study 4).

conservatives are more willing to consider it (Uhlmann, Pizarro, & Tannenbaum, 2009). Furthermore, we tweaked our scenario used in Study 1c to make the innocent bystanders potentially more sympathetic: They were the innkeeper and his family instead of a Syrian translator. If people appeal to inferred MOs only when they see less clear moral signal in others' behavior, then liberals may be less influenced by our visual salience manipulation. But if, as we have seen evidence of in other studies, reliance on inferred MOs merely complements other sources of moral information (e.g., direct judgments of the target's behavior), then we may find that liberals show evidence of seeing more signal of moral character in the behavior itself but no less influence from inferred MOs.

5.1. Method

5.1.1. Participants and design

One thousand one hundred eighty-nine Americans were recruited from Amazon Mechanical Turk. They were randomly assigned to one of four conditions in a 2 (visible target: terrorist or innocent bystanders) × 2 (decision: utilitarian or deontological) full-factorial, between-subjects design. We designed two questions that permitted us to identify random responders, robots, or others who advanced through the study without reading the study materials. Just after completing the moral evaluation measures, we asked participants to identify whether the person they evaluated was looking at the terrorist or innocent bystanders (86% accuracy) and whether he ordered the inn to be bombed or not (92% accuracy). The 963 participants who passed both checks are included in all analyses reported below.

5.1.2. Procedure

We modified the terrorist-inn scenario used in Study 1c in a few ways. One purpose was 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 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 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". From Michael's lookout, the only person he can see through a window is the innkeeper's family. We reminded participants that despite their different vantage points "both Michael and Matt know who all is in the inn." 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 ("It is wrong to kill innocent civilians regardless of the circumstances") and utilitarian ("One must stop the terrorists from proceeding with their terrorist attack, even if that means killing innocent people to stop the worse tragedy") mental occurrents. Next, participants learned about the behavior of only one of the commanders, either Michael (innocent bystanders 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 five moral evaluation items, indicating whether the agent is: a bad (vs. good) person, had a bad (vs. good) conscience, is completely (vs. not at all) in the wrong, had praiseworthy (vs. blameworthy) character, and is an immoral (vs. moral) person. Items were coded so that higher numbers would reflect more positive moral evaluations ($\alpha = 0.89$).

After participants completed the two attention checks, they responded to two questions asking "In general, how do you identify politically?" Responses were offered on two seven-point scales. These were anchored at 1 (*conservative/Republican*) and 7 (*liberal/Democrat*). The two items were highly correlated ($r = 0.78$) and averaged to form a *political orientation composite*. The manipulations appeared not to affect this composite, $Fs < 1$.

5.2. Results and discussion

Participants believed that visual salience would influence the occurrence of the two moral MOs: A 2 (visible target: terrorist or bystanders) × 2 (MO: deontological or utilitarian) interaction emerged, $F(1, 962) = 691.80$, $p < .001$, $\eta_p^2 = 0.418$. Participants inferred that Michael, who was looking at the innocent bystanders, would be more likely to experience the deontological mental occurrent than the utilitarian one, paired $t(962) = 20.75$, $p < .001$, $d = 0.67$. Matt, who was looking at a terrorist, was instead assumed to be experiencing the utilitarian mental occurrent more, paired $t(962) = 14.96$, $p < .001$, $d = 0.48$ (see Table 4 for the relevant descriptive statistics).

We then tested whether moral praise for each decision depended on who was salient in the agent's visual field. The Decision × Visible Target interaction emerged, $F(1, 959) = 4.66$, $p = .031$, $\eta_p^2 = 0.005$. When the innocent family was visually salient, Mike was praised much more for refusing to order the strike than for ordering it, $t(959) = 15.04$, $p < .001$, $d = 1.31$. But when the terrorist was salient (for Matt), this tendency was reduced, $t(959) = 11.42$, $d = 1.09$. That targets were generally judged to be of better character when they refused to bomb the innshows that inferred MOs are not the only contributor to moral evaluation. But to better test whether the agents' visual perspective moderated this effect due to the matching-praise account, we proceeded to test whether matching (but not competing) inferred MOs mediated this effect.

We used a similar analytic strategy to that used in our earlier studies (see Fig. 3). The more the target was thought to have the matching MO, the more he was praised, $F(1, 957) = 86.27$, $p < .001$, $\eta_p^2 = 0.083$. The competing MO had no influence on moral evaluations, $F < 1$. With the two potential mediators included as covariates, the

Decision × Visual Salience interaction remained significant, but only because it flipped in direction, $F(1, 957) = 4.14$, $p = .042$, $\eta_p^2 = 0.004$. 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 matching MO. We formally tested for mediation using Hayes's PROCESS model and found that the matching inferred MO reliably mediated praise judgments, 95% CI [0.1343, 0.2325], whereas the competing inferred MO did not, 95% CI [−0.0207, 0.0425]. Now for the sixth time, only the matching-praise account of inferred moral MOs was supported.

5.2.1. Are inferred MOs most important to those who most see this as a dilemma?

Although people generally agreed that the agent was a better person when he refused to kill the innocent family to eradicate the terrorists, we had suspected that this might depend on participants' own political orientation (Uhlmann et al., 2009). To test whether this was the case, we returned to our first model. We added as predictors political orientation (standardized) as well as the interactions that could be made with our original factors. The originally observed Decision × Visual Salience interaction was significant, $\beta = 0.06$, $t(955) = 2.27$, $p = .023$. As expected, the Decision × Political Orientation interaction was as well, $\beta = 0.13$, $t(955) = 4.90$, $p < .001$. This reflected that liberals were most willing to say that bombing the inn reflected worse character. Were these participants—who saw clearer signal in the behavior—less affected by the visual salience manipulation (which was in essence a manipulation of inferred MOs)? Speaking against this possibility, the three-way interaction did not reach significance; if anything, it trended in the other direction, $\beta = -0.03$, $t < 1$. This both reinforces our point that inferred MOs are not the sole determinant of moral evaluation, but also that they can remain influential even when perceivers see clear signal in the behavior itself.

5.2.2. Replication that removes inferred MO measures

As in Study 3, the disadvantage of measuring both inferred MOs 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 study in which we omitted the IMO measures. Undergraduates (N = 312) at the University of California, Berkeley, saw one of the 4 versions of the terrorist-inn dilemma used in Study 1c (i.e., without the minor modifications included in Study 4). Bolstering confidence in the robustness of our effect, a reliable Decision × Visual Field interaction emerged, $F(1, 308) = 8.55$, $p = .004$, $\eta_p^2 = 0.03$. The commander was given more praise for ordering the strike when the terrorist (as opposed to the innocent translator) was visible (*M*s = 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 (*M*s = 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 inferred MOs 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. What is important for the present purposes is that perceivers assume that visual perspective affects inferred MOs and use this information in assigning moral praise.[6]

6. Study 5: a moral dilemma about risk

In our final study, we employed the medical decision making dilemma from Study 1d. In the dilemma, a hospital director must determine in what sequence to operate on two patients. The inferred MO that it is wrong to knowingly let someone die when there is a chance to save their life would justify a *risky* course of action. The inferred MO that it is wrong to take unnecessary risks with someone's life would justify a risk-averse or *certain* course of action.

The hospital director happened to be able to see only the patient who would benefit from the risky course of action, or only the patient who benefit most from the certain course of action. After verifying in a pretest that the visual salience manipulation did indeed change inferred MOs, we tested in our main study whether the director's vantage point changed how he was evaluated for each course of action. That is, although the visual salience of one patient or another does not change which action is more or less consonant with a moral principle, our preferred MO account predicts that it should change moral evaluations for each action. Furthermore, because in our main study participants were not directed to consider inferred MOs, we can be more confident that any effects on moral evaluation reflect the effects of spontaneous mental state inference.

Finally, in our earlier studies, we measured moral evaluations mostly using bipolar scales. The assumption was that having praiseworthy character is the opposite of having blameworthy character. It is possible that our results stemmed from variation on only one of those poles (how moral the target was perceived to be or how immoral the target was perceived to be). To permit a test for such an asymmetry, we separated judgments of praiseworthy and blameworthy moral character and tested whether our effects significantly differed for the two composites.

6.1. Method

6.1.1. Participants and design. Three hundred thirty-three Americans were recruited from Amazon's Mechanical Turk and paid a nominal amount for their participation. They were randomly assigned to one of four conditions of a 2 (visible target: Emily or Michael) × 2 (decision: risky or certain) between-subjects design.

6.1.2. Procedure. Participants learned about the dilemma introduced in Study 1d, in which Robert—"the director of Healthcare Management at a rural hospital"—must decide how to sequence Michael and Emily's surgeries. Michael's only chance of survival is if his surgery is first. In this way, the (risky) MO that "it is wrong to knowingly let someone die when there is a chance to save their life" would push for prioritizing Michael's surgery. But by prioritizing Michael's surgery, participants learned that Emily's life would be put in danger. Although her survival would be certain if she were attended to right away, her survival became uncertain if Michael were to be operated on first. For this reason, the MO that "it is wrong to risk someone's life when such a risk is not necessary" is the (certain) MO that would justify operating on Emily first. As the director sat deliberating in his office, he looked across the courtyard at a patient wing. Although most of the window

[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 moral mental occurrents 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.

curtains were drawn, one was not. Through the glass, the director could see either Michael or Emily "looking anxious and afraid, holding the hands of [his, her] parents."

Before proceeding to the main study, we first wanted to verify that this visual salience manipulation would change what MOs the director was assumed to have. To allow us to run the pretest entirely within-subjects, respondents (N = 162 Americans on Amazon's Mechanical Turk) were told there were two hospital directors—one for whom Michael was visually salient, one for whom Emily was. Participants rated inferred MOs for each director on 9-point scales anchored at 1 (*not at all*) and 9 (*is strongly occurring to him*). Consistent with hypotheses, we observed a significant Visible Target (Michael or Emily) × MO (risky or certain) interaction, $F(1, 161) = 12.62$, $p = .001$, $\eta_p^2 = 0.07$ (Table 4). The risky IMO was seen as more likely to occur to the director looking at Michael than the one looking at Emily, paired $t(161) = 3.23$, $p = .001$, $d = 0.25$. In contrast, the certain IMO was seen as more likely to occur to the director looking at Emily than the one looking at Michael, paired $t(161) = 2.49$, $p = .014$, $d = 0.20$.

In our main sample, participants learned about the visual perspective and decision of a single hospital director. That is, the director was said to make the risky (Option 1) or certain (Option 2) decision. We did not use the words "risky" or "certain;" instead, the options and accompanying outcome probabilities were presented in a table. Unlike in our earlier studies, we had participants form moral evaluations of the director on positive and negative items separately. Participants were asked, "Given all you know about [the director], to what extent would you say he was: a good person, of good moral conscience, of praiseworthy moral character, and a moral person ($\alpha = 0.93$). Also on 1-to-7 scales, four items assessed the extent to which he was seen as immoral, asking if the target was": a bad person, of bad moral conscience, of blameworthy moral character, and an immoral person ($\alpha = 0.83$). Speaking to these measures' similarity, the moral and immoral composites were strongly negatively correlated, $r(331) = -0.58$, $p < .001$. For the purpose of analyses, we reverse scored our immoral evaluation composite.

6.2. Results

In order to test whether the visual salience manipulation changed how the director was morally evaluated for the two courses of action, we submitted the moral evaluation composites to a 2 (visible target: Emily or Michael) × 2 (decision: certain or risky) × 2 (composite: moral or immoral) mixed-model ANOVA. Only the final factor was measured within-subjects. Although we observed a significant main effect of Decision, $F(1, 329) = 15.05$, $p < .001$, $\eta_p^2 = 0.044$, the Visible Target × Decision interaction emerged as well, $F(1, 329) = 4.80$, $p = .029$, $\eta_p^2 = 0.015$ (Table 4). This interaction was not further qualified by an interaction with composite, $F(1, 329) = 2.69$, $p = .102$, $\eta_p^2 = 0.008$. In other words, the predicted interaction was not driven more by either moral or immoral evaluations.

The main effect of decision suggested participants generally gave higher moral evaluations to the director when he took the risky action (i.e., the one that tried to save both lives). This pattern was clear when it was Michael, the one who stood to gain for the risky action, who was in the director's visual field. In that case, the director was evaluated more positively when he took the risky option ($M = 6.09$) as opposed to the certain one ($M = 5.39$), $t(329) = 4.33$, $p < .001$, $d = 0.67$. When Emily was visually salient, the director was evaluated just as positively for the risky choice ($M = 5.86$) as the certain one ($M = 5.67$), $t(329) = 1.19$, $p > .23$, $d = 0.19$.

7. General discussion

Our studies document a novel means by which one form of mind-reading (mental state inference) unfolds. We empirically distinguish among three accounts of how such inferred content influences moral evaluations, thereby helping to predict when various non-moral, contextual cues (e.g., who is visible to an agent) affect moral evaluations. By our theoretical account, moral character can be conceived of as a person's moral cognitive machinery—the processor that is guided by external inputs in selecting specific courses of action. Although the operations of the machinery are not directly observable, context hints at what mental occurrents are likely present. The machinery works well when a moral mental occurrent effectively pushes for a matching course of action. Because moral MOs may be assumed to occur (or not occur) to people due to non-moral features of the agent and the context (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 character evaluation.

Studies 1a–1d document that inferred MOs help explain which actions do or do not receive praise. Our studies relied on moral dilemmas similar to those used in much previous moral psychology research. These scenarios focused squarely on the details of a choice posed to an agent instead of on outside factors that might be seen to affect agents' MOs. Consistent with the matching-praise moral mental occurrent account, the extent to which an agent was praised for each course of action was mediated by inferences about the matching mental occurrent. There was no support for the direct-information or the competing-blame accounts: Competing inferred MOs neither led to more praise nor blame, as these accounts would have predicted, respectively. This evidence, combined with the consistent finding that there was no significant negative correlation between the extent to which agents were assumed to experience one inferred MO vs. the other, ruled out the inferred-MOs-as-expectations artifactual account. In other words, inferred MOs do not appear to merely identify the perceived wisdom of choosing each course of action.

Studies 2–5 offered experimental tests of our model by varying features that were assumed to shift agents' moral mental occurrents. Study 2 varied whether an agent was rushed in his decision; Study 3 varied whether an agent suffered brain deficits in emotion or reason; Studies 4 and 5 varied who was visually salient to the agent. These manipulations affected inferences about the agents' MOs and, in turn, how much praise the agent received for each course of action. We found consistent evidence that people spontaneously relied on inferred MOs to inform moral evaluations: Although mediation models found consistent support for the matching-praise account alone (Studies 1a–1d, 3–4), consistent effects emerged even when we did not directly measure inferred MOs (and thus did not call special attention to an agent's mental occurrents; Studies 2 and 5, Follow-ups to Studies 3–4).

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., Dweck, 1975; Kelley, 1967)—but also to help them identify the underlying moral 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 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 moral MOs are likely active in an agent's mind, which changes the meaning of the subsequent behavior.

Perceivers' moral evaluations were sensitive to whether a matching moral MO was assumed to be present, not whether a competing MO was. This might seem to suggest that participants were engaging in a positive test strategy: a consideration of information that can support a hypothesis, instead of that which can speak against it. Such a pattern can underlie a confirmation bias (Fairfield & Charman, 2019; Oswald & Grosjean, 2004). Given our ultimate interest is in moral evaluation, the hypothesis being propped up would have to be that the agent who acts in a certain way is indeed a good person. But two aspects of our results suggest that our findings are not a straightforward extension of

confirmation bias. First, when we measured inferred moral mental occurrents, we always did so *before* participants had information about targets' ultimate behavior. In other words, participants were not attempting to bolster that the agent likely did have the matching mental occurrent once they knew how the agent behaved. Second, we think it would be a mistake to conceive of an inference that the competing occurrent belief was present as potentially disconfirming evidence that is being neglected. After all, in each study both mental occurrents that we measured were *moral* ones, and thus not neglected evidence that should straightforwardly call into question a presumption of good character.

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 mental state inferential process may seem inconsistent. A similar apparent contradiction was considered by Simonson (2008), who asked how people's preferences show signs of being constructed in the moment they are asked to report them, even as underlying preferences show clear signs of stability. Simonson's resolution applies equally well to our study of moral psychology: The error is in thinking psychological processes must be characterized in either one way or the other, for in actuality both can apply. Moral evaluation may be shaped by fairly sophisticated processes like inferring MOs 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 inferred MOs are the only influence on moral evaluations. In Studies 1a, 1b, and 1d, inferred MOs partially mediated effects on praise, and in Studies 2–5, residual main effects suggested one action typically elicited more positive moral evaluations than the other despite manipulations designed to change (and often flip) inferences about MOs. The fact that inferred MOs fully mediated the interactive effects of our manipulations on moral evaluation indicates inferred MOs fully account for these manipulations' influence on moral praise. However, the studies in which main effects of decision lingered (even after controlling for inferred MOs) are the circumstances in which features other than inferred MOs 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 MO), participants still thought it was relatively worse to trade off a child's life for more money. That inferred MOs are not the only influence on moral evaluation can also be seen in the fact that our manipulations' effects on inferred MOs tended to be much stronger than their effects on moral evaluations.

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 moral mental occurrents, and thus, the praiseworthiness of particular courses of action?" In our studies, participants' intuitions about inferred MOs conformed to certain patterns that need not (and likely do not) apply in all situations. For example, although participants in Study 2 inferred that deontological occurrents quickly occur to an agent, in other moral dilemmas it is actually utilitarian beliefs that are quick and intuitive (Kahane et al., 2012). Whereas participants in Study 3 inferred a relationship between deontology and emotion, in other contexts it may be 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 occurrents spring to mind and perceivers' assumptions about these patterns. Stated differently, the validity of our findings does not hinge on people having accurate predictions about what contextual features drive MOs. After all, there is disagreement about whether social perceivers—at least those who lack disorders that interfere with mental state inference (Craig, Hatton, Craig, & Bentall, 2004)—are consummate experts at mental state attribution (Zaki & Ochsner, 2011) or merely as good as they need to be (Ickes, 2011). The usefulness of our model depends only on the ability to identify contextual factors that are typically assumed to make certain mental occurrents more or less accessible.

Future research may also explore whether the inferred MO 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 the permissibility of actions (Cushman & Young, 2011), it may be helpful to consider how these same contextual variations shift inferences about the agent's moral mental occurrents. For example, intentionally applying personal force to a victim likely requires that agents hold the victim in their visual field. If this visual perspective is assumed to make the moral MO condemning harm salient (as in Studies 4–5), this could explain why perceivers believe agents should take the deontological action. This suggests that a greater understanding of what influences inferences about MOs may help us to preemptively predict which actions will earn an agent praise.

One may ask whether the current research can be directly extended to understanding immoral or even non-moral (neither moral nor immoral) mental occurrents. That is, if people receive praise to the extent that they are assumed to have accessible a mental occurrent 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 mental occurrents prior to their actions? We suspect immoral mental occurrents may be treated differently. The influence of such occurrents, because they are counternormative, may instead be consistent with the direct-information model of inferred mental occurrents. If a person encounters a charity donation box and has the known mental occurrent, "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. In other cases, non-moral mental occurrents ("I'm terrified, I'm not sure I'm strong enough to do this") might amplify praise when people do *not* act on such occurrents. Extending the inferred MO account to evaluations based on immoral and non-moral mental occurrents would be a worthy task for future research.

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 in understanding whether a particular 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 others' mental occurrents.

Open practices

This article earned Open Materials and Open Data badges for transparent practices. Materials and data are available at https://osf.io/vm6fe/?view_only=5d40d95583ee4893b79d2dc5d4162d45

Acknowledgments

The research reported in this article was supported in part by U.S. National Science Foundation award 1749608 to Clayton R. Critcher.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jesp.2019.103906.

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.

Bartlett, G. (2018). Occurrent states. *Canadian Journal of Philosophy, 48*, 1–17.

Berger, J., 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.

Craig, J. S., Hatton, C., Craig, F. B., & Bentall, R. P. (2004). Persecutory beliefs, attributions and theory of mind: Comparison of patients with paranoid delusions, Asperher's syndrome and healthy controls. *Schizophrenia Research, 69*, 29–33.

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., & Dunning, D. (2014). Thinking about others vs. another: Three reasons judgments about collectives and individuals differ. *Social and Personality Psychology Compass, 8*, 687–698.

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., & Young, L. (2011). Patterns of moral judgment derive from non-moral psychologicalrepresentations. *Cognitive Science, 35*, 1052–1075.

Cushman, F. A., & 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.

De Freitas, J., & Cikara, M. (2018). Deep down my enemy is good: Thinking about the true self reduces intergroup bias. *Journal of Experimental Social Psychology, 74*, 307–316.

De Freitas, J., Cikara, M., Grossmann, I., & Schlegel, R. (2017). Origins of the belief in good true selves. *Trends in Cognitive Sciences, 21*(9), 634–636.

Dweck, C. S. (1975). The role of expectations and attributions in the alleviation of learned helplessness. *Journal of Personality and Social Psychology, 31*, 674–685.

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.

Fairfield, T., & Charman, A. (2019). A dialogue with the data: The Bayesian foundations of iterative research in qualitative social science. *Perspectives on Politics, 17*, 154–167.

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.

Fiedler, K., Harris, C., & Schott, M. (2018). Unwarranted inferences from statistical mediation tests—An analysis of articles published in 2015. *Journal of Experimental Social Psychology, 75*, 95–102.

Goldman, A. (1970). *A theory of human action.* Englewood Cliffs, NJ: Prentice-Hall.

Goldman, A. I. (2001). Desire, intention, and the simulation theory. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.). *Intentions and intentionality: Foundations of social cognition* (pp. 207–224). Cambridge, MA: MIT Press.

Goldman, A. I. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading.* New York: Oxford University Press.

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, 11*(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., & 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.

Helms, N. R. (2019). *Cognition, mindreading, and Shakespeare's characters.* Cham, Switzerland: Palgrave Macmillan.

Helzer, E. G., & Critcher, C. R. (2018). What do we evaluate when we evaluate moral character? In K. Gray, & J. Graham (Eds.). *Atlas of moral psychology* (pp. 99–107). New York: Guilford Press.

Ickes, W. (2011). Everyday mind reading is driven by motives and goals. *Psychological Inquiry, 22*, 200–206.

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

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

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 (Vol. 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., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature, 446*, 908–911.

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, Vol. 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., & 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.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In R. F. Pohl (Ed.). *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*. New York: Psychology Press.

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

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.

Robinson, J. S., Page-Gould, E., & Plaks, J. E. (2017). I appreciate your effort: Asymmetric effects of actors' exertion on observers' consequentialist versus deontological judgments. *Journal of Experimental Social Psychology, 73*, 50–64.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after P-hacking. Meeting of the Society for Personality and Social Psychology.* LA: New Orleans. Retrieved from SSRN https://ssrn.com/abstract_id=2205186.

Tate, C. U. (2015). On the overuse and misuse of mediation analysis: It may be a matter of timing. *Basic and Applied Social Psychology, 37*(4), 235–246.

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.
Uhlmann, E. L., Pizarro, D. A., Tannenbaum, D., & Ditto, P. H. (2009). The motivated use of moral principles. Judgment and Decision Making. *4*, 479–491.
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 judgements 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.
Zaki, J., & Ochsner, K. (2011). Reintegrating the study of accuracy into social cognition research. *Psychological Inquiry, 22*, 159–182.