Two paths to blame: Intentionality directs moral information processing along two distinct tracks

Andrew E. Monroe[1]* and Bertram F. Malle[2]
[1]Appalachian State University, [2]Brown University

*Correspondence to:
Andrew E. Monroe
Department of Psychology
Appalachian State University
222 Joyce Lawrence Ln.
Boone, NC 28608, USA
E-mail: monroeae1@appstate.edu

Abstract

There is broad consensus that features such as causality, mental states, and preventability are key inputs to moral judgments of blame. What is not clear is exactly how people process these inputs to arrive at such judgments. Three studies provide evidence that early judgments of whether or not a norm violation is intentional direct information processing along one of two tracks: if the violation is deemed intentional, blame processing relies on information about the agent's reasons for committing the violation; if the violation is deemed unintentional, blame processing relies on information about how preventable the violation was. Owing to these processing commitments, when new information requires perceivers to switch tracks, they must reconfigure their judgments, which results in measurable processing costs indicated by reaction-time delays. These findings offer support for a new theory of moral judgment (the Path Model of Blame) and advance the study of moral cognition as hierarchical information processing.

Keywords: moral judgment; blame; information processing; intentionality; moral updating

People make moral judgments all the time. Even infants detect moral and immoral behaviors quickly and evaluate the person who performs them (Darley & Shultz, 1990; Hamlin, 2013; Hamlin, Wynn, & Bloom, 2007). Research in moral psychology has extensively documented the inputs to moral judgment, such as causality (Lagnado, Gerstenberg, & Zultan, 2013; Pizarro, Uhlmann, & Bloom, 2003; Shaver, 1985), intentionality (Plaks, McNichols, & Fortune, 2009; Sousa & Swiney, 2016), desires and beliefs (Cushman, 2008; Greene et al., 2009; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002) and foreseeability (Lagnado & Channon, 2008; Young & Saxe, 2009). Yet, comparatively little attention has been paid to understanding the process of making moral judgments—how people go from perceiving an immoral outcome (e.g., a bullet-ridden body on the ground) to considering moral information (e.g., Did a person cause this? Did he do it intentionally?) and finally producing a moral judgment (e.g., How much blame does he deserve?).

Existing models that address process questions typically focus on distinctions between quick, affective judgments and slow, cognitive ones (e.g., Alicke, 2000; Greene, 2013; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt, 2001). These models, however, do not spell out exactly how known inputs (e.g., intentionality, reasons, preventability) are related and how the unfolding process integrates them into the final moral judgment.

For example, Greene (2013; Greene et al., 2001), drawing on a long tradition in of philosophical theory, offers one of the best known process models of moral judgment. His dual-process model distinguishes between two types of judgments: consequentialist and deontological. Greene argues that consequentialist judgments are slow to arise, are cognitively intensive, and respond to considerations of outcome and intent (e.g., how many people are saved vs. killed). Conversely, deontological judgments are said to derive their motivation from people's automatic, affective aversion to actions that cause harm (see Cushman, 2015; Cushman, Gray, Gaffey, & Mendes, 2012; Miller, Hannikainen, & Cushman, 2014 regarding evidence for people's automatic aversion to causing harm).

Greene's two-systems model is well equipped to make predictions about judgments of moral permissibility (the typical dependent variable in tests of the model), but the deontological-consequentialist distinction is too abstract to derive predictions for exactly how judgments of *blame* come about. To assign blame (especially *degrees* of blame), people need to process information about causality, intentionality, the agent's reasons, and so on (Cushman, 2008; Lagnado & Channon, 2008; Malle, Guglielmo, & Monroe, 2014).

Which of the two processing systems—fast, affective deontology or slow, deliberative consequentialism—might take in such blame-relevant information? Greene suggests that deliberative judgments capture such processing. But in fact, recent research shows that considerations of blame-relevant information happens for both paths. For example, considerations of intentionality are sometimes fast, and seemingly automatic (Barrett, Todd, Miller, & Blythe, 2005; Malle & Holbrook, 2012). Decety and Cacioppo (2012), suggest that people begin to make intentionality judgments as fast as 62ms post stimulus and that these intentionality judgments predict subsequent activation in the amygdala (often described as subserving emotion processing; Adolphs, 1999; Decety, Michalska, & Kinzler, 2012). Greene et al. (2009) also showed that perceptions of intentionality moderated people's tendency to give deontological responses in trolley-problem dilemmas, and recent neuroimaging studies suggest that intentionality processing may be gating affective response to violations (Liljeholm, Dunne, & O'Doherty, 2014; Yu, Li, & Zhou, 2015). Thus, by itself, the dual-process model is not fine-grained enough to specify how people process particular pieces of information shown to be integral to moral judgments, such as intentionality, motives, or obligations. The Path Model of Blame (Malle et al., 2014) tries to fill this gap by specifying the temporal order of processing critical inputs to blame judgments.

Theoretical Model and Predictions

The Path Model of Blame describes the typical process of making a blame judgment. The specific information components entering those judgments, and their ordering, are grounded in research from developmental psychology (Baldwin, Baird, Saylor, & Clark, 2001; Woodward, 1998) and adult social cognition (Barrett et al., 2005; Heider, 1958; Malle & Holbrook, 2012; Malle & Knobe, 1997; Scholl & Tremoulet, 2000). The model maintains that blame judgments do not rely on one dominant psychological process (e.g., intuition) or two competing processes (e.g., emotion vs. deliberation). Instead, information acquisition and processing can at times be fast and automatic (e.g., hearing the word "intentionally" in a behavior description or seeing a certain movement configuration as a "reach" for something), and at other times they can be effortful and deliberate (e.g., looking for traces of an agent's causal involvement or inferring the agent's specific reasons for the focal action). Thus, the particular *mode* of processing is secondary to *what* is being processed (Kruglanski & Orehek, 2007), and the model postulates that specific kinds of information are processed in predictable (though not inevitable) phases of judgment formation (see Figure 1). In particular, when social perceivers detect a norm-violating event they infer information about who or what caused the event; if they determine that an agent caused the norm-violating event they infer information about whether the agent brought it about intentionally or unintentionally; if they determine the agent brought it about intentionally they infer information about the agent's reasons for bringing it about, and if they determine the agent brought it about unintentionally they infer information about whether the agent could have prevented the norm violating event (capacity to prevent) and should have prevented it (obligation to prevent).

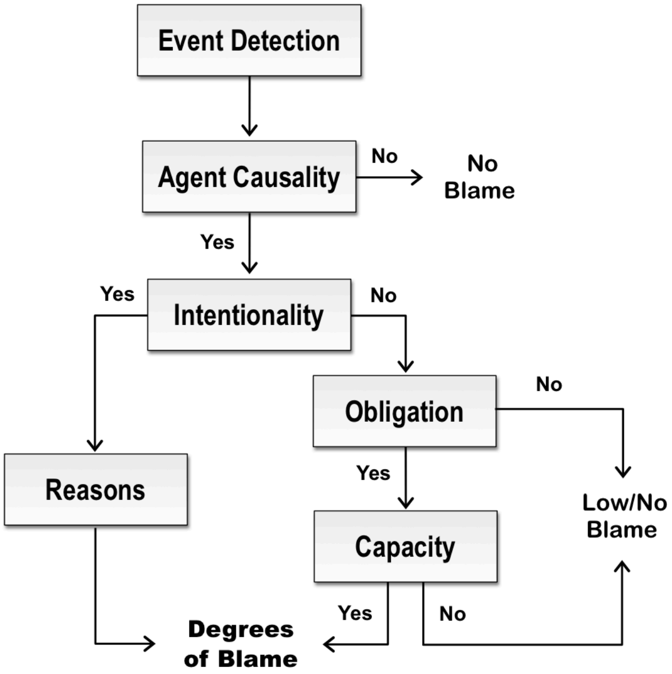

Figure 1. The Path Model of Blame: Concepts and information processing paths. (Reprinted [with permission to be acquired] from Malle et al. (2014), *Taylor & Francis.*

Among the information processed for blame judgments, intentionality (whether the norm-violating event was intentional or unintentional) plays a pivotal role and is therefore the focus of the present investigation.[1] Previous models of blame agree (and empirical evidence confirms) that intentionality amplifies blame. But in addition, the Path Model posits that intentionality *structures* the information processing toward the eventual judgment of blame. Specifically, judgments of intentionality bifurcate information processing into two independent tracks: (1) an intentional track, where blame depends on an agent's reasons for committing the norm violation

[1] The present studies hold constant the information classes of *agent causality* and *obligation to prevent* and manipulate the information classes downstream from intentionality.

(e.g., his goals and beliefs), and (2) an unintentional track, where blame depends on whether the agent had the capacity to counterfactually prevent the norm violation (i.e., could have avoided it). That is, judgments of intentionality configure perceivers' cognitive systems to expect (or seek out) different types of information. When people perceive a behavior as intentional, they expect and more quickly process information about an agent's reasons for acting; when people perceive a behavior as unintentional they expect and more quickly process information about the violations' preventability.

Because of this commitment to one or the other information processing track, switching between tracks is predicted to be cognitively costly. Whenever new information requires switching between tracks, people's cognitive systems must be reconfigured to infer a different type of information, and this step slows information processing (see Gilbert & Shallice, 2002; Monsell, 2003; Wylie & Allport, 2000 for empirical demonstrations of switch costs in non-moral domains).

As stated above, this two-track prediction fills in important informational gaps in dual process models of moral judgment (Greene, 2013; Greene et al., 2001). The Path Model's two-track prediction also differs from nonhierarchical models of moral judgment (e.g., Alicke, 2000). On these models, processing speed is presumably a function of the sufficiency of the information inputs. Every blame judgment is made under uncertainty, and having more information about a norm-violating event (e.g., intentionality, causality, or motives) makes judgments easier (and presumably faster). However, on these models there shouldn't be a processing difference between learning explicit motive information after a likely unintentional event compared with learning the same information after a likely intentional event — in each case, the new information is going to make the judgment easier and faster, as there was no commitment to infer only a particular type of information.

By contrast, the Path Model's prediction draws on research from cognitive psychology on task switching (Allport, Styles, & Hsieh, 1994; Monsell, 2003; Monsell, Yeung, & Azuma, 2000; Rubinstein, Meyer, & Evans, 2001; Wylie & Allport, 2000). Gilbert and Shallice argue that task switching requires "a stagelike executive control process which reconfigures the cognitive system for the upcoming task" (2002, p. 301). Thus, if people do indeed commit to one or the other information processing track toward blame, then when new information requires track-switching, perceivers must engage an executive control process of reconfiguring the system to integrate the new information, slowing the overall judgment. What is involved in this reconfiguration depends on the specifics of the judgment task, which we describe below, followed by more precise predictions.

Experimental Paradigm

We employ a reaction-time paradigm in which people update their initial moral judgments in response to new information. The updating feature is crucial to test the switch-cost hypothesis because all trials should contain new information (and thus require revised processing), but for some of these trials the theory predicts additional processing costs. Participants are presented with an initially sparse event description (e.g., "Henry kicked Fred."). The descriptions were designed to contain the minimal information required to register the event as a moral violation (i.e., an agent, a patient, and harm, see; Gray, Young, & Waytz, 2012), and they were selected so that a majority of people would initially consider the violation as either *likely intentional* or *likely unintentional*, sending the perceiver, as hypothesized, down one track of information processing or the other. After participants make an initial blame judgment, they receive new information that either (a) *matches* the information track implied by the event description (e.g., reason information after a *likely intentional* event) or (b) *switches* information tracks (e.g., *preventability* information after a likely *intentional* event). After receiving this new information, participants are asked to update their initial blame judgment, and we measure the speed with which they update their judgments.

The core prediction of the Path Model is that blame updating is slowed when new information causes perceivers to switch processing tracks compared to when new information matches the processing track participants were already on. More specifically, as soon as perceivers infer intentionality or know it to be present [absent], they consider possible reasons [preventability], settle on a plausible inference, and compute blame. So at the initial stage (after the first sentence), a tentative blame judgment is based on two inferences: one about the event's intentionality and another about specific content: either a plausible reason (if the event is considered intentional) or a plausible level of preventability (if the event is considered unintentional). When new information matches the initial track, at least the (un)intentionality inference is confirmed, and recomputing blame is relatively fast (specific reason/preventability inferences may or may not be confirmed, depending on the particular information condition; see below). When new information belongs to the other track, perceivers must delete two initial inferences (intentionality and the specific content) and process the newly presented content information (e.g., the agent's specific reason) to update their blame judgments. These processes of deletion and reprocessing should slow recomputing blame.

We refer to this prediction as the switch-cost hypothesis and test it in three studies. Study 1 tests this hypothesis with a student sample and text stimuli. Study 2 replicates Study 1 with a community sample and presents all events as audio stimuli, reducing overall reaction times and increasing the precision of processing measurement. Finally, Study 3 tests the alternative hypothesis that switch-costs are due to the surprise effect of track-changing information trials. For all studies, we report all of our manipulations, and each study's sample size was determined prior to data collection.

Study 1

Method

Participants

Participants (*N* = 60) were students recruited from Brown University's subject pool. The sample size for this study was planned prior to beginning data analysis and based on previous reaction time experiments in the lab (e.g., Malle & Holbrook, 2012). Two participants were omitted from the analyses for failing to follow instructions (final *N* = 58). The sample was predominantly female (*N* = 42), and the majority of participants identified as White (57%), with fewer participants identifying as Asian (26%), Black (5%), Latin/Hispanic (2%), or multi-ethnic (7%). The sample had an average age of 19.5 years (*SD* = 1.27).

Procedure

Participants were tested in groups of two to six. After giving informed consent, participants were randomly assigned to individual testing rooms equipped with a desktop computer. The experimenter explained that the task involved reading brief descriptions of behavior on the computer and making judgments using an on-screen click-and-drag slider bar.

Once participants indicated that they understood the task, the experimenter left the room and participants proceeded through a set of on-screen instructions and completed three practice trials. Then they completed 36 experimental trials divided into three blocks of 12, with short breaks between blocks. After completing the experimental trials, participants completed 36 trials of reading-time assessment, a brief demographics questionnaire, and were debriefed.

Materials

Computer task. Each experimental trial consisted of four screens displayed in succession. Participants read a short description of a norm-violating event (screen 1, displayed for three seconds) and made an initial moral judgment ("How much blame does [agent] deserve?"), using a click-and-drag slider bar with endpoints of 0 ("no blame at all") and 100 ("the most blame you would ever give") (screen 2). Immediately afterwards participants read new information about the event and were free to update their initial judgment using the click-and-drag slider bar on the same screen (screen 3). Finally, participants were asked to "write in their own words what happened" (screen 4) as a check of their understanding of the stimulus events. Participants were not allowed to revisit previous judgments or information.

Norm-violating event descriptions. Initial event descriptions were pretested to be either likely *intentional* or likely *unintentional*.[2] For example, "Jim kicked Aaron hard in the shin" was a likely intentional event; "Ted hit a man with his car" was a likely unintentional event. Of the 36 initial event descriptions, 20 were prototypically intentional and 16 were prototypically unintentional (see Appendix A). This split was determined by a clear gap in the bimodal rating distribution of the intentionality pretest.

Information updating. Following their initial moral judgment, participants were presented with one of six possible pieces of new information about the initial norm-violating event (see Appendix B for a complete list of all stimuli). According to the Path Model of Blame (Malle et al., 2014), judgments of intentionality bifurcate the perceiver's processing onto one of two tracks. Intentional events prompt people to search for an agents' (good or bad) reasons, and unintentional events prompt people to search for information about (un)preventability. Accordingly, we constructed six information conditions, three for the intentional track, three for the unintentional track: (1) [that the event was] *intentional* + [the agent had] *morally bad reasons*, (2) *intentional* + *morally good reasons*, and (3) *intentional* (without specified reasons); further, (4) [that the event was] *unintentional* + *preventable*, (5) *unintentional* + *unpreventable*, and (6) *unintentional* (without specified preventability). These six information conditions made the specific content of the new information highly unpredictable and allowed several different pairs of matching/switching trials. For example, for the initial event "Ted hit a man with his car," a participant would read one of the six types of new information as follows:

1) **Intentional-only**: Ted intentionally hit a man with his car.
2) **Intentional + Morally bad reason**: Ted intentionally hit a man with his car because he was in a hurry and did not feel like waiting on the man to cross the street.
3) **Intentional + Morally good reason**: Ted intentionally hit a man with his car because he saw the man had a knife and was chasing a young, frightened woman.
4) **Unintentional-only**: Ted accidentally hit a man with his car.

[2] Prior to Study 1, a group of community members ($N = 40$), recruited from Amazon Mechanical Turk, provided the norming data for the behavior events. Participants were asked: "For the behaviors below, rate whether you think each one is typically intentional or accidental." They responded using a -3 (Definitely accidental) to +3 (Definitely intentional) scale with a midpoint of 0 (Not sure).

5) **Unintentional + Preventable**: Ted accidentally hit a man with his car. Ted didn't check his blind spot before backing up.
6) **Unintentional + Unpreventable**: Ted accidentally hit a man with his car. Ted's brakes failed to work.

New information type was manipulated within-subject across experimental trials (6 replications per type), but for any given stimulus event a participant would see only one of the six possible types. Each stimulus event's six types were distributed across six between-subjects forms and thus presented equally often.

The switching manipulation. As a test of processing costs when people have to switch "tracks," new information about an event either *matched* the implied intentionality of the initial event description (matching trials) or *switched* intentionality (switching trials). For example, a matching trial would be an initial event of "Matt killed Frank" (a likely intentional event) followed by the new information of "Matt intentionally killed Frank because he wanted to collect the insurance money for Frank's death" (Intentional + Morally bad reason). A switching trial would be the same initial event followed by the new information of "Matt accidentally killed Frank. Matt gave Frank expired medicine because he didn't read the expiration label on the box" (Unintentional + Preventable). Each participant completed 18 switching trials and 18 matching trials.

Updating blame judgments. After receiving new information, participants were given an opportunity to update their initial blame judgment. Participants viewed the blame slider bar, with the pointer set at the position of their initial judgment, and had a chance to reposition it if so desired. To ensure that participants did not feel pressured to alter their initial judgments, instructions explicitly stated that they were not required to change their initial judgment. Further, participants were not provided with any instructions regarding having to respond quickly or slowly, and participants were not aware that their responses were being timed. For each trial we recorded the amount of time participants spent updating their judgments (Updating RT).

Reading trials. After participants finished the experimental trials, they read the same 36 descriptions of new information that they had seen in the experimental trials and simply pressed the 'Submit' button as soon as they finished reading each description. We subtracted each of these idiosyncratic reading times from the amount of time participants spent on their earlier updating judgments (their second blame judgment), thus computing reading-time corrected reaction times for blame updating. These corrected scores are used in all subsequent analyses.

Results

Data Preparation

We first trimmed outliers (+/- 2.5 SDs from the trial mean) from the reading-time corrected reaction times (RTs). We also replaced missing RT values when individuals who had valid responses for the vast majority of the design cells would have been omitted entirely from the within-subject ANOVA. We used cell-based sample means to replace 63 missing values within the matrix of 696 potential values (58 participants × 12 cells)[3].

Switch Costs in Moral Updating

[3] These 12 cells were comprised of the six new information conditions × switching vs. matching trials. In all studies, trimmed and missing RT data were replaced with the trial means.

According to the Path Model of Blame, intentionality judgments split information processing into two distinct tracks: an intentional track and an unintentional track. When moral perceivers are on one of these tracks, they expect different types of information to follow (reasons for the intentional track and preventability for the unintentional track), and it is cognitively costly for them to switch between the two (e.g., learning about reasons after initially encountering an unintentional event). To test this switch-cost hypothesis, we compared reaction times across all switching trials to reaction times across all matching trials. A paired-samples t-test confirmed that the total amount of time participants spent updating their blame judgments was longer for switching trials (*M* = 9304 ms, SD = 2074) than for matching trials (*M* = 8869 ms, *SD* = 1848), *t*(57) = 2.37, *p* = .02, *d* = 0.22 [CI = .04, .48].

Follow-up analyses. In addition to the primary test of our hypothesis, we conducted two further tests of the switch-cost hypothesis. In the first analysis we broke down the total updating RT into two stages: a "processing RT" (indexed by the time from the presentation of the judgment screen to participants' first click on the slider to update their judgments) and an "adjustment RT" (the time from participants' first click on the slider to submitting their judgment). One might argue that the processing RT more accurately captures the amount of time necessary for people to "complete the switch" and begin to update their judgments. The total RT also includes processes unrelated to track switching (e.g., judgment fine-tuning), for which the Path Model makes no predictions. Thus by examining each of these two stages we offer a more stringent test of the switch-cost hypothesis.

Examining the processing RT revealed further support for the switch-cost hypothesis. Participants were significantly slower to begin updating their judgments in the switching trials (*M* = 7102 ms, *SD* = 1771) than in the matching trials (*M* = 6508 ms, *SD* = 1576), *t*(57) = 3.59, *p* < .001, *d* = 0.35 [CI = .09, .62]. By contrast, results for the adjustment RT showed that participants did not significantly differ between the switching trials (*M* = 2200 ms, *SD* = 766) and the matching trials (*M* = 2360 ms, *SD* = 748), *t*(57) = 1.69, *p* = .10, *d* = 0.21 [CI = -.47, .05].

Discussion

This study's primary goal was to test whether information processing en route to blame is bifurcated by judgments of intentionality into two tracks in which the perceiver considers different types of blame-relevant information (Malle et al., 2014). One track is for intentional violations (in which case the perceiver considers the agent's reasons for committing the violation); another track is for unintentional violations (in which case the perceiver considers the agent's capacity to prevent the violation). Because, according to the Path Model of Blame, intentionality sets perceivers on a particular track of information search and inference, processing should be slowed when new information requires perceivers to switch tracks. The data confirmed this prediction: across numerous different violations, participants provided slower blame judgments when they had to switch tracks—because new information demanded a reinterpretation of the norm violation with respect to its intentionality and therefore a switch to considering different types of blame-relevant information.

This study has two limitations. First, it relies on a sample drawn from a highly selective student population. Thus, one could argue that the pattern of results stems from participants' willingness to reason more carefully than the general population. Second, whereas the data showed both an overall and an "initial processing" switch cost (how long it took people to read the new information and to initiate their updated blame judgments) they showed no "adjustment" switch cost (how long it took people to settle on their final judgment). Initial processing is arguably the more face-valid of the two processing stages to test a switch cost hypothesis, but the length of this stage (more than 11 seconds when uncorrected for reading time) leaves unclear exactly what processes are unfolding during the integration of initial and new information. One might worry that the information processing demands are actually identical between the switching and matching trials and the observed difference in reaction times is due to people taking slightly more time to read (or re-read) the new information in the switching trials because they find it surprising or counterintuitive. Though we statistically corrected for individual reading times, even more control over this aspect of the cognitive process would be desirable. Study 2 addresses these limitations.

Study 2

Addressing the concern about an elite sample, we recruited a sample made up entirely of community members. In addition, we removed the potential confound of reading processes by converting all of the written stimuli into audio stimuli. The use of audio removes individual differences in reading speed and, more importantly, the concern that presumed switch costs are driven by participants preferentially revisiting information in the switching trials. As in Study 1, the core prediction for this study is that information processing (as measured by updating reaction times) should be slowed when new information demands a switch of processing tracks (e.g., a likely intentional behavior updated with preventability information) compared to when new information matches the processing track participants are already on (e.g., a likely intentional behavior updated with reason information).

Method

Participants

We recruited community participants by advertising a “Paid research study” on Craigslist. Thirty-seven people responded to the ad and participated in the experiment. Prior to the study we had planned to recruit 40 participants (because we expected audio stimulus presentation to reduce RT error variance compared with Study 1); however, sluggish response to recruitment and time constraints led us to conclude data collection at 37 participants. Six participants reported being unfamiliar with using a computer or having difficulty with reading the study instructions and were omitted from the analyses (final *N* =31). Even with these omissions, because of the within-subjects design, we had 551 observations (31 participants x 18 trials) for each of the matching and switching conditions.

The majority of the 31 participants identified as White (74%), with small numbers of participants identifying as Black (3%) Latin/Hispanic (7%) or multi-ethnic (10%). Average age in the sample was 32.3 years (*SD* = 13.1) and represented a diverse range of education.

Procedure and Materials

The procedure was identical to Study 1 except that the initial and new information were presented as audio streams over headphones. The audio stimuli were between two and four seconds long, recorded with neutral affect by a female speaker who had no knowledge of the research hypotheses. After each audio segment, the program immediately advanced to the relevant judgment screen (either initial blame or final blame).

Results

The major goal of this study was to test whether support for the switch cost hypothesis in Study 1 would replicate with audio stimuli and in a community sample. The data confirmed this prediction. Participants were significantly slower updating their moral judgments in the switching trials (*M* = 4325 ms, *SD* = 975) than in the matching trials (*M* = 3871 ms, *SD* = 879), *t*(30) = 3.16, *p* = .004, *d* = .49 [CI = .11 .86]. Additionally, we conducted a follow-up analysis testing the switch cost hypothesis after omitting the "Intentional-only" and the "Unintentional-only" information conditions. This analysis confirmed that participants required significantly more time to update their moral judgments in the switching trials (*M* = 4573 ms, *SD* = 924) compared to the matching trials (*M* = 4006 ms, *SD* = 1183), *t*(30) = 3.25, *p* = .003, *d* = 0.53 [CI = .16 .91].

The results also replicated the processing RT findings (the time to begin adjusting the rating slider). Participants took longer to begin updating their moral judgments for switching trials (*M* = 2910 ms, *SD* = 815) than for matching trials (*M* = 2474 ms, *SD* = 861), *t*(30) = 3.17, *p* = .004, *d* = 0.52 [CI = .14 .89]. However, examining the adjustment RTs for matching trials (*M* = 1398 ms, *SD* = 169) and switching trials (*M* = 1415 ms, *SD* = 292) revealed that participants spent identical amounts of time fine-tuning their judgments after the first click, *t*(30) = 0.35, *p* = .73, *d* = 0.07 [CI = -.42 .28].

Discussion

Study 2 further supported the hypothesis that judgments of intentionality bifurcate moral information processing into two independent tracks and that switching between these tracks has processing costs. Using a community sample, controlling for reading time, and thereby more accurately measuring updating speed, we showed that people were reliably slower at updating their blame judgments when new information demanded a switch of information processing tracks (e.g., unintentional behaviors followed by reason information) than when new information matched the initially established track (e.g., intentional behaviors followed by reasons).

There is a possible concern, however, about the relative level of unexpectedness of matching and switching trials. The moral updating paradigm requires people to integrate old and new information, which involves some degree of expectation formation and violation. This is true for all events, whether switching or matching, because even same-track new information is novel and potentially surprising to participants. However, a skeptic might argue that, regardless of any processes of track switching, matching and switching information contents might differ in mere surprisingness and may thereby account for the switch-cost effect. To address this concern we conducted two follow-up studies to test whether variability of the switch cost effect (the average RT difference between switching and matching trials) can be explained by (1) variability in surprise (the average difference in perceived surprise between switching and matching trials) and (2) by variability in the perceived plausibility of the updating stimuli in the matching versus the switching trials.

Both of the studies (*N* = 36 for each) used a modified updating paradigm, in which participants worked through the same stimulus events as participants in Studies 1 and 2, reading the initial information on one screen followed by the new information on the next screen. However, rather than making blame judgments, participants in the surprise study judged the new information's surprisingness ("Given the first piece of information you read, how surprising is this new information?"), using a click-and-drag slider bar with endpoints 0 (not at all surprising) and 100 (very surprising). Participants in the plausibility study judged the plausibility of the new information ("Given the first piece of information you read, how plausible is this new information?"), also using a click-and-drag slider bar with endpoints 0 (not at all plausible) and 100 (very plausible).

Because each behavior item was presented as a matching trial for some participants and as a switching trial for other participants, we averaged across participants the switching and matching surprise scores for each item and formed 36 switching minus matching difference scores for the surprise judgments (the higher the score, the more surprising the switching version of any given item). Similarly, we formed 36 switching minus matching difference scores for the plausibility judgments (the higher the score the more plausible the switching event). Then, using data from Studies 1 and 2, we computed in the same way 36 average switching and matching RTs and formed a difference score (i.e., a switch-cost score) for each of the 36 items. To test whether the switch cost-effect survived when jointly controlling for the perceived surprisingness and plausibility of the stimuli, we simultaneously regressed switch costs on surprise and plausibility scores. The analysis showed that surprise and plausibility made small but statistically nonsignificant contributions to switch costs, surprise: $t(33) = 1.41$, $p = .17$; plausibility: $t(33) = 1.33$, $p = .19$. The switch costs themselves remained significantly different from zero, $t(33) = 2.45$, $p = .006$.

These data suggest that surprise or plausibility alone cannot explain the switch cost effect. However, the findings so far have relied on a stimulus context in which the ratio of switching to matching trials was 50-50. Switching trials may be generally more surprising merely because they occur in a pool of behaviors in which half of them (the matching trials) confirm an initial intentionality judgment. Study 3 examines this possibility of a mere stimulus context effect by manipulating the base rate of switching trials.

Study 3

We tested the hypothesis that the RT difference between the switching and matching trials was due to a greater general level of unexpectedness (Shannon entropy) of switching trials (i.e., when a switch happens, the system slows processing). To do so we used a between-subjects manipulation of the base rate for switching and matching trials. One condition mirrored the previous two studies' switching-matching base rate (50-50), and one condition made switching trials far more common (and expected) than matching trials (80-20). Therefore, if general unexpectedness explains the RT difference between switching and matching trials, then the effect of slower updating for switching trials should disappear in the 80-20 condition. If, however, the RT difference is due to inherent processing costs when perceivers switch tracks (no matter how typical that may be), then these costs should be present in both the 50-50 and the 80-20 conditions.

Method

Participants

Participants ($N = 40$, 19 females) were students recruited from Brown University's CLPS subject pool. Three participants were omitted from the analyses for failing to follow instructions (final $N = 37$). A plurality of participants identified as White (46%), with fewer participants identifying as Asian (30%), Black (5%), Latin/Hispanic (8%), or multi-ethnic (8%). The sample had an average age of 19.7 years ($SD = 1.43$).

Procedure and Materials

Procedures and materials were identical to Study 2, except that the base rate of switching trials was manipulated between subjects. Participants listened to an initial event description; made a blame judgment; listened to new information, and then had an opportunity to update their initial blame judgment. We measured the amount of time participants spent updating their blame judgments.

In one condition (*N* = 18), the switching-matching rate was 50-50 (18 switching and 18 matching trials), identical to previous studies; in the other condition (*N* = 19), the switching-matching rate was 80-20 (28 switching and 8 matching trials), making switching trials normative and unsurprising. Each cell of the within-subject design had over 150 observations.

Results

We conducted a 2 (base rate: 50-50 vs. 80-20) × 2 (trial type: matching vs. switching) mixed between/within ANOVA to test whether mere surprise explains the RT difference between switching and matching trials, or if the switch-cost effect replicated even when switching trials were common.

Supporting the switch-cost hypothesis, people updated their blame judgments more slowly in switching trials (*M* = 4039 ms, *SD* = 1072) than in matching trials (*M* = 3623 ms, *SD* = 968), *F*(1, 35) = 21.2, *p* < .001, *d* = 0.41 [CI = .07, .74]. There was no main effect of base rate manipulation (*F*[1, 35] = 0.86, *p* = .77, *d* = 0.10 [CI = -.39, .25]) and no base rate by trial type interaction, *F*(1, 35) = 0.76, *p* = .78, *d* = 0.13 [CI = -.34, .27] (see Figure 2). Similarly, the processing RT findings showed that participants took longer to begin updating their moral judgments in the switching trials (*M* = 2928 ms, *SD* = 888) compared to the matching trials (*M* = 2604 ms, *SD* = 826), *F*(1, 35) = 16.5, *p* < .001, *d* = 0.38 [CI = .04, .71]. However, there was no significant effect of the base rate manipulation on processing RTs, (*F*[1, 35] = 0.12, *p* = .73, *d* = 0.11 [CI = -.53, .76]) and no significant base rate by trial type interaction, *F*(1, 35) = 0.39, *p* = .53, *d* = 0.20 [CI = -.44, .85].

Examining the adjustment RTs showed no significant differences between matching trials (*M* = 1046 ms, *SD* = 252) and switching trials (*M* = 1110 ms, *SD* = 320), *F*(1, 35) = 1.72, *p* = .20, *d* = 0.22 [CI = -.10, 55], and no significant effects of the base rate manipulation or the base rate by trial type interaction (*F*s < 1.0, *p*s > .30).

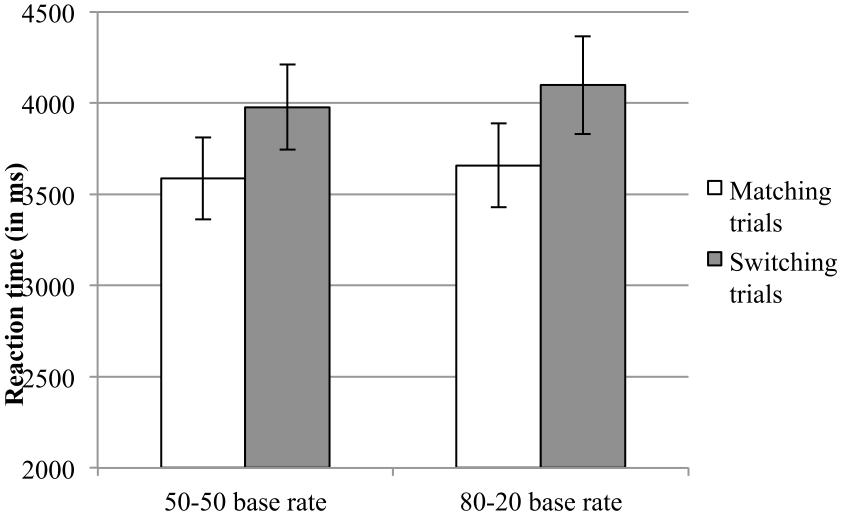

Figure 2. RT data for moral updating across base-rate conditions. Error bars = ±1 SE.

We also examined the possible attenuation of the switching effect. One potential critique of our method is that perhaps it took a number of trials for participants to notice the commonness of switching trials in the 80-20 condition, and only then did switch costs decrease. To test this possibility, we equally divided the experimental trials (in both the 50:50 and the 80:20 condition) into an early, middle, and late phase and tested whether the switching vs. matching RT difference varied across these three phases. We found that overall (across matching and switching trials and across base-rate conditions), participants became significantly faster at updating their moral judgments over the course of the experiment, $F(2,70) = 47.8, p < .0001, \eta^2 = .58$. Critically, however, the switch cost effect was not qualified by any interactions with phase of experiment, Fs < .74 ps > .40. Thus the effect of the switch-cost on reaction times is consistent across phases on the experiment.

Meta Analysis

Three studies offer consistent evidence that intentionality judgments bifurcate moral information processing. However, questions may remain regarding the size of the switch-cost effect and its symmetry across the two possible switches: from intentional to unintentional or from unintentional to intentional. The Path Model of Blame predicts such symmetry, but some authors have claimed a bias for people to see behaviors as intentional (Rosset, 2008), especially when they are negative (Knobe, 2003), and this would suggest a more sizeable switch from intentional to unintentional than the other way around. A meta-analysis of the present three studies would allow us to estimate a more reliable effect size and to examine whether switch costs are symmetric across tracks.

To test the reliability of our findings we derived effect sizes (Cohen's *d*) for comparing switch RTs with match RTs within each of the three studies and computed average effect sizes, using inverse variance weights. In both fixed and random effects models, the average effect size of overall switch costs was $d = 0.32$ (corresponding to 435 ms), 95% CI [0.17; 0.47], $z = 4.2, p < .0001$. No heterogeneity across studies was detected, Q ($df = 2$) = 2.01, $p = .37$. Next we separately analyzed effect sizes for the two types of switches. For intentional-to-unintentional switches, the average effect size was $d = 0.34$, 95% CI [0.17; 0.52], $z = 3.8, p = .0002$; for unintentional-to-intentional switches, the average effect size of switch costs was $d = 0.22$, 95% CI [0.03; 0.40], $z = 2.3, p = .02$. In a homogeneity analysis, these two effects did not differ, Q ($df = 1$) = 0.92, $p = .34$. In fact, in all three studies the reaction times for the two types of switch trials were within 120ms of each other. Thus, switch cost patterns are robust both across studies and across switch types.

General Discussion

Three studies examined the bifurcating role of intentionality in the process of moral judgments of blame. We examined information processing in student samples (Studies 1 and 3) and community samples (Study 2), using both text stimuli (Study 1) and audio stimuli (Studies 2 and 3). The results suggest that early judgments of intentionality split information processing for blame into two tracks: an intentional track where people search for and infer information about an agent's reasons, and an unintentional track where people search for and infer information about preventability. When new information breaks off this initial process and requires people to reconsider intentionality and infer different track-specific information (e.g., reasons rather than preventability), then processing costs become measureable. These processing costs do not arise because the reconsidered new information is any less plausible or more surprising; rather, they reflect the necessity, in such cases, to actually “switch tracks” en route to blame judgments.

Together, the present studies offer evidence that at least some moral judgments are the result of systematic, ordered information processing. This finding fills in some of the details of information processing that prominent models of moral judgment do not address (e.g., Greene, 2013; Haidt, 2001) but others have called for (e.g., Huebner, Dwyer, & Hauser, 2009; Mikhail, 2008). Further, the finding that judgments of intentionality play a key role in moral information processing further adds to recent evidence showing that judgments of intentionality emerge quickly (Decety & Cacioppo, 2012; Malle & Holbrook, 2012) and strongly influence downstream moral and evaluative judgments (Guglielmo & Malle, 2010a, 2010b; Liljeholm et al., 2014; Reeder, Monroe, & Pryor, 2008; Yu et al., 2015).

Additional tests of the Path Model’s bifurcation prediction will of course be necessary to gain confidence in this information processing model. For example, one approach is to examine people’s spontaneous, active search for information en route to making moral judgments, instead of measuring integration of information (as in the current studies). The Path Model’s predictions are straightforward: When people assume a behavior is intentional they should preferentially search for an agent’s reasons for acting, and when they assume a behavior is unintentional they should preferentially search for preventability information (see Guglielmo & Malle, under review).

Another approach would be to measure the accessibility of various kinds of information — intentionality, reasons, and preventability — while people are evaluating norm violations. This could be done by way of lexical decision tasks (relevant information contents would be more quickly identified as genuine words) or Stroop tasks (information contents likely on people’s mind would slow down naming the color in which the corresponding words are displayed). In addition, a probe reaction-time paradigm could assess the likelihood and speed with which people infer the various kinds of information at different stages during moral processing (Malle & Holbrook, 2012).

These variants of the present paradigm may also answer more general questions about the kinds of information processing that underlie blame judgments, such as whether people primarily *infer* information en route to blame or often *assume* such information. To *assume* a piece of information requires only to retrieve it from memory. To *infer* a piece of information requires integrating observed evidence with stored knowledge and selecting the most plausible integration from several possible ones. Inference processes typically take longer and should be more vulnerable to executive function limitations than mere retrieval, so by putting people under time pressure or cognitive load we may be able to identify whether or when the information processing steps toward blame tend to be evidence-based inferences or memory-based assumptions. Moreover, by creating cases in which some people would *assume* certain information (e.g., because they hold a stereotype-based belief about a person) whereas others would *infer* information consistent with the observed evidence we could assess how information processing toward moral judgment is affected by stereotypes about social category memberships (e.g., gender, ethnicity, age).

A path toward integrating affect and cognition in moral judgment

The data reported here show evidence for a detailed information processing prediction of the Path Model of Blame (Malle et al., 2014). But the model also permits the detailed study of affective processes in the emergence of moral judgments—either induced exogenously or measured as an integral part of the judgment process. Whereas prior research on morality and emotion focused on the impact of induced emotion on final judgments of blame or punishment, the Path Model of Blame opens the opportunity to consider exactly how emotion interacts with the information processing elements of moral judgment.

One possibility is that emotion may influence the process of blame by setting default values for various information inputs (e.g., intention, preventability, causality). That is, while the process of moral judgment predicted by the Path Model would remain unaltered—people systematically consider causality, intentionality, reasons, etc.—initial beliefs about these inputs would be biased in ways that intensify or mitigate blame. In support of this possibility, people with clinically elevated levels of anger and aggression tend to view the slights of others as intentional by default (Dodge & Frame, 1982; Dodge, Price, Bachorowski, & Newman, 1990).

Alternatively, emotion may cause people to deviate from the canonical process order for making blame judgments. People may skip over steps or adopt simple heuristic rules (e.g., "if you did it, then you're to blame."). Research on happiness, for example, suggests that inducing feelings of happiness causes people to rely more on stereotypic and heuristic thinking (Bodenhausen, Kramer, & Süsser, 1994; Wegener, Petty, & Smith, 1995). Similarly, emotions associated with feelings of certainty (such as anger or happiness) instill more heuristic thinking (Tiedens & Linton, 2001).

In addition to experimental manipulations of exogenous affect and emotion, we would need other paradigms to reveal the role of affective phenomena in the routine emergence of blame judgments. Detailed multivariate measurement models with fine-grained temporal resolution are needed, combining physiology, implicit and explicit cognitive measures, and facial-bodily expressions, all the way to measures of neural processes with ERP or fMRI methods. That way we might empirically track the relative contribution of affect and cognition — or integrated affective-cognitive information processing— en route to everyday moral judgments.

Applicability to other moral judgments

Another important question for future research is to what extent the switch costs we documented for blame also hold for other moral judgments. Switch costs, according to the Path Model of Blame (Malle et al., 2014), occur because intentionality judgments bifurcate the processing of different kinds of information about *agents* (i.e., their mental states vs. their capacity to prevent negative outcomes). Thus, a potentially useful guide for predicting switch costs is whether the judgment is person-focused or behavior-focused (Uhlmann, Pizarro, & Diermeier, 2015). Person-focused moral judgments (e.g., responsibility, punishment, and ascriptions of moral character) rely on considerations of agents' minds, apply to both intentional and unintentional behaviors, and therefore invite consideration of both reasons (for intentional behaviors) and preventability information (for unintentional behaviors). Indeed, recent research shows that punishment decisions are sensitive to intentionality (Chernyak & Sobel, 2016; Martin & Cushman, 2016), as are considerations of moral responsibility (Plaks et al., 2009; Woolfolk, Doris, & Darley, 2006). As a result, these judgments are candidates for switch costs.

By contrast, behavior-focused judgments (e.g., badness, permissibility, and wrongness) may not be susceptible to switch-costs because perceivers do not as fully consider the agent's mind and intentionality in order to render a judgment. For example badness judgments are quite insensitive to intentionality (Malle, Guglielmo, & Monroe, 2012; Malle et al., 2014), and permissibility and wrongness judgments do not normally apply to unintentional violations such as accidents.[4] At most we might expect wrongness to show one type of switching, namely, when a potentially intentional violation turns out to be unintentional. Suspending one's initial moral wrongness judgment, however, makes the wrongness concept inapplicable rather than requiring consideration of new, other-track information; switch costs are therefore unlikely.

Conclusion

One might consider the switch costs for blame to reflect a cognitive inefficiency. However, we believe they show that people carefully consider intentionality and anticipate the specifically suitable information for intentional violations (i.e., reasons) and unintentional violations (i.e., preventability). Minor switch costs are a small price to pay for evidence-guided information processing. Such information processing is beneficial to social communities because people who carefully process blame are more likely to accurately and appropriately enforce norms violations; and those who commit such violations will experience fairer treatment if they are blamed to appropriate degrees—proportional to the norm violation and the relevant causal-mental factors. Communities would be well served, then, to cultivate evidence-guided blame processing (Voiklis & Malle, in press); and, for the most part, these are the kinds of social communities we live in.

Acknowledgements: This project was partially supported by grants from the Office of Naval Research (ONR), N00014-13-1-0269 and N00014-14-1-0144. The opinions expressed here are our own and do not necessarily reflect the views of ONR. A subset of these data were presented at the Justice and Morality Preconference during the 2015 meeting of the Society for Personality and Social Psychology. We thank the Editor and three anonymous reviewers. Their comments and suggestions greatly improved this manuscript.

References

Adolphs, R. (1999). The human amygdala and emotion. *The Neuroscientist*, *5*, 125–137. doi:10.1177/107385849900500216

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

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilt & M. Moscovitch (Eds.), *Attention and performance XI: Conscious and nonconscious information processing*, Attention and performance series. (pp. 421–452). Cambridge, MA, US: The MIT Press.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants Parse Dynamic Action. *Child Development*, *72*, 708–717. doi:10.2307/1132450

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*, 313–331. doi:10.1016/j.evolhumbehav.2004.08.015

Bodenhausen, G. V., Kramer, G. P., & Süsser, K. (1994). Happiness and stereotypic thinking in social judgment. *Journal of Personality and Social Psychology*, *66*, 621–632. doi:10.1037/0022-3514.66.4.621

[4] A brief search in the *Corpus Of Contemporary American English* showed that in a random sample of 50 uses of "morally wrong," not a single ascription of wrongness referred unambiguously to an unintentional event.

Chernyak, N., & Sobel, D. M. (2016). "But he didn't mean to do it": Preschoolers correct punishments imposed on accidental transgressors. *Cognitive Development*, *39*, 13–20. doi:10.1016/j.cogdev.2016.03.002

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. (2015). From moral concern to moral constraint. *Current Opinion in Behavioral Sciences*, *3*, 58–62. doi:10.1016/j.cobeha.2015.01.006

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

Darley, J. M., & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology*, *41*, 525–556. doi:10.1146/annurev.ps.41.020190.002521

Decety, J., & Cacioppo, S. (2012). The speed of morality: a high-density electrical neuroimaging study. *Journal of Neurophysiology*, *108*, 3068–3072. doi:10.1152/jn.00473.2012

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex*, *22*, 209–220. doi:10.1093/cercor/bhr111

Dodge, K. A., & Frame, C. L. (1982). Social cognitive biases and deficits in aggressive boys. *Child Development*, *53*, 620–635. doi:10.2307/1129373

Dodge, K. A., Price, J. M., Bachorowski, J.-A., & Newman, J. P. (1990). Hostile attributional biases in severely aggressive adolescents. *Journal of Abnormal Psychology*, *99*, 385–392. doi:10.1037/0021-843X.99.4.385

Gilbert, S. J., & Shallice, T. (2002). Task switching: A PDP model. *Cognitive Psychology*, *44*, 297–337. doi:10.1006/cogp.2001.0770

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

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. New York: Penguin Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*, 364–371. doi:10.1016/j.cognition.2009.02.001

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

Guglielmo, S., & Malle, B. F. (under review). *Information-seeking processes in moral judgment.*

Guglielmo, S., & Malle, B. F. (2010a). Enough skill to kill: Intentionality judgments and the moral valence of action. *Cognition*, *117*, 139–150. doi:10.1016/j.cognition.2010.08.002

Guglielmo, S., & Malle, B. F. (2010b). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*, 1635–1647. doi:10.1177/0146167210386733

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

Hamlin, J. K. (2013). Moral Judgment and Action in Preverbal Infants and Toddlers Evidence for an Innate Moral Core. *Current Directions in Psychological Science*, *22*, 186–193. doi:10.1177/0963721412470687

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

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

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

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis*, *63*, 190–194. doi:10.1111/1467-8284.00419

Kruglanski, A. W., & Orehek, E. (2007). Partitioning the domain of social inference: Dual mode and systems models and their alternatives. *Annual Review of Psychology*, *58*, 291–316. doi:10.1146/annurev.psych.58.110405.085629

Lagnado, D. A., & 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

Lagnado, D. A., Gerstenberg, T., & Zultan, R. 'i. (2013). Causal Responsibility and Counterfactuals. *Cognitive Science*, *37*, 1036–1073. doi:10.1111/cogs.12054

Liljeholm, M., Dunne, S., & O'Doherty, J. P. (2014). Anterior insula activity reflects the effects of intentionality on the anticipation of aversive stimulation. *The Journal of Neuroscience*, *34*, 11339–11348. doi:10.1523/JNEUROSCI.1126-14.2014

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In J. P. Forgas, K. Fiedler, & C. Sedikides (Eds.), *Social Thinking and Interpersonal Behavior*, Sydney symposium of social psychology (pp. 313–331). New York, NY US: Psychology Press.

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

Malle, B. F., & Holbrook, J. (2012). Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology*, *102*, 661–684. doi:10.1037/a0026790

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121. doi:dx.doi.org/10.1006/jesp.1996.1314

Martin, J. W., & Cushman, F. (2016). Why we forgive what can't be controlled. *Cognition*, *147*, 133–143. doi:10.1016/j.cognition.2015.11.008

Mikhail, J. (2008). Moral cognition and computational theory. *Moral psychology, Vol. 3: The neuroscience of morality* (pp. 81–92). Cambridge, MA: MIT Press.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion*, *14*, 573–587. doi:10.1037/a0035361

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*, 134–140. doi:10.1016/S1364-6613(03)00028-7

Monsell, S., Yeung, N., & Azuma, R. (2000). Reconfiguration of task-set: Is it easier to switch to the weaker task? *Psychological Research*, *63*, 250.

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

Plaks, J. E., McNichols, N. K., & Fortune, J. L. (2009). Thoughts versus deeds: Distal and proximal intent in lay judgments of moral responsibility. *Personality and Social Psychology Bulletin*, *35*, 1687–1701. doi:10.1177/0146167209345529

Reeder, G. D., Kumar, S., Hesson-McInnis, M. S., & Trafimow, D. (2002). Inferences about the morality of an aggressor: The role of perceived motive. *Journal of Personality and Social Psychology*, *83*, 789–803.

Reeder, G. D., Monroe, A. E., & Pryor, J. B. (2008). Impressions of Milgram's obedient teachers: Situational cues inform inferences about motives and traits. *Journal of Personality and Social Psychology*, *95*, 1–17. doi:10.1037/0022-3514.95.1.1

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, *108*, 771–780. doi:10.1016/j.cognition.2008.07.001

Rubinstein, J. S., Meyer, D. E., & Evans, J. E. (2001). Executive control of cognitive processes in task switching. *Journal of experimental psychology. Human perception and performance*, *27*, 763–797.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*, 299–309. doi:10.1016/S1364-6613(00)01506-0

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness.* New York, NY US: Springer Verlag.

Sousa, P., & Swiney, L. (2016). Intentionality, morality, and the incest taboo in Madagascar. *Cultural Psychology*, *7*, 494. doi:10.3389/fpsyg.2016.00494

Tiedens, L. Z., & Linton, S. (2001). Judgment under emotional certainty and uncertainty: The effects of specific emotions on information processing. *Journal of Personality and Social Psychology*, *81*, 973–988. doi:10.1037/0022-3514.81.6.973

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*, 72–81. doi:10.1177/1745691614556679

Voiklis, J., and Malle, B. F. (in press). Moral cognition and its basis in social cognition and social regulation. In K. Grey and J. Graham (Eds.), *Atlas of Moral Psychology.* New York, NY: Guilford.

Wegener, D. T., Petty, R. E., & Smith, S. M. (1995). Positive mood can increase or decrease message scrutiny: The hedonic contingency view of mood and message processing. *Journal of Personality and Social Psychology*, *69*, 5–15. doi:10.1037/0022-3514.69.1.5

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor’s reach. *Cognition*, *69*, 1–34. doi:10.1016/S0010-0277(98)00058-4

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

Wylie, G., & Allport, A. (2000). Task switching and the measurement of “switch costs.” *Psychological research*, *63*, 212–233.

Young, L., & 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

Yu, H., Li, J., & Zhou, X. (2015). Neural substrates of intention–consequence integration and its impact on reactive punishment in interpersonal transgression. *The Journal of Neuroscience*, *35*, 4917–4925. doi:10.1523/JNEUROSCI.3536-14.2015