October 30, 2015

Moral judgment as information processing: an integrative review

Steve Guglielmo

frontiers in Psychology

REVIEW
published: 30 October 2015
doi: 10.3389/fpsyg.2015.01637

Moral judgment as information processing: an integrative review

*Steve Guglielmo**

Department of Psychology, Macalester College, Saint Paul, MN, USA

How do humans make moral judgments about others' behavior? This article reviews dominant models of moral judgment, organizing them within an overarching framework of information processing. This framework poses two distinct questions: (1) What input information guides moral judgments? and (2) What psychological processes generate these judgments? *Information Models* address the first question, identifying critical information elements (including causality, intentionality, and mental states) that shape moral judgments. A subclass of *Biased Information Models* holds that perceptions of these information elements are themselves driven by prior moral judgments. *Processing Models* address the second question, and existing models have focused on the relative contribution of intuitive versus deliberative processes. This review organizes existing moral judgment models within this framework and critically evaluates them on empirical and theoretical grounds; it then outlines a general integrative model grounded in information processing, and concludes with conceptual and methodological suggestions for future research. The information-processing framework provides a useful theoretical lens through which to organize extant and future work in the rapidly growing field of moral judgment.

Keywords: moral judgment, blame, mental states, intuition, reasoning, emotion, information processing

OPEN ACCESS

Edited by:
Mark Hallahan,
College of the Holy Cross, USA

Reviewed by:
Wing-Yee Cheung,
University of Southampton, UK
Sean Michael Laurent,
University of Oregon, USA

****Correspondence:***
Steve Guglielmo
sguglie@macalester.edu

Specialty section:
This article was submitted to Personality and Social Psychology, a section of the journal Frontiers in Psychology

***Received:** 24 August 2015*
***Accepted:** 12 October 2015*
***Published:** 30 October 2015*

Citation:
Guglielmo S (2015) Moral judgment as information processing: an integrative review. Front. Psychol. 6:1637. doi: 10.3389/fpsyg.2015.01637

Judging the morality of behavior is critical for a well-functioning social group. To ensure fair and effective interactions among its members, and to ultimately promote cooperation, groups and individuals must be able to identify instances of wrongdoing and flag them for subsequent correction and punishment (Boyd and Richerson, 1992; Fehr and Gächter, 2002; DeScioli and Kurzban, 2009; Tooby and Cosmides, 2010; Chudek and Henrich, 2011). Humans are quite adept at levying moral judgments and punishment upon others (Henrich et al., 2006; Boyd et al., 2010). One need only read the news on a given day to discover accusations, and appeals for punishment, of moral misconduct.

The study of morality has a rich history. Early and influential philosophers (Aristotle, 1999/330 BC) and psychologists (James, 1890/1950; Freud, 1923/1960) aimed to understand human morality and its implications for social behavior. More recent investigations have widened this scope of inquiry to examine a host of important questions concerning the evolutionary origins of morality (Hauser, 2006; Krebs, 2008), the emotional underpinnings of moral development and moral behavior (Eisenberg, 2000), the infusion of morality into everyday social interactions (Skitka et al., 2005; Wright et al., 2008), and the instantiation of moral judgment in systems of artificial intelligence (Wallach, 2010; Malle, 2014).

But an understanding of these questions requires an understanding of moral judgments themselves. Perhaps the most fundamental way in which humans categorize and understand behavior is to differentiate between *good* and *bad* (Osgood et al., 1957; Barrett, 2006b); moral judgment is an extension of this basic classification, although it is clearly more varied and complex. The literature has, for example, explored numerous related yet distinct moral judgments, including *responsibility* (Schlenker et al., 1994; Weiner, 1995), *blame* (Shaver, 1985; Alicke, 2000; Cushman, 2008; Guglielmo et al., 2009), and *wrongness* or *permissibility* (Haidt, 2001; Greene, 2007; Mikhail, 2007; Knobe, 2010).

How do humans make moral judgments? All judgments involve information processing, and although the framework of information processing has been widely implemented in models of cognitive psychology (Rosch, 1978; Marr, 1982), it has not been explicitly considered in investigations of morality. Nonetheless, existing models of moral judgment endorse such a framework, even if implicitly. With respect to moral judgment, this framework poses two fundamental questions: (1) What is the input *information* that guides people's moral judgments? and (2) How can we characterize the psychological *processes* that generate moral judgments? Extant models of moral judgment typically examine just one of these questions, with the unfortunate result that we know little about how the questions interrelate. This article critically reviews dominant models by locating them within this guiding theoretical framework, then provides an integrative account of moral judgment and offers suggestions for future research.

OVERVIEW OF THE CURRENT REVIEW

The study of moral judgment has grown rapidly, particularly within the past decade, yielding numerous proposed models of moral judgment. But although existing models have areas of substantial overlap, they are often studied in isolation, and empirical support for a particular aspect of a model is often taken as evidence for the veracity of the model as a whole. Further, these models investigate a suite of different moral judgments—including responsibility, blame, and wrongness, among others—that share commonalities but are not identical. A comprehensive and systematic analysis of moral judgment that assesses existing models in their entirety and in relationship to other models is sorely needed. Such an analysis would enable clarification of the claims of and support for existing models, explication of their areas of agreement and divergence, and organization within a unifying theoretical framework. This article provides such an analysis, critically evaluating existing models on empirical and theoretical grounds while also locating them within an overarching framework that emphasizes the information processing nature of moral judgment.

Existing models of moral judgment can be organized around their two fundamental goals. The first goal is to account for the particular information content that underlies people's moral judgments: the aspects of a behavior, or the agent who performed it, that lead people to hold the agent responsible, blameworthy, and so on. Models that focus on this goal are here referred to as *information models* (Shaver, 1985; Schlenker et al., 1994; Weiner, 1995; Cushman, 2008). These models include a subclass of *biased information models* (Alicke, 2000; Knobe, 2010), which hold that the very perceptions of such information content are driven by prior moral judgments. The second goal is to identify the psychological processes that generate moral judgments, including the extent to which these judgments are driven by intuitive or emotional processes on the one hand, or by deliberative processes on the other. Models that focus on this goal are here referred to as *processing models* (Haidt, 2001; Greene, 2007).

The goals of information models and processing models can be regarded as largely independent of one another. Revealing the importance of particular information features does not thereby establish the relative importance of intuitive or deliberative processes; similarly, revealing the importance of these processes does not thereby establish which information content drives moral judgments (cf. Rosch, 1978). A metaphor helps illustrate the distinction: we can separately examine the directions of travel on the one hand (information models) and the modes of travel on the other (processing models), although we will clearly be most successful by examining them together.

In reviewing the most prominent models within each of these classes, this article has three general aims: specifying the claims of each model; clarifying how the models compare to one another; and evaluating each model on empirical and theoretical grounds. The article then outlines a general information-processing view of moral judgment and highlights a specific recent model that adopts an information-processing approach (Malle et al., 2012, 2014). Finally, the paper offers conceptual and methodological suggestions for future research.

Before proceeding, though, we must first establish the domains in which moral judgment is relevant. Which general kinds of behavior have the capacity to elicit moral judgments? *Harm* and *fairness* are paradigmatic domains of moral judgment (Kohlberg, 1969; Turiel, 1983), but recent work has demonstrated the additional importance of *loyalty*, *authority*, and *purity* domains (Haidt, 2007, 2008; Graham et al., 2009, 2011; Haidt and Graham, 2009). Some scholars have argued, in contrast, that harm represents the single superordinate moral domain (Gray et al., 2012), and others suggest that moral judgments fundamentally reflect concerns about maintaining social relationships (Rai and Fiske, 2011). Despite the promise of a multitude of perspectives, extant research on moral judgment has been dominated by investigations of harm and fairness, which will therefore, by necessity, be the primary focus of the current analysis.

INFORMATION MODELS

Information models specify the features of an agent's behavior that shape people's moral judgments. Early models emphasized the concept of responsibility (Shaver, 1985; Schlenker et al., 1994; Weiner, 1995) and although they have provided noteworthy contributions, the concept of responsibility has proven to be incomplete in capturing the sensitivity of people's moral judgments, as we will see. More recent models, reviewed subsequently, have examined less ambiguous types of moral judgments such as wrongness or blame (Cushman, 2008).

Models of Responsibility

Shaver: Responsibility and Blame

Building upon the seminal work of Heider (1958), Shaver (1985) offers one of the earliest comprehensive psychological accounts of the particular components that underlie moral judgment. Shaver differentiates between *responsibility* and *blame* judgments, asserting that the latter presuppose the former. The heart of the model concerns responsibility judgments, which Shaver (1985, 1996; Shaver and Drown, 1986) argues are guided by five elements: the agent's *causal* contribution; *awareness* of negative consequences; *intent* to cause the event; degree of *volition* (e.g., freedom from coercion); and appreciation of the action's *wrongness*. Indeed, moral evaluations are sensitive to an agent's causal and intentional involvement in a negative action (Darley and Shultz, 1990; Ohtsubo, 2007; Lagnado and Channon, 2008), differentiate between responsibility and blame (Harvey and Rule, 1978), and follow a causality → responsibility → punishment pattern in particular (Shultz et al., 1981).

However, some aspects of the model are puzzling. Shaver (1996, p. 246) suggests that in some cases full responsibility applies yet blame is nullified—namely, when an agent has acceptable *justifications*, which "claim a larger positive social goal for which the intentional harm was produced," or *excuses*, which "claim that the particular consequences were not intended." But justifications seemingly appeal to Shaver's *wrongness* element of responsibility, and excuses seemingly appeal to the *intentionality* element. Thus, justifications and excuses should also weaken responsibility, not just blame. Further, Shaver claims that blame is assigned "after the perceiver *assesses and does not accept*" the offender's justifications and excuses (Shaver and Drown, 1986, p. 701, emphasis added). Although justifications and excuses can moderate blame substantially—socially desirable reasons or motives mitigate blame (Lewis et al., 2012; Piazza et al., 2013), whereas socially undesirable reasons or motives exacerbate blame (Reeder et al., 2002; Woolfolk et al., 2006)—there is no evidence that perceivers necessarily consider these factors prior to assessing blame. The emphasis on withholding blame until evaluating justifications and excuses may be ideal for a prescriptive model of how people *should* assign responsibility and blame but not for a descriptive model of how people actually make these judgments. As it turns out, Shaver's (1985) model is intended to be prescriptive; thus, its explanatory aim differs notably from descriptive models of moral judgment, on which the remainder of this paper will focus.

Weiner: Responsibility and Social Conduct

Weiner (1995) examines two related phenomena: people's judgments of responsibility and their emotional and behavioral reactions to others' behavior. In this model, considerations of controllability drive people's responsibility judgments, which in turn guide their emotional responses (e.g., anger vs. sympathy) and social actions (e.g., retaliation vs. helping) toward others. Weiner, like Shaver, holds that causality is a necessary but not a sufficient condition of responsibility: "the cause must be controllable if the person is to be held responsible" (Weiner, 1995, p. 11). If the cause of a negative outcome is "uncontrollable"—such as a heart attack or a low mental aptitude—responsibility judgments are withheld. Weiner (1995) reviewed a wealth of evidence showing that perceptions of controllability influence people's judgments of responsibility.

Although Weiner (1995) identifies several critical inputs to moral judgment, the model omits one key factor: intentionality. The distinction between intentional and unintentional actions is critical for moral judgment (Darley and Shultz, 1990; Ohtsubo, 2007; Gray and Wegner, 2008; Lagnado and Channon, 2008), but the concept of controllability is too broad to capture this distinction. On Weiner's (1995) model, both intentional and unintentional behaviors will often be "controllable," because the agent could have acted differently. But people's moral judgments distinguish between intentional behavior and negligent behavior, even if the negative consequences are identical (Cushman, 2008), which is reflected in the legal distinction between (intentional) murder and (unintentional) manslaughter. While Weiner's model cannot readily distinguish between intentional and unintentional behavior generally, the notion of controllability (i.e., consideration of the agent's capacity to foresee and prevent the negative outcome) nonetheless succeeds in explaining moral judgments about unintentional behavior specifically.

Schlenker et al.: Triangle Model of Responsibility

Schlenker et al. (1994) propose that responsibility judgments are shaped by the links between a prescription, an event, and an agent's identity. In particular, "people are held responsible to the extent that a clear, well-defined set of prescriptions is applicable to the event (prescription-event link), the actor is perceived to be bound by the prescriptions by virtue of his or her identity (prescription-identity link), and the actor seems to have (or to have had) personal control over the event, such as by intentionally producing the consequences (identity-event link)" (p. 649). The first link resembles Shaver's wrongness element and the third resembles Weiner's controllability element; the second link (prescription-identity) identifies the importance of an agent's obligations in the given situation. Schlenker et al. (1994) provided evidence that each link independently contributed to people's judgments of how responsible a worker was for his or her job performance. However, Schlenker et al.'s (1994) model has the same critical weakness as Weiner's: it omits intentionality.[1] As discussed above, the concept of controllability is too coarse to capture the distinction between intentional and unintentional behavior; although both types of behaviors typically are "controllable," people's moral judgments differentiate markedly between them.

Limitations of Responsibility Models

Extant models of responsibility highlight several components that shape people's moral judgments, including causality, controllability, and obligation. But these models fall short as comprehensive accounts of moral judgments due to their prescriptive emphasis (Shaver, 1985) or their omission of intentionality (Schlenker et al., 1994; Weiner, 1995). A further

[1] According to Schlenker et al.'s (1994) model, intentionality is only incidentally relevant, representing one way in which events may be controllable.

concern is that the concept of responsibility itself has taken on a host of meanings in the literature and is therefore not an ideal candidate for understanding moral judgment. Responsibility sometimes indicates mere causality—for example, Harvey and Rule (1978) examined "whether moral evaluations and causal responsibility are distinct judgmental dimensions," and Critchlow (1985) found that responsibility and causality judgments were similar across a range of behaviors. It can also denote general obligations (e.g., "Who is responsible for cleaning up?"), or it can simply be synonymous with blame (e.g., "Moral responsibility refers to the extent to which the protagonist is worthy of blame"; Shultz et al., 1981, p. 242, emphasis in original). Consequently, responsibility either lacks clear moral content (e.g., when it stands for causality) or is redundant with less ambiguous moral judgments (e.g., blame). Recent models have therefore examined less equivocal moral judgments while nonetheless incorporating key insights from early responsibility models.

Cushman: Causal-intentional Model of Wrongness and Blame

Cushman's (2008) causal-intentional model aims to account for distinct moral judgments of wrongness and blame. The model (Cushman, 2008, p. 364), shown in **Figure 1**, asserts that information about mental states underlies wrongness judgments, whereas mental states and consequences/causality jointly underlie blame judgments. More specifically, Cushman argues that inferences about *beliefs* (whether the agent believed his behavior would cause harm) and *desire*s (whether the agent wanted to cause harm) independently contribute to judgments of both wrongness and blame. The joint presence of these two mental states typically connotes that the behavior in question was intentional (Malle and Knobe, 1997a), and whereas many studies on moral judgment manipulate intentionality, Cushman (2008) examines beliefs and desires independently. In addition to these mental states, Cushman's (2008) model holds that *causes and consequences* (i.e., what actually happens as a result of an agent's action) influence blame. Agents may or may not actually cause harm—regardless of their intent—and blame will track the amount of harm (e.g., as in the differential punishment assigned to actual vs. attempted murder, both of which imply an intention to harm but differ in whether the harm actually occurred).

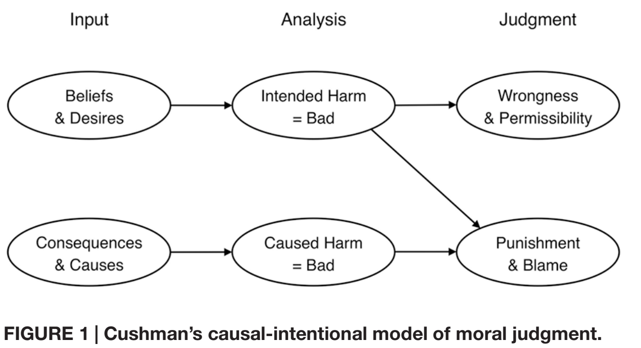

FIGURE 1 | Cushman's causal-intentional model of moral judgment. Reprinted from Cushman (2008) with permission from Elsevier.

Evidence for Cushman's Causal-intentional Model

The importance of causality and intentionality in moral judgment is well established. Blame is greater to the extent that an agent is seen as the cause of a negative event (Lagnado and Channon, 2008), and a substantial body of evidence shows that intentional negative actions are blamed and punished more than unintentional negative actions (Darley and Shultz, 1990; Ohtsubo, 2007; Gray et al., 2012). Further, culpable beliefs, desires, and motives increase blame both among adults (Young and Saxe, 2009; Tannenbaum et al., 2011; Inbar et al., 2012) and among children (Suls and Kalle, 1978; Nelson-Le Gall, 1985; Zelazo et al., 1996).

Cushman (2008) tested the model's more specific claims by independently varying belief, desire, and negative consequences, and then probing wrongness and blame judgments. For example, one vignette described Jenny, who was working in a sculpture class with a partner. Jenny did [not] want to burn her partner (desire present [absent]) and did [not] think that welding a piece of metal would burn her partner (belief present [absent]); Jenny welded the metal, which did [not] burn her partner (consequence present [absent]). Wrongness judgments were influenced by beliefs and desires but were essentially unaffected by consequences. Blame judgments told a different story: mental states continued to be critical, but blame was additionally influenced by consequences. Together, the results suggest that whereas wrongness is guided solely by mental states, blame is guided by mental states and consequences.

Other evidence for Cushman's (2008) model comes from studies on "outcome bias" or "moral luck," the pattern whereby an agent receives more blame for a behavior that *happens to have* bad consequences than for one that does not (or for a behavior whose bad consequences are worse than another's). For example, someone who carelessly backs out of a parking spot would receive more blame if he happened to hit a bystander than if he happened not to. A great deal of evidence reveals such outcome bias effects. For example, Mazzocco et al. (2004) found that while blame was mostly predicted by mental states (negligence, in their studies), it was partially predicted by negative outcomes too. Cushman et al. (2009) showed that punishment of a person's act of rolling a die (which could produce negative, positive, or neutral outcomes) was higher not only when the person intended to cause a bad outcome but also when a bad outcome occurred by chance. These patterns are consistent with Cushman's model, showing that blame is jointly a function of mental states and, to a lesser extent, consequences.

Limitations of Cushman's Model

Cushman highlights the role of consequences and causality, which indeed shape blame. But whereas the model treats these as identical inputs to blame (see **Figure 1**), in truth they refer to very different features. Consequences denote whether a negative outcome in fact occurred; causality denotes whether the agent in question was the cause of this outcome. The distinct roles of these two factors can be illustrated by a pattern in Cushman's (2008) Study 3, which showed that people gave an agent more blame when the agent caused harm than when the harm was caused by someone else. In both cases, the negative consequences were identical; what differed was whether the agent was the cause of the consequences or not. This suggests that an agent's causal role in producing harmful consequences is more important for blame than merely whether such consequences occurred.

One possibility not directly addressed by Cushman's model is that causal and intentional factors influence one another. Appearing to contradict this possibility is Cushman's finding that mental states and consequences had no interaction effect on moral judgments. But Cushman's vignettes *manipulated* mental state and consequence information, making obvious the presence or absence of each one. In contrast, people usually need to make these inferences themselves, and information about one factor will often guide inferences about another. For example, if a person performs an action that she believes will cause harm, people will tend to infer that she wanted to bring about harm (Reeder and Brewer, 1979; Guglielmo and Malle, 2010; Laurent et al., 2015a). Moreover, if an agent causes a negative outcome, people may infer corresponding culpable mental states (Pettit and Knobe, 2009; Young et al., 2010).

Summary of Information Models

Information models seek to characterize the critical information elements that guide people's moral judgments. Extant models have examined a range of different moral judgments, and have identified key distinctions between them. Yet several important consistencies have emerged. Moral judgments stem from identifying the occurrence of a negative event and the causal involvement of an agent. Moreover, perceivers consider whether the agent acted intentionally, as well as the agent's more specific mental states such as desires (including reasons and motives) and beliefs (including foresight and controllability). Notably, the importance of these features emerges early in development. Six-month olds generally dislike those who cause harm (Hamlin et al., 2007) and 8-month olds are sensitive to intentions, preferring an agent who intends to help over one who intends to harm (Hamlin, 2013). Further, 8-month olds prefer that agents respond with harmful behavior to an antisocial other (Hamlin et al., 2011), illustrating a rudimentary understanding that certain reasons or motives may permit an otherwise disfavored negative act. Lastly, children are sensitive to an agent's beliefs and the controllability of behavior, viewing negligent harm as morally worse than purely accidental harm (Darley and Shultz, 1990) and freely chosen harm as worse than behaviorally constrained harm (Josephs et al., 2015).

BIASED INFORMATION MODELS

Biased[2] information models hold that although the critical information elements identified by the preceding models—causality, intentionality, and other mental states—may shape explicit moral judgments such as blame, these elements are themselves directly influenced by more implicit moral judgments about the badness of an outcome or an agent (Alicke, 2000; Knobe, 2010). These models reverse the order of judgment postulated by information models in suggesting that moral judgments can precede, rather than just result from, causal and mental analysis. Although biased information models are not strictly incompatible with the preceding information models (since neither type explicitly denies the existence of the processing order favored by the other type), these two types clearly disagree about which processing order is most prevalent and thus has the most explanatory power with respect to people's moral judgments.

Alicke: Culpable Control Model of Blame

Alicke's (2000) culpable control model specifies the impact of "spontaneous evaluations" on causal and mental judgments (Alicke refers to these judgments as "structural linkage assessments"), as well as on blame. Spontaneous evaluations are affective reactions that arise "in response to information concerning a person's intentions, behaviors, or the consequences they produce" (Alicke, 2000, p. 558). Structural linkage assessments refer to judgments about mental states (e.g., intentions and foresight) and causality, which are of course the key elements identified by information models. Alicke (2000, p. 559, emphasis added) holds that "spontaneous evaluations influence blame attributions both *directly* as well as *indirectly* by means of their effect on more deliberate structural linkage assessments". To facilitate comparison to other models, "structural linkage assessments" are hereafter called "causal-mental judgments."

Clarifying the Predictions of Alicke's Model

Although Alicke does not provide a full graphical depiction of his model, we can construct one from his discussion (Alicke, 2000, pp. 564–568) of five unique combinations between spontaneous evaluations, blame, and causal-mental judgments (e.g., spontaneous evaluations may directly influence blame, but blame may have no further influence on causal-mental judgments, etc.). Three combinations posit direct effects of spontaneous evaluations on blame; one posits an indirect effect (via causal-mental judgments); one posits simultaneous direct and indirect effects. **Figure 2** combines these into a single representation of Alicke's model, whereby the more proposed pathways between pairs of variables (Alicke, 2000, p. 565), the thicker the arrow connecting them. From this construction, we see that the spontaneous evaluations → blame link is the strongest, followed by the spontaneous evaluations → causal-mental judgments link, and

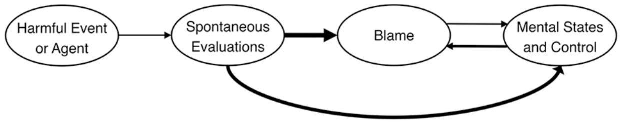

FIGURE 2 | Implied graphical representation of Alicke's culpable control model.

[2] "Bias" often carries normative connotations of error, but this is not the intended meaning here, since the models reviewed in this section disagree about whether their posited patterns reflect judgmental error. The current analysis invokes the more neutral meaning of "bias," merely connoting a particular tendency.

the causal-mental judgments → blame link (the last two of which constitute the indirect effect of spontaneous evaluations on blame). In short, Alicke's model implies that the direct effect of spontaneous evaluations on blame is much larger than the indirect effect.

Once represented explicitly in this way, we see that the model contains every pairwise connection between spontaneous evaluations, blame, and causal-mental judgments. This quality—that the model is "saturated"—makes model evaluation difficult, as saturated models accommodate every relationship and therefore cannot by falsified on statistical grounds.[3] To evaluate Alicke's model fairly, either the direct or the indirect effect of spontaneous evaluations on blame should be omitted, thereby avoiding the problem of model saturation. Emphasizing the direct effect is consistent with the graphical implication of Alicke's model (**Figure 2**) and with the claim that perceivers "search selectively for information that supports a desired blame attribution" (Alicke, 2000, p. 568). In other words, blame precedes and motivates assessments of mental states and causality. Consequently, the strongest evidence for the model will be evidence of a direct effect of spontaneous evaluations on blame; to the extent that the relationship is *indirect* (via causal-mental judgments), this should not be taken as support for blame validation.

Evidence for Alicke's Culpable Control Model

Direct effect

Alicke's (2000) section on Direct Spontaneous Evaluation Effects reviewed a single study that found an effect of outcome negativity on blame that was not mediated by causality ratings (Alicke et al., 1994). It is not clear, though, whether Alicke et al. (1994) assessed ratings of causality, nor are mediation analyses reported. Mazzocco et al. (2004) provide one of the few investigations of the mediational model implied by Alicke's model. In their studies, a protagonist killed an intruder who turned out to be either his daughter's boyfriend or a dangerous criminal. The critical prediction for Alicke's (2000) model is that the direct effect (the outcome → blame path, after accounting for the effect of negligence on blame) should be stronger than the indirect effect (the negligence → blame path, after accounting for the effect of outcome on negligence). However, the results showed the reverse: the indirect effect was significant in all four studies (average r = 0.42), whereas the direct effect was significant in just one study (average r =0.17).[4]

Alicke and Zell (2009) examined whether a protagonist's likeability—a possible measure of spontaneous evaluations—influenced blame. In one study, a socially likeable agent (who was polite to a policeman; volunteered at a homeless shelter) or unlikeable agent (who was rude to a policeman; lied to his boss) accidentally punched and injured an innocent woman. Blame was higher for the unlikeable character than the likeable one, and this effect was mediated by likeability ratings.

Indirect effect

As we have seen, there is little existing evidence for a direct effect of spontaneous evaluations on blame, which is the primary prediction of Alicke's model. We can nonetheless consider the evidence for an indirect effect; if such an effect stems from a motivational bias, whereby people "want" to perceive greater negligence (or causality, etc.), then this pattern may support Alicke's model.

Alicke (1992) found that an unlikeable agent (who was trying to hide cocaine) was judged more causally responsible for his ensuing car accident than was a likeable agent (who was trying to hide a gift). Participants in Mazzocco et al.'s (2004) studies saw an agent as more negligent when his actions had more negative consequences (e.g., the intruder he killed was his daughter's boyfriend vs. a criminal). Similarly, Alicke et al. (1994) found higher ratings of negligence and irresponsibility in the boyfriend vs. criminal condition. In all cases, the claim of Alicke's model is that spontaneous evaluations—triggered by the negativity of the agent and/or the outcome in question—led to enhanced ratings of causality and negligence, which thereby enhance blame.

Limitations of Alicke's Model

The major challenge to Alicke's model is that its primary claim of a direct effect of spontaneous evaluations on blame is not robustly supported. In a possible exception, Alicke and Zell (2009) showed that likeability predicted blame, but the assessed ratings of causality were not included in the mediation models, leaving it unclear whether likability influenced blame directly or indirectly (via causality). Further, the authors asked about the agent's "blameworthiness" in general (rather than for the specific act of injuring the woman), making it possible that the unlikeable agent received greater blame as a result of performing additional negative actions (e.g., being rude, lying).

Evidence consistently shows that the indirect effect from negative outcomes to blame—mediated by causal-mental judgments—is stronger than the direct effect. Could this indirect effect constitute an undue motivational bias? Alicke (2000, p. 566) indeed argues that, "Although a victim's loathsome character is irrelevant for determining legal responsibility (Federal Rules of Evidence, 2009, Rule 404b), there is little doubt that a jury's sympathies and antipathies for the victim influence their verdicts." Interestingly though, while Rule 404b forbids character evidence from informing guilt directly, it *does* permit such evidence to guide mental-state inferences: "Evidence of other crimes, wrongs, or acts. . .may, however, be admissible for other purposes, such as proof of motive, opportunity, intent, preparation, plan, knowledge, identity, or absence of mistake or accident."

This legally permissible pattern of influence may explain how negative character or outcome information, which is generally

[3] Falsification is possible by challenging the temporal or causal relationship between variables, such as by showing that blame guides spontaneous evaluations (in which case the term *spontaneous* evaluations would be a misnomer). This strategy requires manipulation of variables, but blame and causal-mental judgments are measured variables by definition. Further, few studies examine the timing of these judgments, making it difficult to challenge temporal relationships between variables.

[4] Robbennolt's (2000) meta-analysis of outcome bias effects on blame also obtained an average effect size of $r = 0.17$. However, this represents the zero-order outcome bias effect (i.e., without controlling for any related inferences); the residual outcome → blame path would surely reveal a smaller effect size.

more diagnostic than positive information (Reeder and Brewer, 1979; Skowronski and Carlston, 1987, 1989), shapes causal-mental judgments. People might (reasonably) infer that the drug-hiding agent in Alicke's (1992) car accident story was more reckless, impulsive, or apathetic than the gift-hiding agent, and these inferences may help account for the discrepancy in assessments of causality. Similarly, negative outcomes often bring to bear other inferences about foresight or preventability. When an agent mistakenly kills his daughter's boyfriend (as in Alicke et al., 1994), participants might infer that the agent could or should have known about the boyfriend's presence, thus blaming the agent for his unwarranted false belief that the intruder was a criminal. Consistent with this interpretation, Young et al. (2010) showed that people assign substantial blame to agents who act upon false beliefs, regardless of whether they ultimately caused harm. The influence of negative outcome or character information on causal-mental judgments is therefore likely *informational*, not *motivational*, since negative information is a diagnostic indicator of related inferences about dispositions, foresight, and preventability (cf. Uttich and Lombrozo, 2010).

Alicke's model nonetheless raises the important possibility that early affective responses may impact later phases of moral judgment. Future research must be careful to determine whether this link is affective/motivational or informational in nature. If it turns out to be the former, then information models of moral judgment will need to specify how early evaluative responses shape later judgments (e.g., causal-mental judgments and blame).

Knobe: Moral Pervasiveness Model

Knobe's moral pervasiveness model (Pettit and Knobe, 2009; Knobe, 2010) depicted in **Figure 3** (adapted from Phillips and Knobe, 2009) asserts that "initial moral judgments" influence causal-mental judgments. Knobe's (2003a,b) earliest work suggested that initial moral judgments were judgments of blame; more recent specifications view them as akin to judgments of goodness or badness: "people's judgments of good and bad are actually playing a role in the fundamental competencies underlying their concept of intentional action." (Knobe, 2006, p. 221). Knobe's model posits that initial moral judgments influence *all* the components identified by information models, including intentionality (as well as desires, beliefs, or decisions; Knobe, 2010), causality (Knobe and Fraser, 2008) and reasons for acting (Knobe, 2007).

Whereas Alicke's account is that rudimentary initial moral judgments *can* guide causal-mental inferences, Knobe's account is that the very concepts underlying these inferences are fundamentally shaped by moral concerns: "Moral judgment is pervasive; playing a role in the application of *every* concept that involves holding or displaying a positive attitude toward an outcome" (Pettit and Knobe, 2009, p. 593). Thus, both models posit that people have immediate evaluative reactions, which then influence their causal-mental assessments. Alicke holds that this is a *motivational* process of blame-validation, whereby people exaggerate their causal-mental judgments to justify their initial negative evaluations. In contrast, Knobe holds that these effects reflect a *conceptual* influence—by virtue of viewing an action as bad, people directly perceive more culpable causal-mental features.

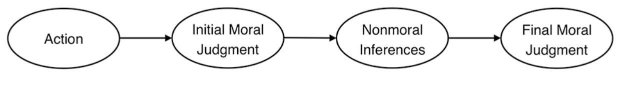

FIGURE 3 | Knobe's moral pervasiveness model.

Evidence for Knobe's Moral Pervasiveness Model

Knobe's model is supported by previously reviewed evidence for (indirect) effects of negativity on blame that are mediated by causal-mental judgments. For example, Mazzocco et al. (2004) showed a strong outcome → blame effect that was mediated by negligence judgments, consistent with Knobe's claim that negativity enhances culpable mental state judgments (and, thereby, blame).

The most widely known evidence for Knobe's model comes from the "side-effect effect" (Leslie et al., 2006), whereby people view negative side effects as more intentional than positive ones. In the original demonstration of the effect (Knobe, 2003a), a CEO adopted a program that increased profits, with a side effect of harming [helping] the environment. The CEO stated, "I don't care at all about harming [helping] the environment," thereby putatively indicating a lack of desire for the side effect. Most people said that harming the environment was intentional but helping was unintentional, a pattern that has emerged across variations in age and vignette content (Leslie et al., 2006; Cushman and Mele, 2008; Mallon, 2008).

Other evidence shows that morality appears to impact a host of other non-moral judgments. People more often judged that the harming CEO, as compared to the helping CEO, *intended* the outcome (Knobe, 2004; McCann, 2005), *knew* about the outcome (Beebe and Buckwalter, 2010), *decided to* bring about the outcome, and was *in favor of* the outcome (Pettit and Knobe, 2009). Moral judgments also appear to influence assessments of causality (Knobe and Fraser, 2008) and freedom (Phillips and Knobe, 2009) in a similar fashion.

Limitations of Knobe's Model

One challenge to Knobe's account is that the harming and helping CEO scenarios differ not only in moral valence but also in the agent's implied attitude toward that outcome. Since people expect others to prevent negative events and foster positive ones, professed indifference about an outcome constitutes evidence of a welcoming attitude when the outcome is negative but not when it is positive. Adjusting these mismatched attitudes by making the harming CEO less welcoming and the helping CEO more welcoming led people to judge the two actions as equally intentional (Guglielmo and Malle, 2010). Moreover, people rarely said the side effect was intentional once given other options of describing the situation; they instead indicated that the CEO knowingly brought about the outcome, and this pattern was identical for the harming and helping scenarios (Guglielmo and Malle, 2010; Laurent et al., 2015b). These findings challenge the claim that moral judgments impact the "fundamental competencies underlying [people's] concept of intentional action" (Knobe, 2006, p. 221).

Knobe's model does not specify what initial moral judgments actually are and therefore what triggers them. The model requires that these judgments are *not* shaped by causal-mental inferences, since such inferences are themselves posited to be guided by initial moral judgments. By virtue of what, then, do initial moral judgments arise? The clearest possibility is that these "judgments of good and bad" are driven by outcomes or consequences. However, several experimental variations reveal low intentionality ratings despite the presence of a bad outcome, such as when the agent "felt terrible" about the outcome (Phelan and Sarkissian, 2008; see also Cushman and Mele, 2008). Even the paradigmatic side-effect effect requires not just the occurrence of a negative outcome but also the agent's *knowledge* that it will occur; when this knowledge is absent, people rarely judge the outcome intentional (Nadelhoffer, 2006; Pellizzoni et al., 2010). Thus, the initial moral judgments of Knobe's model are sensitive at least to the agent's knowledge and attitude (Guglielmo, 2010), challenging the claim that such moral judgments occur prior to and without consideration of an agent's mental states.

A final challenge is that many of the moral pervasiveness patterns likewise emerge for non-moral norm violations. This is because breaking a norm (moral or otherwise) provides diagnostic evidence of the agent's desires, intentions, and causal role (Jones and Davis, 1965; Harman, 1976; Machery, 2008). For example, people judged it more intentional to break rather than conform to a dress code (Guglielmo and Malle, 2010), or to make unconventionally rather than conventionally colored toys (Uttich and Lombrozo, 2010). Puzzlingly, Knobe has sometimes emphasized norm violation in general (Knobe, 2007; Hitchcock and Knobe, 2009), and other times *moral* violation in particular (Pettit and Knobe, 2009; Knobe, 2010). In one striking study that pitted norm violation against morality (Knobe, 2007), the side effect of the CEO's action was either violation of a Nazi law (a good but norm-violating outcome) or conformity to the law (a bad but norm-conforming outcome). People viewed the norm-violating (but good) outcome as intentional far more often (81%) than the norm-conforming (but bad) outcome (30%), demonstrating the supremacy of norm violation over moral concerns.

Summary of Biased Information Models

Biased information models raise the intriguing possibility that causal-mental assessments—which are typically viewed as inputs to moral judgment—are themselves driven by more basic moral judgments. However, the current analysis suggests that this fundamental claim of biased information models is not, at present, well supported. For one, these models have not empirically assessed the operative early moral judgments. Moreover, although negativity impacts non-moral assessments, this pattern can often be accounted for without appealing to motivational or conceptual influences. Norm-violating information provides grounds for related diagnostic inferences. Consequently, the patterns predicted by biased information models emerge even for non-moral norm violations, and the patterns for moral violations become far weaker when controlling for the relevant diagnostic information.

PROCESSING MODELS

The models reviewed so far are concerned primarily with the information components that underlie moral judgments. A distinct set of models—here called processing models—has a different emphasis, instead focusing on the psychological processes that are recruited when people determine whether a behavior is immoral or worthy of blame. Although many possible forms of processing might be examined, the literature has typically examined two putatively competing types: intuitive or emotional processes on the one hand, and deliberative or reason-based processes on the other.

Haidt: Social Intuitionist Model of Moral Judgment

Haidt's (2001, p. 815) Social Intuitionist Model, shown in **Figure 4**, asserts that "moral judgment is caused by quick moral intuitions and is followed (when needed) by slow, ex post facto moral reasoning" (p. 817). This statement contains two distinct claims about the intuitive nature of moral judgment. One is a "negative" claim that reasoning usually does not precede, but rather follows from, moral judgment. This claim, shown in **Figure 4** as the *post hoc reasoning* link, challenges the long tradition of reason-based moral judgment models (Kant, 1785/1959; Piaget, 1932/1965; Kohlberg, 1969; Turiel, 1983). The second, "positive," claim is that intuitions or emotional responses directly cause moral judgments (the *intuitive judgment* link).

The *eliciting situation* element of Haidt's model denotes the kinds of situations that are apt to generate moral intuitions and, therefore, moral judgments. Recent research on these "taste buds" of morality (Haidt and Joseph, 2007) suggests that there are five broad moral domains: *harm*, *fairness*, *ingroup*, *authority*, and *purity* (Graham et al., 2009; Haidt and Graham, 2009; Haidt and Kesebir, 2010). It remains to be seen whether the fundamental links in Haidt's model between intuition, judgment, and reasoning are true

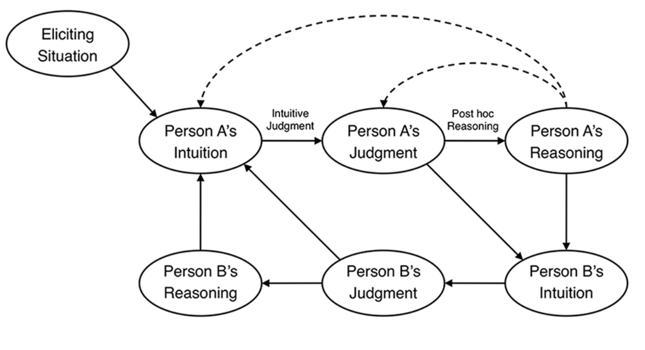

FIGURE 4 | Haidt's Social Intuitionist Model of moral judgment. Reprinted from Haidt (2001) with permission from APA.

for each of these five moral domains; most evidence for the model, as we will see, comes from studies examining purity.

Close inspection reveals that Haidt emphasizes a different type of moral judgment than that examined by information models. Information models assume or stipulate that the moral judgment process begins with the identification of a negative event (e.g., a particular harmful outcome), and thus causal-mental judgments are relevant only insofar as they tie an agent to the event. In contrast, Haidt's model arguably assesses how people determine what constitutes a negative event in the first place. Studies of Haidt's model always hold constant the agent's causal and intentional involvement, so observed differences in moral judgments can be ascribed not to these factors but to whether perceivers viewed the behaviors as negative.

Evidence for Haidt's Social Intuitionist Model

Haidt's (2001) model can be supported by two distinct lines of evidence: one corresponding to the *post hoc* reasoning claim that moral reasoning follows moral judgment, and one to the intuitive judgment claim that intuitive or emotional responses directly guide moral judgments.

Post hoc reasoning

Reasoning processes are sometimes deployed to obtain confirmation for favored conclusions, rather than to discover truth. Kunda (1990) illustrated a host of domains where such motivated reasoning occurs. Strikingly, the vast majority of these domains concern self-relevant judgments—for example, people are inclined to seek, believe, and remember information that depicts themselves as smarter, healthier, and more socially desirable (Kunda, 1990; Mercier and Sperber, 2011). But judgments are typically defined as moral if they have "disinterested elicitors," thus lacking immediate self-relevance (Haidt, 2003). Consequently, to evaluate whether *post hoc* reasoning drives moral judgments, we must consider cases in which the judgments have no direct self-relevance.

In such cases, people's moral judgments can indeed influence subsequent reasoning processes in a motivated manner. When people see an issue in moral terms, they view tradeoffs about the issue as impermissible or taboo (Tetlock, 2003), and their judgments fall prey to various framing effects (Ritov and Baron, 1999; Sunstein, 2005; but see Connolly and Reb, 2003; Tanner and Medin, 2004). Moral judgments can also bias judgments of procedural justice, whereby people view judicial proceedings as more fair to the extent the outcomes are consistent with their own moral views (Skitka and Houston, 2001; Skitka, 2002). In general, these studies illustrate that motivated reasoning can work in the service of moral judgments, buttressing judgments that perceivers have already made. But the critical claim of Haidt's model involves the process of arriving at moral judgments *themselves*.

Perhaps the most compelling method of evaluating Haidt's claim that reasoning follows moral judgments is to jointly probe these judgments and the supporting reasons that people provide for them. Using this method, studies have shown that people sometimes judge behaviors wrong but seemingly cannot provide justificatory reasons, illustrating a phenomenon dubbed "moral dumbfounding" (Haidt et al., unpublished). Participants in Haidt et al.'s (unpublished) study read stories designed to depict disgusting yet harmless actions (e.g., consensual incest; eating a disease-free human cadaver), and although many people judged the actions to be wrong, they sometimes directly stated that they could not explain why. Haidt and Hersh (2001) reported similar results for harmless sexual behaviors (homosexual sex, unusual masturbation, and incest). Participants assigned a moderate amount of moral condemnation, and dumbfounding was observed among 49% of conservatives (and 33% of liberals).

Cushman et al. (2006) adopted a related approach by assessing people's justifications for three principles—the *action*, *intention*, and *contact* principles[5]—that underlie their moral judgments. Participants had to justify patterns of judgments that conformed to the principles (e.g., that an action story was morally worse than an omission story), whereby sufficient justifications cited "a factual difference between the two cases and either claimed or implied that it was the basis of his or her judgments." (Cushman et al., 2006, p. 1084). Any other justifications can be considered insufficient and are akin to instances of moral dumbfounding. Cushman et al. (2006) reported a sizeable proportion of dumbfounding for the intention principle (68%), but dumbfounding was less prevalent for the contact (40%) and action principles (19%).

Intuitive judgment

The second key claim of Haidt's model is that intuitions or emotions directly influence moral judgments. Participants in Haidt et al. (1993) read stories describing harmless actions that were disgusting (e.g., having sex with a dead chicken, then cooking and eating it) or disrespectful (e.g., cleaning the bathroom with a cut up national flag), and their reported negative affect better predicted their moral judgments than did their judgments of harm. Similarly, Haidt and Hersh (2001) and Haidt et al. (unpublished) showed that wrongness judgments were better predicted by "gut feelings" or negative affect than by harm judgments.

Wheatley and Haidt (2005) hypnotized participants to feel disgusted by certain key words, then had them read moral violations, half of which contained the hypnotic disgust word. Participants rated the actions as more morally wrong when the hypnotic disgust word was present. Schnall et al. (2008) report similar findings: participants who were induced to feel disgusted (e.g., with a fart spray or disgusting film clip) made more severe moral judgments than control participants, although this was true only for people highly conscious of their own physical sensations. Eskine et al. (2011) found that people judged behaviors as more morally wrong after first drinking a bitter beverage, as opposed to a sweet or neutral one (but this pattern obtained only among conservative participants).

[5] The action principle holds that harm caused by action is worse than harm caused by omission; the intention principle holds that intended harm is worse than harm brought about as a side-effect; the contact principle holds that harm caused by physical contact is worse than harm caused without physical contact.

Limitations of Haidt's Model

The evidence for Haidt's model may not be widely generalizable to many types of moral violations or intuitions. Although Haidt's definition of intuition appeals to a broad evaluative distinction between good and bad, most attempts to manipulate affective-based intuitions have focused on disgust specifically (Wheatley and Haidt, 2005; Schnall et al., 2008; Eskine et al., 2011). Similarly, most scenarios used to test Haidt's model have involved disgust-based violations (e.g., Haidt et al., 1993; Haidt and Hersh, 2001; Haidt et al., unpublished). Widespread focus on disgust may overstate the role of intuition and the presence of dumbfounding. Disgust is elicited by the mere occurrence of a norm violation, whereas other moral emotions—such as anger—respond to the agent's intentions (Russell and Giner-Sorolla, 2011a). People thus have difficulty justifying feelings of disgust but not feelings of anger (Russell and Giner-Sorolla, 2011b), suggesting that moral dumbfounding may be far less prevalent in non-purity domains. Indeed, when examining harmful behaviors, Cushman et al. (2006) observed dumbfounding in a majority of participants for just one of three moral judgment principles (the other dumbfounding rates were as low as 19%).

Even when focusing specifically on disgust, recent evidence provides a strong challenge to the claim that moral judgment is driven primarily by intuition or emotion. In a meta-analysis of over 50 studies, Landy and Goodwin (2015) report that the effect of induced disgust on moral judgment is verifiable but small ($d = 0.11$), and it disappears once correcting for publication bias. Furthermore, the paradigmatic cases of putatively harmless purity violations (e.g., Haidt et al., unpublished) are typically *not* perceived as harmless, which thereby explains people's persistence in deeming them wrong (Gray et al., 2014; Royzman et al., 2015).

Lastly, studies of Haidt's model typically ask participants whether behaviors are *wrong*, rather than *morally wrong* (Haidt et al., 1993; Haidt and Hersh, 2001; Schnall et al., 2008; Haidt et al., unpublished). These judgments, however, are distinct. Inbar et al. (2009) found that 45% of people said there was something "wrong with couples French kissing in public," which likely reflects not judgments of *moral* wrongness, but rather judgments of social-conventional violation.[6] Consistent with this suggestion that wrongness may not always have moral connotations, people are more willing to describe negative actions as "wrong" than as "morally wrong" (O'Hara et al., 2010). Importantly, this is not true for blame: people did not differentially describe negative actions as "blameworthy" versus "morally blameworthy" (O'Hara et al., 2010). Whereas blame has unambiguously moral connotations, wrongness does not.

Greene: Dual Process Model of Moral Judgment

Greene's (2007, 2013) model asserts that moral judgments are driven not just by intuitive/emotional processes but also by conscious reasoning processes. This dual process distinction has been proposed as a domain-general account of human cognition (Epstein, 1994; Sloman, 1996; Slovic et al., 2004; but see Keren and Schul, 2009). Critically, Greene's (2007) model , shown in **Figure 5**, posits that these two processes underlie different types of moral judgment: "deontological judgments, judgments that are naturally regarded as reflecting concerns for rights and duties, are driven primarily by intuitive emotional responses," whereas "consequentialist judgments, judgments aimed at promoting the greater good, are supported by controlled cognitive processes that look more like moral reasoning" (Paxton and Greene, 2010, p. 513).[7]

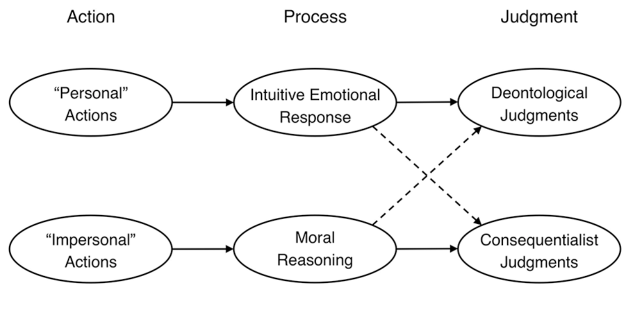

FIGURE 5 | Greene's dual-process model of moral judgment.

Evidence for Greene's Dual-process Model

Greene's model was inspired by a pair of moral dilemmas in which a runaway trolley is on course to kill five innocent workers. In the switch scenario, the hypothetical intervention is flipping a switch to divert the trolley onto a side track, killing a single worker tied to the tracks. In the footbridge scenario, the intervention is pushing a large man over a footbridge, stopping the trolley, and killing the man. Although both actions save five people and kill one, most people deem the switch intervention to be permissible and thus consistent with consequentialism but the footbridge intervention to be impermissible and thus inconsistent with consequentialism (Foot, 1967; Thomson, 1985; Petrinovich et al., 1993; Greene et al., 2001; Hauser et al., 2007). The explanation, according to Greene's (2007, p. 43) model, is that "people tend toward consequentialism in the case in which the emotional response is low and tend toward deontology in the case in which the emotional response is high."

Initial evidence for this model came from a seminal fMRI study by Greene et al. (2001) that compared "personal" dilemmas like *footbridge*, wherein the action involved direct bodily harm, to "impersonal" dilemmas like *switch*. Brain regions associated with emotional processing exhibited greater activation for personal than impersonal dilemmas, whereas regions associated with working memory showed greater activation for impersonal than personal dilemmas. People also took longer to judge personal actions appropriate than inappropriate, suggesting that it takes additional time to override the dominant emotionally aversive response.

[6] Moreover, the question wording in this and other studies ("Is there anything wrong with. . .?") sets a low threshold for assent and may thus elicit artificially high endorsement.

[7] The model also posits that the emotion → consequentialism connection and the reasoning → deontology connection—depicted in **Figure 5** as dashed lines—are possible but rare.

If emotion underlies deontological judgments specifically, then counteracting people's negative emotional responses should increase the acceptability of personal actions. Indeed, participants judged the *footbridge* action (but not the *switch* action) to be more appropriate after watching a funny video (Valdesolo and DeSteno, 2006). Patients with damage to the VMPFC, which is critical for healthy emotional functioning, have dulled physiological responses when considering harmful actions (Moretto et al., 2010) and are therefore more likely than controls to judge personal actions appropriate (Ciaramelli et al., 2007; Koenigs et al., 2007). In contrast, control participants show strong emotional aversion to engaging even in simulated harmful behavior, which predicts their rejection of hypothetical personal actions (Cushman et al., 2012).

If conscious reasoning underlies consequentialist judgments specifically, then taxing people's cognitive processing capacities should impact these judgments. Consistent with this prediction, Greene et al. (2008) showed that whereas the frequency and speed of deontological judgments were unchanged by cognitive load, consequentialist judgments were slower with cognitive load than without. Relatedly, Conway and Gawronski (2013) found that cognitive load selectively weakened consequentialist (but not deontological) judgments. These findings are difficult to explain on Haidt's model—if judgments are driven by immediate intuitive responses, then cognitive load should not affect the speed or content of these judgments. Moreover, participants in Greene et al.'s (2008) study made consequentialist judgments about personal actions 60% of the time, which also presents a challenge for Haidt's model, as it suggests substantial deliberative reasoning despite the highly emotional content of these personal actions.

Limitations of Greene's Model

Greene's model may overstate the role of emotion in moral judgment by often probing first-person judgments (e.g., "Is it appropriate for you to. . ."), rather than third-person judgments. People respond more deontologically when considering their own actions (Nadelhoffer and Feltz, 2008). Thus, heightened emotional responses may be driven partially by the personal implications of the action (e.g., possible punishment, impression management), rather than purely by features of the action itself (cf. Mikhail, 2008). In fact, Borg et al. (2006) have noted that several brain regions implicated by Greene et al. (2001, 2004) are likewise involved in self-referential processing.

Further, the personal/impersonal distinction is coarse and perhaps inaccurate (Mikhail, 2007; McGuire et al., 2009), as it is not clear which features differentiate these categories, nor whether people consistently respond to them in the predicted fashion. McGuire et al. (2009) reanalyzed Greene et al.'s (2001) response time findings and showed that the differences were driven by a small subset of outlier personal dilemmas, which were uniformly (and quickly) judged inappropriate. Greene (2009) now agrees that the criteria distinguishing personal from impersonal actions are inadequate but notes that the veracity of the dual-process model does not depend on this. The model's key claim (Greene, 2009) is that emotional and deliberative processes lead, respectively, to deontological and consequentialist judgments, however, these processes are elicited initially. This is true, but it seems to weaken Greene's model, as it cannot predict the differential elicitation of these distinct processes.

A final challenge regards the utility of the distinction between deontological and consequentialist judgments. Recent evidence indicates that the supposedly consequentialist judgments revealed by classic moral dilemmas are more closely linked to egoist concerns than to concerns about the greater good (Kahane et al., 2015; see also Bartels and Pizarro, 2011). These findings add to a growing concern that moral dilemma scenarios may fail to adequately capture everyday moral judgment (Bloom, 2011; Bauman et al., 2014).

Summary of Processing Models

Processing models seek to describe the psychological processes that give rise to moral judgments. Haidt argues that intuition alone drives most moral judgments, whereas Greene argues that both intuition and reasoning are critical. We can better understand these discrepant claims and findings by invoking the principles identified by information models. Studies of Haidt's model primarily examine cases in which an agent acts intentionally, without apparent exculpatory justification; the key question for perceivers is therefore whether the act itself was negative. This norm-violation detection is often intuitive, and since the other information elements are held constant—causality and intentionality present, justification absent—no further information processing is required and moral judgments also appear intuitive. In contrast, studies of Greene's model primarily examine cases in which an agent performs an intentional action that is indisputably negative, such as killing an innocent person; the key question for perceivers is therefore whether the action is justified by its positive consequences. These studies show that some actions are more easily justified (e.g., those not involving direct physical harm) and that reasoning often drives this process of considering justifications. Taken together, intuition is prominent when detecting initial norm violations, and conscious reasoning is prominent when weighing these early intuitive responses against potential countervailing considerations. As such, intuition and reasoning are both critical for moral judgment, but their relevance emerges in different ways and at different stages of the judgment process.

INTEGRATION AND CONCLUSION

This article has reviewed dominant models of moral judgment, organizing them in a theoretical framework of information processing that has been widely influential in models of cognitive psychology (Rosch, 1978; Marr, 1982) but neglected in models of morality. This framework aims to specify the information elements that shape moral judgments and the psychological processes that bring these judgments to bear. Information models address the first aim, identifying the particular information features that guide moral judgments (Shaver, 1985; Schlenker et al., 1994; Weiner, 1995; Cushman, 2008). These models examine a variety of moral judgments (e.g., blame, wrongness, responsibility) and emphasize different elements, but they coalesce around several key points. Central to moral judgments are variations in causality, intentionality, and mental states more generally, including beliefs, desires, and reasons or motives. Unintentional negative behaviors often receive substantial moral condemnation, particularly when they are preventable or controllable. Moral judgments are therefore guided by information about both the outcome and the mind.

A related set of models—biased information models—hold that the elements identified by information models are themselves guided by more basic moral judgments (Alicke, 2000; Knobe, 2010), such that negative events lead perceivers to make more culpable causal-mental judgments. Alicke offers a motivational explanation for this pattern: negative events trigger blame-validation, whereby perceivers inflate causal-mental judgments to justify initial negative feelings. Knobe offers a conceptual explanation: negative events fundamentally shape the way that people perceive mental states and causality. But these models face a central unresolved issue. Negative events, as instances of norm violations, often provide diagnostic evidence of an agent's beliefs, motives, and intentions. Future research must therefore clarify whether such effects are attributable to morality per se or to concomitant informational elements.

Processing models specify the psychological processes that generate moral judgments, and existing models have primarily been interested in a dichotomy between intuition and reasoning (Haidt, 2001; Greene, 2007). These models endorse a central, and sometimes exclusive, role of intuition. To some degree, this should be unsurprising, as any judgment can be traced to a first principle that cannot be further justified. Just as people would have difficulty justifying their dislike of the color yellow ("it's just ugly"), they will likewise have difficulty justifying why certain actions—such as committing incest or causing physical harm—constitute moral violations ("they're just wrong") (cf. Mallon and Nichols, 2011). Intuition is therefore prominent in detecting initial norm violations, or determining that *something* bad happened. Moral judgments themselves will also be intuitive when critical information elements concerning intentionality, justifications, and mental states are unambiguous and constant. In contrast, when these elements are equivocal or conflicting (e.g., when there is a potential justification for an initial negative event), moral judgments are additionally reliant on deliberate reasoning.

An Information Processing Model of Moral Judgment

Moral judgments, like any others, fundamentally involve information processing, but existing models have typically examined either the information or the processing aspect of these judgments. A successful integrative model will be one that examines the relevant psychological processes as they relate not merely to eventual moral judgments themselves but to constitutive information elements. The subsequent sections examine how the insights of processing models apply to two distinct elements of information models—norm-violation detection and causal-mental analysis—and then discuss a recent a recent model, the Path Model of Blame (Malle et al., 2014), that adopts an integrative information processing approach.

Processes of Norm-violation Detection

For people to levy a moral judgment, they must first detect that a negative event has occurred—that some norm has been violated. Such norm-violation detection usually occurs quickly and triggers affective or evaluative responses (Ito et al., 1998; Van Berkum et al., 2009). Studies of Haidt's model best exemplify the intuitive nature of this detection, showing that people easily, and sometimes without conscious justifications, classify particular behaviors as instances of moral violations (Haidt, 2001; Haidt and Hersh, 2001).

The critical findings of biased information models further support the intuitive basis of norm-violation detection. These models offer two key claims: first, that people identify negative events rapidly, in the form of spontaneous evaluations (Alicke, 2000) or initial moral judgments (Knobe, 2010); and second, that event negativity directly influences causal-mental judgments. Although the current analysis has challenged the second claim, the first claim is undisputed and strongly supported. This process of identifying an initial negative event or norm violation is also a stated or assumed aspect of information models (Shaver, 1985; Schlenker et al., 1994; Weiner, 1995; Cushman, 2008).

Processes of Causal and Mental Analysis

Identifying a negative event is only the first step en route to a moral judgment. It subsequently triggers an explanatory search for the causes of and reasons for the event (Malle and Knobe, 1997b; Wong and Weiner, 1981); and as several models have demonstrated, moral judgments are shaped by these causal-mental considerations (Cushman, 2008; Guglielmo et al., 2009; Gray et al., 2012), such as whether the event was intentional and what the agent's more specific mental states were (beliefs, reasons, or motives).

The processes used to infer causality and mental states are varied and complex, including covariational and counterfactual reasoning, perspective taking, projection, and stereotyping (Hilton, 1990; Ames, 2004; Sloman et al., 2009; Waytz et al., 2010; Alicke et al., in press). These processes are triggered when the causal and mental features of the event at hand are (partially) ambiguous or conflicting, as in most naturalistic instances of moral judgment. In these cases, deliberative processes will often drive causal-mental analysis and, thus, moral judgments themselves. Studies of Greene's model illustrate this pattern, showing that conscious reasoning substantially guides moral judgment when strong positive justifications conflict with highly negative norm violations. In contrast, when causal-mental features are unambiguous or non-conflicting, as in most studies of Haidt's model, there is little need for deliberate reasoning; norm-violation detection and moral judgment become inseparable and largely intuitive.

Consequently, there is no compelling evidence that moral judgments are inherently either intuitive or deliberative. Which process dominates will depend on the nature and strength of the information regarding causality, intentionality, and mental states; but regardless of which process dominates, this causal-mental information is nonetheless considered (cf. Kruglanski and Gigerenzer, 2011). Ambiguous or conflicting information elicits deliberative processing, as when, for example, evaluating a genuine moral dilemma in which multiple courses of action involve different outcomes, tradeoffs, or motives for acting; unequivocal or non-conflicting information elicits intuitive processing, as when evaluating a single course of action for which someone's motives are clearly specified or easily assumed (Monin et al., 2007). Researchers typically choose the strength and ambiguity of this information in service of particular theoretical perspectives. Subtle linguistic differences can help illustrate the point: "Smith was dead" leaves unspecified a perpetrator's causal and intentional involvement (likely triggering more deliberate analysis of these features), whereas "Smith was murdered" obviates the need for deliberate analysis of causality and intentionality (although it may trigger analysis of the agent's motives). Casting further doubt on attempts to characterize moral judgment as either intuitive or deliberative is the fact that even when judgments appear to be intuitive, this may actually reflect the automatization of prior conscious reasoning (Pizarro and Bloom, 2003; Mallon and Nichols, 2011).

The Path Model of Blame

A recent model, the Path Model of Blame (Malle et al., 2014; see **Figure 6**), adopts an explicit information processing view of moral judgment by considering the distinct processes of norm-violation detection and causal-mental analysis, and by specifying how information acquisition and integration underlie blame judgments. The model asserts that blame is initiated by the detection of a negative event or outcome (personal injury, environmental harm, and so on), which is typically an intuitive process. Perceivers then consider various information components en route to blame, but they do so in a particular processing order, which can manifest via either intuitive or deliberative processing. Perceivers assess the causality of the negative event in question and then, if it was agent-caused, they consider whether it was intentional. From there, blame unfolds via different paths: if the event is perceived to be intentional, perceivers consider the agent's reasons or motives for acting; if perceived to be unintentional, perceivers consider the agent's obligation and capacity to prevent the event.

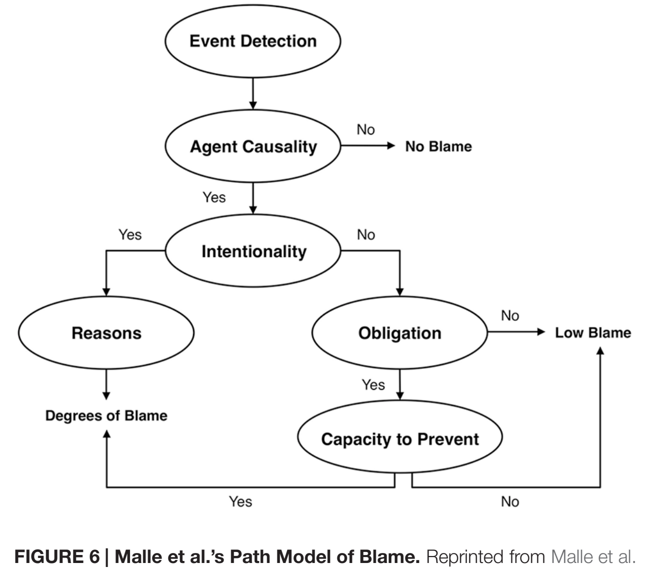

FIGURE 6 | Malle et al.'s Path Model of Blame. Reprinted from Malle et al. (2014) with permission from Taylor and Francis Ltd.

The Path Model has notable similarities with several information models, particularly in recognizing the importance of the specific features of causality (Shaver, 1985; Weiner, 1995; Cushman, 2008), intentionality (Shaver, 1985; Cushman, 2008), reasons (Shaver, 1985), and preventability (Schlenker et al., 1994; Weiner, 1995). Like Cushman's (2008) model, the Path Model also makes explicit that unintentional negative behavior can receive substantial blame. However, the Path Model extends previous models by specifying a processing hierarchy of information features, by identifying separate paths to blame depending on intentionality, and by clarifying how both intuitive and deliberative processes can shape blame. Recent evidence supports the information processing structure of the Path Model. In particular, when people find out about negative events and have an opportunity to acquire additional information, they do so in the order that the model posits, and this holds true even when they face strong time pressure and thus must rely on intuitive processing (Guglielmo and Malle, under review).

THE FUTURE OF MORAL PSYCHOLOGY: DIRECTIONS AND SUGGESTIONS

Conceptualizing moral judgment in a framework of information processing facilitates a synthesis of previous research, helping to clarify the claims of existing models and illustrate their interconnections. Such a framework can likewise help guide future research, particularly by focusing on the affective basis of moral judgment, by diversifying the stimuli and methodologies used to study moral judgment, and by remaining grounded to the descriptive and functional questions of how and why our moral judgments operate as they do, rather than the normative questions of whether they operate correctly.

Affect and Emotion

There is much debate concerning role of emotion in moral judgment. Researchers do not consistently disentangle intuitive judgment from emotion-influenced judgment; and though evidence for the former is relatively strong, evidence for the latter is weaker and has many possible theoretical interpretations (Chapman and Anderson, 2011; Pizarro et al., 2011; Landy and Goodwin, 2015). Emotionally arousing actions are often deemed permissible, and those lacking emotional salience are often judged immoral (Haidt et al., 1993; Greene, 2007; Koenigs et al., 2007). Moreover, even when considering highly emotional stimuli, greater deliberation (Pizarro et al., 2003a; Bartels, 2008) or weaker sensitivity to one's bodily states (Schnall et al., 2008) considerably dulls the effects of emotion on moral judgments. Much additional research is needed—using a wider range of populations, stimulus items, and measures of emotion—before it becomes clear how, and to what extent, emotional mechanisms impact moral judgment (Huebner et al., 2009).

Importantly, any effect of emotion on moral judgment can arise only after causal and mental analysis (cf. Mikhail, 2007). If moral emotions stem from "negative feelings about the *actions or character* of others" (Haidt, 2003, p. 856, emphasis added), then they are predicated upon preceding causal-mental analysis. But negative *affect* may arise prior to such analysis, setting the process of moral judgment in motion. Negative events elicit rapid affective or evaluative responses (Ito et al., 1998; Van Berkum et al., 2009) and trigger processes of explanation and sense-making (Malle and Knobe, 1997b; Wong and Weiner, 1981). Thus, negative affect may lead perceivers to analyze agents' causal and mental contribution, which thereby can elicit specific emotions such as anger (Russell and Giner-Sorolla, 2011a; Laurent et al., 2015c). In this way, negative affect motivates causal-mental analysis, rather than a search for blame-consistent information specifically. Knowing simply *that* a negative event has occurred is not enough for moral judgment (or moral emotion); people need to know *how* it occurred. And to make this determination, they appeal to the causal-mental structure of the event.

This conceptualization, whereby people interpret their negative affect within an explanatory framework prior to experiencing emotion, is consistent with cognitive appraisal theories of emotion (Barrett, 2006a; Barrett et al., 2007). On these accounts, "core affect" arises from the constant valuation of environmental stimuli (e.g., concerning harmfulness or helpfulness) and leads to emotion via the application of a conceptual framework that categorizes and explains the affect (Barrett, 2006a). In the context of moral judgment, causal-mental analysis provides the conceptual framework, appraising negative affect and thus giving rise to emotional experience and moral judgment.[8]

Judgment Timing and Information Search

One domain in which the predictions from various models are decisively testable is that of timing. Many models assume, at least implicitly, that people make certain judgments before others. Both Cushman (2008) and Malle et al. (2014) posit that causality and mental state judgments precede blame. Knobe's (2010) model predicts that initial moral judgments (e.g., about goodness or badness) precede mental state judgments, though the latter may precede full-fledged blame. Alicke's (2000) model suggests that blame (in the form of spontaneous evaluations) should occur prior to judgments about causality and mental states. Testing these predictions about timing can further clarify the way in which moral judgments unfold and can adjudicate between claims made by existing models.

The claims of several models also have implications for perceivers' search for information. Some models imply that, when assessing negative events, perceivers will try to actively acquire information about an agent's causal involvement and mental states, as these most strongly guide blame (Cushman, 2008; Malle et al., 2014). Recent evidence supports such patterns of information seeking behavior (Guglielmo and Malle, under review). Alicke's model, in contrast, might predict that sufficiently negative events will elicit blame and perceivers will rarely seek additional information about mental states (unless they have to justify their blame judgments). Processing models imply that when people are emotionally engaged, they may fail to notice or search for consequentialist information (e.g., how many people will be saved as a result of pushing the man off the footbridge).

Domains, Contexts, and Measurement of Moral Judgment

In addition to attending to the integration of information and processing models, the study of morality will likewise benefit from further diversity and integration. Scholars have long focused on moral domains of harm and fairness, but Haidt (2007, 2008) and Graham et al. (2009, 2011) have emphasized the psychological relevance of various additional domains. Comparisons between moral domains are becoming more prevalent (Horberg et al., 2009; Young and Saxe, 2011; Chakroff and Young, 2015) and may soon yield conclusions about the extent to which existing models are widely, or narrowly, supported across domains.

Although moral judgments are typically studied intrapersonally—as cognitive judgments in the mind of a social perceiver—they undoubtedly serve important interpersonal functions (Haidt, 2001; McCullough et al., 2013; Malle et al., 2014). Moral judgments respond to the presence of social audiences (Kurzban et al., 2007), elicit social distancing from dissimilar others (Skitka et al., 2005), and trigger attempts to modify others' future behavior (Cushman et al., 2009). Given that moral cognition ultimately serves a social regulatory function of guiding and coordinating social behavior (Cushman, 2013; Malle et al., 2014), further forging the connections between intrapersonal moral judgments and their interpersonal manifestations will be a critical direction for future research.

The measurement of moral judgment will also require detailed comparison and integration. Existing models primarily examine a single type of judgment—such as responsibility, wrongness, permissibility, or blame—and although all such judgments of course rely on information processing, they nonetheless differ in important ways (Cushman, 2008; O'Hara et al., 2010; Malle et al., 2014). Wrongness and permissibility judgments typically take intentional actions as their object of judgment (Cushman, 2008). Thus, judging that it is wrong (or impermissible) to X implies that it is wrong to *intentionally* X; it usually makes little sense to say that unintentionally X-ing is wrong. In contrast, responsibility and blame take both intentional and unintentional actions as their object of judgment. Thus, one can be judged responsible (Schlenker et al., 1994) or blameworthy (Cushman, 2008; Young and Saxe, 2009) even for purely unintentional negative behavior. Furthermore, because blame takes into account an agent's reasons for acting, those who commit negative actions for justified reasons—such as self defense (Piazza et al., 2013)—can be

[8] Negative affect itself also requires appraisal—at minimum, that the event in question is negative.

deemed fully responsible yet minimally blameworthy (McGraw, 1987). Since these various moral judgments differ with respect to the amount and type of information they integrate, future work can further differentiate them by assessing both the temporal sequence of these judgments, and their sensitivity to different information features.

Finally, in reflecting the overwhelming preponderance of existing research, this review has focused on negative moral judgments. But what is the information processing structure of positive moral judgments? Relatively few studies have directly compared negative and positive moral judgments, although those that have done so reveal that these judgments are not mere opposites. Consistent with general negativity dominance effects (Baumeister et al., 2001; Rozin and Royzman, 2001), positive moral judgments are less severe than negative ones (Cushman et al., 2009; Goodwin and Darley, 2012), and certain categories of events—including outcomes that are unintended yet foreseen—elicit substantial blame when negative but essentially no praise when positive (Knobe, 2003a; Guglielmo and Malle, 2010). Since perceivers expect, by default, that others will try to foster positive outcomes and prevent negative ones (Pizarro et al., 2003b; Knobe, 2010), earning praise is more difficult than earning blame. Moreover, people often perceive that positive behavior is driven by ulterior motives (Tsang, 2006), which can quickly erode initial positive impressions (Marchand and Vonk, 2005). Thus, whereas positive and negative moral judgments share some information processing features—including sensitivity to intentionality and motives—the former are weaker and less broadly applicable.

Beyond Bias

Claims of people's deviation from normative or rational models of behavior abound in the psychological literature. As Krueger and Funder (2004) have shown, bias is often implied both by pattern X and by pattern not X, leaving it near impossible to discover unbiased behavior. As one example, viewing oneself more favorably than others constitutes a bias (self-enhancement), as does viewing oneself less favorably (self-effacement).

The emphasis on bias, and its supposed ubiquity, similarly exists in the moral judgment literature. Haidt (2001, p. 822) notes that "moral reasoning is not left free to search for truth but is likely to be hired out like a lawyer by various motives," and many theorists appear to agree with this portrayal of biased judgment. The problem, however, is that opposing patterns of judgment are taken as evidence of such bias. The designation "outcome bias" implies that relying on outcome information connotes bias. To avoid biased judgment, perceivers should ignore outcomes and focus on the contents of the agent's mind. In contrast, consequentialist accounts hold that "consequences are the *only* things that ultimately matter" (Greene, 2007, p. 37), which implies that perceivers should substantially—or even exclusively—rely on outcome information.

We have therefore doomed perceivers to be inescapably biased. Whatever judgments they make (e.g., whether using outcome information fully, partially, or not at all), they will violate certain normative standards of moral judgment. It is time, then, to move beyond charges of bias (cf. Bennis et al., 2010; Elqayam and Evans, 2011; Krueger and Funder, 2004). Future research will be more fruitful by focusing not on normative questions of how "good" or "correct" moral judgments are but on descriptive and functional questions: How do moral judgments work? And why do they work this way?

CONCLUSION

This paper advanced an information-processing framework of morality, asserting that moral judgment is best understood by jointly examining the information elements and psychological processes that shape moral judgments. Dominant models were organized in this framework and evaluated on empirical and theoretical grounds. The paper highlighted distinct processes of norm-violation detection and causal-mental analysis, and discussed a recent model—the Path Model of Blame (Malle et al., 2014)—that examines these in an explicit information processing approach. Various suggestions for future research were discussed, including clarifying the roles of affect and emotion, diversifying the stimuli and methodologies used to assess moral judgment, distinguishing between various types of moral judgments, and emphasizing the functional (not normative) basis of morality. By remaining cognizant of the complex and systematic nature of moral judgment, exciting research on this topic will no doubt continue to flourish.

REFERENCES

Alicke, M. D. (1992). Culpable causation. *J. Pers. Soc. Psychol.* 63, 368–378. doi: 10.1037/0022-3514.63.3.368

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychol. Bull.* 126, 556–574. doi: 10.1037/0033-2909.126.4.556

Alicke, M. D., Davis, T. L., and Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Soc. Cogn.* 12, 281–308. doi: 10.1521/soco.1994.12.4.281

Alicke, M. D., Mandel, D. R., Hilton, D. J., Gerstenberg, T., and Lagnado, D. A. (in press). Causal conceptions in social explanation and moral evaluation: a historical tour. *Perspect. Psychol. Sci.*

Alicke, M. D., and Zell, E. (2009). Social attractiveness and blame. *J. Appl. Soc. Psychol.* 39, 2089–2105. doi: 10.1111/j.1559-1816.2009.00517.x

Ames, D. R. (2004). Inside the mind reader's tool kit: projection and stereotyping in mental state inference. *J. Pers. Soc. Psychol.* 87, 340–353. doi: 10.1037/0022-3514.87.3.340

Aristotle (1999/330 BC). *Nicomachean Ethics,* trans. T. Irwin. Indianapolis, IN: Hackett.

Barrett, L. F. (2006a). Solving the emotion paradox: categorization and the experience of emotion. *Pers. Soc. Psychol. Rev.* 10, 20–46. doi: 10.1207/s15327957pspr1001_2

Barrett, L. F. (2006b). Valence is a basic building block of emotional life. *J. Res. Pers.* 40, 35–55. doi: 10.1037/a0024081

Barrett, L. F., Mesquita, B., Ochsner, K. N., and Gross, J. J. (2007). The experience of emotion. *Annu. Rev. Psychol.* 58, 373–403. doi: 10.1146/annurev.psych.58.110405.085709

Bartels, D. M. (2008). Principled moral sentiment and the flexibility of moral judgment and decision making. *Cognition* 108, 381–417. doi: 10.1016/j.cognition.2008.03.001

Bartels, D. M., and Pizarro, D. A. (2011). The mismeasure of morals: antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition* 121, 154–161. doi: 10.1016/j.cognition.2011.05.010

Bauman, C. W., McGraw, A. P., Bartels, D. M., and Warren, C. (2014). Revisiting external validity: concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Soc. Pers. Psychol. Compass* 8, 536–554. doi: 10.1111/spc3.12131

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., and Vohs, K. D. (2001). Bad is stronger than good. *Rev. Gen. Psychol.* 5, 323–370. doi: 10.1037/1089-2680.5.4.323

Beebe, J., and Buckwalter, W. (2010). The epistemic side-effect effect. *Mind Lang.* 5, 474–498. doi: 10.1111/j.1468-0017.2010.01398.x

Bennis, W. M., Medin, D. L., and Bartels, D. M. (2010). The costs and benefits of calculation and moral rules. *Perspect. Psychol. Sci.* 5, 187–202. doi: 10.1177/1745691610362354

Bloom, P. (2011). Family, community, trolley problems, and the crisis in moral psychology. *Yale Rev.* 99, 26–43. doi: 10.1111/j.1467-9736.2011.00701.x

Borg, J. S., Hynes, C., Van Horn, J., Grafton, S., and Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: an fMRI investigation. *J. Cogn. Neurosci.* 18, 803–817. doi: 10.1162/jocn.2006.18.5.803

Boyd, R., Gintis, H., and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science* 328, 617–620. doi: 10.1126/science.1183665

Boyd, R., and Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* 3, 151–227.

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

Chapman, H. A., and Anderson, A. K. (2011). Varieties of moral emotional experience. *Emot. Rev.* 3, 255–257. doi: 10.1177/1754073911402389

Chudek, M., and Henrich, J. (2011). Culture-gene coevolution, norm-psychology and the emergence of human prosociality. *Trends Cogn. Sci.* 15, 218–226. doi: 10.1016/j.tics.2011.03.003

Ciaramelli, E., Muccioli, M., Làdavas, E., and di Pellegrino, G. (2007). Selective deficit in personal moral judgment following damage to ventromedial prefrontal cortex. *Soc. Cogn. Affect. Neurosci.* 2, 84–92. doi: 10.1093/scan/nsm001

Connolly, T., and Reb, J. (2003). Omission bias in vaccination decisions: where's the "omission"? Where's the "bias"? *Organ. Behav. Hum. Decis. Process.* 91, 186–202. doi: 10.1016/S0749-5978(03)00057-8

Conway, P., and Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *J. Pers. Soc. Psychol.* 104, 216–235. doi: 10.1037/a0031021

Critchlow, B. (1985). The blame in the bottle: attributions about drunken behavior. *Pers. Soc. Psychol. Bull.* 11, 258–274. doi: 10.1177/0146167285113003

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

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

Cushman, F., Gray, K., Gaffey, A., and Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion* 12, 2–7. doi: 10.1037/a0025071

Cushman, F., and Mele, A. (2008). "Intentional action: two-and-a-half folk concepts?," in *Experimental Philosophy*, eds J. Knobe and S. Nichols (New York, NY: Oxford University Press), 171–188.

Cushman, F., Young, L., and Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment: testing three principles of harm. *Psychol. Sci.* 17, 1082–1089. doi: 10.1111/j.1467-9280.2006.01834.x

Cushman, F. A. (2013). "The role of learning in punishment, prosociality, and human uniqueness," in *Signaling, Commitment and Emotion: Psychological and Environmental Foundations of Cooperation*, Vol. 2, eds R. Joyce, K. Sterelny, B. Calcott, and B. Fraser (Cambridge, MA: MIT Press).

Darley, J. M., and Shultz, T. R. (1990). Moral rules: their content and acquisition. *Annu. Rev. Psychol.* 41, 525–556. doi: 10.1146/annurev.ps.41.020190.002521

DeScioli, P., and Kurzban, R. (2009). Mysteries of morality. *Cognition* 112, 281–299. doi: 10.1016/j.cognition.2009.05.008

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annu. Rev. Psychol.* 51, 665–697. doi: 10.1146/annurev.psych.51.1.665

Elqayam, S., and Evans, J. S. (2011). Subtracting "ought" from "is": descriptivism versus normativism in the study of human thinking. *Behav. Brain Sci.* 34, 233–248. doi: 10.1017/S0140525X1100001X

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *Am. Psychol.* 49, 709–724. doi: 10.1037/0003-066X.49.8.709

Eskine, K. J., Kacinik, N. A., and Prinz, J. (2011). A bad taste in the mouth: gustatory disgust influences moral judgment. *Psychol. Sci.* 22, 295–299. doi: 10.1177/0956797611398497

Federal Rules of Evidence (2009). *Rule 404b*. Ithaca, NY: Cornell Law School.

Fehr, E., and Gächter, S. (2002). Altruistic punishment in humans. *Nature* 415, 137–140. doi: 10.1038/415137a

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxf. Rev.* 5, 5–15.

Freud, S. (1923/1960). *The Ego and the Id*. New York, NY: Norton.

Goodwin, G. P., and Darley, J. M. (2012). Why are some moral beliefs perceived to be more objective than others? *J. Exp. Soc. Psychol.* 48, 250–256. doi: 10.1016/j.jesp.2011.08.006

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *J. Pers. Soc. Psychol.* 96, 1029–1046. doi: 10.1037/a0015141

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *J. Pers. Soc. Psychol.* 101, 366–385. doi: 10.1037/a0021847

Gray, K., Schein, C., and Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: automatic dyadic completion from sin to suffering. *J. Exp. Psychol. Gen.* 143, 1600–1615. doi: 10.1037/a0036149

Gray, K., and Wegner, D. M. (2008). The sting of intentional pain. *Psychol. Sci.* 19, 1260–1262. doi: 10.1111/j.1467-9280.2008.02208.x

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

Greene, J. (2013). *Moral Tribes: Emotion, Reason, and the Gap between Us and Them*. New York, NY: Penguin Press.

Greene, J. D. (2007). "The secret joke of Kant's soul," in *Moral Psychology: The Neuroscience of Morality: Emotion, Disease, and Development*, Vol. 3, ed. W. Sinnott-Armstrong (Cambridge, MA: MIT Press), 35–79.

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: a reply to McGuire, Langdon, Coltheart, and Mackenzie. *J. Exp. Soc. Psychol.* 45, 581–584. doi: 10.1016/j.jesp.2009.01.003

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition* 107, 1144–1154. doi: 10.1016/j.cognition.2007.11.004

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

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science* 293, 2105–2108. doi: 10.1126/science.1062872

Guglielmo, S. (2010). Questioning the influence of moral judgment. *Behav. Brain Sci.* 33, 338–339. doi: 10.1017/S0140525X10001755

Guglielmo, S., and Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Pers. Soc. Psychol. Bull.* 36, 1635–1647. doi: 10.1177/0146167210386733

Guglielmo, S., Monroe, A. E., and Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry* 52, 449–466. doi: 10.1080/00201740903302600

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

Haidt, J. (2003). "The moral emotions," in *Handbook of Affective Sciences*, eds R. J. Davidson, K. R. Scherer, and H. H. Goldsmith (Oxford: Oxford University Press), 852–870.

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

Haidt, J. (2008). Morality. *Perspect. Psychol. Sci.* 3, 65–72. doi: 10.1111/j.1745-6916.2008.00063.x

Haidt, J., and Graham, J. (2009). "Planet of the Durkheimians, where community, authority, and sacredness are foundations of morality," in *Social and Psychological Bases of Ideology and System Justification*, eds J. Jost, A. C. Kay, and H. Thorisdottir (New York, NY: Oxford), 371–401.

Haidt, J., and Hersh, M. A. (2001). Sexual morality: the cultures and emotions of conservatives and liberals. *J. Appl. Soc. Psychol.* 31, 191–221. doi: 10.1111/j.1559-1816.2001.tb02489.x

Haidt, J., and Joseph, C. (2007). "The moral mind: how five sets of innate intuitions guide the development of many culture-specific virtues, and perhaps even modules," in *The Innate Mind*, Vol. 3, eds P. Carruthers, S. Laurence, and S. Stich (New York, NY: Oxford), 367–391.

Haidt, J., and Kesebir, S. (2010). "Morality," in *Handbook of Social Psychology*, 5th Edn, eds S. Fiske, D. Gilbert, and G. Lindzey (Hoboken, NJ: Wiley), 797–832.

Haidt, J., Koller, S. H., and Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *J. Pers. Soc. Psychol.* 65, 613–628. doi: 10.1037/0022-3514.65.4.613

Hamlin, J. K. (2013). Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition* 128, 451–474. doi: 10.1016/j.cognition.2013.04.004

Hamlin, J. K., Wynn, K., and Bloom, P. (2007). Social evaluation by preverbal infants. *Nature* 450, 557–559. doi: 10.1038/nature06288

Hamlin, J. K., Wynn, K., Bloom, P., and Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proc. Natl. Acad. Sci. U.S.A.* 108, 19931–19936. doi: 10.1073/pnas.1110306108

Harman, G. (1976). Practical reasoning. *Rev. Metaphys.* 29, 431–463.

Harvey, M. D., and Rule, B. G. (1978). Moral evaluations and judgments of responsibility. *Pers. Soc. Psychol. Bull.* 4, 583–588. doi: 10.1177/014616727800400418

Hauser, M. (2006). *Moral Minds: How Nature Designed Our Universal Sense of Right and Wrong*. New York, NY: Harper Collins.

Hauser, M., Cushman, F., Young, L., Jin, R. K.-X., and Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind Lang.* 22, 1–21. doi: 10.1111/j.1468-0017.2006.00297.x

Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York, NY: Wiley.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., et al. (2006). Costly punishment across human societies. *Science* 312, 1767–1770. doi: 10.1126/science.1127333

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychol. Bull.* 107, 65–81. doi: 10.1037/0033-2909.107.1.65

Hitchcock, C., and Knobe, J. (2009). Cause and norm. *J. Philos.* 106, 587–612. doi: 10.5840/jphil20091061128

Horberg, E. J., Oveis, C., Keltner, D., and Cohen, A. B. (2009). Disgust and the moralization of purity. *J. Pers. Soc. Psychol.* 97, 963–976. doi: 10.1037/a0017423

Huebner, B., Dwyer, S., and Hauser, M. (2009). The role of emotion in moral psychology. *Trends Cogn. Sci.* 13, 1–6. doi: 10.1016/j.tics.2008.09.006

Inbar, Y., Pizarro, D. A., and Cushman, F. (2012). Benefiting from misfortune: when harmless actions are judged to be morally blameworthy. *Pers. Soc. Psychol. Bull.* 38, 52–62. doi: 10.1177/0146167211430232

Inbar, Y., Pizarro, D. A., Knobe, J., and Bloom, P. (2009). Disgust sensitivity predicts intuitive disapproval of gays. *Emotion* 9, 435–439. doi: 10.1037/a0015960

Ito, T. A., Larsen, J. T., Smith, N. K., and Cacioppo, J. T. (1998). Negative information weighs more heavily on the brain: the negativity bias in evaluative categorizations. *J. Pers. Soc. Psychol.* 75, 887–900. doi: 10.1037/0022-3514.75.4.887

James, W. (1890/1950). *The Principles of Psychology*, Vol. 2. New York, NY: Dover.

Jones, E. E., and Davis, K. E. (1965). "From acts to dispositions: the attribution process in person perception," in *Advances in Experimental Social Psychology*, ed. L. Berkowitz (New York, NY: Academic Press), 219–266.

Josephs, M., Kushnir, T., Gräfenhain, M., and Rakoczy, H. (2015). Children protest moral and conventional violations more when they believe actions are freely chosen. *J. Exp. Child Psychol.* doi: 10.1016/j.jecp.2015.08.002 [Epub ahead of print].

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., and Savulescu, J. (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition* 134, 193–209. doi: 10.1016/j.cognition.2014.10.005

Kant, I. (1785/1959). *Foundation of the Metaphysics of Morals,* trans. L. W. Beck. Indianapolis, IN: Bobbs-Merrill.

Keren, G., and Schul, Y. (2009). Two is not always better than one: a critical evaluation of two-system theories. *Perspect. Psychol. Sci.* 4, 533–550. doi: 10.1111/j.1745-6924.2009.01164.x

Knobe, J. (2003a). Intentional action and side effects in ordinary language. *Analysis* 63, 190–193. doi: 10.1093/analys/63.3.190

Knobe, J. (2003b). Intentional action in folk psychology: an experimental investigation. *Philos. Psychol.* 16, 309–324. doi: 10.1080/09515080307771

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

Knobe, J. (2006). The concept of intentional action: a case study in the uses of folk psychology. *Philos. Stud.* 130, 203–231. doi: 10.1007/s11098-004-4510-0

Knobe, J. (2007). Reason explanation in folk psychology. *Midwest Stud. Philos.* 31, 90–106. doi: 10.1111/j.1475-4975.2007.00146.x

Knobe, J. (2010). Person as scientist, person as moralist. *Behav. Brain Sci.* 33, 315–329. doi: 10.1017/S0140525X10000907

Knobe, J., and Fraser, B. (2008). "Causal judgment and moral judgment: two experiments," in *Moral Psychology: The Cognitive Science of Morality: Intuition and Diversity*, Vol. 2, ed. W. Sinnott-Armstrong (Cambridge, MA: MIT Press), 441–447.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., et al. (2007). Damage to the prefrontal cortex increases utilitarian moral judgments. *Nature* 446, 908–911. doi: 10.1038/nature05631

Kohlberg, L. (1969). "Stage and sequence: the cognitive-developmental approach to socialization," in *Handbook of Socialization Research and Theory and Research*, ed. D. A. Goslin (Chicago: Rand McNally), 347–480.

Krebs, D. L. (2008). Morality: an evolutionary account. *Perspect. Psychol. Sci.* 3, 149–172. doi: 10.1111/j.1745-6924.2008.00072.x

Krueger, J. I., and Funder, D. C. (2004). Towards a balanced social psychology: causes, consequences, and cures for the problem-seeking approach to social behavior and cognition. *Behav. Brain Sci.* 27, 313–327. doi: 10.1017/S0140525X04000081

Kruglanski, A. W., and Gigerenzer, G. (2011). Intuitive and deliberative judgments are based on common principles. *Psychol. Rev.* 118, 97–109. doi: 10.1037/a0020762

Kunda, Z. (1990). The case for motivated reasoning. *Psychol. Bull.* 108, 480–498. doi: 10.1037/0033-2909.108.3.480

Kurzban, R., DeScioli, P., and O'Brien, E. (2007). Audience effects on moralistic punishment. *Evol. Hum. Behav.* 28, 75–84. doi: 10.1016/j.evolhumbehav.2006.06.001

Lagnado, D. A., and Channon, S. (2008). Judgments of cause and blame: the effects of intentionality and foreseeability. *Cognition* 108, 754–770. doi: 10.1016/j.cognition.2008.06.009

Landy, J. F., and Goodwin, G. P. (2015). Does incidental disgust amplify moral judgment? A meta-analytic review of experimental evidence. *Perspect. Psychol. Sci.* 10, 518–536. doi: 10.1177/1745691615583128

Laurent, S. M., Clark, B. A. M., and Schweitzer, K. A. (2015a). Why side-effect outcomes do not affect intuitions about intentional actions: properly shifting the focus from intentional outcomes back to intentional actions. *J. Pers. Soc. Psychol.* 108, 18–36. doi: 10.1037/pspa0000011

Laurent, S. M., Nuñez, N. L., and Schweitzer, K. A. (2015b). The influence of desire and knowledge on perception of each other and related mental states, and different mechanisms for blame. *J. Exp. Soc. Psychol.* 60, 27–38. doi: 10.1016/j.jesp.2015.04.009

Laurent, S. M., Nuñez, N. L., and Schweitzer, K. A. (2015c). Unintended, but still blameworthy: the roles of awareness, desire, and anger in negligence, restitution, and punishment. *Cogn. Emot.* doi: 10.1080/02699931.2015.1058242 [Epub ahead of print].

Leslie, A. M., Knobe, J., and Cohen, A. (2006). Acting intentionally and the side-effect effect. *Psychol. Sci.* 17, 421–427. doi: 10.1111/j.1467-9280.2006.01722.x

Lewis, A., Bardis, A., Flint, C., Mason, C., Smith, N., Tickle, C., et al. (2012). Drawing the line somewhere: an experimental study of moral compromise. *J. Econ. Psychol.* 33, 718–725. doi: 10.1016/j.joep.2012.01.005

Machery, E. (2008). The folk concept of intentional action: philosophical and experimental issues. *Mind Lang.* 23, 165–189. doi: 10.1111/j.1468-0017.2007.00336.x

Malle, B. F. (2014). "Moral competence in robots?," in *Sociable robots and the Future of Social Relations: Proceedings of Robo-Philosophy 2014*, eds J. Seibt, R. Hakli, and M. Nørskov (Amsterdam: IOS Press), 189–198.

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2012). "Moral, cognitive, and social: the nature of blame," in *Social Thinking and Interpersonal Behavior*, eds J. Forgas, K. Fiedler, and C. Sedikides (Philadelphia, PA: Psychology Press), 311–329.

Malle, B. F., Guglielmo, S., and Monroe, A. E. (2014). A theory of blame. *Psychol. Inq.* 25, 147–186. doi: 10.1080/1047840X.2014.877340

Malle, B. F., and Knobe, J. (1997a). The folk concept of intentionality. *J. Exp. Soc. Psychol.* 33, 101–121. doi: 10.1006/jesp.1996.1314

Malle, B. F., and Knobe, J. (1997b). Which behaviors do people explain? A basic actor-observer asymmetry. *J. Pers. Soc. Psychol.* 72, 288–304.

Mallon, R. (2008). Knobe versus Machery: testing the trade-off hypothesis. *Mind Lang.* 23, 247–255. doi: 10.1111/j.1468-0017.2007.00339.x

Mallon, R., and Nichols, S. (2011). Dual processes and moral rules. *Emot. Rev.* 3, 284–285. doi: 10.1177/1754073911402376

Marchand, M. A. G., and Vonk, R. (2005). The process of becoming suspicious of ulterior motives. *Soc. Cogn.* 23, 242–256. doi: 10.1521/soco.2005.23.3.242

Marr, D. (1982). *Vision*. New York, NY: Freeman.

Mazzocco, P. J., Alicke, M. D., and Davis, T. L. (2004). On the robustness of outcome bias: no constraint by prior culpability. *Basic Appl. Soc. Psychol.* 26, 131–146. doi: 10.1080/01973533.2004.9646401

McCann, H. J. (2005). Intentional action and intending: recent empirical studies. *Philos. Psychol.* 18, 737–748. doi: 10.1080/09515080500355236

McCullough, M. E., Kurzban, R., and Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behav. Brain Sci.* 36, 1–15. doi: 10.1017/S0140525X11002160

McGraw, K. M. (1987). Guilt following transgression: an attribution of responsibility approach. *J. Pers. Soc. Psychol.* 53, 247–256. doi: 10.1037/0022-3514.53.2.247

McGuire, J., Langdon, R., Coltheart, M., and Mackenzie, C. (2009). A reanalysis of the personal/impersonal distinction in moral psychology research. *J. Exp. Soc. Psychol.* 45, 577–580. doi: 10.1016/j.jesp.2009.01.002

Mercier, H., and Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behav. Brain Sci.* 34, 57–74. doi: 10.1017/S0140525X10000968

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends Cogn. Sci.* 11, 143–152. doi: 10.1016/j.tics.2006.12.007

Mikhail, J. (2008). "Moral cognition and computational theory," in *Moral Psychology: The Neuroscience of Morality: Emotion, Disease, and Development*, Vol. 3, ed. W. Sinnott-Armstrong (Cambridge, MA: MIT Press), 81–92.

Monin, B., Pizarro, D. A., and Beer, J. S. (2007). Deciding versus reacting: conceptions of moral judgment and the reason-affect debate. *Rev. Gen. Psychol.* 11, 99–111. doi: 10.1037/1089-2680.11.2.99

Moretto, G., Làdavas, E., Mattioli, F., and di Pellegrino, G. (2010). A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *J. Cogn. Neurosci.* 22, 1888–1899. doi: 10.1162/jocn.2009.21367

Nadelhoffer, T. (2006). Desire, foresight, intentions, and intentional actions: probing folk intuitions. *J. Cogn. Cult.* 6, 133–157. doi: 10.1163/156853706776931259

Nadelhoffer, T., and Feltz, A. (2008). The actor-observer bias and moral intuitions: adding fuel to Sinnott-Armstrong's fire. *Neuroethics* 1, 133–144. doi: 10.1007/s12152-008-9015-7

Nelson-Le Gall, S. A. (1985). Motive-outcome matching and outcome foreseeability: effects on attribution of intentionality and moral judgments. *Dev. Psychol.* 21, 332–337. doi: 10.1037/0012-1649.21.2.332

O'Hara, R. E., Sinnott-Armstrong, W., and Sinnott-Armstrong, N. A. (2010). Wording effects in moral judgments. *Judgm. Decis. Mak.* 5, 547–554.

Ohtsubo, Y. (2007). Perceiver intentionality intensifies blameworthiness of negative behaviors: blame-praise asymmetry in intensification effect. *Jpn. Psychol. Res.* 49, 100–110. doi: 10.1111/j.1468-5884.2007.00337.x

Osgood, C. E., Suci, G. J., and Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.

Paxton, J. M., and Greene, J. D. (2010). Moral reasoning: hints and allegations. *Top. Cogn. Sci.* 2, 511–527. doi: 10.1111/j.1756-8765.2010.01096.x

Pellizzoni, S., Girotto, V., and Surian, L. (2010). Beliefs about moral valence affect intentionality attributions: the case of side effects. *Rev. Philos. Psychol.* 1, 201–209. doi: 10.1007/s13164-009-0008-1

Petrinovich, L., O'Neill, P., and Jorgensen, M. (1993). An empirical study of moral intuitions: toward an evolutionary ethics. *J. Pers. Soc. Psychol.* 64, 467–478. doi: 10.1037/0022-3514.64.3.467

Pettit, D., and Knobe, J. (2009). The pervasive impact of moral judgment. *Mind Lang.* 24, 586–604. doi: 10.1111/j.1468-0017.2009.01375.x

Phelan, M. T., and Sarkissian, H. (2008). The folk strike back; or, why you didn't do it intentionally, though it was bad and you knew it. *Philos. Stud.* 138, 291–298. doi: 10.1007/s11098-006-9047-y

Phillips, J., and Knobe, J. (2009). Moral judgments and intuitions about freedom. *Psychol. Inq.* 20, 30–36. doi: 10.1080/10478400902744279

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

Piazza, J., Russell, P. S., and Sousa, P. (2013). Moral emotions and the envisaging of mitigating circumstances for wrongdoing. *Cogn. Emot.* 27, 707–722. doi: 10.1080/02699931.2012.736859

Pizarro, D., Inbar, Y., and Helion, C. (2011). On disgust and moral judgment. *Emot. Rev.* 3, 267–268. doi: 10.1177/1754073911402394

Pizarro, D. A., and Bloom, P. (2003). The intelligence of the moral intuitions: comment on Haidt (2001). *Psychol. Rev.* 110, 193–196. doi: 10.1037/0033-295X.110.1.193

Pizarro, D. A., Uhlmann, E., and Bloom, P. (2003a). Causal deviance and the attribution of moral responsibility. *J. Exp. Soc. Psychol.* 39, 653–660. doi: 10.1016/S0022-1031(03)00041-6

Pizarro, D., Uhlmann, E., and Salovey, P. (2003b). Asymmetry in judgments of moral blame and praise: the role of perceived metadesires. *Psychol. Sci.* 14, 267–272. doi: 10.1111/1467-9280.03433

Rai, T. S., and Fiske, A. P. (2011). Moral psychology is relationship regulation: moral motives for unity, hierarchy, equality, and proportionality. *Psychol. Rev.* 118, 57–75. doi: 10.1037/a0021867

Reeder, G. D., and Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychol. Rev.* 86, 61–79. doi: 10.1037/0033-295X.86.1.61

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., and Trafimow, D. (2002). Inferences about the morality of an aggressor: the role of perceived motive. *J. Pers. Soc. Psychol.* 83, 789–803. doi: 10.1037/0022-3514.83.4.789

Ritov, I., and Baron, J. (1999). Protected values and omission bias. *Organ. Behav. Hum. Decis. Process.* 79, 79–94. doi: 10.1006/obhd.1999.2839

Robbennolt, J. K. (2000). Outcome severity and judgments of "responsibility": a meta-analytic review. *J. Appl. Soc. Psychol.* 30, 2575–2609. doi: 10.1111/j.1559-1816.2000.tb02451.x

Rosch, E. (1978). "Principles of categorization," in *Cognition and Categorization*, eds E. Rosch and B. B. Lloyd (Hillsdale, NJ: Erlbaum), 27–48.

Royzman, E. B., Kim, K., and Leeman, R. F. (2015). The curious tale of Julie and Mark: unraveling the moral dumbfounding effect. *Judgm. Decis. Mak.* 10, 296–313.

Rozin, P., and Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Pers. Soc. Psychol. Rev.* 5, 296–320. doi: 10.1207/S15327957PSPR0504_2

Russell, P. S., and Giner-Sorolla, R. (2011a). Moral anger, but not moral disgust, responds to intentionality. *Emotion* 11, 233–240. doi: 10.1037/a0022598

Russell, P. S., and Giner-Sorolla, R. (2011b). Social justifications for moral emotions: when reasons for disgust are less elaborated than for anger. *Emotion* 11, 637–646. doi: 10.1037/a0022600

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., and Doherty, K. (1994). The triangle model of responsibility. *Psychol. Rev.* 101, 632–652. doi: 10.1037/0033-295X.101.4.632

Schnall, S., Haidt, J., Clore, G. L., and Jordan, A. H. (2008). Disgust as embodied moral judgment. *Pers. Soc. Psychol. Bull.* 34, 1096–1109. doi: 10.1177/0146167208317771

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York, NY: Springer.

Shaver, K. G. (1996). Too much of a good thing? Commentary on "Searching for order in social motivation." *Psychol. Inq.* 7, 244–247. doi: 10.1207/s15327965pli0703_9

Shaver, K. G., and Drown, D. (1986). On causality, responsibility, and self-blame: a theoretical note. *J. Pers. Soc. Psychol.* 50, 697–702. doi: 10.1037/0022-3514.50.4.697

Shultz, T. R., Schleifer, M., and Altman, I. (1981). Judgments of causation, responsibility, and punishment in cases of harm-doing. *Can. J. Behav. Sci.* 13, 238–253. doi: 10.1037/h0081183

Skitka, L. J. (2002). Do the means always justify the ends, or do the ends sometimes justify the means? A value protection model of justice reasoning. *Pers. Soc. Psychol. Bull.* 28, 588–597. doi: 10.1177/0146167202288003

Skitka, L. J., Bauman, C. W., and Sargis, E. G. (2005). Moral conviction: another contributor to attitude strength or something more? *J. Pers. Soc. Psychol.* 88, 895–917. doi: 10.1037/0022-3514.88.6.895

Skitka, L. J., and Houston, D. A. (2001). When due process is of no consequence: moral mandates and presumed defendant guilt or innocence. *Soc. Justice Res.* 14, 305–326. doi: 10.1023/A:1014372008257

Skowronski, J. J., and Carlston, D. E. (1987). Social judgment and social memory: the role of cue diagnosticity in negativity, positivity, and extremity biases. *J. Pers. Soc. Psychol.* 52, 689–699. doi: 10.1037/0022-3514.52.4.689

Skowronski, J. J., and Carlston, D. E. (1989). Negativity and extremity biases in impression formation: a review of explanations. *Psychol. Bull.* 105, 131–142. doi: 10.1037/0033-2909.105.1.131

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychol. Bull.* 119, 3–22. doi: 10.1037/0033-2909.119.1.3

Sloman, S. A., Barbey, A. K., and Hotaling, J. M. (2009). A causal model theory of the meaning of cause, enable, and prevent. *Cogn. Sci.* 33, 21–50. doi: 10.1111/j.1551-6709.2008.01002.x

Slovic, P., Finucane, M., Peters, E., and MacGregor, D. G. (2004). Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. *Risk Anal.* 24, 1–12. doi: 10.1111/j.0272-4332.2004.00433.x

Suls, J., and Kalle, R. J. (1978). Intention, damage, and age of transgressor as determinants of children's moral judgments. *Child Dev.* 49, 1270–1273. doi: 10.2307/1128777

Sunstein, C. R. (2005). Moral heuristics. *Behav. Brain Sci.* 28, 531–542. doi: 10.1017/S0140525X05000099

Tannenbaum, D., Uhlmann, E. L., and Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *J. Exp. Soc. Psychol.* 47, 1249–1254. doi: 10.1016/j.jesp.2011.05.010

Tanner, C., and Medin, D. L. (2004). Protected values: no omission bias and no framing effects. *Psychon. Bull. Rev.* 11, 185–191. doi: 10.3758/BF03206481

Tetlock, P. E. (2003). Thinking the unthinkable: sacred values and taboo cognitions. *Trends Cogn. Sci.* 7, 320–324. doi: 10.1016/S1364-6613(03)00135-9

Thomson, J. J. (1985). The trolley problem. *Yale Law J.* 94, 1395–1415. doi: 10.2307/796133

Tooby, J., and Cosmides, L. (2010). "Groups in mind: the coalitional roots of war and morality," in *Human Morality and Sociality: Evolutionary and Comparative Perspectives*, ed. H. Høgh-Olesen (New York, NY: Macmillan), 191–234.

Tsang, J. (2006). The effects of helper intention on gratitude and indebtedness. *Motiv. Emot.* 30, 198–204. doi: 10.1007/s11031-006-9031-z

Turiel, E. (1983). *The Development of Social Knowledge: Morality and Convention*. Cambridge: Cambridge University Press.

Uttich, K., and Lombrozo, T. (2010). Norms inform mental state ascriptions: a rational explanation for the side-effect effect. *Cognition* 116, 87–100. doi: 10.1016/j.cognition.2010.04.003

Valdesolo, P., and DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychol. Sci.* 17, 476–477. doi: 10.1111/j.1467-9280.2006.01731.x

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., and Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychol. Sci.* 20, 1092–1099. doi: 10.1111/j.1467-9280.2009.02411.x

Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics Inf. Technol.* 12, 243–250. doi: 10.1007/s10676-010-9232-8

Waytz, A., Gray, K., Epley, N., and Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends Cogn. Sci.* 14, 383–388. doi: 10.1016/j.tics.2010.05.006

Weiner, B. (1995). *Judgments of Responsibility: A Foundation for a Theory of Social Conduct*. New York, NY: Guilford Press.

Wheatley, T., and Haidt, J. (2005). Hypnotic disgust makes moral judgments more severe. *Psychol. Sci.* 16, 780–784. doi: 10.1111/j.1467-9280.2005.01614.x

Wong, P. T. P., and Weiner, B. (1981). When people ask "why" questions, and the heuristics of attributional search. *J. Pers. Soc. Psychol.* 40, 650–663. doi: 10.1037/0022-3514.40.4.650

Woolfolk, R. L., Doris, J. M., and Darley, J. M. (2006). Identification, situational constraint, and social cognition: studies in the attribution of moral responsibility. *Cognition* 100, 283–301. doi: 10.1016/j.cognition.2005.05.002

Wright, J. C., Cullum, J., and Schwab, N. (2008). The cognitive and affective dimensions of moral conviction: implications for attitudinal and behavioral measures of interpersonal tolerance. *Pers. Soc. Psychol. Bull.* 34, 1461–1476. doi: 10.1177/0146167208322557

Young, L., Nichols, S., and Saxe, R. (2010). Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. *Rev. Philos. Psychol.* 1, 333–349. doi: 10.1007/s13164-010-0027-y

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

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

Zelazo, P. D., Helwig, C. C., and Lau, A. (1996). Intention, act, and outcome in behavioral prediction and moral judgment. *Child Dev.* 67, 2478–2492. doi: 10.2307/1131635

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.