Moral Updating

By

Andrew E. Monroe

B.S., Illinois State University, 2005

M.S., Illinois State University, 2007

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in the Department of Cognitive, Linguistic,

and Psychological Sciences at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2013

This dissertation by Andrew E. Monroe is accepted in its present form

By the Department of Cognitive, Linguistic and Psychological Sciences as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Bertram F. Malle, Advisor

Recommended to the Graduate Council

Date_____ _____

Joachim I. Krueger, Reeder

Date_____ _____

Fiery Cushman, Reeder

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

Andrew E. Monroe

CURRICULUM VITAE

Department of Cognitive, Linguistic and Psychological Sciences, Box 1821
Brown University, Providence, RI, 02912
401.863.2544 - Andrew_Monroe@Brown.edu

Education

Brown University, Providence, Rhode Island
Ph.D. Candidate, Department of Cognitive, Linguistic and Psychological Sciences
Expected, May 2013

Illinois State University, Normal, Illinois
Master of Science, Psychology, 2007

Illinois State University, Normal, Illinois
Bachelor of Science, Psychology and Economics, 2005
Summa Cum Laude, Phi Beta Kappa

Awards and Honors

Big Questions in Free Will Grant, Brown University (with Bertram Malle), $270,488, 2011
Brown University Dissertation Fellowship, 2010-2011
Society for Philosophy and Psychology, Poster Prize Semi-Finalist, 2011
Mellon Fellowship: Free Will and Morality, Brown University, $8000, 2010
Graduate Education Committee Travel Award, University of Oregon, $500, 2008
Outstanding Graduate Assistant Award, Illinois State University, $200, 2007

Publications

Monroe, A. E., Malle, B. D., & Dillon, K. D. (*under review*). Is the soul to blame?

Monroe, A. E., Guglielmo, S., & Malle, B. F. (2012). Morality goes beyond mind perception. *Psychological Inquiry, 23,* 179-184.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In J. Forgas, K. Fiedler, and C. Sedikides (Eds.), *Social thinking and interpersonal behaviour* (14th Sydney Symposium of Social Psychology). Philadelphia, PA: Psychology Press.

Pryor, J. B., Reeder, G. D., & Monroe, A. E. (2012). The infection of bad company: Stigma-by-association. *Journal of Personality and Social Psychology, 102,* 224-41.

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

Monroe, A. E. & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology, 9,* 211-224.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52,* 449-466.

Pryor, J.B., Reeder, G.D., Monroe, A. E. & Patel, A (2008). Stigma and pro-social behavior; Are people reluctant to help stigmatized persons? In S. Stürmer & M. Snyder (Eds.), *Psychology of Helping: New Directions in Intergroup Prosocial Behavior* (pp 59-80). London: Blackwell Publishers.

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.

Presentations

Monroe, A. E. Dillon, K. D. & Malle, B. F. (June, 2012) *Free will, the soul, and moral blame.* Talk presented at the Society for Psychology and Philosophy Conference, Boulder, CO.

Monroe, A. E. & Malle, B. F. (November, 2011) *The collapse of free will and morality?* Invited talk. Yale University, New Haven, CT.

Monroe, A. E. (June, 2011). *Mind and morality: Motive inferences guide moral judgments.* Talk presented at the Institute for Prospective Cognition, Illinois State University, Normal, IL.

Monroe, A. E., & Reeder G. D. (March, 2010). *Motive-Matching: Perceptions of Intentionality for Coerced Aggression.* Yale University, New Haven, CT.

Monroe, A. E., Reeder, G. D., & Pryor J. B. (May, 2008). *Perceptions of Coerced Aggressive Acts: The Case for the Savvy Perceiver.* Talk presented at the Midwestern Psychological Association Conference, Chicago, IL.

Monroe, A. E., & Reeder, G. D. (May, 2007). *As Sheep not Wolves: A Re-examination of the Attribution of Obedience within the Milgram Paradigm.* Talk presented at the Midwestern Psychological Association, Chicago, IL.

Teaching Experience

Invited Guest Lectures

Introduction to Psychology, Brown University (Spring 2011; Spring 2012)
Experimental Philosophy, Brown University (Spring 2012)
Social Psychology, Suffolk University (Spring 2012)

Brown University

TEACHING ASSISTANT: *Quantitative Methods in Psychology* (Spring 2010), *Social Psychology*, (Fall 2009),

TRAINING: Sheridan Center Professional Development Seminar, 2010 – 2011, Sheridan Center Teaching Seminar, 2009 – 2010

University of Oregon

INSTRUCTOR: *Research Methods in Psychology* (Summer 2008; Fall 2008), *Social Psychology* (Summer 2008)

TEACHING ASSISTANT: *Cognitive Science* (Spring 2008), *Research Methods in Psychology* (Spring 2008), *Mind and Behavior* (Fall 2007)

Illinois State University

TEACHING ASSISTANT: *Research Methods in Psychology* (2005-2007), *Quantitative Methods in Psychology* (2004-2005)

Acknowledgements

I am grateful for the support of many who have contributed to this research. I would especially like to thank my advisor, Bertram Malle for his guidance and patience throughout the last five years. Your mentorship has refined (and continues to refine) my writing and my approach to conducting scientific research. I would also like to thank my committee members: Joachim Krueger and Fiery Cushman for their encouragement throughout this project as well as their thoughtful comments on the many drafts of the dissertation proposal and manuscript.

Finally, I would like to thank Kristen Goldsmith for so many things – moving across the country with me, being a constant source of laughter in difficult times, pushing me to work harder, and supporting me when I could not work anymore.

Table of Contents

List of Figures

Chapter 1

Moral Updating

Making moral judgments is commonplace in everyday life. From a very young age people can (and do) make moral evaluations of others (Darley & Shultz, 1990). Research in moral psychology has focused on uncovering the inputs to moral judgment and examining how a variety of these are relevant for moral judgments: desire and beliefs (Cushman, 2008; Monroe & Reeder, 2011; Woolfolk, Doris & Darley, 2006; Young & Saxe, 2009), intentions (Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen, 2009), foreknowledge (Guglielmo & Malle, 2010a; Lagnado & Channon, 2008), causality (Pizarro, Uhlmann & Bloom, 2003; Plaks, McNichols & Fortune, 2009), and affective reactions (Haidt, 2007; Inbar & Pizarro, 2009; Pizarro, Inbar & Helion, 2011; Schnall, Haidt, Clore & Jordan, 2008; Wheatley & Haidt, 2005).

For example, imagine that a doctor prescribes a medication to a patient, which, as it turns out, contains an ingredient that the patient is allergic to and because of which the patient becomes sick and dies. In this scenario, people's moral judgments will turn on the doctor's intentions, desires, or if she could have foreseen the fatal consequence of her actions. Examples like this have informed the field of moral psychology for decades. Researchers provide people with various moral outcomes (e.g., physical harm to others) and manipulate the type of information that accompanies the outcomes (e.g., intention to harm, foreknowledge, desire, amount of harm inflicted).

If this typical research strategy parallels the process by which people normally make moral judgments, it makes certain implicit assumptions about what information is

available in the social world. Specifically, it assumes that the information necessary for moral judgments is always immediately available and that it does not change over time. However, such cases capture only a fraction of the universe of moral judgment. Oftentimes, people's moral judgments of blame change in response to new information, but existing research in moral psychology has yet to examine the *process* by which these judgments change and are eventually updated.

To illustrate, consider the recent case of Congressman Weiner, which unfolded over the course of several weeks: A lewd photo of Weiner was leaked onto the Internet. Initially, the explanation for the leaked photo was that a hacker had gained access to his private files and maliciously released the photo. As in the doctor scenario above, one can easily imagine people's moral sentiments toward Congressman Weiner – they might withhold blame entirely (and instead feel sympathy for him) or they might blame him only slightly for not better securing his personal files. However, the story (and the process of moral judgment) does not end there. Later, it was revealed that Weiner himself sent the photo, and moreover, he had sent many similar photos to various women over the past few years, despite being married.

The example highlights the dynamic nature of people's moral judgments. Public approval of Weiner tracked the information that was available at several distinct time points. Initially (when the photo was first leaked) approval of Weiner was impressive – 71% of district voters wanted him to stay in office. Later, with the revelation of all information, approval fell; only 41% of voters said that he should stay in office (CNN, 2011). The current project seeks to understand moral judgments in such contexts, where people form an initial judgment of an agent, encounter new information, and then adjust

their judgment[1] – a process I label *moral updating*.

New information about people and events is becoming available and changing at a pace never before seen in human history. In the age of YouTube, the 24-hour news cycle, Twitter, and Facebook, an understanding of how people make and then update their moral evaluations is important for a complete account of moral judgment. Moral theories that assume the information necessary for moral judgments is always immediately available and does not change over time will be in danger of describing only a limited sphere of the larger moral universe. This work is a first step towards describing not only how people make moral judgments, but how those judgments may change as well.

This research will address three questions pertaining to the phenomenon of moral updating: First, how frequently do people update their moral judgments, and how do updated judgments compare to single moral judgments more commonly studied in the literature? Second, how can we best describe the way people change their blame judgments (if they change at them all)? Do people have a general motivation to blame that causes them to "anchor" on early judgments and make only token changes in response to new information? Or is blaming a process of assigning meaning to an event, during which, people adjust blame in proportion to new information? And third, how do people process new mental state information (e.g., reasons, intentions, preventability)? Is processing new information like reading off a list, where the ability to process one piece of information is independent of all the other items on the list, or is there a structure to

[1] A skeptic of moral updating might argue that what people actually do is make two distinct moral judgments – one at time A with information X, and a second, independent judgment at time B with information Y. Such a proposal is possible; however it appears to run counter to the considerable body of research in judgment and decision making that shows that people use early judgments as informational inputs for later judgments (c.f., Tversky & Kahneman, 1974).

moral judgment in which processing is selectively facilitated or impeded based on preceding information?

Explaining moral updating requires understanding the relevant inputs to moral judgment and how people process new information and integrate it with previous knowledge. The field of moral psychology has extensively studied the first issue. Thus a review of this literature will inform predictions about how blame may change in response to altering the content of those inputs. I consider four moral judgment models prominent in the literature: Weiner's (1995) Responsibility model, Alicke's (2000) Culpable Control model, Haidt's Social Intuitionist model, and the Path Model of Blame (Malle, Guglielmo & Monroe, 2012). I review the models' theoretical assumptions, examine the empirical work supporting them, and explore their application to the question of updating. To the extent that a model claims to describe the phenomenon of moral judgment, one might expect the model to offer some (even oblique) predictions about how people's judgments would respond to new information. Further, in places where the models arrive at differential predictions for updating, this project will test their competing claims.

Review of moral judgment models

Weiner (1995): Attributions of moral responsibility

Weiner (1995) proposed a model in which attributions of responsibility (and blame) depend on two criteria: causality and controllability. According to the model, people render a responsibility judgment by: (1) determining whether an agent (as opposed to the environment) *caused* the outcome, (2) deciding whether the cause of the outcome is *controllable*, and (3) considering mitigating circumstances. Though not explicitly argued, these three steps are logically implied to be sequential steps in a process of moral

judgment (see Figure 1). A perceiver must first resolve that an agent caused a negative outcome before she can determine if the agent was able to control that outcome, and consideration of mitigating circumstances is irrelevant if the agent did not cause or could not control the outcome in the first place.

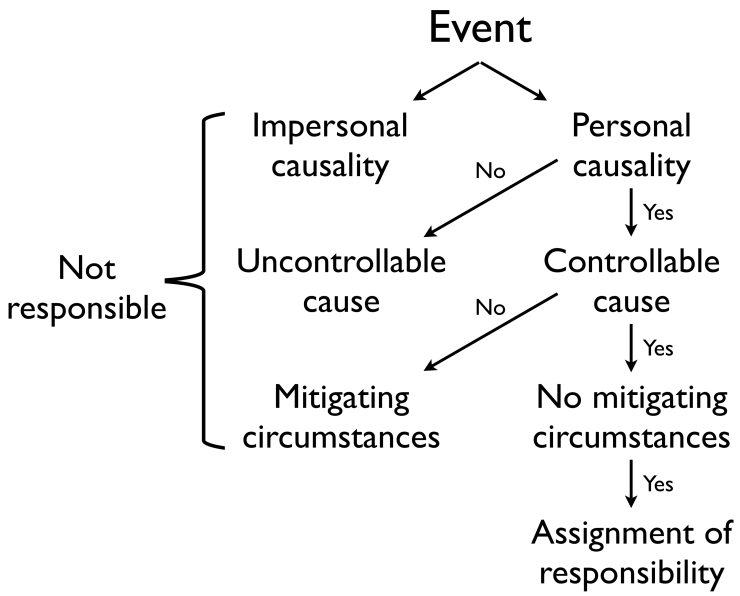

Figure 1. Weiner's model of responsibility.

Blame enters the model as a consequence of a responsibility judgment – inferences of responsibility produce emotional reactions (e.g. anger) and judgments of blame. In the following I focus on the model's predictions for *blame* rather than responsibility. I do this because responsibility is a poorly defined concept and because Weiner repeatedly equates responsibility with blame. Gailey and Falk (2008) showed that the term "responsibility" subsumes several distinct constructs: wrongfulness, causality, foreknowledge, and intentionality . Further, across multiple research programs, responsibility is equated with blame (Shaw & Sulzer, 1964), intentionality (Fincham & Bradbury, 1992), moral responsibility (Fincham & Shultz, 1981), obligation (Hamilton,

1986), and moral eligibility (Oshana, 2001), and it is sometimes unclear which of these definitions Weiner (1995) adopts in his model.

For example, in his discussion of mitigating circumstances, Weiner (1995) appears to equate responsibility and blame. He argues that justificatory reasons (e.g., hitting a car in order to keep it from veering off the road) decrease responsibility judgments. However, justifications typically do not deny responsibility. Instead, they quite clearly indicate that an agent *was* responsible for an outcome (e.g., "Yes I shot the man.") but reduce *blame* by offering socially acceptable reasons for the action (e.g., "He was attacking my children"). Thus, justifications speak to the question of how much *blame* an agent deserves, not whether he was technically responsible for the negative outcome. Therefore, because Weiner does not offer a definition of responsibility, and because his use of justifications (central to questions about moral updating) does not apply to responsibility, I derive the model's predictions concerning judgments of blame.

The role of causality and controllability for blame.

Weiner (1995) proposes that observers progress through three sequential steps on their way to making a moral judgment. The first step in the model is determining whether a negative outcome was personally or impersonally caused; that is, whether the outcome was caused by an agent or by an external, environmental cause (Weiner, 1992). Consistent with attribution researchers (e.g. Gilbert & Malone, 1995; Kelley, 1971; Ross, 1977), in the absence of strong disconfirming evidence (which is presumed to be rare), the model predicts that perceivers will attribute causality to agents, rather than environmental causes. If however, observers decide that the event was caused by something in the environment, rather than by a causal agent, then people exit the process

of responsibility attribution and no responsibility is assigned; otherwise, people progress to the second step of the model: answering the question of controllability.

Controllability in Wiener's model refers to the ability to, intentionally, cause an outcome. For an outcome to be controllable it must be amendable to "volitional alternation" (e.g., an agent has the skill to carry out an action, or knew that the action had certain consequences) (Weiner, 2001, p. 333). For example, an agent is not responsible for accidentally hitting a passerby during an epileptic seizure. Nor is a child responsible for failing a calculus exam, as in these cases the agents lack the ability to either regulate or carry out actions.

Causality and controllability are jointly necessary for people to render a blame judgment. If either is denied, blame for an agent should be severely reduced (Weiner, 1992; 1995; 2003). This makes intuitive sense with regard to causality. If there is no causal agent, no blame is attributed. People do not blame a falling rock for denting their car (though they may still be angry). However, if an agent carelessly (let alone intentionally) dents the car, blame will likely follow.

Similarly, Weiner (1995) argues that people do not assign blame for uncontrollable outcomes. "Internal causes that are not controllable, such as a physical handicap, can be considered as good as external causes in that the actor is not held personally responsible" (Weiner, Amirkhan, Folkes, & Verette, 1987, p. 316). Uncontrollable outcomes are those where the "cause [of the outcome] is located within the person, but cannot be controlled" (Weiner, 1995, p. 7). Uncontrollability tracks the whether an agent could have avoided the outcome. If an outcome was due to a lack of skill, foreknowledge, or (in some cases) self-restraint then, according to the model,

perceivers conclude that the cause of the outcome was uncontrollable and blame should be withheld.

Mitigating circumstances

The last step of the model is to consider potential mitigating circumstances (if any) and to pronounce a moral judgment of the agent. Mitigating information may take many forms – incapacity, infancy, excuses, coercion, lack of skill or foreknowledge – but they are usually theorized to alter perceiver's moral judgments by either denying agent causality (e.g., "I didn't hit the car; it was a rock."), controllability (e.g., "I couldn't stop my car; the brakes were out.") or by providing a justificatory reasons (e.g., "I was trying to prevent you from hitting that child in the road."). In one study Weiner, Perry, and Magnussen, (1988) support this claim by showing that people have very different reactions to an AIDS patient based on whether contracting the diseases was controllable or not. In one condition, the researchers described the patient contracting AIDS due to a promiscuous, and unprotected, sex life. In another condition, the patient is described as contracting AIDS through a routine blood transfusion, where the blood happened to be infected. Participants reported more anger, less willingness to help, and a stronger belief that the patient is responsible for his plight when he contracted the disease in a controllable manner (i.e., leading a promiscuous sex life) compared to when contracting the disease was uncontrollable (i.e., contracting it from a blood transfusion).

In a similar vein, Weiner et al. (1987) tested the effectiveness of various excuses for failing to fulfill a social obligation (e.g., being late to an appointment). In a first study, the researchers asked participants to describe a recent event where they gave an explanation for missing a social appointment (e.g., a party or a date). They found that

people were most likely to offer explanations that were external and uncontrollable (e.g., My car broke down) rather than internal or controllable explanations (e.g., I didn't want to come). A subsequent study manipulated the content of explanations for a confederate's late arrival to a study and measured participants' self-reports of emotional reactions to the confederate and desire to interact with the confederate in the future. The confederate posed as a subject arriving late to the research study and offered either an explanation referring to an uncontrollable cause (e.g., an exam that ran over time), an explanation referring to a controllable cause (e.g., chatting with a friend), or no explanation for her tardiness. Results showed that participants had less negative emotional responses and were more willing to interact with the confederate in the future when she gave an uncontrollable explanation compared to the other two explanations conditions (Weiner et al., (1987). Weiner, Figueroa-Munoz, and Kakihara (1991) later showed that the reduction in negative sentiment is accomplished (at least in part) by reducing perceived causality, controllability, and the perceived likelihood that the agent will repeat the offense in the future.

Updating blame on Weiner's (1995) model

Weiner (1995) argues that mitigating information must have its effect on moral judgment by either denying causality or personal control. The model's key prediction for making and updating moral judgments is that causality and controllability are presumed to be jointly necessary for blame. "Although responsibility can be minimized by shifting from internal to external causality, internal causes that are not controllable, such as a physical handicap, can be considered as good as external causes in that the actor is not held personally responsible" (Weiner, et al., 1987, p. 316). The implications of this

prediction are twofold. First, negating either of these nodes of the model should move perceivers to the left side of the model (see Figure 1) and to a judgment of no responsibility and no blame. Second, controllability and causality are argued to be the only relevant inputs to moral judgment. Other types of information such as character information (Alicke, 1992) that do not speak to the controllability or causality of an outcome should not affect blame. Justificatory reasons are included in the mitigating circumstances step of the model; however, these also affect blame by relocating the cause of an outcome or by denying controllability. Weiner (1995) gives the example of a man who lost a business deal because of a lack of effort. However the man has a justification: He was caring for his sick family member. Weiner argues that this explanation relocates the causality for the failure "the cause of the loss is the sick person" (p. 9) and responsibility for the man is therefore withheld.

A possible third path to mitigation is by ameliorating negative emotional reactions. While the first two paths for blame mitigation are explicit in the model, the third is implied given the causal relationship between evoked emotion and blame specified in the model.

It is likely that the three techniques have diverging consequences for responsibility judgments, but may result in similar blame judgments. The first two mitigation techniques would be effective strategies for reducing responsibility judgments; ameliorating negative affect however, does nothing to deny responsibility, but insofar as a copious literature has shown that increasing negative affect intensifies blame (e.g., Inbar & Pizarro, 2009; Pizarro, et al., 2011; Schnall, et al., 2008) it seems reasonable to hypothesize that decreasing negative affect would lessen blame. However, softening

emotional reactions in Weiner's model is most likely to be a side effect of the first two strategies rather than an independent strategy for reducing blame, since affect is hypothesized to be a result of a responsibility judgment.

In describing the process of moral judgment, Weiner uses the metaphor of a courtroom: "life may be considered a courtroom where dramas related to transgression are played out. The observer is a scientist in determining causal judgments, but then acts in a Godlike manner by reaching moral conclusions regarding right and wrong, good and evil" (Weiner, 2001, p337-338). One critique raised against the model is that it is a prescriptive, rather than a descriptive, model of blame. Among the critics, Alicke (2000) and Haidt (2001) have asserted that the model describes only how a perfectly rational perceiver might make blame judgments, and that the model neglects affective and cognitive factors that may bias (or possibly drive) moral judgment. These authors proposed alternative theories, which focus on the role of affect (Alicke, 2000) and moral intuitions (Haidt, 2001) for moral judgment.

Alicke (2000): The culpable control model

Alicke's culpable control model allots a central role to affective reactions ("spontaneous evaluations" in the model) for determining blame judgments. He proposes that spontaneous evaluations (SE's) bias the process of moral judgment in two ways. First, they directly cause blame. That is, detecting a norm-violation arouses negative affect (e.g., anger, disgust), which in turn directly causes blame. Second, SE's indirectly shape blame judgments by biasing people's evaluations of three criteria for blame ("structural linkage assessments" or SLA's in the model). The SLA's describe the relationship between an agent's mental states, her behavior, and the outcome. Alicke labels these: causal control, volitional behavior control, and volitional outcome control.

Causal control pertains to an agent's unique causal contribution to an outcome. Volitional behavior control indicates whether the agent's actions were intentional and free. Volitional outcome control indicates whether the agent foresaw an outcome and desired it.

Alicke formulates blame criteria such as causation, intention and foresight in terms of a superordinate concept: personal control. He argues that this conceptualization allows for blame to be incrementally exacerbated or mitigated by increasing or reducing personal control[2] (Alicke, 2000). He contrasts this with previous models that adopt a more dichotomous stance (e.g. Shaver, 1985; Weiner, 1995).

The direct path from affect to blame

The direct path from SE's to blame operates via unconscious, affective reactions that cause blame independent of any rational evaluation of blame criteria. "People use outcome information as a basis for ascribing blame and they then justify their attributions by altering their judgments of the a priori criteria" (Alicke, Davis, & Pezzo, 1994, pp. 283-284). Spontaneous evaluations are driven by a host of possible factors, such as the visceral or heinous nature of an offense (Kassin & Garfield, 1991) an agent's race, gender, reputation, likability, or attractiveness (Alicke & Davis, 1989; Alicke, Weigold & Rogers, 1990; Wissler & Saks, 1985), or the tendency to over-attribute human causes for events (Gilbert & Malone, 1995; Jones, 1979; Kelley, 1971). These evaluations are thought to be ubiquitous and unconscious (Alicke, 2000; 2008).

---

[2] Two potential challenges arise from this conceptualization. First, it is unclear how Alicke scales concepts like causation, intention and foresight into personal control. Second, it is unclear whether this is meant to be a descriptive claim about what people do cognitively when they make judgments, or whether this is just a theoretical simplification.

On his model, behaviors that evoke more negative affect are blamed more compared with less evocative behaviors (holding all else constant). In support of this view Alicke and his colleagues (Alicke & Davis, 1989; Mazzocco, Alicke, & Davis, 2004) report several studies demonstrating that outcome information biases blame judgments. In these studies, Alicke and colleagues describe a scenario where a man, who is home alone, hears suspicious loud noises coming from his daughter's bedroom. The man goes to investigate, discovers a shadowy intruder, and reacting quickly, attacks and fatally injures the intruder. The key manipulation in these studies is the outcome of the man's actions. In the "bad outcome" condition the intruder happens to be his daughter's boyfriend; in the "good outcome" condition however, the intruder turns out to be a burglar who had been robbing houses in the neighborhood. Participants blamed the man more, found him to be more negligent, and recommended a more severe sentence in the bad outcome condition than in the "good" outcome condition. Mazzocco, et al. (2004) argues that these results demonstrate that "outcomes were spontaneously evaluated, and these evaluations fed directly into attributions of blame." (p. 145).

Further, Alicke (2000) argues that the direct path from spontaneous evaluations to blame allows extralegal factors (e.g., character and likability information), insofar as they elicit negative affect, to bias moral judgment. In one study, Alicke and Zell (2009) describe a scenario where a male protagonist gets into a bar fight with another man during the fight, accidentally hits a woman in the face causing a permanent scar and extensive dental damage. Prior to the depiction of the bar fight, participants are provided with information that either portrays the agent positively (e.g., he volunteers at a homeless shelter; he cleans his friend's house) or negatively (e.g., he lies about having to

miss work; he steals some beer from a friend). The authors measured participants' impressions of the agent's likability and how much blame the agent deserved for injuring the woman during the bar fight. As might be expected, the likable agent received less blame than the dislikable one; and furthermore, the link between the agent portrayal (positive or negative) and blame was mediated by perceived likability (Alicke & Zell, 2009).

Alicke and his colleagues explain their results by arguing that people's experience of negative affect due to negative outcomes (Alicke & Davis, 1989; Mazzocco, et al., 2004) or dislikable perpetrators (Alicke & Zell, 2009) directly causes people's blame judgments. However, while these studies show that people's blame judgments may be susceptible to information about character or outcome negativity, the role of affect in *producing* these effects is untested. Affective reactions are never directly measured in any of these studies, nor is their hypothesized meditating role for moral judgment ever tested.

The indirect path from affect to blame

In addition to directly causing blame, SE's are hypothesized to indirectly produce blame by biasing people's evaluations of SLA's. As Mazzocco, et al. (2004) write, "perceivers spontaneously evaluate action scenarios based on such features as outcomes and actor characteristics. This spontaneous evaluation, in turn, can color judgments about three different types of linkages that are concerned with rational criteria for blame ascription" (p.146).

Typically, this is discussed (and studied) in terms of the effect of negative spontaneous evaluations biasing people's estimations of the structural links to moral

judgment, thereby enhancing blame; however, the model also allows for reducing blame by weakening the SLA's (Alicke, 1990). For example the model accommodates reductions in blame due to accidents (Pizarro, et al., 2003; Young & Saxe, 2009), impulsivity (Pizarro, Uhlmann, & Salovey, 2003), or duress (Reeder, Monroe & Pryor, 2008; Woolfolk, Dorris & Darley, 2006), by proposing that these explanations weaken one (or more) of the structural linkages and therefore reduce blame.

By contrast, negative spontaneous evaluations are thought to cause people to exaggerate an agent's perceived control over an event, the presence of blameworthy mental states (desires and intentions), or his ability to foresee harmful outcomes, or to lower the threshold by which people judge these criteria (Alicke, 2000; 2008). In support of this view, Alicke (1992) showed that people view an agent as being more causal and more blameworthy for an unintentional bad outcome (a car crash) when he is described as rushing home because he wants to hide a vial of cocaine (which Alicke presumes causes spontaneous negative evaluations of the agent) compared to when he is rushing home to hide an anniversary present for his parents.

Alicke, Buckingham, Zell and Davis (2008, Study 3) replicated this general finding using a different scenario. In this study, participants read a story about an agent who is informed that his mother is about to pass away. Upon hearing the news he immediately rushes to the hospital; however, before arriving he stops to either buy cocaine from a friend or to assist a young mother and her child who had been injured in a car accident. In both conditions, his mother dies before he makes it to the hospital, and participants are asked how much blame the agent deserves, if they have a favorable or unfavorable impression of the agent, and how causal he was for failing to make it to the

hospital before his mother died. Unsurprisingly, participants blamed the agent more and viewed him as more causal when the reason for failing to make it to the hospital on time was buying cocaine compared to helping an injured mother and son. Alicke et al., (2008) take these data as evidence that people's negative reaction to the dislikable agent biases their estimations of his causal contribution for his lateness upwards, driving up blame judgments. One challenge to this interpretation however, is that negative information also provides particularly diagnostic evidence of a person's moral character and mental states (e.g., desires, motives) (Reeder & Brewer, 1979; Skowronski & Carlston, 1989). Therefore, the results may arise from an informational impact of negativity, rather than a motivational one.

Blame validation and updating moral judgments

What might a blame validation model propose for updating moral judgments? Alicke (2000) proposes that while people may be sensitive to many of the factors specified in rational models of moral judgment (e.g., Shaver, 1985; Weiner, 1995) he claims that the process of moral judgment is fundamentally biased by an affectively driven blame validation process. "The phase 'culpable control' reflects the assumption that the desire to hold someone culpable slants perceptions of the control evidence in a direction that validates blame" (Alicke, 2008, p. 183). Thus, people's early affective reactions to norm-violating events cause them to inflate an agent's causal contribution to an event (Alicke, 1992) or his ability to foresee outcomes and intend them (Alicke, et al., 1990) in order to justify a blame judgment.

Such inflations serve a motivational goal – to hold someone accountable for a misdeed. Thus, one might expect that this motivation to hold others accountable and the

desire to maintain existing moral judgments would disincline people to accept new, mitigating information about an event or an agent. Additionally, negative spontaneous reactions may suppress the effect of mitigating explanations. To the extent that a norm-violating event produces a negative spontaneous reaction, these reactions dampen blame reduction by causing people to strongly anchor on their initial judgment and to adjust away only slightly. By contrast, such a blame validating process should intensify the effect of new information that is aggravating or exacerbating (e.g., John killed his uncle so that he would inherit a large sum of money).

Haidt (2001): The social intuitionist model

Alicke argues that affect biases reasoned moral judgments. Haidt (2001) in his Social Intuitionist model (SIM, Figure 2) goes beyond this claim and proposes that affect is not just a biasing factor; instead, affect-heavy intuitions are the machinery of moral judgment. Conscious reasoning, by contrast, only serves to justify intuitive judgments.

Drawing on the large body of work on automaticity in social cognition (Bargh, 1994; Bargh, Chaiken, Raymond & Hymes, 1996; Bargh & Chartrand, 1999; Greenwald & Banaji, 1995), Haidt's SIM proposes a model of moral judgment (Figure 2) in which a negative moral event triggers an intuition that culminates in the realization of a moral judgment (pathway 1). Moral intuitions are proposed to be a specific type of cognition that arise fast, are affectively laden, and do not require attentional resources. Further, the process that gives rise to moral intuitions is hypothesized to operate outside conscious awareness. For example, when people hear about a man who finds his pet dog dead in the road and decides to cook it for dinner, people 'know' that it is wrong to eat the deceased pet but may not be aware of why that belief arises for them or be able to rationally justify their position. Therefore, once a judgment has been made, it engages a motivated

reasoning process (pathway 2). The goal of this reasoning is not to discover the truth of about an event, but it acts as an intuitive prosecutor, seeking arguments to support the already articulated intuitive judgment.

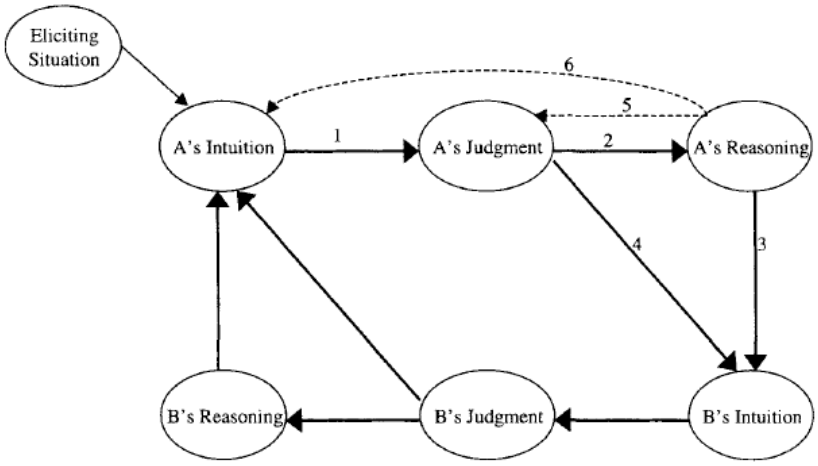

Figure 2. The social intuitionist model of moral judgment.

Haidt's SIM relies on two novel claims that sharply differ from other moral judgment models. First, he proposes that intuition, instead of reasoning, is (nearly) entirely responsible for moral judgment. Second, he claims that conscious reasoning functions primarily as a *post hoc* justification for moral intuitions.

Intuitions drive moral judgment

Haidt's first claim proposes that moral reasoning is grounded in "gut-level" responses rather than in reason (Haidt, 2001; 2007). Disgust, in particular, is supposed to signal the presence of moral violations and serve as information for moral intuitions. Thus, by dint of experiencing a pang of disgust, morally neutral actions can be brought into the moral sphere and their moral violations becomes harsher. In one study, Wheatley and Haidt (2005) tested this hypothesis by hypnotizing participants to briefly experience disgust when reading particular trigger words (e.g., often) and then asking participants to

read six moral violation vignettes (e.g., Cousin incest, eating one's dead dog, shoplifting), half of which contained the hypnotic disgust word. Participants are subsequently asked to judge how disgusting and morally wrong the actions were. As predicted, people rated the vignettes containing the hypnotic trigger words as more disgusting and more morally wrong than the vignettes lacking the hypnotic trigger word[3].

A second study replicated this effect; participants rated vignettes containing the disgust-triggering word as slightly more morally wrong ($M = 68.6$) than scenarios in which the trigger word was absent ($M = 73.4$). More compellingly, the authors also included a non-moral scenario, which described a student council member picking topics for a group discussion. When the disgust word was present participants rated the action as slightly morally wrong, but as not at all wrong when the word was absent.

In a related study, Schnall, et al. (2008) induced disgust by exposing participants to a foul smell (Studies 1 and 2), having them work in a dirty room (Study 3), or having them watch a disgusting video (Study 4). Across studies, Schnall et al. (2008) found that participants induced to feel disgust made more severe moral judgments than participants in a control condition. Other work (e.g., Horberg, Oveis, Keltner, & Cohen, 2009; Inbar, Pizarro, Knobe, & Bloom, 2009) has demonstrated that individual differences in disgust sensitivity predict moral sentiment. These studies show that people who are more easily disgusted are more likely to view certain acts (e.g., eating a dead pet, homosexual sex) as immoral compared with people who are not as easily disgusted.

These data are said to highlight the role of intuition as driving people's moral judgments. However, some researchers have taken issue with this interpretation of the

[3] Only two of the six vignettes (cousin incest and bribery) showed an effect of the disgust manipulation on participants' moral judgments; the effect was not significant for the other four vignettes.

data (e.g., Pizarro et al., 2011). They contend that these studies demonstrate that disgust serves as a *moral amplifier*, rather than serving to independently moralize actions. On this view, affectively valenced intuitions cannot be solely responsible for moral judgment. There must also be some type of reasoning that tracks social and moral norms and flags violations of these norms.

Reasoning is (mostly) post-hoc

Haidt's second claim is that reasoning functions primarily as a post-hoc justification for intuitive judgments by showing that people can be "morally dumbfounded" (Haidt, Björklund, & Murphy, n.d; Haidt & Hersh, 2001; Haidt, Koller & Dias, 1993). That is, when presented with certain moral cases, people label them as wrong but are unable to explain why. Consider the following case:

> "Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love?" (Haidt, 2001, p. 814).

Many people answer that it was wrong for Julie and Mark to have sex, but when pressed to justify their stance, many are unable to come up with viable arguments as to why this is the case. Haidt (2001; 2007) argues that this type of inarticulateness is moral intuition in action. People experience a flash of insight whereby they simply know that an act is wrong, and moral reasoning is merely a post hoc construction produced to justify

the intuition (Haidt, 2001). This claim is buttressed by two studies in which affective responses were superior to perceptions of harmfulness in predicting people's moral judgments (i.e., wrongness) for social taboos, such as cleaning a toilet with the American flag, and sexual morality issues (e.g. homosexuality) (Haidt, et al., 1993; Haidt & Hersh, 2001).

Updating and competing intuitions for moral judgment

Though the SIM does not explicitly discuss updating, it allows for competing intuitions and overriding previous judgments. One method discussed in the model is where people "by sheer force of logic" override their intuitive response (Haidt, 2001, p. 819). However, Haidt argues that this is extremely rare and occurs primarily in unique populations (e.g., philosophers). A second route to altering moral judgments, which Haidt hypothesizes is much more common, is by producing new intuitions (Pathway 6). According to the SIM new information about an event or person can act as a new eliciting situation, which in turn produces new intuitions that conflict with previous ones. Thus, one can build updating into the SIM by adding a direct link between new information and intuitions (See Figure 3). In this sense, updating information serves as a new eliciting situation, and produces new affective intuitions (e.g. sympathy or pity), which may support or compete with previously held intuitions.

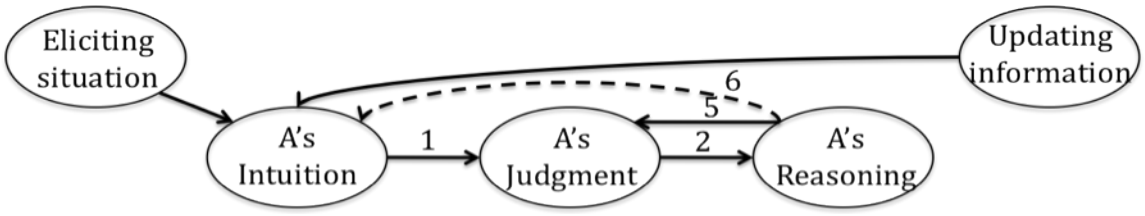

Figure 3. The SIM with updating information (adapted from Haidt, 2001)

Greene et al.'s (2001) famous trolley dilemma illustrates one occasion where people have competing moral intuitions vying for supremacy. Participants are presented with a scenario in which a man sees a runaway trolley barreling towards five workers. The five workers can be saved if the train is stopped, but doing so requires sacrificing another person. In one condition (the "switch" dilemma) the man can divert the train to a sidetrack (with the single person on it) by pulling a lever. In the "footbridge" dilemma, the man must push an innocent bystander off a footbridge and the weight of his body will stop the trolley, again saving the five workmen. In both conditions, participants are asked to evaluate whether it would be morally permissible for the man to sacrifice the bystander in order to save the five workmen.

Whether people find it permissible to sacrifice one life to save five others depends on which dilemma they face. People are generally willing to sacrifice one person to save five if they can do so at a distance by pulling a lever. However, people are much less willing to endorse sacrificing one person to save five in the footbridge case, where the agent must push the bystander to his death (Greene, et al., 2001).

These "trolley cases" highlight the competition between differing moral intuitions. Greene et al. (2001) argues that people have an immediate negative reaction in the footbridge case (e.g., "it is wrong to kill"), which results in a determination of impermissibility. However, many people also have the competing intuition that killing the bystander serves the morally justifiable goal of saving the lives of five innocent workmen, and is therefore permissible. Greene et al., (2001) argues that the different answers to the footbridge dilemma (compared to the "switch" case) are a result of strong

emotional, intuitive processes that override the rational calculation of saving five lives by sacrificing one.

In support of such processes, Greene et al., (2001) cite greater activation in several brain areas reportedly associated with emotional processing (e.g. medial frontal gyrus, posterior cingulate gyrus, bilateral angular gyrus) in the footbridge condition than in the switch condition. Moreover, participants who respond that pushing the bystander is permissible have longer reaction times to the permissibility question and increased activation in working memory areas of the brain than participants who reject pushing. One interpretation of these data is that participants in the footbridge case have competing intuitions – one says "Do not push!" whereas the other says "Saving five lives is better than one," and it takes time and thinking to resolve the competition.

Further, the strength of these intuitions can be manipulated by selecting patient populations with deficits in emotional processing or by inhibiting people's experience of negative emotion. One patient study showed that participants with ventromedial prefrontal lesions (an area associated with emotional processing) were more likely to say pushing the man from the footbridge was permissible than were matched control subjects (Koenigs et al., 2007). Valdesolo and DeSteno (2006) showed a similar set of findings in a nonpatient sample using mood induction. The authors hypothesized that the experience of positive emotion would dampen the connection between negative affective reaction and the act of pushing, resulting in increased permissibility judgments. The data bore out this pattern; participants who experienced a positive induction prior to judging the trolley case were more likely to endorse pushing the man off of the footbridge compared to participants who received a neutral mood induction.

This work also highlights a potential challenge for updating—at least updating in response to mitigating information. If people who witness a moral transgression experience a negative, intuitive reaction that pushes for a moral response (e.g., "Don't!", "that's bad"; Greene, 2009; Haidt, 2001), then attempts to mitigate blame must to override these intuitive responses. This could hamper mitigation in two ways. First, the initial intuitive response may act as an anchor that people must attempt to adjust away from. And second, if the initial intuition is overridden or discarded, the affective tinge of the intuition may still be present. As a result, the residual emotion contained in the intuition may set an upper limit to the amount blame that can be reduced.

Updating in response to exacerbating information, however, should not be held back by from existing intuitions, as both the original and new intuitions (as triggered by the new information) lead perceivers toward the same conclusion. Indeed, Haidt's (2001) intuitive prosecutor should readily accommodate information of this sort and increase blame quickly and easily.

The path model of blame

The Path Model of Blame diverges from the previous three models by specifying a comprehensive set of information contents that are relevant for a moral judgment of blame and the process people step through en route to the judgment. According to the model (Figure. 4), moral judgments are based on information components that are processed in a logical and temporal hierarchy. The model describes the typical process people step through en route to making a moral judgment, and the specific criteria for those judgments and their ordering are grounded in research developmental (e.g., Baldwin, Baird, Saylor, & Clark, 2001; Woodward, 1998) and adult social cognition

(e.g., Barrett, Todd, Miller, & Blythe, 2005; Heider, 1958; Malle, 2004; Malle & Holbrook, 2012).

The process of moral judgment begins with the detection of a norm-violating event (e.g., discovering a large scratch on the side of your car) and proceeds with establishing whether an agent caused the event and, if so, whether the agent caused the event intentionally. Critical to the model is the process-guiding role of intentionality. The intentionality judgment bifurcates the processing stream, shifting people onto one of two information paths. When a behavior is *intentional*, the model predicts that moral evaluations will depend on the agent's mental states, specifically the *reasons* an agent had for acting. For unintentional behaviors, moral judgments depend on an assessment of an agent's obligation to prevent harm and her capacity to prevent the harm. Below I discuss in detail each concept and its role in the emergence of blame.

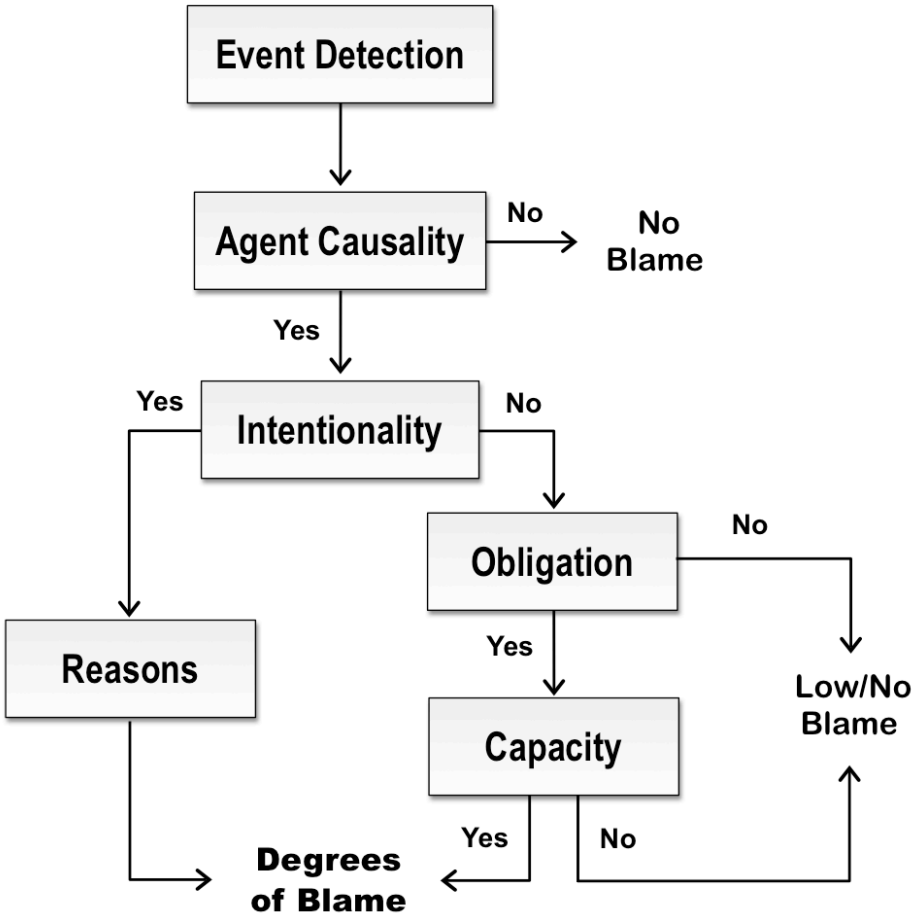

Figure 4. The Path Model of Blame

Negative event detection

In order to begin the process of making a moral judgment, perceivers must first detect a negative event. That is, an event that violates a norm of behavior. Graham, Haidt, and Nosek (2009) suggest that moral judgments arise in response to violations within distinct moral domains, including harm, fairness, authority, purity, and ingroup loyalty. The path model does not differentiate between these different types of norms. Also, these norms need not be objective, consensual, or moral. As long as an observer *perceives* a norm to be violated, the moral judgment process will be activated.

Agent causality

Social perceivers blame people, not physics (though they may still be angry at the latter). So for negative outcomes (e.g., a broken window, a dead person) to lead to blame, perceivers must establish whether an agent caused the outcome (Shaver, 1985; Sloman, Fernbach, & Ewing, 2009). But what do people categorize as an "agent," specifically a "moral agent"? One minimal requirement is the presence of a mind (Gray, Waytz & Young, 2012). Also, one might stipulate that moral eligibility requires the ability to understand and remember norms along with the ability to choose to follow those norms (Guglielmo, Monroe & Malle, 2009). If such abilities are absent (e.g., in infancy or in certain mental or physical illnesses) blame will either not be assigned or be decisively mitigated, both in everyday life and in the law (Alicke, 1990; Robinson & Darley, 1995, chap. 5).

In many situations, causal agency takes on a dichotomous value of Yes (a morally eligible agent caused the event) or No (no morally eligible agent caused the event). For situations in which there is distributed causality (e.g., coercion or wayward causation) the concept can take a continuous value. But any nonzero value of causal agency should suffice to activate the subsequent concept of intentionality.

Intentionality

The theory postulates that an agent's causal involvement falls into two fundamentally different categories—intentional and unintentional (Heider, 1958; Malle, 1999; Reeder, 2009b; White, 1995). People quickly perceive intentionality in everyday situations (Malle & Holbrook, 2012), often perceptually (Scholl & Tremoulet, 2000) or as part of scripts (Schank & Abelson, 1977), but they may also more carefully consider if

an agent acted intentionally if actions are ambiguous (Guglielmo & Malle, 2010a, 2010b; Monroe & Reeder, 2011) or if the weight of the judgment demands it (Reeder, 2009a).

Similar to other models of moral judgment, the path model of blame asserts that intentionality amplifies blame, which is already well supported in the literature (Darley & Shultz, 1990; Lagnado & Channon, 2008; Ohtsubo, 2007; Young & Saxe, 2009). But the model makes the novel claim that intentionality judgments bifurcate the perceiver's processing of norm-violating events (see Figure 4). People are sensitive to very different information when encountering intentional as opposed to unintentional negative events, and these differential tracks of information processing are described next.

Intentional track: Reasons

For intentional behaviors (the left path in Fig. 1), perceivers consider the agent's reasons for acting. The agent's specific reasons for intentional immoral actions modulate the severity of blame. Providing morally good reasons for harm (e.g. a doctor who amputates an infected limb to save a patient) will mitigate blame, while morally bad reasons (e.g., selfish or vengeful reasons) will exacerbate it (Howe, 1991; Reeder, Kumar, Hesson-McInnis, & Trafimow, 2002; Tetlock, Self, Singh, 2010). Recently, Greene, Cushman, Stewart, Lowenberg, Nystrom, and Cohen (2009) found that inferences about motives shape people's moral judgments in trolley cases. When the switch-thrower's motives are described as malicious, participants are much less likely to accept his actions as morally permissible; however, when sacrificing one workman is described as a side-effect of a goal to save the lives of the others, acceptance increases (Greene, et al., 2009)

An example from the legal literature highlights the powerful role of justifications in moderating blame. In People v. Goetz, Goetz was charged with attempted murder and first-degree assault for shooting four teenagers who had asked him for money on the New York subway. Having recently been a victim of a mugging, Goetz feared the four teenagers were planning on beating and robbing him. Consequently, he drew a concealed pistol and fired multiple shots at the youths, injuring all four. At the trial, Goetz related his previous mugging experience and his belief that the teenagers intended to harm him. He was acquitted of both attempted murder and first-degree assault. Apparently the jury agreed that Goetz's decision to shoot the four teenagers was justified by his belief that he was in danger.

Unintentional track: Obligation and capacity to prevent

In the case of unintentional negative events (the right path in Fig. 1), blame depends on two conditions: Obligation (i.e., should the agent have prevented the event) and capacity (i.e., could the agent have prevented the event). Capacity to prevent harm comes in two varieties: the cognitive capacity to foresee a negative outcome (i.e., knowledge or foreseeability) and the physical capacity to avert the event. Denial of either of these capacities, or the obligation to prevent an outcome, will reduce blame.

Most studies of moral judgment hold obligation constant, typically presenting stories in which the agents unquestionably have an obligation to prevent negative events. Consequently, there is little direct evidence for the impact of obligation on blame. When obligations have been empirically examined, however, they have exerted considerable influence. In one study Hamilton (1986) reported that people in higher positions of a

social hierarchy are subject to stronger obligations for preventing negative outcomes and are blamed more for those outcomes when they do occur.

The impact of the cognitive capacity element has been widely demonstrated in adults as well as children from age four on (e.g., Nelson-Le Gall, 1985; Shaw & Sulzer, 1964). Agents who caused a norm-violating event that they foresaw received blame whereas agents who caused a norm-violating event they did not foresee remained blameless (holding physical capacity constant). For example, people's judgments of blame for a CEO (Knobe, 2003) who adopts a profit making policy with an environment-harming side-effect stem (in part) from the CEO's cognitive capacity to foresee possible harm. Even though the CEO did not intend to harm the environment, he nevertheless *knew* that adopting the program would cause harm and he is therefore blameworthy (Guglielmo & Malle, 2010a). Critical for the notion of capacity, therefore, is not only which particular factors have caused the negative outcome, but which alternative options were available for preventing the outcome. Indeed, Alicke et al. (2008) show blame for a doctor unintentionally killing a patient is dependent upon whether the doctor pursued all reasonable alternative actions that could have prevented the death.

Updating moral judgments

Perceivers do not always have sufficient information to make a confident blame judgment. The Path Model allows for new information to continuously be revealed as perceivers refine and update their moral judgments. In some cases, descriptions of norm-violating events are quite sparse and people must search for additional information to fill in the steps of the model. In other cases, the description of a norm violation provides a rich body of information. For example, the statement "Travis killed Conrad" indicates

that there was a norm violation (a killing), a causal agent (Travis), and that the act was intentional.

In either case, acquiring additional information may cause people to step back in the canonical process order and update the value of an earlier concept. For example, when new information contradicts previous beliefs (e.g, "Travis killed Conrad by mistakenly giving him expired medicine.") the model predicts that perceivers must step back and even switch from the intentional to the unintentional path of the model, possibly engaging in a search for new information (e.g., "Could Travis have known the medicine was expired?"). Relative to continuing within a track, switching tracks is presumed to be cognitively costly and requires additional time to process information in order to update a blame judgment (Allport, Styles & Hsieh, 1994).

Moral judgment in summary

Thus far I have reviewed the moral judgment literature, focusing on four models of moral judgment, their supporting empirical evidence where available, and their perspective on updating. Below, I attempt to derive specific predictions from each of the models for three questions pertaining to the phenomenon of moral updating: How often do people update their moral judgments, and how do updated judgments compare to single moral judgments more commonly studied in the literature? How can we best describe the way people change their blame judgments ? Do people adopt a blame validation model of updating where they anchor on early judgments and make only token changes in response to new information? Or is blaming a process of assigning meaning to an event, during which, people adjust blame in proportion to new information? And lastly, is information processing during updating similar to reading off a list, where processing one piece of information is independent of any other information present, or is

there a structure to moral judgment in which processing is selectively facilitated or impeded based on preceding information?

Predictions for moral updating

In the current project I test how people update their moral judgments in light of new information. I compare four moral judgment models: Weiner's (1995) model of responsibility, Alicke's (2001) culpable control model, Haidt's (2001) Social intuitionist model (SIM), and the Path Model of Blame (Guglielmo, et al., 2009; Malle, et al., 2012) and their ability to account for the updating phenomenon. Below I review the models' predictions about (1) the frequency and sufficiency of updating, (2) people's processing speed for updating moral judgments, and (3) the patterns of blame change in response to updating information.

The frequency and sufficiency of updating

Some research in moral psychology suggests that people should be disinclined to update their moral judgments (e.g., Alicke, 2000; Haidt, 2001). According to these views, people take the role of an intuitive prosecutor who is motivated to defend her judgments rather than striving for accuracy. Instead of updating accurately, people are affectively biased by their early judgments and therefore adopt more lenient criteria for blame in order to legitimize them (Alicke, 2000), or people engage in post-hoc reasoning, seeking out information that confirms their initial intuitive judgment (Haidt, 2001).

In contrast, a considerable body of research in judgment and decision making, perspective taking, and person perception argues that while updating is commonplace, doing so accurately is cognitively effortful and typically insufficient (Chapman & Johnson, 1994; Epley & Gilovich, 2001; 2006; Epley, Keysar, Van Boven & Gilovich,

2004; Gilbert Pelham & Krull, 1988; Quattrone, 1982; Strack & Mussweiler, 1997; Trope, 1986; Tversky & Kahneman, 1974).

These two lines of research suggest a strong and a weak hypothesis for updating. The strong hypothesis predicts that the frequency of moral updating would be low, since people are motivated to defend their existing evaluations and updating is cognitively difficult. The weak hypothesis predicts that while updating may be common, it should be characterized by a process of anchoring and (minimal) adjustment. Therefore, if one were to compare blame judgments made under updating conditions (i.e., updated judgments from complete information after initial judgments from incomplete information) with blame judgments immediately from complete information, the two would diverge, with the updated judgments anchored on their initial judgments and insufficiently adjusted for complete information.

By contrast, the Path Model asserts that the process of blaming answers a why question; it resolves uncertainty by filling a gap in understanding and restoring coherence and control (Hilton, 2007; Malle, 2004). Thus, detecting a negative event almost inevitably comes with an attempt to find its cause and its explanation. Accordingly, the Path Model predicts that: (1) the frequency of updating should be high as people attempt to generate meaning for an event, and (2) contrasting with an anchoring and adjustment hypothesis, blame change should be proportional to new information, regardless of whether the information is available all at once or if it is initially incomplete and subsequently filled in. Thus, in the scenario described above, if one compared blame judgments made under updating conditions with blame judgments immediately from

complete information, the Path Model would predict similar blame judgments across the two contexts.

Weiner (1995) argues that people rationally step through the process of moral judgment, and if new information becomes available people should use it to modify their judgments accordingly. Thus, for updating frequency, I treat Weiner's model and the Path Model as offering identical predictions.

Predictions for blame intensity

Weiner (1995) identifies two jointly necessary criteria for blame: causality and controllability. He argues that people assign blame based on the presence or absence of these criteria. This leads to two critical predictions about updated blame judgments. First, blame should be (roughly) dichotomous and track the joint presence of causality and controllability. Denying *either* of these criteria should result in blame being severely reduced. Second, information that does not pertain to causality and controllability (e.g., character information, intentions, obligations) should not affect blame judgments. Thus, his model would predict that blame should be low (or at 0) in response to information labeling actions as uncontrollable (e.g., unintentional, morally good reasons, unpreventable outcomes), but high in all other conditions.

Updating in Alicke's (2000) culpable control model is accomplished by strengthening (or weakening) the links connecting the agent to the negative outcome. However, spontaneous evaluations are hypothesized to bias the evaluation of these links by engaging a practice of blame validation. This practice of blame validation highlights information that legitimizes or intensifies judgments of blame, while minimizing mitigating information. Two predictions fall out from this model. First, because information processing is primarily an effort to validate the existing blame judgments,

this predicts that when new information is positive (e.g., a morally good reason for an action) people should anchor on their initial judgment, discounting the new information and making only token changes to blame. The second prediction is that when new information is negative (e.g., a morally bad reason for an action) people's blame validation system should highlight this information resulting in large increases in blame.

Haidt's (2001) SIM proposes a radical recomputation – new information triggers new intuitions, which compete with previous intuitions to produce new judgments. If a new, contrasting intuition is sufficiently strong, it can override the previous intuition and produce a new intuitive judgment. In some cases however, new (morally bad) information will reinforce previous intuitions, rather than compete with them. Haidt (2000) proposes that people are "intuitive prosecutors" seeking out information to justify their moral intuitions. Thus, new negative information should be highlighted by Haidt's intuitive prosecutor and produce large increases in blame.

By contrast positive information must compete with previous intuitions, and even if it successfully overrides the previous intuition, research on affect and moral judgment suggests that the affective tinge of the original (blameworthy) intuition will not dissipate completely (e.g., Greene, et al., 2001; Haidt, et al., 1997; Wheatley & Haidt, 2005; Schnall, et al., 2005). More likely, the original intuition remains, though diminished, and traces of the intuition contaminate latter intuitions and judgments resulting in restrained blame change. Thus this account appears to predict that, similar to Alicke's (2000) model, blame change in response to positive information will exhibit a pattern of anchor and adjusting, while negative information will produce large increases in blame.

The Path Model predicts that blame change and updated blame judgments will track the new causal-mental information (e.g., intentions, reasons, preventability) people receive. Instead of anchoring and adjusting, the model makes two predictions for blame change. First, blame change should respond to new causal-mental information regardless of its valence. For example, the decrease in blame in response to a morally good reason for an action should approximate the increase in blame in response to a morally bad reason. Second, blame change should track the particular content of the new information. Information specifying that a norm violating event was unintentional but preventable should reduce blame least, with unintentionality information further reducing blame (because it leaves preventability unspecified). Blame should be reduced most when events are updated with a morally good reason (i.e., a justification) or indications that the event was unintentional and unpreventable. Further, people should increase blame in response to information indicating that an agent intentionally caused harm, and more so when the agent also had a morally bad reason for acting.[4]

Predictions for processing speed

While researchers may disagree about the particular way that people make moral judgments, there is broad consensus that features such as causality, mental states, and intentions matter for moral judgment (Alicke, 2000; Cushman, 2008; Guglielmo & Malle, 2010a; Malle, 2006; Pizarro, et al., 2003; Weiner, 1995; 2005; Woolfolk, 2006; Young & Saxe, 2009). What is not clear from previous research is whether these different types of causal-mental information are processed independently, and processing speed is therefore dependent on non-moral factors such as complexity or quantity, or if there is a structure

[4] This second prediction would be a strong test of the model as people may assume morally bad reasons for an agent's actions if they know the action was intentional.

to the process of moral judgment, and this structure that facilitates or slows processing of new information based on its relationship to other information (e.g., intentionality).

Most models of moral judgment (e.g., Alicke, 2000; Weiner, 1995) do not make predictions about how quickly information is processed. Instead, they tacitly assume that information is processed identically regardless of its content. For example, Weiner (1995) does not differentiate between types of information (e.g., preventability, foreseeability, capacity). He only grants (some of) their effects on causality and controllability. Similarly, Alicke (2000) discusses how new information alters the relative strength of the structural linkage assessments for moral judgment (e.g., from the mind to action), but is mute regarding processing speed of different causal-mental information components. Thus in the absence of differential processing assumptions, their models predict equal processing speed, or processing speed that is a function of known (nonmoral) factors, such as quantity, complexity, and uncertainty.

Haidt's (2001) model implies a form of radical recomputation, where new information triggers entirely new intuitions, independent of previous ones. And given that all moral judgment involves the immediate intuitive approval or disapproval of an action or agent: "One sees or hears about a social event and one instantly feels approval or disapproval" (Haidt, 2001, p. 818), the same immediate process should happen in light of any evidence — complete or incomplete. Thus, one might expect that new information (regardless of its content) is processed quickly and at equal speeds.

In contrast to the models above, the Path Model (Malle et al., *in prep*) predicts that updating is differentially speeded depending on the interplay between existing and new information. The model proposes that the intentionality of a behavior produces two

"tracks" to blame (intentional and unintentional) and switching between these tracks (even if no explicit judgment has been made) is cognitively costly. Thus, if people encode an ambiguous event as intentional and later information indicates that the event was accidental, people will be slower to render a blame judgment compared to when they do not have to revise their beliefs about the intentionality of the action.

Moreover, if the negative event in question is encoded as an intentional action people expect information pertaining to the reasons explaining why the agent acted; and by default people expect that bad intentional actions are committed for morally bad reasons (e.g., Frank killed his wife for the insurance money) (Brown & Kleeck, 1989). Thus, because violated expectations involve additional processing demands people should be slower at processing good reasons compared to bad reasons. Alternatively, if the negative event is encoded as accidental people expect information about preventability; and given the ease with which people can generate counterfactuals, their default may be to expect that unintentional events are preventable (Goldinger, Kleider, Azuma & Beike, 2003). Thus, for unintentional events, people should be slower processing unpreventable explanations relative to preventable explanations.

Taken together there are two ways to violate people's expectancies: (1) switching intentionality tracks (e.g., Intentional --> Unintentional) or (2) providing unexpected causal-mental information (e.g., morally good reasons for intentional, bad actions). When either of these expectations is violated I predict that people incur a cost to their processing speed. Further, additional expectancy violations (e.g., violating both intentionality and mental state expectancies) should exacerbate processing speed costs.

Outline of experimental studies

In five studies I test a new question in moral psychology: How do people update

their judgments of blame? I examine how different types of causal-mental information are processed by perceivers and affect blame judgments. Additionally, I examine how the process of moral updating may be bounded by taxing cognitive resources. The first two studies test the frequency of updating, the effect of new information on processing speed, and changes in blame judgments in a lab setting using text stimuli with a student sample (Study 1) and a community sample (Study 2). Study 3 replicates and extends Study 1 and 2's findings using audio stimuli. Studies 4 and 5 examine how putting people under cognitive load by asking them to produce a series of random taps (Study 4) or by making them socially accountable for their judgments (Study 5) might alter the way people update.

In addition, these studies use updating to test the predictions of several moral judgment models (Alicke, 2000; Haidt, 2001; Malle, et al., 2012; Weiner, 1995) as they apply to updating. Study 1 tests my three questions and the models' predictions by presenting people with text descriptions of norm-violating events, asking for an initial blame judgment, and then presenting new information that gives participants an opportunity to update their initial blame judgment.

Chapter 2

Study 1: Moral Updating with Text Stimuli

Method

Participants

Participants ($N = 60$) were students recruited from Brown University's Cognitive, Linguistic and Psychological Sciences subject pool. Two participants were omitted from the analyses because they failed 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 ($SD = 1.27$). The sample's political attitudes tended towards liberal, $M = 6.5$ ($SD = 1.54$) on a 1-9 scale (1 = Extremely conservative; 5 = moderate; 9 = Extremely liberal), and they were below the midpoint on religiousness ($M = 1.2$, $SD = 1.02$) on a 0-4 scale (0 = not at all religious; 4 very religious).

Procedure and Materials

Participants were tested in groups of two to six people. When participants arrived at the lab they were greeted by the experimenter and shown to an individual testing room. The experimenter explained the experimental task – reading brief descriptions of behavior and making judgments using an on-screen click-and-drag slider bar – loaded the experiment, and left the room. After providing informed consent, participants read a set of on-screen instructions, which described the experimental task and how to use the slider bar. Then participants proceeded through three practice trials, which allowed them to practice reading the sentences and using the slider bar for their judgments. Once

participants finished the practice trials, they completed 36 experimental trials, 36 comprehension trials, and a brief demographics questionnaire.

Each experimental trial consisted of four screens displayed in succession. On the first screen participants read a brief description of a norm violating event. On the second screen, participants made an initial moral judgment about the agent described in the event (e.g., How much blame does Jay deserve?). Then, on the third screen, participants read new information that updated their knowledge about the event (e.g., Jay intentionally tracked mud all over the floor) and had an opportunity to update their initial moral judgment. To ensure that participants attended to all of the stimuli, they were asked to "write in their own words what happened" on the fourth screen of each trial. Participants were not allowed to go back and revisit previous judgments or information.

Initial information. On the first screen of each trial participants read a description of a norm violating event (e.g., Jay tracked mud all over his mother's freshly cleaned floor) (see Appendix A for a complete list). All of the norm violating events included three pieces of information: (1) that a social or moral norm had been violated, (2) the identity of a specific agent who violated the norm, and (3) information about the intentionality of the norm violation. The experiment manipulated this latter piece of information (within subjects). In 20 of the 36 trials the event was intentional, and in 16 of the 36 trials the event was unintentional. The intentionality of the norm violating event was established in pretesting prior to Study 1[5]. Using this pretesting data, I classified each of the norm violating events as either prototypically *intentional* or prototypically

[5] A separate 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).

unintentional. "Killing," for example, was classified as a prototypically intentional event, whereas "burning a dinner" was classified as a prototypically unintentional event. See Appendix 1 for a complete list of the norm violating events and their intentionality ratings.

Making initial moral judgments. Descriptions of the norm violating events were displayed for three seconds. After that time elapsed the program advanced to the second screen of the trial where participants were asked to make a moral judgment of the agent. Participants made their moral judgments using a click-and-drag slider bar, displayed on the computer screen (from 0, no blame at all, to 100, the most blame you would ever give). The slider bar's initial cursor position was always at zero. For each initial judgment I measured three variables: (1) the time from the presentation of the judgment request to when the participant first clicked the cursor (the processing RT), (2) the time from the participant's first click to submitting the response (the change RT), and (3) participants' *completed* blame judgment (i.e., the final position of the cursor).

Updating information. Following the initial moral judgment, participants read new information about the norm-violating event. The new information revealed either (a) whether the agent acted intentionally or unintentionally; or (b) the valence of her reasons (morally good or morally bad); or (c) whether the agent could have (reasonably) prevented the outcome (see Appendix B for a complete list of the new information). These six possible types of new information were manipulated within-subjects across experimental trials, and counterbalanced across six forms so that each type of information was presented an equal number of times. For example, for the event "Ted hit a man with his car" a participant would read one of the six types of information described below:

1) **Intentional** (e.g., Ted intentionally hit a man with his car.)
2) **Intentional + Morally bad reason** (e.g., 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** (e.g., 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** (e.g., Ted accidentally hit a man with his car.)
5) **Unintentional + Preventable** (e.g., Ted accidentally hit a man with his car. Ted didn't check his blind spot before backing up.)
6) **Unintentional + Unpreventable** (e.g., Ted accidentally hit a man with his car. Ted's brakes failed to work.)

Each participant received six repetitions of each of the different types of information; participants were never exposed redundant information. That is, while participants read about six morally good reasons across the 36 experimental trials, each specific reason was only presented once.

The switching manipulation. Combined with the intentionality of the initial information, the content of the new information provided a way to test a core prediction of the Path Model – the effect of switching between intentionality tracks on information processing. In a given trial the initial information described either a prototypically intentional or a prototypically unintentional event. The new information provided later in the trial (Screen 3) either matches the intentionality of the initial information, or it switches intentionality (e.g., from intentional to unintentional). For example, if the initial information was: “Matt killed Frank” (a prototypically intentional event) and the new

information was that “Matt intentionally killed Frank for the insurance money” (Intentional + Morally bad reason) this would constitute a *non-switching trial*; however, if instead the new information was “Matt accidentally killed Frank. Matt gave Frank expired medicine because he didn't read the expiration label on the box.” (Unintentional + Preventable) this would constitute a *switching trial*.

Updating judgments. After participants read the updating information, they received back their initial moral judgments and were invited to modify them (“Knowing this new information, now how much blame does X deserve?”). Participants used a click-and-drag slider bar identical to the one used for the initial moral judgment. Importantly, the starting value for the cursor for the updating judgments was equal to the final position of participants’ initial judgment.

I measured three variables for each updated moral judgment: (1) the processing RT (i.e., the time from the beginning of the screen to when the participants first clicked the cursor), (2) the change RT (i.e., the time from participants’ first click to when they submitted their judgment), and (3) participants’ completed blame judgment (i.e., the final position of the cursor on the sliding scale after updating). Also, I calculated a blame change score for each trial by subtracting participants *initial* blame judgment from their *updated* blame judgment.

The new information was displayed on the same screen as the updating moral judgment scale. To control for the time participants spent reading the updating information, we added a series of comprehension trials (described below) at the end of the experiment to provide a measure of participants’ idiosyncratic reading time for the updating information.

Finally, on the last screen of a trial, participants were asked to briefly describe the event in their own words ("Describe in your own words what X did."). Then, when they submitted their open-ended response, participants began the next trial.

Comprehension trials. Following the completion of the 36 experimental trials, participants completed 36 comprehension trials. These trials served as a measure of participants' reading time for the updating information. Participants were asked to read brief event descriptions and to press the 'Submit' button as soon as they felt they understood what they read. The comprehension trials were matched with the experimental trials, so a participant saw the same updating information in the comprehension trials as she had seen in the experimental trials.

After the comprehension trials participants completed a brief demographics questionnaire and were thanked and debriefed.

Results

Data preparation

Prior to data analysis I computed a reading-time corrected processing RT by subtracting each comprehension trial reaction time from the processing RT for the matched experimental trials. This corrected RT is used in all of the analyses that refer to "processing speed." In this and all subsequent studies, I trimmed reaction time outliers from the dataset. Reaction times greater than two standard deviations of the mean were eliminated from the data. I also replaced missing values when individuals who had valid responses for the majority of the design cells would have been omitted from the within-subject ANOVA. To retain the 58 valid participants, I used cell-based sample means to

replace 63 missing values within the matrix of 696 potential values (58 participants x 12 cells) [6].

Frequency of updating

A first question addressed by this study is: "How often do people update their moral judgments?" A strong hypothesis by Alicke (2000) and Haidt predicts that updating should be rare, even when people have an opportunity to. The data show however, that people updated in the vast majority of cases (88.5%), and this pattern was not significantly correlated with participants' political orientation (r = .086, p = .52) or religiosity (r = -.03, p = .84)[7]. Similarly, the frequency of updating within each updating information condition was also consistent with Study 1.

Attesting to the Path Model's two tracks for blame, updating was more frequent when new information prompted people to switch from one intentionality track to the other (e.g., an intentional event updated with unintentional + unpreventable information), $F(1,285) = 18.8$, $p < .001$, $d = .34$ (See Figure 5). Paired samples t-tests showed that people updated more frequently in response to intentionality information ($p < .05$, d = .82), morally bad reasons ($p < .05$, $d = .55$), and unintentional + unpreventable information ($p < .05$, $d = .23$) when the new information prompted switching intentionality tracks compared to when new information did not call for switching tracks (e.g., an intentional event updated with intentional + morally good reasons). This effect was marginally significant for morally good reasons as well ($p = .054$, $d = .34$); however

[6] These 12 cells were comprised of the six new information conditions x switching trials (switching vs. non-switching). Replacing these cells does not change the cell means; however, it does slightly lower their standard deviations.

[7] Given the restricted range, I do not report the correlations with age or education.

the pattern was not significant for unintentional + preventable information ($p > .2$, $d = .50$) and reversed for unintentionality information ($p = .042$, $d = .50$).

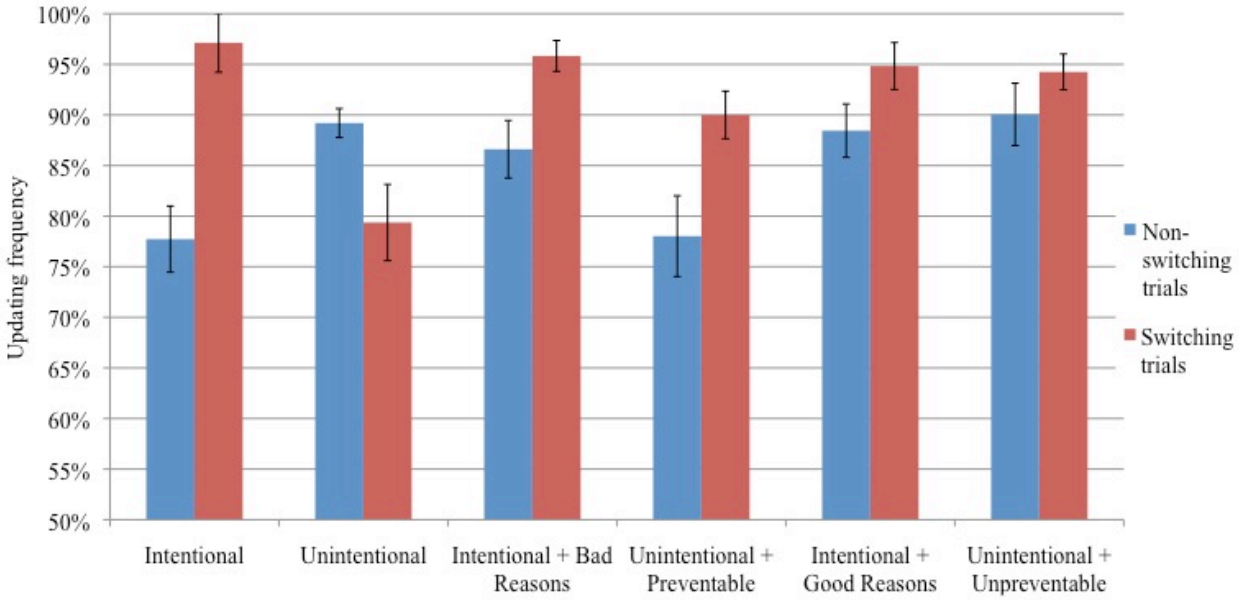

Figure 5. Frequency of updating by information condition for switching and non-switching trials

Updating blame judgments

The second critical hypothesis tested by this study pertains to how new causal mental state information changes the intensity of people's blame judgments. To answer this question I examined people's initial blame judgments, the change in blame during updating, and participants' final blame judgments.

As expected, people's initial blame judgments tracked the intentionality of the event descriptions, $t(57) = 9.01$, $p < .001$, $d = .62$. Unintentional events received less blame ($M = 54.1$, $SD = 12.9$) than intentional actions ($M = 62.5$, $SD = 14.0$).

I conducted a 2 (switching) x 6 (information condition) within-subjects ANOVA to test the effects of new information and switching on the magnitude of change in blame during updating. Blame change was strongly predicted by the type of new information

participants received, attesting to the systematic ways in which people update blame, $F(5,285) = 25.7$, $p < .001$. (see Figure 6) As predicted, blame increased in response to information about the event being intentional and information about morally bad reasons. Blame decreased in the other four conditions. This reduction was smallest in response to information about the event being unintentional but preventable, and it was larger in response to information merely about the event being unintentional. Yet larger blame reduction occurred after learning that the intentional action was committed for a morally good reason and the strongest blame reduction after learning that the event was unintentional and unpreventable (see Figure 6).

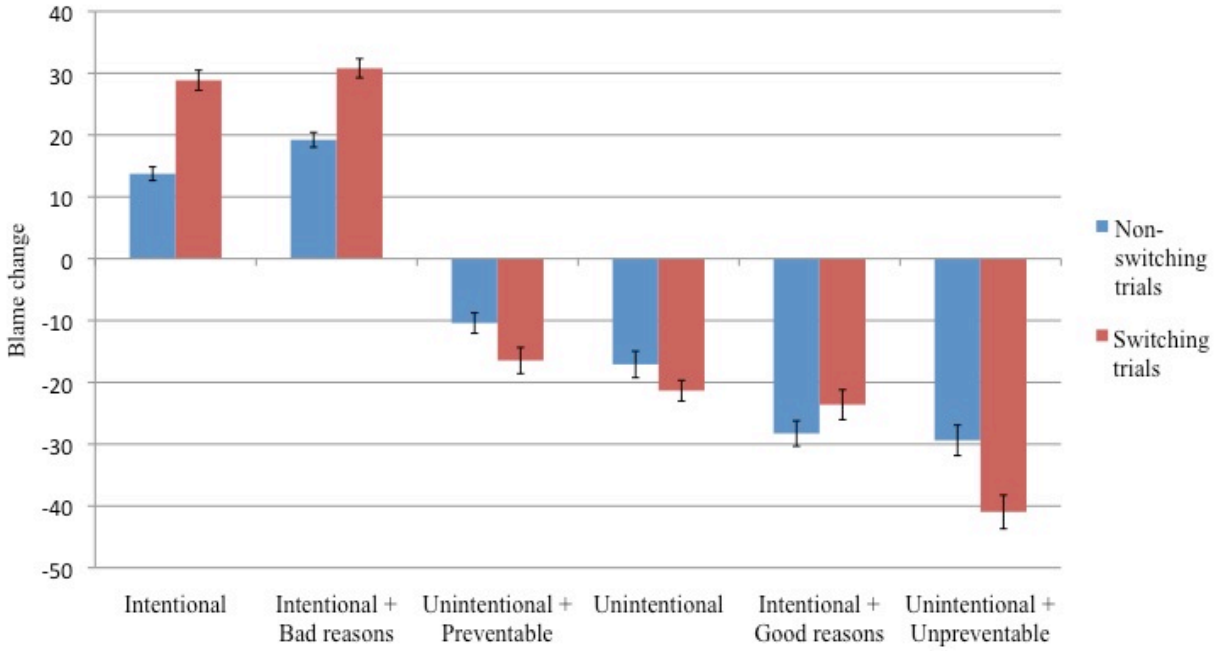

Figure 6. Blame change by updating information condition and intentionality switching

Additionally, switching intentionality had a significant effect on the magnitude of blame change ($F[1,285] = 45.1$, $p < .001$), supporting the view that intentionality serves as a moral amplifier. Switching from the unintentional to the intentional track had the effect of increasing blame intensity (holding causal-mental state information constant);

whereas switching in the opposite direction further decreased blame intensity. Lastly, there was a significant switching by information interaction (F[5,285] = 9.37, $p < .001$), showing that switching differentially amplified the effects of causal-mental information. Specifically, intentionality switching amplified blame change (up or down) for all of the information conditions ($Fs = 4.6$, $ps < .05$), with two exceptions: unintentional information ($p = .053$) and morally good reasons ($p = .136$).

People's updated blame judgments closely tracked the different new information types, $F(5, 285) = 232.4$, $p < .001$. Blame was highest in response to intentional harm ($M = 79.9$ $SD = 15.2$) and intentional harm with morally bad reasons ($M = 83.0$, $SD = 13.5$). Final blame was lower for causing unintentional but preventable harm ($M = 47.9$, $SD = 16.1$), unintentionally causing harm ($M = 38.1$, $SD = 19.0$), and blame was lowest when there was a morally good reason for the action ($M = 28.7$, $SD = 14.7$), or when harm was unintentional and unpreventable ($M = 23.5$, $SD = 12.5$).

Blame updating: Processing speed

A key difference between Alicke (2000), Haidt (2001) and Weiner's (1995) models of moral judgment and the Path Model (Malle et al., 2012) is that the former three models hypothesize that the speed of processing of different causal-mental information is undifferentiated – people have a list of relevant moral criteria and updating is a function of the number of items people must revisit. Since, participants always receive two pieces of information in any given trial: intentionality and (either) reasons or preventability, these "list" models predict no difference between switching and non-switching trials. Contrastingly, the Path Model specifies two independent tracks to blame – an intentional and an unintentional track – and posits that switching between these tracks results in a

cost to processing speed. Thus, the Path Model predicts that participants' processing RT should be higher in trials that switch between these hypothesized tracks compared to non-switching trials.

To test these competing hypotheses, I aggregated reaction times for all the switching trials (e.g., an unintentional event description followed by reason information) across the different new information conditions and compared them to the aggregate of the non-switching trials (e.g., an intentional event description followed by reason information). A paired-samples t-test showed that participants were significantly slower to make a first response in the switching trials (M = 7.10 sec, SD = 1.77) than in the non-switching trials (M = 6.51 sec, SD = 1.58), $t(57) = 3.59$, $p < .001$, $d = .35$. This pattern also held for the total amount of time participants spent updating their judgments; they took longer to update their judgments in the switching trials (M = 9.30 sec, SD = 2.07) than in the non-switching trials (M = 8.87 sec, SD = 1.85), $t(57) = 2.37$, $p < .05$, $d = .22$. The pattern tended to be reversed for the change RT (i.e., the time between participants' first click and when they submitted their judgment), but not significantly so, $t(57) = -1.69$, $p = .10$, $d = .19$. Participants spent slightly more time adjusting their judgments in the non-switching trials (M = 3.43 sec, SD = 1.39) than in the switching trials (M = 3.15 sec, SD = 1.46),

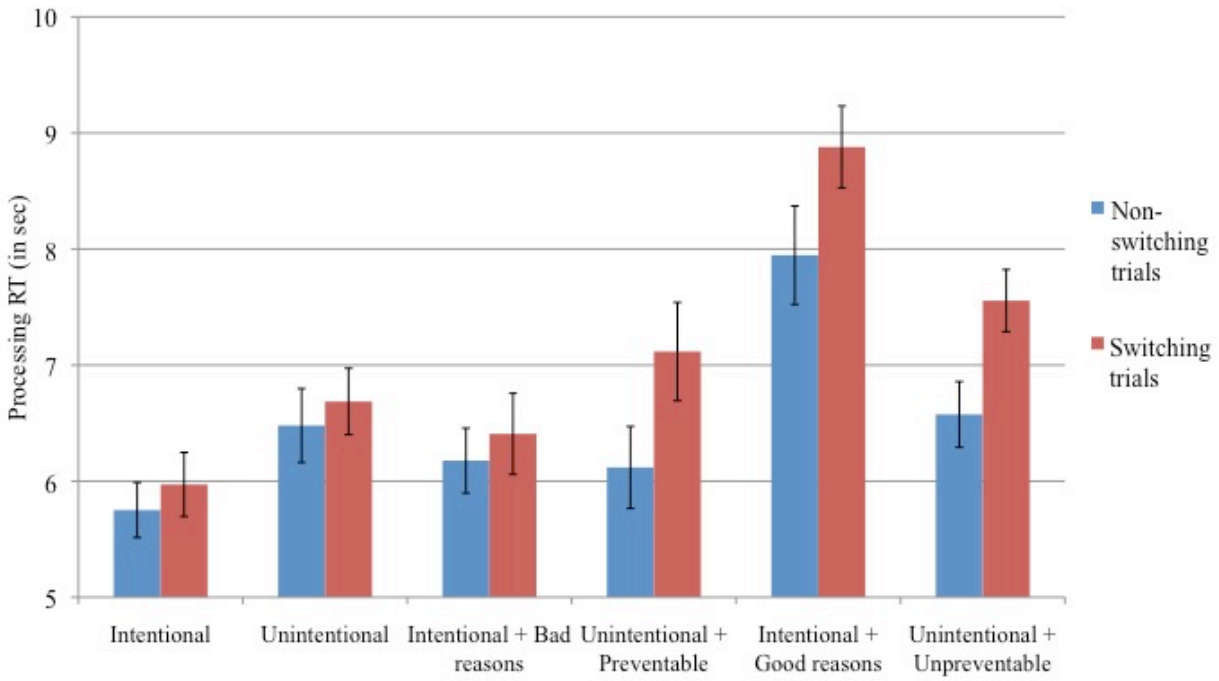

Figure 7. Processing RT's by intentionality switching and updating information conditions.

This pattern held for many of the individual information conditions as well (See Figure 7). People were slower in the switching trials to update their moral judgments after learning about a morally good reason, preventability, or unpreventability information (*Fs* > 4.9*, ps* < .05) compared to non-switching trials. However, while processing RT's were slightly longer for switching trials relative to non-switching trials for morally bad reasons or intentionality information (either affirming or denying intentionality), this effect was not significant (*Fs* < 1).

Study 1 Discussion

In this study I tested how people update initial moral judgments in response to new causal-mental information. In addition, I compared the predictions of four moral judgment models (Alicke, 2000; Haidt, 2001; Malle, et al., *in prep*; Weiner, 1995) as they pertain to the frequency of updating, information processing speed, and changing blame

intensity. The results largely favored the Path Model of Blame (Malle, et al., *in prep*) but also demonstrated limited support for the other three moral judgment models (Alicke, 2000; Haidt, 2001; Weiner's (1995)). I discuss the main findings of the study below.

Updating was much more frequent than might have been expected from previous research. Participants updated their judgments in the majority of cases (88.5%); indeed, for some types of information, participants always updated (or "all participants updated"). This result contrasts with a strong view of models that propose updating is limited by blame validation or post-hoc rationalization (e.g., Alicke; 2000; Haidt, 2001), but it fits with models according to which, blame is a process of assigning meaning to an event (Malle, et al., in prep; Weiner, 1995).

One critique of this study might be that the design artificially inflated the rates of updating. In this study, participants were provided with new information rather than having to search for it, which might mirror the naturalistic environment more closely. However, it was not the case that participants updated in response to any type of new information. Rather the pattern of people's updating frequency appears to point to violations of expectation as a possible trigger for updating. Information that participants viewed as surprising (e.g., morally good reasons for an initially negative intentional action or switching intentionality tracks) drew more updating; information that participants may have assumed from the event description (e.g., an agent's morally bad reasons for a negative intentional action) resulted in less updating. This seems to indicate that participants are updating selectively, rather than merely changing their blame judgments in response to any new information.

This study also revealed new insights into the process of updating and the relationship between intentionality and mental state information. The field of moral psychology has largely focused on how manipulated inputs to blame affect people's judgments. Little research to date has examined how people process these inputs, or how specifying values for one type of mental state information (e.g., intentionality) alters the processing of subsequent mental state information (e.g., reasons and preventability).

These results imply a two-track structure for people's process of blaming, which the updating paradigm can uniquely violate (i.e., by switching between tracks). These violations then reveal the structure's undergirding logic by demonstrating where people are surprised or have (more) difficulty making judgments. Here it seems that the underlying structure is built upon the perceived intentionality of the event. People encode events as either intentional or unintentional (Malle & Knobe 1999; Malle & Holbrook, 2012), and based on that inference different types of information are expected to follow. If people encode an event as intentional, they process information about reasons much faster than information about preventability (which constitutes a switch of tracks). Conversely, if people encode an event as unintentional, they process information about the preventability of the outcome more quickly than information about reasons.

Interestingly, this study revealed a second assumption people hold about norm violating events and the mental states that underlie them. The processing RT showed that (regardless of intentionality) people were slower to update judgments following information about *morally good reasons* compared with *morally bad reasons*. Similarly, people were slower to update their moral judgments in response to information about unpreventable events compared with preventable events. These data show possible

evidence for a "blameworthiness assumption." That is, people believe that intentional, bad actions are generally committed for bad reasons and harmful accidents are generally preventable. Prima facie, this seems consistent with a blame validation model (Alicke, 2000). However, at least in this first study, these assumptions had effects only on people's processing speed, not on their pattern of blame change or their final blame judgments. In fact, the pattern of blame change systematically tracked the mental state content in the new information, increasing in response to information that indicated intentionality or revealed morally bad reasons and decreasing in response to information that denied intentionality, denied preventability, or offered morally good reasons. So, while it may take people longer to reason from their expectation to reality, they still take reality into account when making judgments of blame.

While the fact that blame judgments responded to mental state information may be uncontroversial, the specific pattern of results is not well accounted for by most models of moral judgment. Insofar as Alicke (2000) and Haidt (2001) predict a pattern of anchoring and (minimal) adjustment, the blame change data not fit these predictions well. The pattern of blame change showed that participants made sweeping changes to their initial blame judgments in response to the updating information. This was not only true when the information was exacerbating, as the models would predict, but also when the information was mitigating. Weiner's (1995) model however, was partially supported by the blame change data. While judgments did conform to the dichotomous prediction stipulated by the model, blame judgments were sensitive to information that underlies judgments of controllability: intentionality, preventability, and justifying reasons. Indeed, some of the largest reductions in blame resulted from unpreventability information.

A possible counter to my interpretation of Alicke's (2000) model as failing to account for the data would be that in actuality the data fit the model perfectly well – people received information that called for adjustment of the hypothesized structural linkage assessments and they used it accordingly to ramp blame up or down. However, this interpretation appears to have shaky footing. On Alicke's model, because participants initially made a negative moral judgment of an agent, they should engage their blame validation mechanism that discounts mitigating information and augments exacerbating information. The blame change data do not support this view. Averaging across switching trials, information about intentionality (M = 21.3) and unintentionality (M = -19.2) produced similarly sized changes in blame judgments. Reason information displayed an almost identical pattern; morally bad reasons increased blame (M = 25.0) by the same amount that morally good reasons decreased blame (M = -26.0). Moreover, the unintentional + unpreventable condition (M = -35.2) showed the largest effect of any of the condition and reduced blame by 40% more than morally bad reasons increased it.

However, it is true that Study 1 only partially tests the claim that updating is best captured by anchoring and insufficient adjustment. Although participants clearly make full use of the scale to adjust their moral judgments, these adjustments may fail to approximate the judgments one might give in the absence of an anchor (i.e., the initial moral judgment). Study 2 addresses this concern by comparing blame judgments in an updating condition to a control condition where participants make only a single moral judgment in light of full information.

An additional limitation of this previous study is that participants took a long time to begin updating their judgments (over 6 seconds on average). While the patterns of

processing speed fit well with the predictions derived form the Path Model of Blame, the amount of time people take to update their judgments and the large amount of variance in those reaction times raises important questions about whether the pattern of results is replicable and what processes occur within participants' long response times

Chapter 3

Study 2: Replication of Moral Updating with Community Participants

Given the long reaction times and associated variability in Study 1, I wanted to replicate the effects for frequency of updating, processing speed, and blame change. Moreover, I wanted to test whether the pattern of updating observed in Study 1 would replicate for participants sampled from the community rather than from the college population. Finally, Study 2 addresses the claim from the judgment and decision making literature and both Alicke's (2000) and Haidt's (2001) moral judgment models that updating should be characterized as a process of anchoring and (minimal) adjustment. Study 2 tests this claim by comparing sets of blame judgments in an updating condition to a control condition in which participants make only a single moral judgment in light of full information.

Method

Participants

Participants were recruited from the local community using a Craigslist ad for a "Paid research study." Participants who responded to the ad were invited into the lab to participate in the experiment. In total, 107 people participated. Eleven participants were omitted from the analyses because they failed to follow instructions or had difficulty reading the stimulus sentences.

Of the remaining participants ($N = 96$), 45 were female. The majority of participants identified as White (65%), with fewer participants identifying as Black (15%), multi-ethnic (8%), Latin/Hispanic (6%), or Asian (2%). The sample had an average age of 32.5 years ($SD = 12.4$) and represented a diverse range of education: 49%

(N = 47) reported finishing high school as their highest level of education; 14.6% (N =14) attained a 2-year degree; 24% (N = 23) attained a 4-year degree; and 11.4% (N = 11) attained a Master's degree or higher. The sample's political attitudes tended towards liberal, averaging M = 6.4 (SD = 1.90) on a 1-9 scale (1 = Extremely conservative; 5 = moderate; 9 = Extremely liberal), and they were slightly below the midpoint on religiousness (M = 1.4, SD = 1.16) on a 0-4 scale (0 = not at all religious; 4 very religious).

Procedure and Materials

Updating condition. Fifty-eight subjects participated in the updating condition. This condition was an exact replication of Study 1 in terms of stimuli, dependent variables, and procedure. Participants were run in small groups (2-6 participants per testing session). Upon arriving at the lab they were greeted by the experimenter and shown to an individual testing room. The experimenter briefly explained the experimental task, loaded the experiment, and left the room. Participants read the same set of instructions as in Study 1, practiced the task, and then completed 36 experimental trials and 36 comprehension trials. After completing all of the experimental and comprehension trials, participants completed a brief demographics section and were paid and thanked for their participation.

Control condition. In order to test the hypothesis that updating is characterized by anchoring and adjusting, I included a control condition where participants were asked to make only a single moral judgment in light of full information. Thirty-eight subjects participated in this condition. The control condition was a modified version of the updating condition; the stimuli and dependent variable were identical across the two

conditions. The only difference between the two conditions was that participants in the control condition did not make initial and final judgments. Instead, participants in the control condition were presented with the complete description of the event, including causal-mental state information and then asked to make a blame judgment. I then compared the pattern of blame judgments from these participants to the final blame judgment of the participants from the updating paradigm.

Below I first report the frequency, blame and reaction time results from the updating condition alone. Then, I report the results for participants' final blame judgments, comparing between the updating and the control conditions to test the prediction that updating is characterized by anchoring and adjusting.

Results

Frequency of updating

I replicated Study 1's findings for frequency of updating. Participants updated their judgment in the majority of the experimental trials (74.4%), though the overall rate of updating was somewhat lower than in Study 1. This pattern was not significantly correlated with participants' age (r = .11, p = .41), political orientation (r = .18, p = .24), or religiosity (r = -.13, p = .35); but it was significantly correlated with participants' level of education (r = .33, p = .011). Participants with higher levels of education updated a greater percentage of their judgments than participants with lower levels of education, which may explain the difference between the updating frequency in this study and Study 1. On average, Study 1 participants attained a higher level of education than Study 2 participants.

Also, replicating Study 1's support for the Path Model's two tracks for blame, updating was more frequent when new information prompted people to switch from one

intentionality track to the other, $F(1,285) = 8.35$, $p < .01$, $d = .15$ (See Figure 8). However, paired samples t-tests showed that the effects of switching on individual information conditions were much weaker than in Study 1. Updating was more frequent in switching than in non-switching trials when new information indicated that an action was intentional ($p < .05$, $d = .41$) or when new information described an agent's morally bad reasons for acting ($p < .05$, $d = .68$). Trials with information about an agent's morally good reasons for acting showed a similar, though weaker pattern ($p = .10$, $d = .30$). However, the effects of switching on updating frequency were not significant for preventability information ($p > .5$, $d = .04$) and information indicating that an action was unintentional ($p = .14$, $d = .34$). The reversed pattern for the effect of switching on unintentional information replicates the pattern from Study 1. This may reflect that when people encode an event as intentional, and merely find out that the action was unintentional they may not believe that this new information warrants updating – instead they assign blame on strict causality. However, when an event is unintentional, information affirming that assumption seems to elicit more willingness to change an earlier blame rating. Alternatively, it may reflect that participants doubt the new information in the switching cases. In these cases they may not be confident that events are truly unintentional and therefore refrain from updating.

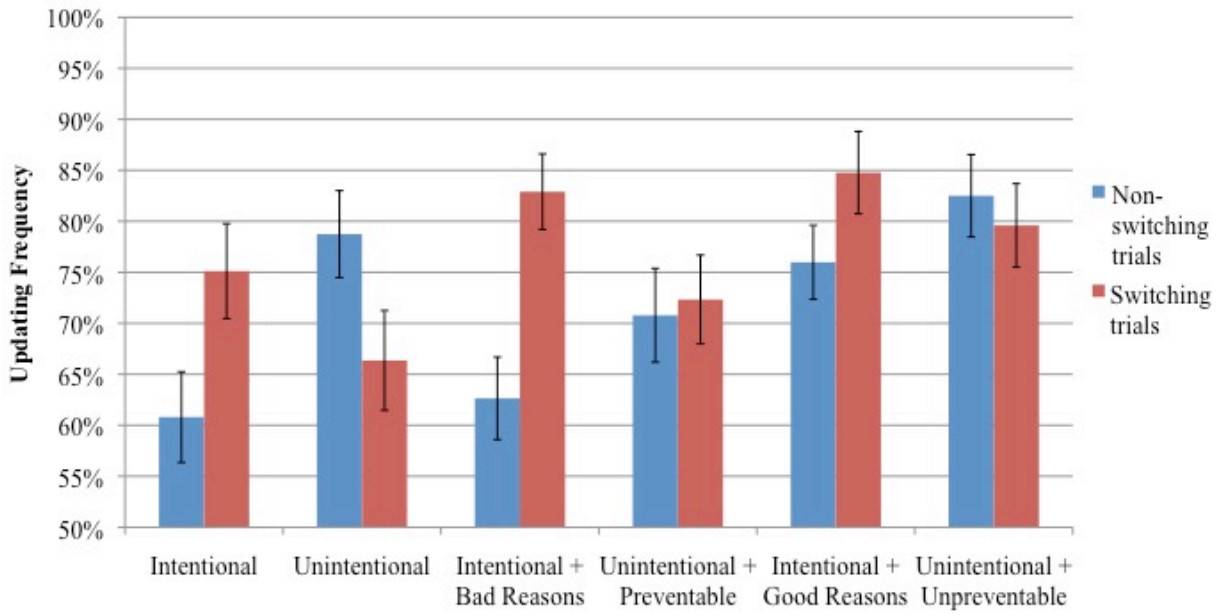

Figure 8. Frequency of blame by updating information condition and switching trials.

Updating blame judgments

Finally, I examined changes in people's response to updating information. As in Study 1, initial blame judgments tracked the intentionality of the event descriptions, $t(57) = 6.58$, $p < .001$, $d = .52$. Unintentional actions received less blame ($M = 56.6$, $SD = 18.2$) than intentional actions ($M = 65.2$, $SD = 14.9$).

As in Study 1, a 2 (switching) x 6 (information condition) within-subject ANOVA tested the effects of new information and switching on the magnitude of blame change. Replicating the previous study, blame change was strongly predicted by the type of updating information participants received, $F(5, 285) = 8.41$, $p < .001$. Blame increased in response to information affirming that an agent acted intentionally and to information about morally bad reasons. Blame was reduced in the other four conditions. This reduction was smallest when new information indicated the event was unintentional but preventable, and it was larger in response to unintentionality information alone. Blame

reduction was larger still after learning that an intentional event was committed for a morally good reason, that the event was unintentional and unpreventable (see Figure 9).

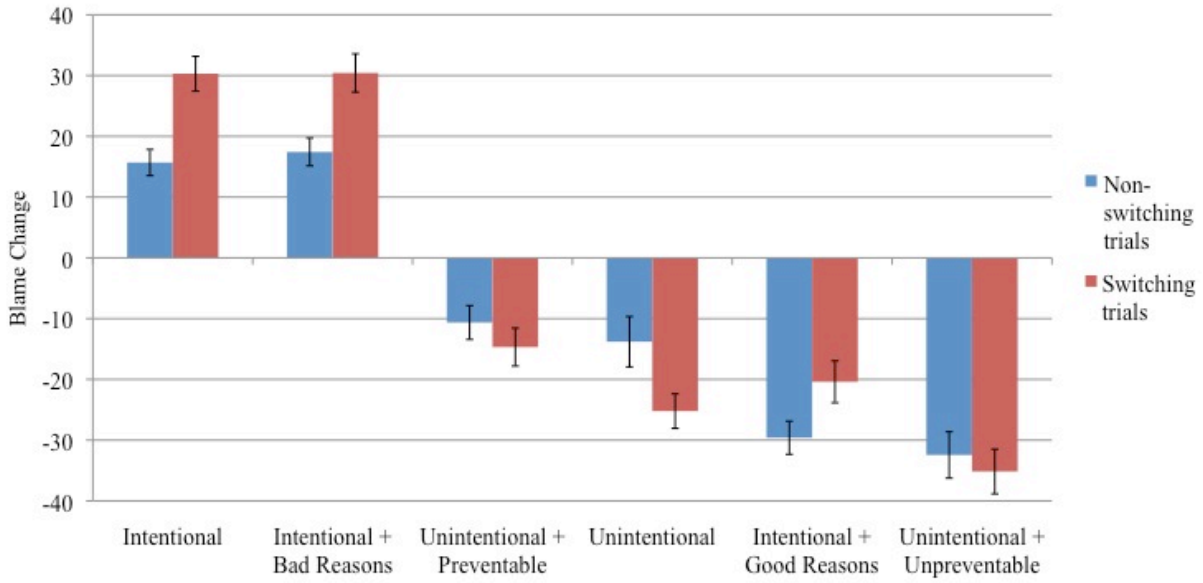

Figure 9. Blame change by updating information and switching.

In addition, switching tracks amplified blame change, (F[1, 285) = 16.6, p < .001). As in Study 1, switching from the unintentional to the intentional track resulted in heightened blame intensity; switching from the intentional to the unintentional track led to reduced blame intensity. However, this effect was not identical across information conditions, as there was a significant switching by information interaction, F(5, 285) = 5.37, p < .001). Blame change was larger for switching trials relative to the non-switching trials in response to information about morally bad reasons, intentionality alone, and unintentionality alone (Fs = 13.5 ps < .001). Preventability information showed a similar trend, but this was not significant (ps > .2) Morally good reasons, by contrast, showed an opposite pattern such that blame change was smaller here when people switched tracks upon hearing that a previously negative unintentional event was actually done for good reasons. This finding may be explained by opposing forces on blame inherent in only this

condition. That is, when new information causes people to switch from the unintentional to the intentional track they may initially increase blame; however, upon considering the morally good reason for the intentional action (i.e., justification) people are pushed to reduce blame. These opposing forces may lead to smaller amounts of blame change overall in this condition.

Lastly, I examined people's final blame judgments (i.e., blame judgment post-updating). A within subjects ANOVA showed a significant effect of the type of updating information $F(5, 285) = 160.2$, $p < .001$. Final blame judgments were highest after reading about someone who had morally bad reasons for acting ($M = 86.5$, $SD = 12.8$) or about someone who intentionally caused harm ($M = 84.6$, $SD = 15.1$), followed by reading about someone who caused unintentional but preventable harm ($M = 48.9$, $SD = 21.3$) or someone who unintentionally caused harm ($M = 41.1$, $SD = 22.9$). Blame was lowest when the harm was unintentional and unpreventable ($M = 30.7$, $SD = 17.8$) or when there was a morally good reason for the action ($M = 28.7$, $SD = 17.3$).

Blame updating: Processing speed

As in Study 1, we predicted that switching between the intentional and unintentional tracks to blame carries a cognitive cost, which results in slower processing of causal-mental information en route to updating a moral judgment. As in study 1, I aggregated all of the RT's for the switching and non-switching trials and compared them using a paired samples t-test. The results again supported the Path Model's hypothesis. Participants were significantly slower to make a first response in the switching trials (M =8.57 sec, $SD = 2.64$) than in the non-switching trials ($M = 7.60$ sec, $SD = 2.30$), $t(57) = 4.70$, $p < .001$, $d = .4$. The pattern was reversed for the change RT (i.e., the time between

participants' first click and when they submitted their judgment). Participants spent slightly more time adjusting their judgments in the non-switching trials (M = 3.43 sec, SD = 1.39) than in the switching trials (M = 3.15 sec, SD = 1.46), $t(57) = 2.00$, $p = .050$, d = .19. Overall however, participants still spent more time updating their judgment in the switching trials (M = 11.7 sec, SD = 3.15) than in the non-switching trials (M = 11.0 sec, SD = 2.74), $t(57) = 2.99$, $p < .01$ $d = .23$.

This pattern held for many of the individual information conditions as well (See Figure 10). People were slower to process information in the switching trials than in the non-switching trials when they received information either about an agent's reasons (good or bad) or unpreventability information ($Fs > 6.8$, $ps < .05$). People showed similar trends of slower processing for preventability or intentionality information (both affirming and denying intentionality), but these effects were not significant ($Fs < 1$).

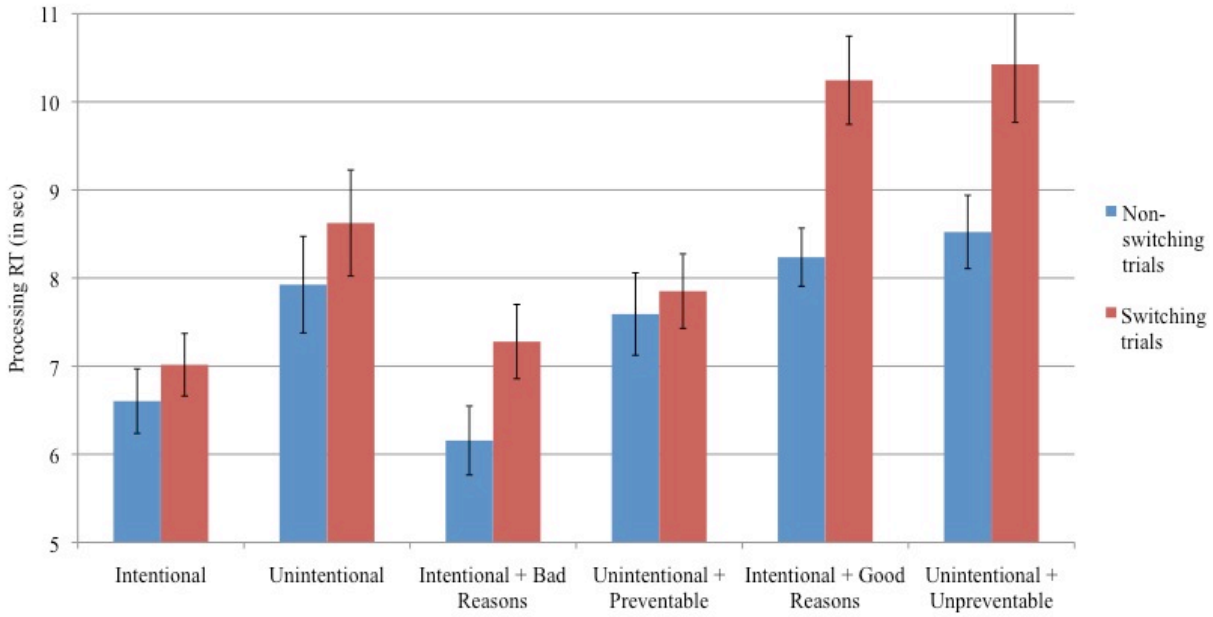

Figure 10. Processing RT's for individual information conditions by switching.

Is updating motivated anchoring and adjustment?

In addition to replicating the effects of Study 1 with a community sample, this study directly tests the Alicke (2000) and Haidt's (2001) blame validation hypothesis. This hypothesis predicts that final blame judgments in the updating condition should diverge in particular ways from a control condition. Specifically, these models predict that blame judgments in the updating condition should be *higher* than judgments in the control condition. The effect of mitigating information (e.g., morally *good* reasons) should be restrained by anchoring on initial judgments; contrastingly, exacerbating information (e.g., morally *bad* reasons) should be emphasized because of blame validation.

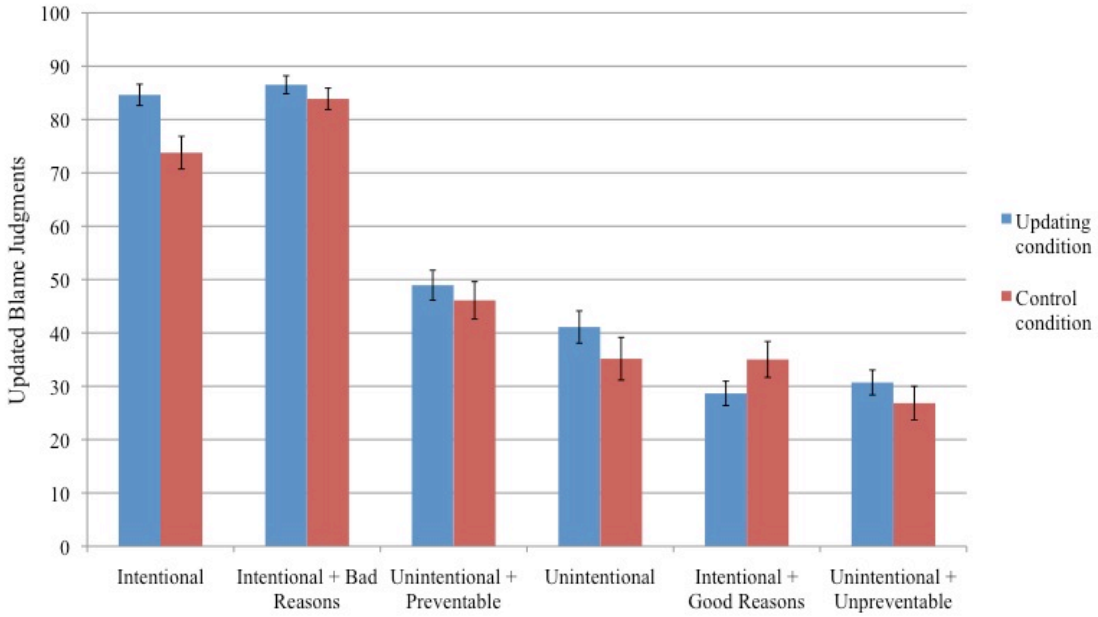

Figure 11. Study 2 updated blame judgments in updating and control conditions.

A 2 (study) x 6 (information condition) mixed model ANOVA showed that participants in the two conditions arrived at strikingly similar blame judgments (Figure 11). In both conditions, final blame judgments varied as a function of mental state information, $F(5, 470) = 245.6$, $p < .001$). But contrary to the blame validation

hypothesis, there was no significant main effect of condition, $F(1,94) = 1.46$, $p = .23$. There was, however, a significant information by condition interaction, $F(5, 470) = 3.24$, $p < .01$. Descriptively, the pattern is driven by higher blame in the updating condition than the control condition for intentionality information and lower blame for morally good reasons in the updating condition than in the control condition. When people received an affirmative update of an initially intentional event (compared with receiving the intentionality information right away) they arrived at a higher final judgment of blame than in the control condition. Contrastingly, when people received an affirmative update of an initially intentional event but with the *further* information of good reasons they arrived at a lower final judgment of blame than in the control condition. While the former result is congruent with a blame validation model, the predicted effect for morally bad reasons did not manifest. Additionally the latter result is opposite its predictions, and overall, this appears to be poor evidence for a blame validation model.

One critique of my interpretation of the comparison between the control and updating conditions is that I collapsed across trials where people may have anchored on low or high initial judgments of blame, and that this led me to erroneously conclude that there is no effect of blame validation when in fact there is. Therefore, I conducted two additional comparisons between the updating and control conditions. I examined people's final updated blame judgments for trials where the initial information would provide people with a low anchor for their blame judgment (e.g., burning a birthday meal) and for trials where the initial information would provide people with a high anchor for their blame judgment (e.g., killing). The low-anchor trials did not show a significant effect of condition (updating vs. control) ($F[1,96] = 1.83$, $p = .18$) and is therefore not discussed

further. The high anchor trials however, showed evidence of anchoring and insufficient adjustment (See Figure 12). There was a significant effect of condition in these trials, $F(1,96) = 11.6$, $p < .01$, $d = 1.22$. Final updated blame judgments were higher in the updating condition compared to the control condition, though the systematic effect of information was preserved ($F[5, 480] = 237.1$, $p < .001$), and there was no significant condition by information interaction ($F < 1$).

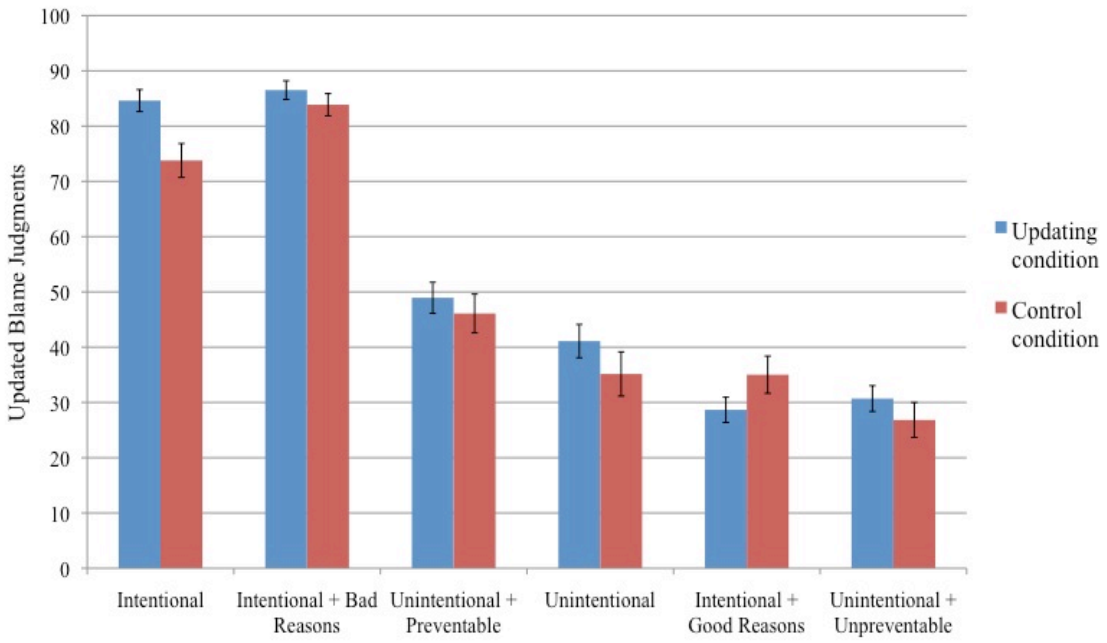

Figure 12. Updated blame judgments in the updating and control conditions (high anchor only).

Study 2 Discussion

This study largely replicated the results of Study 1. Consistent with the previous study, community participants updated much more frequently than expected from previous research (74.4%), contrasting with blame validation or post-hoc rationalization model (e.g., Alicke; 2000; Haidt, 2001). Further, the reaction time results lend additional support to the two-track structure of blame proposed by the Path Model (Malle, et al., in

prep). When people interpreted an event as being intentional, they processed subsequent information about reasons much faster than information about preventability, and the same pattern held for preventability information when people encoded an event as unintentional. The pattern of blame change and final blame judgments across these two studies demonstrates the powerful impact causal-mental state information has on moral judgment and they systematic way people use such information to update their moral judgment in contrast to the predictions of blame validation (Alicke, 2000; Haidt, 2001).

This study more explicit tested the blame validation hypothesis by comparing blame judgments from an updating condition to a control condition where a separate sample of community participants received all of the information at once and then made a moral judgment. Overall, the pattern of updated blame judgments across the two samples was remarkably similar. Whether people made a single judgment, or an initial judgment followed by updating, they arrived at comparable final evaluations of the agents. However, when the results were broken down into low- and high-blame anchor trials, the high anchor trials showed evidence for blame validating anchoring an adjustment. In these trials, initial blame judgments appeared to provide an anchor that people found it difficult to adjust down from. Despite this, people's updated blame judgments still tracked the particular causal-mental information provided by the new information. Moreover, the fact that anchoring and adjustment emerged only when actions produced a high initial blame judgment appears limiting for the blame validation as an account of moral judgment. While this study shows evidence for blame validation in moral updating, it also shows that the blame validation relies on highly evocative behaviors (e.g., incest, killing) to generate traction for the model. Updating moral judgments for everyday moral

infractions (e.g., minor physical harm, properly damage, broken social contracts) is best described by the systematic use of causal-mental information, not blame validation.

While this study successfully replicated the results of Study 1, it did not address the issue of long reaction times. Indeed, many of the reaction times were longer in this study than in Study 1. Study 3 addresses this concern.

Chapter 4

Study 3: Moral Updating with Audio Stimuli

One limitation in the two previous studies was the use of text stimuli, which produced long reaction times. Because of this, it is not entirely clear what processes are unfolding during the time when the new information and the updating judgment are available. An alternative explanation to the Path Model's switching account suggests that the processing demands are identical between the switching and non-switching trials, and the observed difference in reaction times is due to people rereading the new information more often in the switching trials.

Study 3 addresses this concern by converting the text stimuli into audio stimuli. This change accomplishes two goals. It removes the possibility that variations in processing RT's are caused by extraneous factors (e.g., participants' rereading information or variability in reading speed). Also, it provides a cleaner measure of participants' processing RT's on their way to updating their moral judgments. The study accomplishes this by immediately presenting the updating judgment screen after the conclusion of the new information audio, eliminating the ambiguity created by variations in participants' reading time.

With these changes, Study 3 seeks to replicate three key results from the previous studies. First, it tests whether the frequency of updating remains high when new information is presented aurally, instead of being displayed as text and in conjunction with the invitation to update a previous judgment. Second, it examines whether the Path model's prediction of two distinct paths to blame replicates by testing whether

information that necessitates intentionality switching is processed more slowly than information that does not require switching. Finally, this study seeks to replicate the systematic use of causal-mental state information in moral updating, which in the previous two studies was resistant to hypothesized anchoring and adjustment effects (Alicke, 2000; Haidt, 2001).

Method

Participants

As in Study 2, participants were recruited from the local community by advertising a "Paid research study" on Craigslist. Thirty-seven people responded to the ad and participated in the experiment. Six participants were omitted from the analyses for failing to follow instructions (final N =31).

Of the remaining 31 participants, 11 were female. The majority of participants (74%, N = 23) identified themselves as white, with small numbers of participants identifying as Black (3%, N = 1) Latin/Hispanic (7%, N = 2,) or multi-ethnic (10%, N = 3). The sample had an average age of 32.3 years (SD = 13.1) and represented a diverse range of education: 51.6% (N = 16) reported finishing high school as their highest level of education; 6.5% (N = 2) attained a 2-year degree; 35.5% (N = 11) attained a 4-year degree; and 6.5% (N = 2) attained a Master's degree or higher. The sample's political attitudes were slightly liberal, averaging M = 6.29 (SD = 1.65) on a 1-9 scale (1 = Extremely conservative; 5 = moderate; 9 = Extremely liberal), and they were below the midpoint on religiousness (M = 1.19, SD = 1.08) on a 0-4 scale (0 = not at all religious; 4 very religious).

Procedure and Materials

Study 3 replicated the general structure of the previous studies, with important changes as noted. As in previous studies, participants were run in small groups of two to six people at a time. All participants received a verbal explanation of the experimental task prior to beginning the task. Then participants were ushered into an individual testing room, given additional instructions, and allowed to begin the experiment. Participants read a brief set of instructions describing the task and how to use the slider bar for their judgments. Following these instructions, participants proceeded through three practice trials, 36 experimental trials, and completed a demographic questionnaire.

In contrast to the previous experiments, the screen where participants learn new information and update their initial judgment was broken into two separate screens for a total of five screens per trial (See Figure 13). Thus, the procedure for a single trial was that participants listened to a brief description of a norm violating event (Screen 1), made their initial moral judgment (Screen 2), heard new information about the event (Screen 3), updated their initial moral judgment (Screen 4), and then typed a description of the event (Screen 5). All of the event descriptions and new information were presented aurally using sound-blocking headphones. The stimuli were recorded using a female voice with neutral affect. The presentation of the Screens 1 and 3 were synchronized with the length of the audio they presented. The only visual information displayed on these screens was the text "Listen to this information." The program automatically advanced to the next screen when the audio concluded. Participants were not allowed to go back and revisit previous judgments or information, eliminating the concern that differences in processing RT's were driven by participants differentially revising the new information.

Participants made their initial blame judgments using the click-and-drag slider bar from Studies 1 and 2 (Endpoints: 0, no blame at all – 100, the most blame you would ever give). The slider bar's initial cursor position was always at zero. For each initial judgment I measured three variables: (1) the time from the presentation of the judgment request to when the participant first clicked the cursor (the processing RT), (2) the time from the participant's first click to submitting the response (the change RT), and (3) participants' completed blame judgment (i.e., the final position of the cursor).

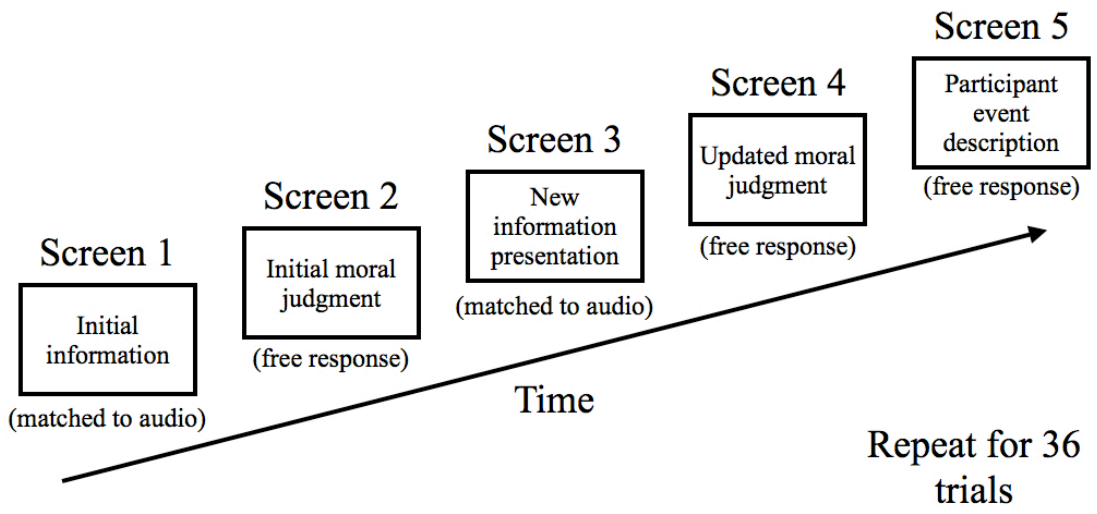

Figure 13. Study 3 updating paradigm.

After participants submitted their initial moral judgment, they listened to new information about the norm-violating event (manipulated within-subjects). When the audio concluded, the program automatically advanced to the updating judgment (Screen 4). The content of the new information was identical to the previous studies. New information revealed details about whether the agent acted intentionally vs. unintentionally, about the valence of her reasons (e.g., good or bad), and about whether the agent could have (reasonably) prevented the outcome. Counterbalancing across forms was also identical to the previous studies.

After hearing the new information, participants were allowed to update their initial judgment. Participants were shown their initial judgment and allowed to adjust the cursor up or down or to submit their judgment without making any changes. The slider bar was identical to the one they used for their initial moral judgments, and for each updated judgment I measured participants' processing RT, the change RT, and participants' updated blame judgment. As in the previous studies, I calculated a blame change score for each trial by subtracting participants *initial* blame judgment from their *updated* blame judgment.

At the end of each trial participants were asked to briefly describe the event by typing ("Describe in your own words what X did"). After participants completed all 36 experimental trials they responded to a brief demographics questionnaire and were thanked and debriefed.

Results

Frequency of updating

The high frequency of updating replicated the previous studies' results. Participants updated in the majority of trials (77.5%), and updating frequency was not correlated with participants' age ($r = .17$, $p = .36$), level of education ($r = .25$, $p = .17$), or religiosity ($r = -.13$, $p = .48$). Updating frequency was significantly correlated with participants' political orientation ($r = .48$, $p < .05$). Participants who identified as more liberal updated their judgments more frequently than participants who identified as more conservative. However, in contrast to the previous studies, the frequency of updating was not affected by the new information or by switching ($Fs < 1$).

Updating blame judgments

The major goal of this study was to test whether people's updated moral judgments replicated the systematic pattern of change produced in Studies 1 and 2, or if people anchor and adjust their judgments when they are not able to reread the new information. As in the previous studies, people's initial blame judgments strongly tracked the intentionality semantically implied by the verb of the event description, $t(30) = 4.64$, $p < .001$, $d = .54$. People's blame judgments were less intense for events implied to be unintentional ($M = 53.6$, $SD = 20.4$) than events implied to be intentional ($M = 64.4$, $SD = 19.2$).

The change in blame intensity from the initial to the final judgment was tested using a 2 (switching) x 6 (information) within-subjects ANOVA. The type of new information participants received strongly predicted blame change, $F(5,150) = 4.9$, $p < .001$ (see Figure 14). As in the previous studies people systematically updated their judgments of blame in response to new information. Blame increased in response to information indicating that the event was intentional, and even more so when the agent also had morally bad reasons for acting intentionally. Blame decreased in the other four conditions. The pattern for blame reduction from the previous studies was likewise preserved. Blame was reduced least in response to information indicating the event was unintentional but preventable; blame was reduced further in response to information about the event being unintentional, or in response to an agent's morally good reasons for acting, and blame reduction was strongest in response to information about the event being unintentional and unpreventable (see Figure 14). This result contrasts with the blame validation hypothesis. People updated their blame judgments in response to both mitigating and exacerbating information. Further, the magnitude of change was

equivalent regardless of whether the change was in the direction of reducing blame or increasing it.

Additionally, switching from events' implied intentionality value to the opposite intentionality value explicitly stated in new information had the effect of amplifying blame change in response to new mental state information, $F(1, 150) = 8.19$, $p < .01$. Lastly, there was a significant switching by information interaction, $F(5,150) = 4.2$, $p < .001$. Generally, intentionality switching amplified blame change across the information conditions; however, information about morally good reasons showed an opposite pattern. As in the previous studies, the effect of morally good reasons and switching intentionality tracks tended limit one another, though it is unclear from these results whether this is an effect of intentionality limiting the effect of morally good reasons on blame change or if it is an effect of the strong mitigation of good reasons turning around the otherwise strong increase in blame shown for intentionality-only information.

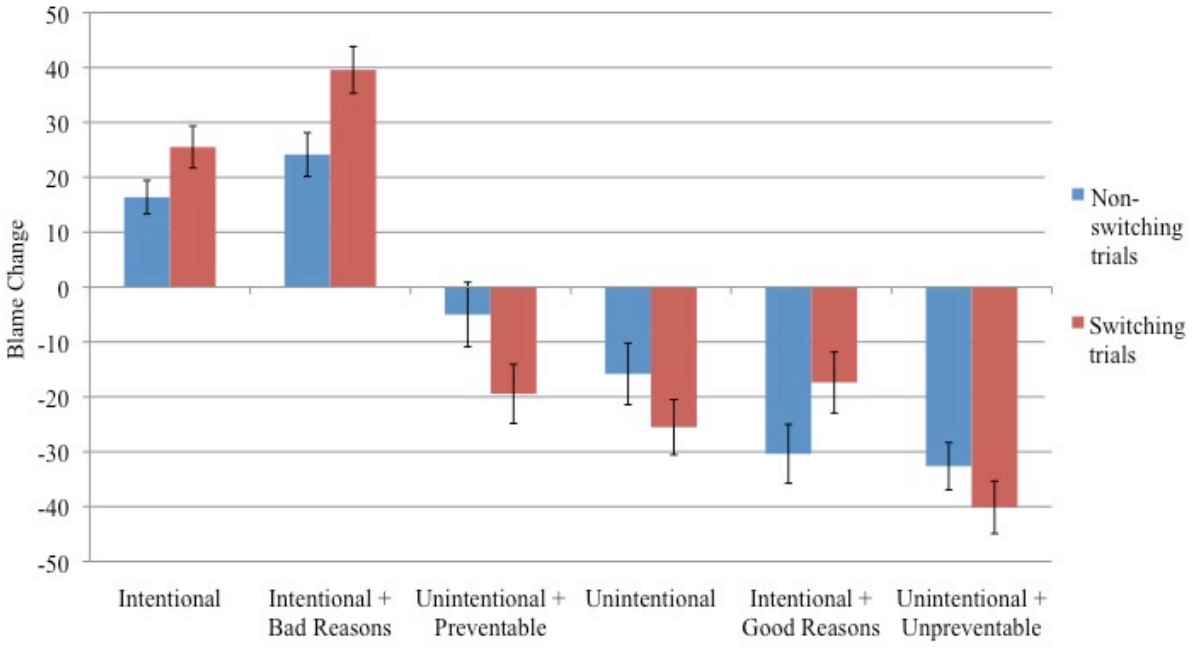

Figure 14. Blame change across information conditions and switching trials.

Lastly, I examined people's final blame judgments (i.e., blame judgment post-updating). A within subjects ANOVA showed a significant effect of information on people's final updated blame judgments $F(5, 150) = 73.3$, $p < .001$. Blame was highest after reading about someone who had morally bad reasons for acting ($M = 89.7$, $SD = 12.6$) or about someone who intentionally caused harm ($M = 79.9$, $SD = 21.2$), followed by reading about someone who caused unintentional but preventable harm ($M = 46.1$, $SD = 19.9$) or someone who unintentionally caused harm ($M = 38.0$, $SD = 26.1$). Blame was lowest when there was a morally good reason for the action ($M = 30.9$, $SD = 18.5$), or when the harm was unintentional and unpreventable ($M = 25.1$, $SD = 193$).

Blame updating: Processing speed

The critical test of the Path model's two-track prediction is that switching intentionality tracks is cognitively costly, which is reflected in longer processing RT's in switching trials relative to non-switching trials. I tested this hypothesis by aggregating the reaction times for the switching trials and compared them with those from non-switching trials. Replicating the previous studies, participants were significantly slower processing new information on their way to updating their moral judgments in the switching trials ($M =$2.91 sec, $SD = .82$) relative to the non-switching trials ($M = 2.47$ sec, $SD = .86$), $t(31) = 3.17$, $p < .01$, $d = .52$. This pattern held for many of the individual information conditions as well (See Figure 15). People were slower processing information about an agent's reasons (both morally good and morally bad) for acting or information that cast the event as unintentional and unpreventable ($Fs > 4.3$, $ps < .05$) in the switching trials compared with non-switching trials. While trending in the predicted direction, processing speed in response to intentionality information (either affirming or denying intentionality)

or information about unintentional but preventable outcomes did not show an effect of switching (*Fs* < 1).

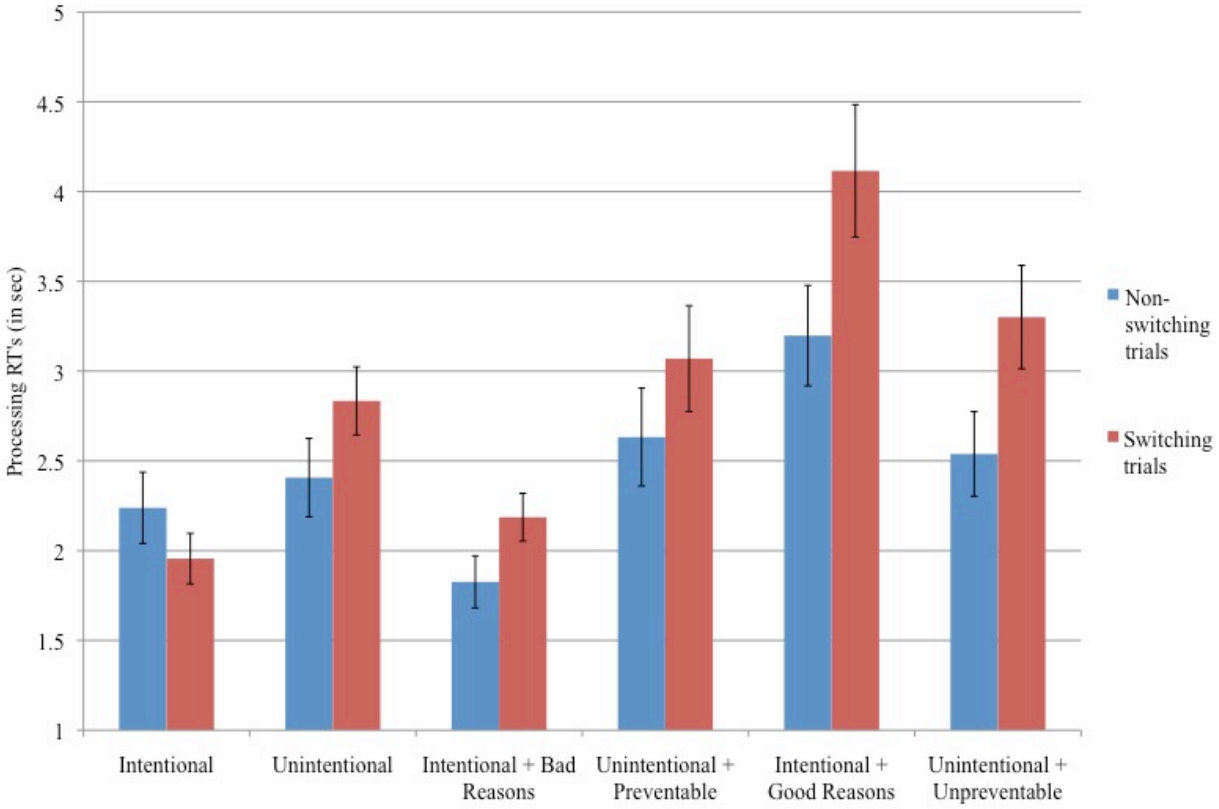

Figure 15. Processing RT's by intentionality switching and updating information conditions.

Participants' change RT (i.e., the time between the first click and submitting the updated judgment) did not differ between switching (M = 1.40 sec, SD = .17) and non-switching trials (M = 1.41 sec, SD = .29), $t(31) = .349$, $p = .73$, $d = .04$. But the switching pattern held for the total amount of time participants spent updating their judgment, $t(31) = -3.16$, $p < .01$, $d = .49$. As expected, participants were slower overall in switching trials (M = 3.87 sec, SD = .88) relative to non-switching trials (M = 4.33 sec, SD = .98).

Interim Discussion

In three studies I tested how people update their moral judgments in response to

new causal-mental information. These studies examined the phenomenon of moral updating using text stimuli (Studies 1 and 2) and audio stimuli (Study 3), using both student (Study 1) and community samples (Studies 2 and 3). Several novel findings emerged from these studies.

First, updating is more common than previously hypothesized, and it was more frequent when new information violated people's expectations about intentionality or their assumptions about the typical causes of negative events (e.g., having morally good reason for intentionally causing harm). The result of expectation-violation increasing updating is predicted by models that describe blame as a process of assigning meaning to events (Malle, et al., in prep; Weiner, 1995). On these models, revealing information that could not be easily inferred from the initial event description should prompt updating as people attempt to understand *why* an event happened and how to respond. The result is not predicted, however, by models that conceptualize moral updating (or moral judgment more generally) as being limited by blame validation or post-hoc rationalization (e.g., Alicke; 2000; Haidt, 2001). These models predict that people should be sensitive to information that validates a previous blame judgment but should discount information that contradicts it. The blame change and updated blame judgment data in the previous three studies do not support such a view.

The second main finding of the preceding studies is that people show a reliable difference in their processing speed for new information when the new information constitutes an intentionality switch. This finding uniquely supports the Path Model of Blame (Malle, et al., in prep), which first hypothesizes that there are two distinct tracks of blame (intentional and unintentional), and second, that switching between these tracks

carries a cognitive cost. The other prominent moral judgment models (Alicke, 2000; Haidt, 2001; Weiner, 1995) lack strong processing hypotheses. Instead, they predict that processing speed is a function of factors such as quantity, complexity, or uncertainty, which were held constant between the switching and non-switching conditions.

The third main finding is that people are systematic in their use of new causal-mental information to update their blame judgment. They do not anchor and adjust. Rather, consistent with previous research on single moral judgments, people's updated judgments respond to affirmations or denials of intentionality, the level of justification of an agent's reasons, and whether the negative event in question was preventable (Cushman, 2008; Guglielmo & Malle, 2010a; Greene, et al., 2009; Lagnado & Channon, 2008; Monroe & Reeder, 2011; Young & Saxe, 2009). Further, these studies go beyond previous research by uniquely showing that the effect of information on blame depends on what precedes it. Across the three studies, blame change was higher when people switched tracks (in either direction) than when people used new causal-mental information to update a judgment within a track.

All of the present studies examined people's private blame judgments in cognitively optimal conditions. Yet, when people make blame judgments they do so while performing other social and cognitive tasks. Thus, one could critique the previous studies as failing to reflect the process of moral updating as it occurs in everyday life where people are rarely able (or motivated) to think carefully about their judgments. Such distracting conditions may curtail updating efforts or eliminate the systematic way people process and use new causal-mental information. Study 4 addresses this critique and investigates the whether moral updating is bounded by cognitive load.

Chapter 5

Study 4: Moral Updating under Cognitive Load

Study 4 addresses the critique that the updating findings described above are dependent on cognitively optimal conditions rarely found in everyday social life. It tests this possibility by examining if and how people update their moral judgments when under cognitive load. The experiment manipulated cognitive load within subjects using a blocked design. In the block of cognitive load trials, participants were asked to produce a series of random taps with their index finger, a procedure that specifically targets executive function load, not secondary mechanisms such as the phonological loop (Stuyven, Van der Goten, Vandierendonck, Claeys, & Crevits, 2000). In the block of load-free trials participants made their moral judgments as usual.

This experiment examined whether moral updating depends on effortful cognition, or whether people update their moral judgments under conditions that favor automatic processing. Specifically, the study tested three predictions. First, if moral updating relies on effortful, consciously controlled reasoning, then cognitively loading participants should reduce the overall frequency of updating. However, if updating is common and easily performed in everyday life, people should update their moral judgments in response to new information at similar rates even when they are under cognitive load. Second, this study tested whether cognitive load disrupts the systematic way people use new information to update the intensity of their blame judgments. It is possible that people's default method for updating is best characterized by blame validation; however, when people have ample time and cognitive resources they are able to override this tendency. If this is true, then cognitively loading people should disrupt

the systematic pattern of blame change and blame updating in favor of blame validation. If however, if updating does not rely on effortful cognition, then the pattern of blame change and blame updating found in the previous studies should be preserved.

Third, The Path Model predicts that people easily step through the process of moral judgment; however, switching intentionality tracks requires stepping "backwards" in the default processing order, which is hypothesized to be cognitively costly. Following from this hypothesis, I predict that cognitive load should selectively interfere with people's ability to switch between intentionality tracks, further slowing processing in switching trials while leaving non-switching trials unaffected.

Method

Participants

Participants were recruited from the local community using an ad posted on Craigslist. 42 people responded and participated in the experiment. Nine participants were omitted from the analyses for failing to follow instructions ($N = 5$) or for failing to tap randomly ($N = 4$) during the cognitive load trials (final $N = 33$).

Of the remaining participants, 13 were female. The majority of participants (70%) identified as White, with the remaining participants identifying as Black (15%), multi-ethnic (9%), or Native American (3%). The sample had an average age of 34.2 years ($SD = 11.1$) and represented a diverse range of education: 46% ($N = 15$) reported finishing high school as their highest level of education; 15% ($N = 5$) attained a 2-year degree; 24% ($N = 8$) attained a 4-year degree; and 9% ($N = 3$) attained a Master's degree or higher. The sample's political attitudes were moderate, averaging $M = 5.16$ ($SD = 1.91$) on a 1-9 scale (1 = Extremely conservative; 5 = moderate; 9 = Extremely liberal),

and they were slightly below the midpoint on religiousness (M = 1.18, SD = 1.21) on a 0-4 scale (0 = not at all religious; 4 very religious).

Materials and Procedure

The experiment manipulated cognitive load within-subjects using a blocked design. Participants completed 18 control trials followed by 18 cognitive load trials (specific behaviors were counterbalanced across blocks). The procedure for the control trials was identical to Study 3 – participants listened to a description of a norm-violating event, made an initial moral judgment, heard new information about the event, were invited to update their moral judgment, and finally they typed their description of the event. In the cognitive load trials, participants listened to the event description and made their initial moral judgments normally; however, during the presentation of the new information participants were instructed to produce a series of random taps with their left hand (see Figure 16).

The cognitive load manipulation. After participants completed the 18 control trials they saw a screen instructing them to exit the testing room and inform the experimenter that they had finished the first half of the experiment. At this point the experimenter explained the cognitive load task. The task, labeled the "tapping task," was adapted from Stuyven et al. (2000). Participants were asked to produce a series of random taps with their left hand. They were instructed to tap at least once every second and to try to be as random as possible. Then, with the experimenter present, participants listened to two audio examples of a random series of taps and then practiced producing random taps on their own. The experimenter gave verbal feedback on their tapping (e.g., "That was good, but really concentrate on being as random as you can.").

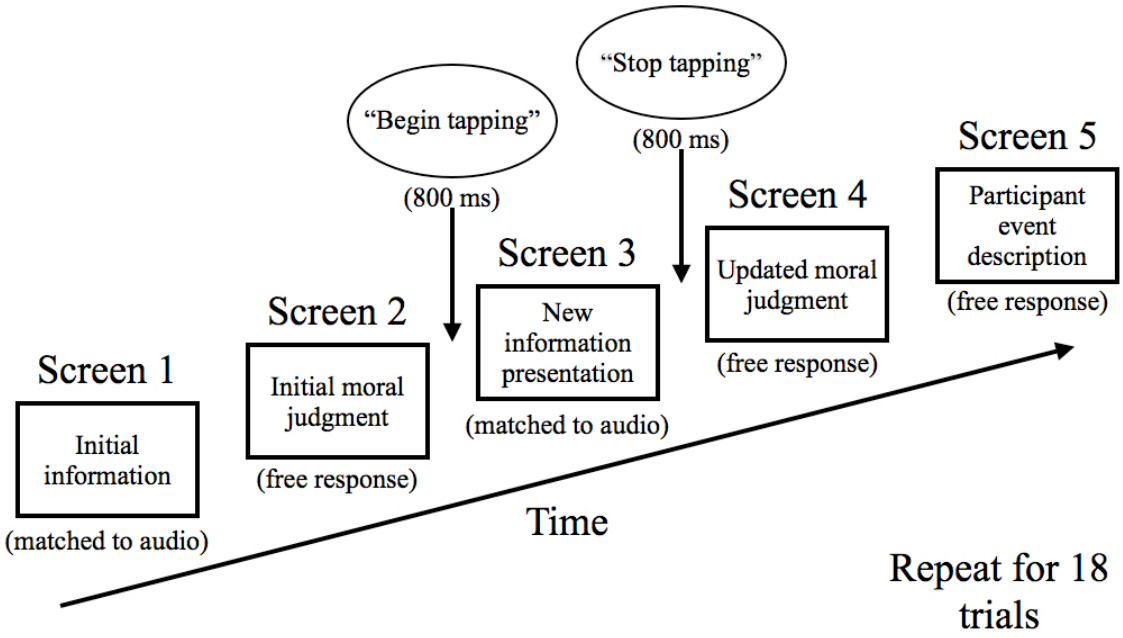

Figure 16. Experimental paradigm for the cognitive load trials of Study 4.

After practicing, the experimenter left the room and participants were allowed to complete the cognitive load trials. As in the control trials, participants listened to the event description and made their initial moral judgments. Prior to the presentation of the new information, participants were cued to begin tapping by a male voice instructing them to "Begin tapping" (700 ms)[8]. After the presentation of the new information, the same male voice instructed participants to "Stop tapping" (800ms). Immediately after the participants were instructed to stop tapping, they were asked to make their updated judgments, and finally to type their description of the event.

Results

This study tested the effects of a cognitive load manipulation on the process of moral updating. Specifically, this study tested three questions: (1) Is the overall frequency

[8] I recorded participants' tapping behavior during the experiment. Participants who reliably failed to tap or who merely tapped rhythmically in the majority of the cognitive load trials were omitted from the analyses.

of updating preserved even when participants are placed under cognitive load? (2) Does cognitive load eliminate the systematic way people use new information to update the intensity of their blame judgments? (3) Does cognitive load interfere with participants' ability to switch between intentionality tracks?

The data answer the first question in the affirmative. Results show that the frequency of updating was consistent across conditions. The frequency of updating in the cognitive load trials (78.6%) was comparable to the control trials (80.9%), $t(32) = 1.04$, $p > .3$, $d = .13$.

To answer the second question I examined people's initial, change, and updated blame judgments. As in previous studies, the intentionality of the event description drove initial blame judgments, $F(1,32) = 8.77$, $p < .01$, $d = .25$. Neither cognitive load ($F[1,32] = 2.20$, $p > .15$) nor the load by intentionality interaction ($F[1,32] = .23$, $p > .6$.) were significant, indicating that the cognitive load manipulation did not affect people's initial judgments of blame. Intentional behaviors ($M = 62.8$, $SD = 19.6$) received higher initial blame judgments relative to unintentional behaviors ($M = 57.8$, $SD = 20.6$).

Blame change was strongly predicted by the type of new information participants were given, $F(5, 160) = 42.1$, $p < .001$ (See Figure 17). People largely maintained their systematic use of causal-mental information across the control and the load conditions; however, information about intentionality and unintentionality largely failed to shift people's judgments of blame. People increased the intensity of their blame judgment in response to information about an agent's morally bad reasons for acting, and decreased blame in the remaining three conditions. Mirroring previous studies, blame reduction was smallest in response to information indicating that an event was unintentional though

preventable; it was larger when the outcome was both unintentional and unpreventable; and blame was reduced most when the new information provided a morally good reason for acting.

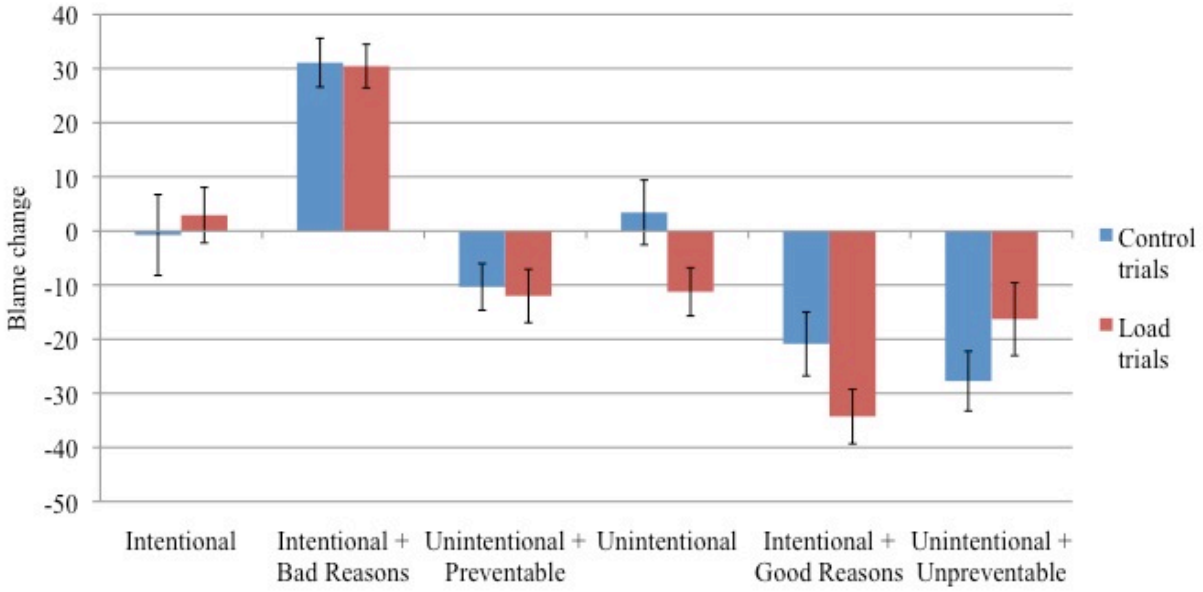

Figure 17. Blame change in cognitive load and control trials across information conditions.

Cognitive load did not independently affect blame change, (F[1, 32] = 2.2, p = .14), but there was a significant information by cognitive load interaction, F(5, 160) = 4.92, p < .001. Relative to control trials, cognitive load produced larger reductions in people's blame judgments in response to morally good reasons and information about unintentionality (ps < .01). Conversely, blame change was larger in the control trials than in the cognitive load trials for information indicating the event was unintentional and unpreventable (p < .05).

Final blame judgments reflected similar effects (See Figure 18). Information had a large effect on participants updated judgments of blame, F(5, 160) = 47.3, p < .001. Cognitive load did not have a significant effect, F(1, 32) = .31, p > .5. However, there

was a significant load by information interaction, $F(5, 160) = 4.52$, $p < .001$. Irrespective of cognitive load, people arrived at similar final blame judgments in the intentional-only, intentional with morally bad reasons, and unintentional-preventable information conditions. The cognitive load manipulation resulted in higher updated blame judgments in the unintentional-unpreventable information condition ($p < .001$) and trended toward slightly lower updated blame judgments in the unintentional-only ($p = .1$) and intentional but morally good reasons ($p = .06$) information conditions.

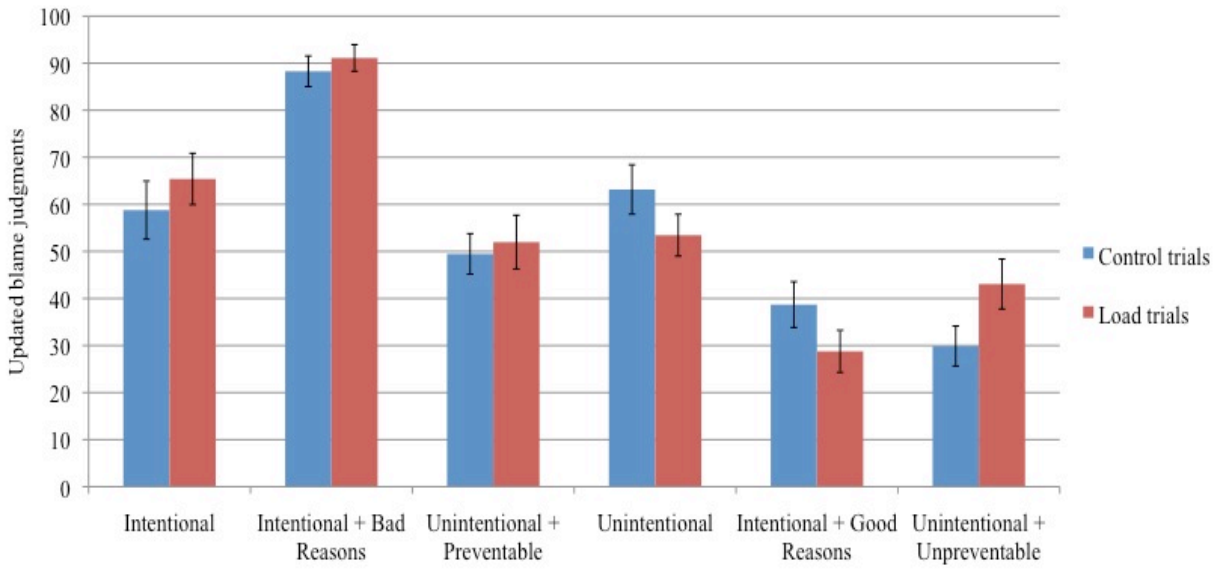

Figure 18. Final, updated blame in cognitive load and control trials across information conditions.

The reaction time data speak to the last question. A 2 (cognitive load) x 2 (switching) within subjects ANOVA revealed an effect of switching, ($F[1, 32] = 12.1$, $p < .001$), but no effect of cognitive load, $F(1, 32) = .48$, $p > .4$. As in previous studies, people were slower overall when processing new information in the switching trials compared to the non-switching trials, though people were equally fast processing new information in the load and the control trials.

While there was no main effect of cognitive load on processing speed, there was a significant load by switching interaction, $F(1, 32) = 7.10$, $p < .01$. I replicated the effect of switching on processing RT's in the control trials – processing speed was slowed in the switching trials relative to the non-switching trials ($t[32] = 4.80$, $p < .001$, $d = .74$); however, the cognitive load manipulation eliminated this effect, $t(32) = 1.04$, $p > .3$, $d = .20$ (See Figure 19). Interestingly, the load manipulation did not slow processing in the switching trials; rather, the load manipulation slowed processing of the non-switching trials relative to control ($t[32] = 2.37$, $p < .05$, $d = .44$), making them comparable to the switching trials.

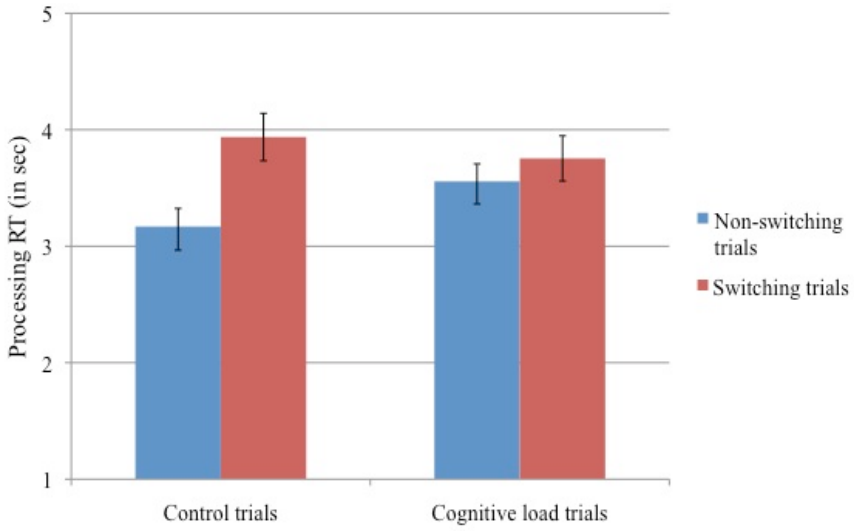

Figure 19. Processing RT's for control and cognitive load trials.

A closer examination of the processing RT data showed that the effect of cognitive load on processing in the non-switching trials was primarily due to slower processing of new information involving unintentional outcomes (see Figure 20). People were slower processing unintentional-only ($p = .06$, $d = .46$), unintentional-preventable ($p = .06$, $d = .44$), and unintentional-unpreventable information ($p = .13$, $d = .32$). This may suggest that people attempted to adopt a more systematic processing strategy to

compensate for the effect of the cognitive load on their ability to accurately process new information.

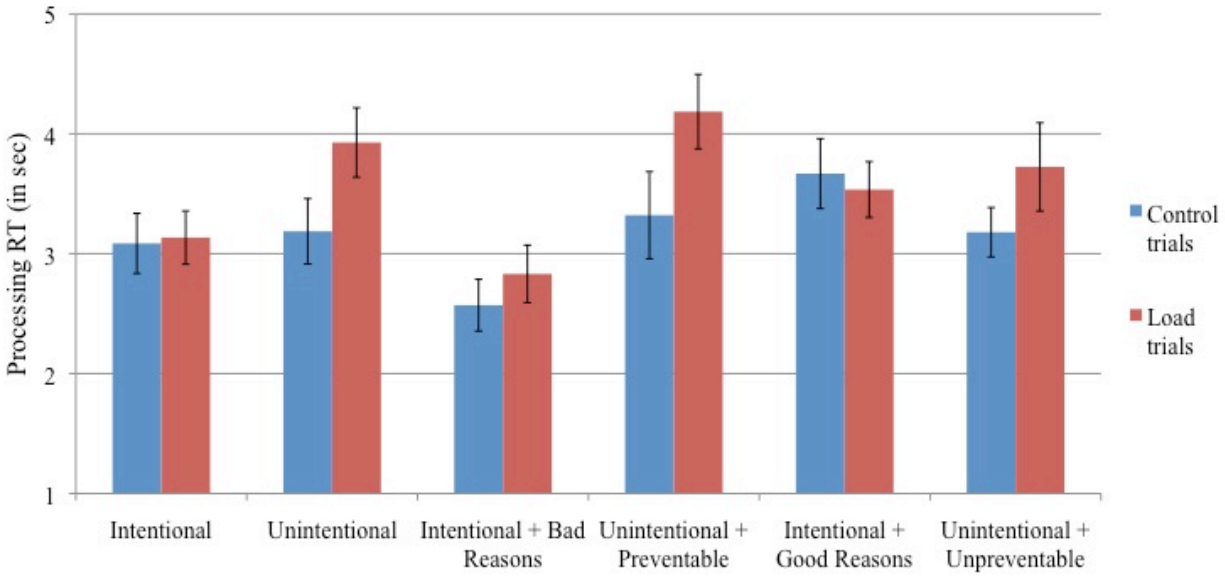

Figure 20. Processing RT's for non-switching trials across cognitive load and control trials.

Study 4 Discussion

Study 4 tested the effects of cognitive load as a possible boundary condition of moral updating. Overall, the results, with some exceptions, hew closely to the previous studies and support the view that people do update their moral judgments even when under cognitive load. The frequency data most clearly support this account, as people's propensity to update their judgments did not differ between the control and the cognitive load conditions. Similarly, blame change and updated judgments were sensitive to new causal-mental state information, though in contrast to the previous experiment, intentionality information (either affirming or denying intentionality) had little effect on blame change in either control or load trials. However, the critical question in this study pertained to blame change. Specifically, whether the introduction of a cognitive load

manipulation would disrupt the systematic way people use new information, favoring a blame validation process instead. The data do not support such a conclusion. Despite the muted effects of intentionality, there were no significant effects of cognitive load on blame change and updated blame judgments.

The processing RT data provided surprising finding. The control trials replicated the general pattern from the previous studies of people incurring a processing cost when switching intentionality. However, the cognitive load manipulation eliminated this effect. Processing speed for new information was equal across switching and non-switching trials, which does not support the Path Model's prediction of selective interference in the switching trials. However, the results also do not support the alternative account put forward by Alicke (2000), Haidt (2001) and Weiner's (1995) models as processing RT's were not significantly longer under cognitive load than in the control trials.

One possible explanation for these findings is that the load manipulation caused people to adopt a more systematic processing strategy in response to the cognitive load manipulation. Switching intentionality triggers this systematic process on its own, but in the non-switching trials, people fought the effect of the cognitive load manipulation by more thoroughly processing information, especially information that they might not be generally very confident about (e.g., unintentional outcomes). This would explain why people were selectively slower in the non-switching trials, and why this effect was driven by people's slower processing of information involving unintentional outcomes. Evaluating whether an outcome was accidental or preventable requires considering possible counterfactuals. Was it *really* an accident? Could the person have known what would happen? It is possible that cognitive load limits people's ability to carefully reason

about such cases. The blame change data offer limited support for this conclusion. Blame change in response to information about unpreventable and preventable outcomes did not differ from each other when participants were under cognitive load.

Together, the reaction time and blame change data suggest that the cognitive load manipulation may have caused people to adopt a more systematic processing strategy to information that is usually less effortful to process. Additionally, while people fought the effects of cognitive load by extending their processing of new information, they were not completely successful. I test these hypotheses using a different form of cognitive load in Study 5.

Chapter 6

Study 5: Moral Updating when Socially Accountable for Judgments

Thus far I have examined the phenomenon of updating under cognitive optimal conditions using text (Studies 1 and 2) and audio stimuli (Study 3) as well as testing a potential bounding condition of updating: cognitive load (Study 4). Results across these studies show that updating is a robust phenomenon that appears to operate even under conditions favoring automatic processing (i.e., cognitive load). However, Study 4 showed two effects that were not common to the previous studies. First, in both the control and updating conditions, people did not update their moral judgments in response to information about agents acting intentionally or unintentionally. This contrasts with the results from Studies 1-3, where people showed a consistent pattern of reducing blame in response to new information that indicated an action was unintentional and increasing blame in response to information labeling an action as intentional. Second, contrary to my prediction regarding the effect of load on reaction times for switching and non-switching trials, the cognitive load manipulation slowed non-switching trials, while switching trials remained unaffected. I interpreted this effect as participants fighting against the load manipulation and taking additional time to "check" the easier judgments in an effort to arrive at the correct judgment.

Given these two surprising findings, I conducted an additional study to test the effects of a different type of cognitive load (social accountability) on updating. Meting out appropriate moral judgments may be viewed (by the blamer) as an exercise in impression management because blaming is itself governed by social norms. Moral

judgment is, at least at times, a social act with the goal of teaching rules and modifying future behavior (Cushman, in press; Duff, 1986), and to accomplish these goals, blame must be expressed socially. However, such public judgments can carry social costs (e.g., ostracism, reactive aggression), especially if a judgment is viewed as unjust. That is, while people may arrive at their blame judgment in their heads as usual, making a public declaration of blame carries with it social costs. Blaming too harshly or too leniently may be viewed as unfair and may itself result in blame. Thus, when they publicly blame others, people must also attempt to monitor what an "acceptable" judgment will be to onlookers, which we use as a manipulation of cognitive load.

Study 5 had two main goals: I wanted to replicate the systematic pattern of blame change and updated blame judgments from Studies 1-3, specifically showing that people use information about intentionality to update the intensity of their blame judgments. Additionally, this study will test the blame validation hypothesis that updating relies on effortful cognition people will revert to blame validation for their blame change and updated blame judgments. Second, I wanted to test the hypothesis that people are at an RT advantage in the non-switching trials when cognitive conditions are optimal; however, placing people under cognitive load makes the non-switching trials more difficult, eliminating the RT advantage. If this hypothesis is correct it would predict the usual pattern of slower RT's for switching trials relative to non-switching trials in the control block, but equal RT's for switching and non-switching trials when participants are cognitively loaded.

Method

Participants

Participants were recruited from the local community using an ad posted on Craigslist. 40 people responded to the ad and participated in the experiment. Two participants were omitted from the analyses for failing to follow instructions (final *N* =38).

Twelve of the 38 participants were female. The majority of participants (63%) identified themselves as White, with fewer participants identifying as Hispanic (16%), multi-ethnic (10%), Black (8%), or Middle Eastern (3%). The sample had an average age of 32.6 years (*SD* = 12.5) and represented a diverse range of education: 55% (*N* = 21) reported finishing high school as their highest level of education; 10% (*N* = 4) attained a 2-year degree; 16% (*N* = 6) attained a 4-year degree; and 10% (*N* = 4) attained a Master's degree or higher. The sample's political attitudes were moderate, averaging *M* = 5.4 (*SD* = 1.86) on a 1-9 scale (1 = Extremely conservative; 5 = moderate; 9 = Extremely liberal), and they were slightly below the midpoint on religiousness (*M* = 1.63, *SD* = 1.30) on a 0-4 scale (0 = not at all religious; 4 very religious).

Materials and Procedure

The experiment manipulated whether participants were socially accountable for their judgments using a within-subjects blocked design. Participants completed 18 control trials followed by 18 accountable judgment trials. The procedure for the control trials was identical to Studies 3 and 4 (control trials) – participants listened to a description of a norm violating event, made an initial moral judgment, heard new information about the event, were invited to update their moral judgment, and finally typed their description of the event. For the accountable judgment trials, participants

listened to the event description and made their initial moral judgments normally; however, prior to the presentation of the new information they were informed that the their next judgment would be shared with the experimenter. With this in mind, participants listened to the new information, were invited to update their judgment, and then provided a spoken justification for their updated judgment (See Figure 21).

The social accountability manipulation. After participants completed the 18 control trials they saw a screen instructing them to exit the testing room and inform the experimenter that they had finished the first half of the experiment. At this point the experimenter explained the social accountability task for the second half of the experiment. Each of the participant's second judgments (i.e., their updated judgment) would be transmitted to the experimenter who was evaluating the judgments for their "fairness." Additionally, participants were informed that instead of typing a description of the event as they had done in the previous 18 trials, they would now be asked to justify each of their second judgments by speaking their justification into the microphone attached to their headset.

After explaining the task and verifying that participants understood that only their second judgments in each trial would be shared, the experimenter left the room and participants were allowed to complete the 18 accountable judgment trials. The procedure for the initial event description and initial moral judgment were identical to the control trials. However, prior to the presentation of the new information, an audio file of a male voice informed them that the "The next judgment will be shared with the experimenter" (2000ms). Participants then listened to the new information and were subsequently invited to update their initial judgment. After they submitted their judgment a screen

displayed the text "Response received" accompanied by audio of the same (800ms). Lastly, participants were asked to speak a justification for their updated judgment into the microphone in their headset.

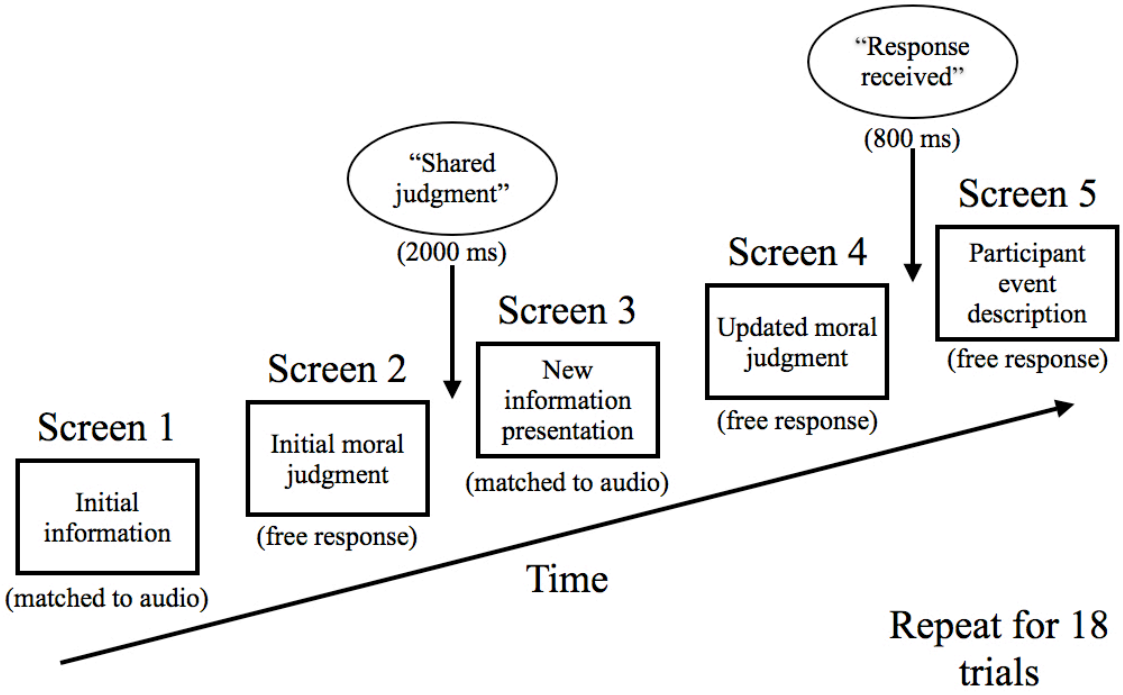

Figure 21. Experimental paradigm for the accountability judgment trials in Study 5.

Results

This experiment how people update their moral judgments when distracted by being socially accountable for their judgment. Following, from Study 4, this study tests the hypothesis that while people are at an RT advantage in non-switching trials when cognitive conditions are optimal, placing them under cognitive load makes the non-switching trials more difficult, eliminating the RT advantage. Additionally, this study sought to replicate the systematic pattern of blame change and updated blame judgments

from Studies 1-3, specifically showing that people use information about intentionality to update the intensity of their blame judgments.

A 2 (judgment type: public vs. private) x 2 (switching) within-subjects ANOVA tested the first hypothesis. Overall, people were slower processing new information in the accountable judgment trials compared to the control judgment trials, $F(1,37) = 15.3$, $p < .001$, $d = .49$. Switching, however, did not by itself affect processing speed, ($F[1,37] = .76$, $p > .3$, $d = .11$), but it did significantly interact with judgment type, $F(1,37) = 6.69$, $p < .05$, $d = .3$. Study 5 replicated the effect of switching on processing RT's for control judgment trials. People took longer to update their judgments of blame in the switching trials ($M = 3.13$ $SD = .86$) than in the non-switching trials ($M = 2.81$ $SD = .76$) ($p < .05$, $d = .39$). In the accountable judgment trials, however, people took equally long to process new information across switching ($M = 3.28$ $SD = .70$) and non-switching ($M = 3.42$ $SD = .76$) trials ($p > .1$, $d = .19$). This replicates the effect from Study 4 showing that people attempt the cognitive load manipulation makes the non-switching trials more difficult and eliminates the RT advantage found in the control judgment trials (See Figure 22).

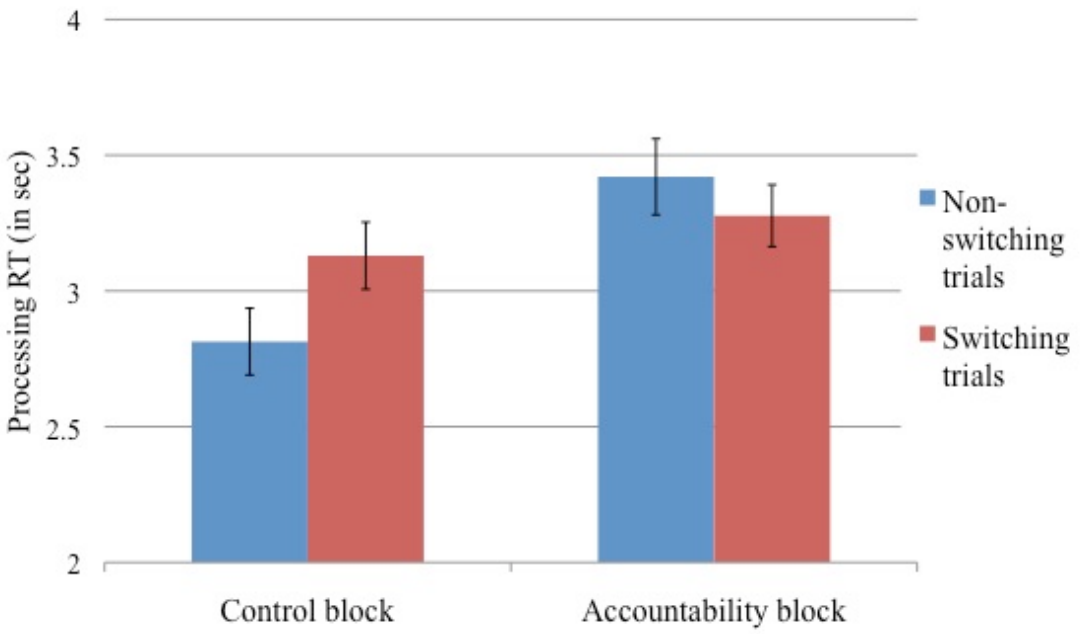

Figure 22. Processing RT's for switching and non-switching trials across accountability and control blocks.

The second hypothesis pertains to the pattern of blame change and people's updated blame judgments. Before testing this hypothesis, I verified that people differentiated between the intentional and unintentional event descriptions, and that the public judgment manipulation did not affect initial blame judgments. A 2 (intentionality) x 2 (judgment type) within subjects ANOVA confirmed that, as in previous studies, the people viewed intentional norm violating events (M = 63.6 SD = 17.1) as more blameworthy than unintentional violations (M = 58.8 SD = 18.9), $F(1,37) = 5.34$, $p < .05$. Further, neither judgment type ($F[1,37] = 1.50$, $p > .2$), nor the interaction with intentionality had any impact on updating judgments, $F(1,32) = 1.74$, $p > .2$).

A 2 (judgment type: control vs. accountable) x 6 (information) within-subjects ANOVA showed that blame change was strongly predicted by the new information, ($F[5, 185] = 125.0$, $p < .001$), but it was not affected by judgment type ($F[1, 37] = 2.24$, $p >$

.1), nor by the interaction between information and judgment type, $F(5, 185) = .26$, $p > .9$. Regardless of whether people were accountable for their judgments or not, they maintained their careful use of the new information update their blame judgments (See Figure 23).

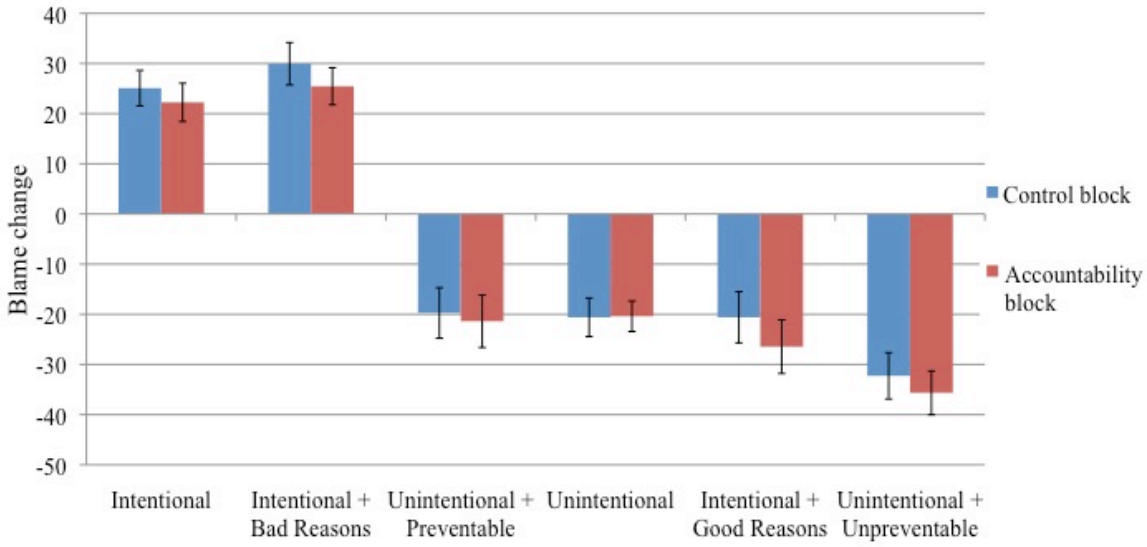

Figure 23. Blame change in response to new information for accountability and control judgments.

Updated blame judgments likewise tracked the new information participants were given $F[5, 185] = 155.7$, $p < .001$), while being unaffected by making public judgments ($F[1, 37] = .29$, $p > .5$), or the interaction between information and judgment type, $F(5, 185) = .34$, $p > .8$. Averaging across judgment type, blame was highest in response to morally bad reasons ($M = 90.7$, $SD = 13.0$) and intentionality information ($M = 85.6$, $SD = 17.3$). Updated blame decreased below the midpoint of the scale in the other four conditions. Of the four, updated blame was reduced most in response to information about unintentional and unpreventable outcomes ($M = 25.1$, $SD = 18.5$), and less so for unintentionality information alone ($M = 35.5$, $SD = 17.5$) and morally good reasons ($M =$

36.4, *SD* = 23.3), and the magnitude of blame reduction was smallest for unintentional and preventable outcomes (*M* = 43.5, *SD* = 23.1).

Study 5 Discussion

This study tested the effects of loading people's processing of new information by making them socially accountable for their judgments on moral updating. . The processing RT results show that people spend more time processing new information when they are accountable for their judgments than when their judgments are private. Moreover, the social accountability manipulation eliminated the RT advantage people show in the control judgment trials. This replicated the result from Study 4 and indicates that people adopt a more systematic processing strategy for non-switching trials when under cognitive load.

People's practice of updating their moral judgments illustrated an important similarity between socially accountable and private judgments of blame. In both contexts people use causal-mental information in a systematic way to update their moral judgments of blame. Neither private, nor public judgments are governed by a pattern of asymmetric blame validation (Alicke, 2000) or post-hoc rationalization (Haidt, 2001). Instead, people use new information to explain the behavior and their moral judgments fall out from those explanations.

However, the lack of any effect of judgment type on blame updating is puzzling. A possible explanation is that making judgments does not alter the processing strategy people employ to make judgments – they already employ a thoughtful processing strategy to new information during the private judgment block. Instead, making public judgments raises the confidence threshold they must meet in order to pronounce their judgment. Thus people may "check in" more often on their processing, especially for information

that they usually process relatively quickly. This, checking in does not alter the final conclusions, because the inputs and the process are the same, but it makes people slower to begin updating their judgments.

Chapter 7

General Discussion

The five studies reported above examined a new question in moral psychology: How do people update existing moral judgments. Specifically, these studies addressed three questions pertaining to the phenomenon of updating: (1) How frequently do people update their moral judgments, and how do updated judgments compare to single moral judgments more commonly studied in the literature? (2) How can we best describe the way people change their blame judgments? Do people have a general motivation to blame that causes them to "anchor" on early judgments and make only token changes in response to new information? Or is blaming a process of assigning meaning to an event, during which people adjust blame in proportion to new information? (3) How do people process new mental state information? Is processing new information like reading off a list, where the ability to process one piece of information is independent of all the other items on the list, or is there a structure to moral judgment in which processing is selectively facilitated or impeded based on preceding information?

I investigated these questions using student (Study 1) and community samples (Studies 2-5), text stimuli (Studies 1 and 2) and audio stimuli (Studies 3-5). Additionally, I tested the effects of cognitive load (Study 4) and making public judgments of blame (Study 5) on people's tendency to update, the magnitude of change to their judgments, and their processing of new information. These studies reinforced the critical role of causal and mental state information for moral judgment (e.g., Cushman, 2008; Greene, et

al., 2009; Guglielmo & Malle, 2010a; Monroe & Reeder, 2011; Young & Saxe, 2009) as well as demonstrating several novel findings pertaining to the phenomenon of updating.

These studies go beyond previous research on single moral judgments by testing how people make judgments and process information in a richer context where information unfolds over time. This arguably more closely mirrors naturalistic moral judgment and the context in which the practice of making moral judgments and punishing emerged, where updating (specifically forgiving transgressions) was adaptive for maintaining social bonds and facilitating social cooperation (Burnette, McCullough, Van Tongeren & Davis, 2012; McCullough, Kurzban, & Tabak, 2010; Petersen, Sell, Tooby, & Cosmides, 2010).

In addition to answering three questions about updating, this project examined the empirical predictions of four prominent moral judgment models and their ability to describe the phenomenon of updating. In places where disagreements between the models arose, I compared the models' competing predictions. I summarize the main empirical findings of this project below and I discuss the implications for Alicke (2000), Haidt (2001), Malle et al., (2012), and Weiner's (1995) moral judgment models.

Summary of empirical findings and their implications for moral judgment theories

The first novel finding in this package of studies was that updating was more common than previously hypothesized. Across five studies people updated their initial judgments 80% of the time. Moreover, updating frequency was consistent across samples and across manipulations of cognitive load (tapping and social accountability). This result is consistent with models that describe blame as a process of assigning meaning to events (Malle et al., in prep; Shaver, 1985; Weiner, 1995), a process in which new information helps answer a pressing *why* question (Malle, 2004). Further, it is consistent with research

on judgment and decision making, which portrays updating as commonplace (Epley & Gilovich, 2001; 2006).

However, the results contrast with views that conceptualize moral updating (and moral judgment more generally) as being limited by a blame validation process (e.g., Alicke; 2000; Haidt, 2001). On these models, people are motivated to validate their judgments of blame and should therefore largely refrain from updating, except in cases where new information would provide evidence supporting their existing blame judgments. The data point in the opposite direction. Updating across studies was high. Moreover, Studies 1 and 2 showed that updating was more frequent when new information violated people's expectations. In particular, people updated more frequently when new information countered their previous beliefs about an event's intentionality. For example, people were more likely to update their moral judgments in response to morally bad reasons if the initial event was an unintentional compared to an intentional event. This finding supports the Path Model's prediction of two distinct tracks to blame (Malle, et al., 2012). When the description of the initial event led people to encode the event as intentional (i.e., putting people on the intentional track to blame), new information about an agent's morally bad reasons was viewed as unsurprising and caused people to update less. However, when the description of the event caused people to encode an event as unintentional (i.e., putting people on the unintentional track to blame), new information about the agent's morally bad reasons was viewed as surprising and resulted in increased updating.

People were also more likely to update their initial judgments if new information violated their general assumptions about the mental states that typically underlie negative

events – for example, their assumption that agents have morally bad reasons for committing negative intentional actions (Brown & Kleeck, 1989; Goldinger, et al., 2003). This assumption is consistent with a blame validating perspective of updating (Alicke, 2000); the data from these studies suggest that while people may start with a blame validation assumption, their processing of the new information and their resulting blame judgments are largely independent of such assumptions. Instead, people update selectively – in response to new information that shows early assumptions or beliefs to be incorrect.

Blame change and updated moral judgments

One of the most robust findings in this series of studies was people's systematic use of causal-mental information to update their blame judgment. Consistent with previous research on single moral judgments, people's updated judgments respond to affirmations or denials of intentionality, the level of justification of an agent's reasons, and whether the negative event in question was preventable (Cushman, 2008; Guglielmo & Malle, 2010a; Greene, et al., 2009; Lagnado & Channon, 2008; Monroe & Reeder, 2011; Young & Saxe, 2009). While the fact that people's blame judgments responded to the mental state information appears uncontroversial, the current studies make two novel contributions to the literature.

First, these studies demonstrate that people's blame judgments depend not only on the specific causal-mental information they receive, but also on the relationship of that information to what preceded it. Across studies, the magnitude of people's changes to their initial blame judgments was larger in "switching trials" (where new information caused people to "switch" intentionality tracks) compared with "non-switching trials"

(with one exception). That is, given a particular piece of information – for example, an agent's morally bad reasons – people changed their blame judgments more (in this case they increased blame) when the initial event was unintentional than when it was intentional. This effect was strongest in Studies 1-3, but it persisted even when participants were under cognitive load or when they were making public judgment of blame. This was true for all of the information conditions except morally good reasons. Here the effect was reversed. People changed their judgments less when the initial behavior was unintentional than when it was intentional, which implies that the justificatory reasons and intentionality information pulled people in different directions (intentionality for increasing blame and justificatory reasons for decreasing blame). These results highlight the important role of intentionality in the process of ascribing blame. Not only is it an important input for moral judgment (e.g., Gray & Wegner, 2008; Greene, et al., 2009; Lagnado & Channon, 2008), but it also *structures* the information people expect and the way they use subsequent information about reasons and preventability in their moral judgments.

The second novel contribution of these studies is that they test the competing predictions of several moral judgment models, a comparison that has thus far been absent in the literature. These studies mostly clearly support the Path Model's predictions, but also demonstrate limited support for Weiner's (1995) responsibility model. The results largely did not support Alicke (2000) and Haidt's (2001) prediction that people would update consistent with a blame validation model.

Despite the evidence that people may initially hold blame validating assumptions about behavior, there was little evidence for the more specific claim that people would

make large changes to blame in response to incriminating information (e.g., causing harm intentionally or acting for bad reasons) making only small, token changes in response to mitigating information (e.g., justifying reasons, unintentional and unpreventable outcomes). Instead, people changed their moral judgment in response to new causal-mental information regardless of its whether it was incriminating or mitigating, and people's updated blame judgments largely conformed to a systematic pattern consistent with the Path Model (Malle, et al., 2012). Blame was highest in response to morally bad reasons and intentionality-only information and lower for unintentional but preventable outcomes, lower still for unintentional-only information, and lowest for morally good reasons and unintentional and unpreventable information.

Additionally, Study 2 directly tested the blame validation hypothesis by comparing updated blame judgments to a control condition where participants were provided with all of the available information at once. Overall, people arrived at remarkably similar blame judgments across the two conditions and for all of the different types of information. However, when this effect was broken down into high and low anchors, the high anchor trials showed support for blame validating anchoring and adjustment. Despite this, people's updated blame judgments still tracked the particular causal-mental information provided by the new information. Moreover, this result indicates that blame validation relies on highly evocative behaviors to produce traction for the model.

One question that emerges from this finding is why, given the rich history of the anchoring and adjustment literature, was these so little anchoring and adjustment? One deflationary explanation is that people were actually making two independent judgments

and because the information that the judgments depended on was identical, the judgments were also identical. However, as discussed earlier, this account does not seem to be correct. First, the initial and final blame judgments were significantly correlated ($r = .30$, $p < .05$). Moreover, such an explanation is unable to account for the robust pattern of reaction time findings showing that people suffer an RT disadvantage when new information necessitates stepping back up in the process of blame and switching tracks. A more likely possibility is that the weight of the judgment and the fluid nature of new information caused participants to more carefully attend to the content of the new information and therefore override previous anchors (Simmons, LeBoeuf & Nelson, 2010).

Together, the blame updating results from these studies demonstrate that there is a structure to blame updating and this structure uniquely depends on intentionality information. Additionally, these studies clearly support an information-processing view of moral judgment, where rational is taken as being responsive to new causal-mental information, in contrast to views that characterize moral judgment as reactive and driven by affect. Even though I do not want to dismiss affect's role in moral judgment, these results undermine models (e.g., Alicke, 2000; Haidt, 2001) according to which affective reactions to moral violations engulf the psychological field.

Information processing en route to updating

The present studies are among the first to examine how people process information en route to moral judgments. The reaction time data offer one of the more striking and reliable findings in this set of studies. People showed a robust effect of a processing RT cost when new information switched the intentionality of an event

compared to when information matched people's intentionality assumption produced by the initial information. Moreover, this effect is not merely due to people's "surprise" at having to integrate new information that counters expectations. Such an explanation would suggest that the RT cost I observe in my data is due to the cognitive cost associated with integrating contradictory information into an existing representation, which would be common to any model of information processing, and therefore the studies reveal nothing new about the process of moral judgment or updating.

While this is certainly *part* of the explanation of the RT data, it is not the *complete* explanation. A meta-analysis of the data from Studies 1-3 shows that people demonstrate (strong) preparedness for reason information following intentional events ($t[146] = 3.23$, $p < .001$, $d = .34$), and (somewhat weaker) preparedness for preventability information following unintentional events ($t[146] = 1.60$, $p = .1$, $d = .19$), even once the RT cost of revising beliefs about intentionality (e.g., intentional to unintentional or unintentional to intentional) is accounted for (See Figure 24).

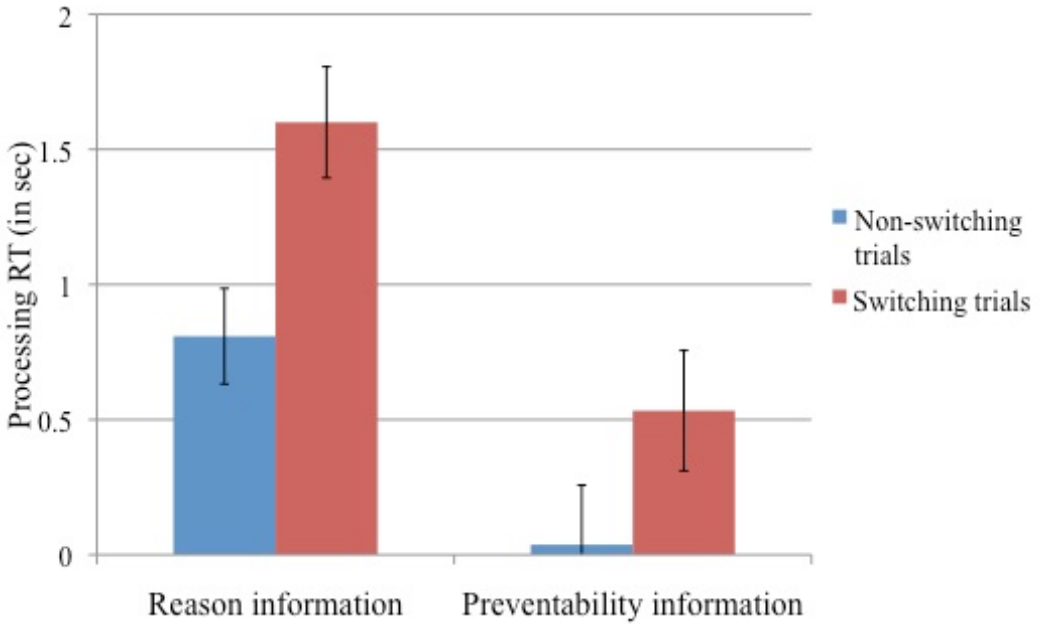

Figure 24. Reaction times for reason and preventability information with intentionality switching partialed out.

Cognitive load (Study 4) and accountability (Study 5) selectively slowed processing in the non-switching trials, thereby equalizing the processing speed for switching and non-switching trials. These results appear puzzling until examined alongside people's updated judgments. In both of these studies, people arrived at very similar blame judgments across cognitive load, public judgment and control blocks. Thus it appears that people are fighting against the manipulations by rechecking information that they usually process quickly in order to make sure they get the answer right.

Additionally, these studies demonstrated that some types of information are reliably processed faster than others. People were fastest processing information about intentionality-only and about morally bad reasons, and they were slowest processing information about morally good reasons. This result suggests differential information processing requirements of good and bad reasons. When learning about an actor's morally bad reasons, people may need to consider little additional information, as such explanations fit familiar scripts for action. Also, in some cases simply having bad motives for an action is perceived as immoral (Reeder, Vonk, Ronk, Ham & Lawrence, 2004). By contrast, morally good reasons require more complex reasoning. People may try to integrate the negative action with the good reason, which may conflict with familiar scripts for action and raises the question of the sufficiency of the justification. Despite this difference however, the magnitude of people's updated blame judgments in these studies attest to a remarkable attention to justificatory reasons even when under cognitive load. It would be interesting to consider whether people's careful appreciation of

justifying reasons is preserved if the amount of time they have to make a judgment is limited. One possibility is that people have a two-stage process for integrating reasons into their moral judgments, similar to some attribution theories (e.g., Trope, 1986). On this type of model people would first maximize blame under script considerations, and then (assuming time) they would correct downward based on an agent's justifying reasons. If this model is correct, limiting updating time would interfere with the second process and result in anchoring on initial judgments. If however, moral judgment is a process of assigning meaning to events, then reasons for actions (justifying or incriminating) directly influence how people conceptualize the action. Limiting updating time may make judgments more variable, but it would not change the systematic way people use reason information.

These data show possible evidence for a "blameworthiness assumption." People were fastest processing new information when it supported increasing blame (e.g., intentionality, morally bad reasons, or preventability of unintentional event) and slower processing information that supported reducing blame (e.g., unpreventability of intentional event, morally good reasons). These data in conjunction with the frequency data appear to indicate that people are ready for bad information; they expect it, and information that runs counter to these assumptions is surprising. However, such trends in updating frequency and processing speed did not affect blame judgments, which is the key prediction of a blame validation model (e.g., Alike, 2000). To the contrary, the magnitude of blame change due to mitigating information (e.g., absence of intentionality or presence of morally good reasons) was on par with the magnitude of blame change due to exacerbating information (e.g., presence of intentionality or of morally bad reasons).

Overall, the pattern of processing RT's observed across the studies supports the view that there is a structure to the process of assigning and updating blame. This stands in contrast to many current moral judgment models (Alicke, 2000; Haidt, 2001; Weiner, 1995), which imply that different types of causal-mental information are processed independently, and processing speed is therefore dependent on non-moral factors such as complexity or quantity. These studies show that the underlying structure of blame is built (at least in part) upon the perceived intentionality of the event. People quickly encode events as either intentional or unintentional (Malle & Knobe 1997; Malle & Holbrook, 2012), and based on that inference, different types of information are expected to follow. This finding uniquely supports the Path Model of Blame (Malle, et al., in prep), which predicts that intentionality bifurcates the processing stream en route to blame. Therefore, if people encoded an event as intentional, they process information about reasons much faster than if they encoded the event as unintentional. Conversely, if people encoded an event as unintentional, they process information about the preventability of the event more quickly than if they encoded the event as intentional. As it currently stands, no other moral judgment model postulates such a processing structure and can explain these data as parsimoniously.

Limitations and future directions

In these studies, updating a moral judgment occurred after participants were presented with new information. One critique of this design is that it may have artificially inflated people's willingness to update. It is possible that if people were allowed to guide their own search for updating information they would search in a manner more amendable to a blame validation model. They might refuse to search and simply maintain their initial blame judgment; alternatively they might seek out information that confirms

their intuitive judgments (Haidt, 2001) or that supports their spontaneous affective reactions (Alicke, 2000). A future study could address this critique by incorporating an information search into the updating paradigm. Instead of providing people with the new information after their initial judgments, they would be given the option to search for new information and then to update their judgments based on their information search. Additionally, to more closely mirror a naturalistic process, this search could be left unstructured. Participants would simply ask questions to find out new information before making their updated judgment. This type of design would more directly address the question of the natural frequency of updating and elucidate the particular types of information that people find most important for blame.

A second limitation on the interpretation of the cognitive load studies (Studies 4 & 5) is that they do not directly speak to some of the predictions of Alicke (2000) and Haidt's (2001) models. While conditions of cognitive load might be assumed to favor their models, it is possible that cognitive load might impair participants' ability to produce intuitions or integrate affect into their judgments. Thus, one could argue that demonstrating that cognitive load does not favor the predictions of their models does not necessarily count as evidence against the models. Instead, a future study could more explicitly test these models' account of updating by testing the role of affect in moral judgment. One way to test this would be to vary the extremity of norm violating actions and measure people's affective responses in addition to their moral judgments. Then one could enter the affective reaction as a competing prediction of people's moral judgments. A blame validation model would predict that once people's affective reactions are accounted for, information about reasons or preventability would only weakly predict

moral judgments. Alternatively, one could test one part of Alicke's (2000) model by including "extra-legal" factors such as character information or likability as sources of new information in the updating paradigm. Alicke's model would predict that, while irrelevant for blame, extra-legal information produces (usually negative) spontaneous evaluations and will thereby affect people's judgments of blame.

My studies of moral updating have primarily supported what most people would refer to as "rational" models of moral judgment; however, there is an established literature demonstrating the powerful effect of emotion on moral judgments (e.g., Inbar & Pizarro, 2009; Schnall, et al., 2008; Wheatley & Haidt, 2005). It seems highly unlikely that updating moral judgments is immune to the effect of emotion in a way that single moral judgments are not. An interesting future direction of this line of research would be to examine *how* affect influences moral updating judgments.

One possibility is that affective reactions like anger or disgust simply turn up the moral volume (Pizarro, et al., 2011). On this hypothesis, negative affect may cause people to judge actions more harshly, but it leaves the process of arriving at the judgment untouched; moral judgments should still be largely determined by reasons (for intentional actions) and by the obligation and capacity to prevent an outcome (for unintentional events). This hypothesis would be compatible with the Path Model (Malle, et al., 2012) and Weiner's (1995) responsibility model. An alternative hypothesis is that affect actually changes the process of moral judgment, or that it changes the way people evaluate the inputs to moral judgment. For example, it is possible that strong negative affect (e.g., anger or disgust) causes people to skip over steps in the process to blame or adopt simple heuristic rules (e.g., "if you did it, then you're to blame."). A future study

could include a between subjects manipulation of negative affect (e.g., anger) prior to participants receiving new information and pit these hypotheses against each other. If the "moral volume" hypothesis is correct, updated blame should be higher in the negative affect condition relative to control, and increases in blame should be similar across each of the information conditions. Moreover, according to this hypothesis, emotion should not affect processing RT's since it leaves the process of blame updating unaffected. If however, emotion alters the process of arriving at a blame judgment, then one would predict an increase in people's final, updated blame judgments in the emotion condition compared to the control condition, and it would predict that changes to blame should vary across types of information. Blame should show large increases in response to exacerbating information and small changes in response to mitigating information. Additionally, the negative emotion manipulation should shorten processing RT's as people presumably consider the information less carefully.

Finally, all of the studies reported here focused on blame judgments of a distant 3rd person. A blame validating motive may well intrude when people are *personally* involved in their blame judgments – for example, when they themselves are the victim or personally connected to the victim. It is possible that in these situations people's affective responses become better predictors of their processing of new information and change in moral judgments. To the extent that being victimized produces anger people may adopt heuristic rules for blaming (e.g., "You caused it; therefore you are to blame), or they may engage in post-hoc rationalization (Haidt, 2001) in order to justify their judgments of blame and possible refusals to update in response to mitigating information.

Alternatively, personal involvement carries with it possibility for the negotiation of blame, and this is an area where the updating paradigm can inform research outside of the immediate domain of moral judgment. Specifically, these studies and the Path Model of Blame can be exported to inform research on apologies and forgiveness. Forgiveness is defined in the literature as "becoming decreasingly motivated to retaliate … [or] to maintain estrangement from the offender" (McCullough, Worthington & Rachal, 1997, p. 322), and in this sense forgiveness can be construed as a form of moral updating. The work here and the Path model would suggest three things. First, it offers a new way to conceptualize the measurement of forgiveness. Most of this research uses vignette or autobiographical studies where participants are either asked to imagine that they are a character in a story (Darby & Schlenker, 1982; Eaton & Struthers, 2006; McCullough, et al., 1997; Ohtsubo & Watanabe, 2009; Scher & Darley, 1997) or to report on a past incident in which they were harmed (Baumeister, Stillwell & Wotman, 1990). The researchers manipulate the complexity of a hypothetical apology or other compensatory behavior and then measure participants' willingness to forgive. Adopting the updating paradigm would provide a more accurate measure of one of the definitions of forgiveness: decreasing motivation to retaliate. This paradigm would allow researchers to obtain an initial measure of forgiveness (e.g., initial motivation to retaliate, willingness to forgive, or desire to maintain the relationship), present participants with an apology, and then measure changes in forgiveness.

Second, research on apologies and forgiveness largely lacks an overarching theory for what constitutes an effective apology, what types of explanations produce forgiveness, or how the severity of the infraction may moderate the effectiveness of an

apology. The Path Model suggests that for apologies to produce forgiveness, the content of an apology must be tailored to the perceived intentionality of the infraction. Offering reasons for accidents or preventability information for intentional actions will be of limited use. Also, merely refuting intentionality may not be sufficient for forgiveness. As shown in these studies, doing so may reduce blame by switching intentionality tracks (e.g., from intentional to unintentional), but such an apology leaves open questions of preventability. Instead, explanations that focus on the morally acceptable reasons for an agent's intentional action or her inability to prevent a harmful outcome will produce the largest amounts of forgiveness..

Updating in conclusion

In the above studies I tested a new question in moral psychology: how do people update their moral judgments? These studies showed that arriving at blame judgments and updating those judgments follows a conceptual and informational structure, which has intentionality at its core. People's reaction times in response to new information showed that intentionality bifurcates the process of assigning blame and prepares people to receive different types of information – reasons for intentional events, and preventability information for unintentional events. Additionally, while people appeared ready to learn incriminating new information, these expectations rarely influenced the way people changed their blame judgments. Instead, updating and moral judgment more generally, were better described as an effort to explain behavior that systematically responds an agent's specific causal contributions and mental states to make and revise moral judgments.

The realm of moral judgment is richer than making single judgments. People rarely make a single judgment about a wrongdoer, or have all the relevant information for

a judgment at once. More often blame is negotiated, new information is revealed, and judgments are *updated*. My hope is that these studies will motivate future work going beyond uncovering the inputs to moral judgment to examining how information moral judgments evolve as information unfolds.

References

Alicke, M. D. (1990). Incapacitating conditions and alterations of blame. *Journal of Social Behavior and Personality, 5,* 651–664.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology, 63,* 368-378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*, 556-574.

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

Alicke, M. D., Buckingham, J., Zell, E., & Davis, T. (2008). Culpable Control and Counterfactual Reasoning in the Psychology of Blame. *Personality and Social Psychology Bulletin, 34,* 1371-1381.

Alicke, M. D., & Davis, T. L. (1989). The role of a posteriori victim information in judgments of blame and sanction. *Journal of Experimental Social Psychology, 25,* 362-377.

Alicke, M. D., Davis, T. L., & Pezzo, M. V. (1994). A posteriori adjustment of a priori decision criteria. *Social Cognition, 12,* 281–308.

Alicke, M. D., Weigold, M. F., & Rogers, S. L. (1990). Inferring intentions and responsibility from motives and outcomes: Evidential and extraevidential judgments. *Social Cognition, 8,* 286–305.

Alicke, M. D., & Zell, E. (2009). Social Attractiveness and Blame. *Journal of Applied Social Psychology, 39,* 2089–2105.

Allport, A., Styles, E. A., & Hsieh, S. (1994). Shifting intentional set: Exploring the dynamic control of tasks. In C. Umilta` & M. Moscovitch (Eds.), *Attention and performance XV* (pp. 421–452). Cambridge MA: MIT Press.

Baird, C. H. & Gautam, P. (2011). From social media to social customer relationship management. Strategy AND Leadership, 39, 30-37.

Baldwin, D. A., Baird, J. A., Saylor, M. M., & Clark, M. A. (2001). Infants parse dynamic action. *Child Development, 72,* 708–717.

Bargh, J. (1994). The four horsemen of automaticity: Awareness, efficiency, intention, and control in social cognition. In J. R. S. Wyer & T. K. Srull (Eds.), *Handbook of social cognition,* 2nd ed (pp. 1-40). Hillsdale. NJ: Erlbaum.

Bargh, J. A., Chaiken, S., Raymond, P., & Hymes, C. (1996). The automatic evaluation effect: Unconditionally automatic activation with a pronunciation task. *Journal of Experimental Social Psychology, 32,* 185-210.

Bargh, J. A., & Chartrand, T. L. (1999). The unbearable automaticity of being. *American Psychologist, 54,* 462 - 479.

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.

Baumeister, R. F., Stillwell, A.,& Wotman, S. R. (1990). Victim and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger. *Journal of Personality and Social Psychology, 59,* 994-1005.

Brown, R., & Kleeck, M. H. V. (1989). Enough said: Three principles of explanation. *Journal of Personality and Social Psychology, 57,* 590–604.

Burnette, J. L., McCullough, M. E., Van Tongeren D. R. & Davis, D. E. (2012). Forgiveness Results From Integrating Information About Relationship Value and Exploitation Risk. *Personality and Social Psychology Bulletin, 38,* 345-356.

Carlsmith, K. M., Darley, J. M. & Robinson, P. H. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. Journal of Personality and Social Psychology. 83, 284-299.

Chapman, G. B. & Johnson, E. J. (1994) The limits of anchoring, *Journal of Behavioral Decision Making, 7,* 223–42.

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

CNN. (2011) Weiner apologizes for lying, 'terrible mistakes,' refuses to resign. Retrieved September 20, 2011 from http://www.cnn.com/2011/POLITICS/06/06/new.york.weiner/

Darby, B. W.,& Schlenker, B. R. (1982). Children's reactions to apologies. *Journal of Personality and Social Psychology, 43,* 742-753.

Darley, J. M. & Shultz, T. R. (1990). Moral rules: Their content and acquisition. *Annual Review of Psychology, 41,* 525-556.

De Brigard, F., Mandelbaum, E., & Ripley D. (2008). Responsibility and the Brain Sciences. *Ethical Theory and Moral Practice, 12,* 511-524.

Duff, A. (1986). *Trials and punishments*. Cambridge, England: Cambridge University Press.

Eaton, J., & Struthers, C. W. (2006). The reduction of psychological aggression across varied interpersonal contexts through repentance and forgiveness. *Aggressive Behavior, 32,* 195–206.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science, 12,* 391–396.

Epley, N., & Gilovich, T. (2006). The Anchoring-and-Adjustment Heuristic: Why the Adjustments Are Insufficient. *Psychological Science, 17*, 311–318.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87,* 327–339.

Fincham, F. D., & Bradbury, T. N. (1992). Assessing attributions in marriage: The Relationship Attribution Measure. *Journal of Personality and Social Psychology, 62,* 457-468.

Fincham, F. D., & Shultz, T. R. (1981). Intervening causation and the mitigation of responsibility for harm. *British Journal of Social Psychology, 20,* 113-120.

Finkel, N. J., Maloney, S. T., Valbuena, M. Z., & Groscup, J. (1996). Recidivism, proportionalism, and individualized punishment. *American Behavioral Scientist, 39,* 474–487.

Gailey, J. A., & Falk, R. F. (2008). Attribution of responsibility as a multidimensional concept. *Sociological Spectrum*, *28*, 659–680.

Gilbert, D.T., & Malone P.S. (1995). The Correspondence Bias. *Psychology Bulletin, 117,* 21-38.

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology, 54,* 733–740.

Goldinger, S. D., Kleider, H. M., Azuma, T., & Beike, D. (2003). "Blaming the victim" under memory load. *Psychological Science, 14,* 81-85.

Graham, S., Weiner, B., Benesh-Weiner, M. (1995).An attributional analysis of the development of excuse giving in aggressive and nonaggressive African American boys. *Developmental Psychology, 31,* 274-284.

Gray, K. & Wegner, D. M. (2008) The sting of intentional pain. *Psychological Science, 19,* 1260–1262.

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

Greene, J. (2005). Emotion and cognition in moral judgment: evidence from neuroimaging, in *Neurobiology of Human Values*, J.P. Changeux, A.R. Damasio, W. Singer, and Y. Christen, Eds., Springer-Verlag, Berlin.

Greene, J.D. (2009) The cognitive neuroscience of moral judgment, in *The Cognitive Neurosciences IV*, M.S. Gazzaniga, Ed. MIT Press, Cambridge, MA.

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

Greene, J.D., 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.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition. *Psychological Review, 102,* 4-27.

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

Guglielmo, S., & Malle, B. F. (2010b). Enough skill to kill: Intentionality judgments and the moral valence of action. Cognition, 117, 139–150.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry, 52,* 449-466.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108,* 814–34.

Haidt, J. (2007). The new synthesis in moral psychology. *Science 316,* 998 -1001.

Haidt, Björklund, & Murphy (n.d.) Moral Dumbfounding: When intuition finds no reason.

Haidt, J., & Hersh, M. (2001). Sexual morality: The cultures and reasons of liberals and conservatives. *Journal of Applied Social Psychology, 31,* 191-221.

Haidt, J., Koller, S., & Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology, 65,* 613-628.

Hamilton, V. L. (1986). Chains of command: Responsibility attribution in hierarchies. *Journal of Applied Social Psychology, 16,* 118-138.

Hilton, D. J. (2007). Causal explanation: From social perception to knowledge-based causal attribution. In A. W. Kruglanski & E. T. Higgins (Eds.), *Social psychology: Handbook of basic principles* (2nd ed., pp. 232–253). New York, NY: Guilford Press.

Horberg, E. J., Oveis, C., Keltner, D., & Cohen, A. B. (2009). Disgust and the moralization of purity. *Journal of Personality and Social Psychology, 97,* 963-976.

Howe, E. S. (1991). Integration of mitigation, intention, and outcome damage information, by students and circuit court judges. *Journal of Applied Social Psychology, 21,* 875–895.

Inbar, Y., & Pizarro, D. A. (2009). Grime and Punishment: How disgust influences moral, social, and legal judgments. *The Jury Expert, 21,* 11-23.

Inbar, Y., Pizarro, D.A., Knobe, J., & Bloom, P. (2009) Disgust sensitivity predicts intuitive disapproval of gays. *Emotion, 9,* 435-439.

Jones, E. E. (1979). The rocky road from acts to dispositions. *American Psychologist, 34,* 107-117.

Kassin, S., & Garfield, D. (1991). Blood and guts: General and trialspecific effects of videotaped crime scenes on mock jurors. *Journal of Applied Social Psychology, 21,* 1459-1472.

Kelley, H. H. (1971). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds). *Attribution: Perceiving the causes of behavior* (pp. 1-26) Morristown, NJ: General Learning Press.

Knobe, J. (2003). Intentional action and side effects in ordinary language. *Analysis, 63,* 190 -194.

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.

Lagnado, D. & Channon, S. (2008) Judgments of cause and blame: The effects of intentionality and foreseeability. *Cognition 108, 3,* 754-770

Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart Strings and Purse Strings: Carryover Effects of Emotions on Economic Decisions. *Psychological Science, 15,* 337-341.

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

Malle, B. F. (2006). The relation between judgments of intentionality and morality. *Journal of Cognition and Culture, 6,* 61-86.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. In J. Forgas, K. Fiedler, and C. Sedikides (Eds.), *Social thinking and interpersonal behaviour* (14th Sydney Symposium of Social Psychology). Philadelphia, PA: Psychology Press.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (in prep). A theory of Blame.

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.

Mazzocco, P. J., Alicke, M. D., & Davis, T. L. (2004). on the robustness of outcome bias: No constraint by prior culpability. *Basic and Applied Social Psychology, 26,* 131-146.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2010). Evolved mechanisms for revenge and forgiveness. In M. Mikulincer & P. R. Shaver (Eds.), *Human aggression and violence: Causes, manifestations, and consequences* (pp. 221-239). Washington, DC: American Psychological Association.

McCullough, M. E., Worthington, E. L., Jr., & Rachal, K. C. (1997). Interpersonal forgiving in close relationships. *Journal of Personality and Social Psychology, 73,* 321-336.

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

Newman, A. A. (2011). Brands Now Direct Their Followers to Social Media. *New York Times*, August 4, B3.

Oshana, M. (2001). Responsibility: Philosophical aspects. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social & Behavioral Sciences* (pp. 13279-13283). Oxford: Pergamon.

Ohtsubo, Y. & Watanabe E. (2009). Do sincere apologies need to be costly? Test of a costly signaling model of apology. *Evolution and Human Behavior, 30,* 114–123

Petersen, M. B., Sell, A., Tooby, J., & Cosmides, L. (2010). Evolutionary psychology and criminal justice: A recalibrational theory of punishment and reconciliation. In H. Høgh-Oleson (Ed.), *Human morality and sociality: Evolutionary and comparative perspectives* (pp. 72-131). New York, NY: Palgrave Macmillan.

Pizarro, D.A., Inbar, Y., & Helion, C. (2011). On disgust and moral judgment. *Emotion Review, 3,* 267–268.

Pizarro, D.A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology, 39,* 653-660.

Pizarro, D.A., Uhlmann, E., & Salovey, P. (2003). Asymmetries in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science, 14,* 267 – 272.

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.

Quattrone, G. A. (1982). Overattribution and unit formation: When behavior engulfs the field. *Journal of Personality and Social Psychology, 42,* 593-607.

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

Reeder, G. D. (2009b). Mindreading: Judgments about intentionality and motives in dispositional inference. *Psychological Inquiry, 20,* 1–18.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review, 86,* 61–79.

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.

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

Ross L. (1977). The intuitive psychologist and his shortcomings. In. L Berkowitz (Ed.), *Advances in experimental social psychology,* 10, (pp. 173-220), San Diego, CA: Academic Press.

Scher, S. J. & Darley, J. M. (1997). How Effective Are the Things People Say to Apologize? Effects of the Realization of the Apology Speech Act. *Journal of Psycholinguistic Research, 26,* 127-140.

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

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

Shaw, M. E., & Sulzer, J. L. (1964). An empirical test of Heider's levels in attribution of responsibility. *Journal of Abnormal and Social Psychology*, *69*, 39–46.

Showers, C.,& Cantor, N. (1985). Social cognition: A look at motivated strategies. *Annual Review of Psychology, 36,* 275-305.

Simmons, J. P., LeBoeuf, R. A. & Nelson, L. D. (2010). The Effect of Accuracy Motivation on Anchoring and Adjustment: Do People Adjust From Provided Anchors? *Journal of Personality and Social Psychology, 99,* 917-932.

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105,* 131–142.

Sloman, S. A., Fernbach, P., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. Bartels, C. Bauman, L. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making* (pp. 1–26). Boston, MA: Academic Press.

Strack, F., & Mussweiler, T. (1997). Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology, 73,* 437–446

Stuyven, E., Van der Goten, K., Vandierendonck, A., Claeys, K., & Crevits, L. (2000). The effect of cognitive load on saccadic eye movements. *Acta Psychologica*, *104*, 69-85.

Tetlock, P. E. (1983). Accountability and complexity of thought. *Journal of Personality and Social Psychology, 45,* 74-83.

Tetlock, P. E., & Kim, J. (1987). Accountability and judgment in a personality prediction task. *Journal of Personality and Social Psychology, 52,* 700-709.

Tetlock, P. E., Self, W. T. & Singh, R. (2010). The punitiveness paradox: When is external pressure exculpatory – And when a signal just to spread blame? *Journal of Experimental Social Psychology 46*, 388-395.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review, 93,* 239-257.

Tversky, A.; Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science 185,* 1124–1131.

Valdesolo, P., & DeSteno, D. (2006). Manipulations of emotional context shape moral judgment. *Psychological Science, 17,* 476–477.

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063-1070.

Weiner, B. (1992). Excuses in everyday interaction. In M. L. McLaughlin, M. J. Cody, & S. R. Read (Eds.), *Explaining one's self to others* (pp. 131-146). Hillsdale, N. J.: Erlbaum.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct.* New York, NY: Guilford Press.

Weiner, B. (2001). Responsibility for social transgressions: An attributional analysis. In B. F. Malle, L. J. Moses, and D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 33l-344). Cambridge, MA: MIT Press.

Weiner, B. (2003). A naïve psychologist examines bad luck and the concept of responsibility. *The Monist, 86,* 164-180.

Weiner, B. (2005). Motivation from an attribution perspective and the social psychology of perceived competence. In Elliot, A. J., & Dweck, C. S. (Eds.), *Handbook of Competence and Motivation* (pp. 73-84). New York, NY: Guilford.

Weiner, B., Amirkhan, J., Folkes, V. S., & Verette, J. A. (1987). An attributional analysis of excuse-giving. *Journal of Personality and Social Psychology, 52,* 316- 324.

Weiner, B., Figueroa-Munoz, A., & Kakihara, C. (1991). The goal of excuses and communication strategies related to causal perceptions. *Personality and Social Psychology Bulletin, 17,* 4-13.

Weiner, B., Perry, R. B., & Magnussen, J. (1988). An attributional analysis of reactions to stigmas. *Journal of Personality and Social Psychology, 55,* 738–748.

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

Wissler, R. L. &. Saks, M. J. (1985). On the inefficacy of limiting instructions: When jurors use prior conviction evidence to decide on guilt. *Law and Human Behavior, 9,* 37-48.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69,* 1–34.

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.

Young, L., & Saxe, R. (2009). Innocent intentions: A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia, 47,* 2065-2072.

Appendix A: Initial event descriptions

Stimuli sentences	Intentionality rating
Sven broke the Wright's front window.	2.30 (*SD* = 1.34)
Charles gave Wayne a black eye.	4.40 (*SD* = 0.98)
Veronica started a forest fire.	2.18 (*SD* = 1.47)
Tom put a large dent in Andrea's car.	2.15 (*SD* = 1.21)
Ryan set the Andersons' house on fire.	3.05 (*SD* = 1.65)
Matt killed Frank.	4.10 (*SD* = 1.35)
Eric broke Monica's arm.	2.65 (*SD* = 1.29)
Paul gave Brittany a deep knife wound in her leg.	4.42 (*SD* = 1.45)
Trent broke the lock on the shed in the Smiths' backyard.	5.05 (*SD* = 1.04)
Jim kicked Aaron hard in the shin.	5.00 (*SD* = 1.13)
When filing her tax returns, Amanda misreported her income.	3.70 (*SD* = 1.40)
Katherine spray-painted what looked like graffiti on Jane's door.	5.65 (*SD* = 0.70)
Jean walked into the kitchen and startled Chuck.	1.40 (SD = 1.37)
Alice broke Jane's pearl necklace.	2.00 (*SD* = 1.06)
Tommy left the restaurant without leaving the waiter a tip.	4.60 (*SD* = 1.22)
Neil fried the motherboard on Allen's computer.	1.51 (*SD* = 1.27)
David bumped into a female classmate in the hall, knocking her down.	1.78 (*SD* = 1.19)
Jie tracked mud all over his mother's freshly cleaned floor.	2.10 (*SD* = 1.39)
Yolanda broke her grandmother's heirloom vase.	1.45 (*SD* = 1.08)
Ophelia burnt Roland's birthday meal.	0.88 (*SD* = 1.07)
Drew gave a customer incorrect change.	1.45 (*SD* = 1.20)
Ted hit a man with his car.	1.58 (*SD* = 1.24)
Brianna, a daycare worker, fed an infant something that made him sick.	1.05 (*SD* = 1.24)
Barry, an agent for the CIA, divulged a sensitive government secret to a foreign spy.	4.20 (*SD* = 2.07)

Stimuli sentences	Intentionality rating
Marissa, who is not handicapped, took the very last handicapped parking space in the lot.	5.02 (*SD* = 1.42)
Randy pulled out some of Eve's hair.	4.00 (*SD* = 1.65)
By endorsing a new company program Bob damaged a local wildlife preserve.	3.88 (*SD* = 1.65)
Lisa shot Tom in the arm.	3.70 (*SD* = 1.44)
Peter answered a phone call while at dinner with a date.	4.42 (*SD* = 1.45)
Don accessed Bert's private email.	5.50 (*SD* = 0.91)
While standing in line John hit the woman behind him with his elbow.	1.32 (*SD* = 1.21)
Fred told his friend the wrong wedding date.	1.92 (*SD* = 1.49)
Emma kicked a dog.	4.38 (*SD* = 1.70)
Ava sunbathed on a public beach nude.	5.42 (*SD* = 1.08)
Liz took a t-shirt out of a store without paying.	4.62 (*SD* = 1.21)
Jacob overlooked his old girlfriend at a party.	5.45 (*SD* = 0.81)

Appendix B: Updating information

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Sven intentionally broke the Wright's front window because he wanted to frighten the Wrights and get them to move out of the neighborhood.	Sven intentionally broke the Wright's front window because a burglar was in the house and Sven wanted to set off the alarm.	Sven accidentally broke the Wright's front window. He was playing baseball too close to the house and accidentally hit a ball through the window.	Sven accidentally broke the Wright's front window. He was playing football in a neighborhood park. The wind caught the ball and it broke the window.	Sven intentionally broke the Wright's front window.	Sven accidentally broke the Wright's front window.
Charles intentionally gave Wayne a black eye because he didn't like Wayne.	Charles intentionally gave Wayne a black eye because Wayne assaulted Charles' girlfriend.	Charles accidentally gave Wayne a black eye. They were playing catch. Charles threw the ball to Wayne when he was not expecting it.	Charles accidentally gave Wayne a black eye. They were playing catch. Charles threw the ball to Wayne but Wayne miscalculated its path.	Charles intentionally gave Wayne a black eye.	Charles accidentally gave Wayne a black eye.
Veronica intentionally started a forest fire because she loves attention and wanted to get on the local news channel.	Veronica intentionally started a forest fire because she is a firefighter and wanted to prevent a larger fire from breaking out.	Veronica accidentally started a forest fire. She did not properly extinguished her campfire; she left several small embers which reignited and spread.	Veronica accidentally started a massive forest fire. She properly extinguished her campfire; however, a small ember on the bottom reignited and spread.	Veronica intentionally started a forest fire.	Veronica accidentally started a forest fire.
Tom intentionally put a large dent in Andrea's car. Andrea broke up with him and he wanted to get back at her.	Tom intentionally put a large dent in Andrea's car. He was trying to prevent it from sliding off the icy road.	Tom accidentally put a large dent in Andrea's car. He tried to stop his car, but he pressed the gas pedal instead of the brakes.	Tom accidentally put a large dent in Andrea's car. He slipped as he was walking past the car and his shoulder hit the car.	Tom intentionally put a large dent in Andrea's car.	Tom accidentally put a large dent in Andrea's car.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Ryan intentionally set the Andersons' house on fire because he didn't like that their house was so big.	Ryan intentionally set the Andersons' house on fire because their house was actually a giant meth lab.	Ryan accidentally set the Andersons' house on fire. He was burning leaves too close to the house and not paying very close attention to the fire which spread to the house.	Ryan accidentally set the Andersons' house on fire. He was burning leaves at a safe distance from the house, but due to a strong wind gust it spread to the house.	Ryan intentionally set the Andersons' house on fire.	Ryan accidentally set the Andersons' house on fire.
Matt intentionally killed Frank because he wanted to collect the insurance money for Frank's death.	Matt intentionally killed Frank because Frank picked up a gun and was about to shoot him.	Matt accidentally killed Frank. Matt gave Frank expired medicine because he didn't read the expiration label on the box.	Matt accidentally killed Frank. Matt unwittingly gave Frank expired medicine because the box did not have an expiration label.	Matt intentionally killed Frank.	Matt accidentally killed Frank.
Eric intentionally broke Monica's arm because he wanted to prevent Monica from playing in an important basketball game.	Eric intentionally broke Monica's arm. Monica's arm was not healing correctly and Eric needed to reset the bone.	Eric accidentally broke Monica's arm. They were sparring and Eric didn't pull his punch in time.	Eric accidentally broke Monica's arm. They were sparring and Eric tripped and fell on her arm.	Eric intentionally broke Monica's arm.	Eric accidentally broke Monica's arm.
Paul intentionally gave Brittany a deep knife wound in her leg because he wanted to hurt her and scar her beautiful legs.	Paul intentionally gave Brittany a deep knife wound in her leg because she was bitten by a snake and Paul was trying to remove the poison.	Paul accidentally gave Brittany a deep knife wound in her leg. While practicing spinning a kitchen knife in the air it fell out of his hand and landed in Brittany's calf.	Paul accidentally gave Brittany a deep knife wound in her leg. While he was putting away a kitchen knife, Brittany startled him and he dropped the knife.	Paul intentionally gave Brittany a deep knife wound in her leg.	Paul accidentally gave Brittany a deep knife wound in her leg.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Trent intentionally broke the lock on the Smiths' shed because he wanted to steal some of the Smith's power-tools.	Trent intentionally broke the lock on the Smiths' shed because his dog was locked in the shed and wanted out.	Trent accidentally broke the lock on the Smiths' shed. The lock was obviously old and fragile.	Trent accidentally broke the lock on the Smiths' shed. The lock looked new but was corroded inside.	Trent intentionally broke the lock on the Smiths' shed.	Trent accidentally broke the lock on the Smiths' shed.
Jim intentionally kicked Aaron because he thought it would be funny to hurt Aaron and watch him hobble around.	Jim intentionally kicked Aaron because he wanted to prevent Aaron from telling a hurtful rumor about a friend.	Jim accidentally kicked Aaron. Jim was was not paying attention and didn't notice Aaron in front of him.	Jim accidentally kicked Aaron. Jim had a migrane and could not see clearly in front of him.	Jim intentionally kicked Aaron.	Jim accidentally kicked Aaron.
Amanda intentionally misreported her income on her taxes because she wanted to have some extra money to spend on herself.	Amanda intentionally misreported her income on her taxes because she thought her tax was too low and wanted to pay her fair share.	Amanda accidentally misreported her income on her taxes. She did not carefully review her tax information.	Amanda accidentally misreported her income on her taxes. Her employer provided her with an incorrect W2.	Amanda intentionally misreported her income on her taxes.	Amanda accidentally misreported her income on her taxes.
Katherine intentionally spray-painted what looked like graffiti on Jane's door because she wanted to deface Jane's new front door.	Katherine intentionally spray-painted what looked like graffiti on Jane's door because it was Halloween and Katherine was helping decorate.	Katherine accidentally spray-painted what looked like graffiti on Jane's door. Katherine thought she was spraying sealant on the door, but the can was clearly marked as red spray paint.	Katherine accidentally spray-painted what looked like graffiti on Jane's door. The spray-can was mislabeled, and Katherine thought she was spraying sealant on the door.	Katherine intentionally spray-painted what looked like graffiti on Jane's door.	Katherine accidentally spray-painted what looked like graffiti on Jane's door.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Jean walked into the kitchen and intentionally startled Chuck. Jean wanted to really scare Chuck.	Jean walked into the kitchen and intentionally startled Chuck. Jean wanted to surprise Chuck with a gift.	Jean walked into the kitchen and accidentally startled Chuck, though she saw that Chuck had earphones plugged in and had music on.	Jean walked into the kitchen and accidentally startled Chuck. She didn't know that Chuck had earphones in and couldn't hear her come in.	Jean walked into the kitchen and intentionally startled Chuck.	Jean walked into the kitchen and accidentally startled Chuck.
Alice intentionally broke Jane's pearl necklace because she was jealous of how pretty the necklace was and how happy it made Jane.	Alice intentionally broke Jane's pearl necklace because Jane had stolen the necklace and Alice wanted to prevent her from selling it for profit.	Alice accidentally broke Jane's pearl necklace. Alice pulled too hard on the necklace when she was looking at it.	Alice accidentally broke Jane's pearl necklace. The necklace turned out to be old and broke at the slightest touch.	Alice intentionally broke Jane's pearl necklace.	Alice accidentally broke Jane's pearl necklace.
Tommy intentionally left the restaurant without leaving the waiter a tip because he didn't want to waste money on being nice.	Tommy intentionally left the restaurant without leaving the waiter a tip because the waiter was rude and insulted Tommy's family.	Tommy accidentally left the restaurant without leaving the waiter a tip. He assumed it had been added to the bill but did not check.	Tommy accidentally left the restaurant without leaving the waiter a tip. The bill indicated that the tip was added to the bill, but it was not.	Tommy intentionally left the restaurant without leaving the waiter a tip.	Tommy accidentally left the restaurant without leaving the waiter a tip.
Neil intentionally fried the motherboard on Allen's office computer because he didn't want Allen to meet a big deadline at work.	Neil intentionally fried the motherboard on Allen's office computer to prevent Allen from releasing a harmful computer virus at work.	Neil accidentally fried the motherboard on Allen's computer. Neil forgot to ground his tools when working on the computer and the motherboard shorted out.	Neil accidentally fried the motherboard on Allen's computer. Neil grounded his tools before working on the computer, but the computer was miswired and shorted out.	Neil intentionally fried the motherboard on Allen's office computer.	Neil accidentally fried the motherboard on Allen's computer.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
David intentionally bumped his classmate because he wanted her to stagger and look silly.	David intentionally bumped his classmate to prevent her from stepping in dog poop.	David accidentally bumped his classmate. David was not watching where he was walking.	David accidentally bumped into a his classmate. David is blind and couldn't see his classmate.	David intentionally bumped his classmate, knocking her down.	David accidentally bumped his classmate, knocking her down.
Jie intentionally tracked mud all over his mother's freshly cleaned floor because he was in a bad mood and wanted to annoy his mother to make himself feel better.	Jie intentionally tracked mud all over his mother's freshly cleaned floor because he wanted to demonstrate the new cleaning device he bought for her.	Jie accidentally tracked mud all over his mother's freshly cleaned floor. He didn't check to see if his shoes were muddy.	Jie accidentally tracked mud all over his mother's freshly cleaned floor. He cleaned his shoes on the doormat, but it was dirty.	Jie intentionally tracked mud all over his mother's freshly cleaned floor.	Jie accidentally tracked mud all over his mother's freshly cleaned floor.
Yolanda intentionally broke her grandmother's heirloom vase because she didn't want her older sister to inherit it.	Yolanda intentionally broke her grandmother's heirloom vase because her grandmother asked her to get rid of it.	Yolanda accidentally broke her grandmother's heirloom vase. She was helping her grandmother move and packed the vase with thin paper.	Yolanda accidentally broke her grandmother's heirloom vase. She was helping her grandmother move and a mover bumped her from behind.	Yolanda intentionally broke her grandmother's heirloom vase.	Yolanda accidentally broke her grandmother's heirloom vase.
Ophelia intentionally burnt Roland's birthday meal because she wanted to really hurt Roland's feelings by ruining his birthday.	Ophelia intentionally burnt Roland's birthday meal because she wanted an excuse to take him to a new restaurant she knew he wanted to visit.	Ophelia accidentally burnt Roland's birthday meal. She set the oven incorrectly, so the dinner cooked at 450 instead of 350.	Ophelia accidentally burnt Roland's birthday meal. The oven temperature was miscalibrated, so the dinner cooked at 450 instead of 350.	Ophelia intentionally burnt Roland's birthday meal.	Ophelia accidentally burnt Roland's birthday meal.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Drew intentionally gave a customer the wrong change because he wanted to keep the extra money for himself.	Drew intentionally gave a customer the wrong change so that the difference could be donated to a local children's hospital.	Drew accidentally gave a customer the wrong change. Without looking at the bills he gave back $1 instead of a $10.	Drew accidentally gave a customer the wrong change. The cash register malfunctioned and he gave back $1 instead of a $10.	Drew intentionally gave a customer incorrect change.	Drew accidentally gave a customer incorrect change.
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.	Ted intentionally hit a man with his car because he saw the man had a knife and was chasing a young, frightened woman.	Ted accidentally hit a man with his car. Ted didn't check his blind spot before backing up.	Ted accidentally hit a man with his car. Ted's brakes failed to work.	Ted intentionally hit a man with his car.	Ted accidentally hit a man with his car.
Brianna intentionally made the infant sick because the baby was crying a lot, annoying Brianna, and she wanted to get rid of the baby.	Brianna intentionally made the infant sick. The baby had just eaten something poisonous and Brianna needed to make the baby throw it up.	Brianna accidentally made the infant sick. She forgot that the food contained peanuts, which she knew the baby was allergic to.	Brianna accidentally made the infant sick. It turned out that the food contained peanuts, which the baby was allergic to.	Brianna intentionally made the infant sick.	Brianna accidentally made the infant sick.
Barry intentionally divulged a government secret to a foreign spy because he wanted to make a lot of money selling secrets.	Barry intentionally divulged a government secret to a foreign spy. Barry is a double agent and was providing false information.	Barry accidentally divulged a government secret to a foreign spy. Barry was told by his superiors that they were unsure if the spy was on their side.	Barry accidentally divulged a government secret to a foreign spy. Barry was told by his superiors that the spy was definitely on their side.	Barry intentionally divulged a government secret to a foreign spy.	Barry accidentally divulged a government secret to a foreign spy.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Marissa intentionally took the very last handicapped parking space because she wanted to park as close as possible to the store.	Marissa intentionally took the very last handicapped parking space because she didn't want her grandmother to have to walk very far.	Marissa accidentally took the very last handicapped parking space. The sign was concealed but the handicap marking on the ground was clearly visible.	Marissa accidentally took the very last handicapped parking space. There was no sign and the handicap marking on the ground had worn away.	Marissa intentionally took the very last handicapped parking space.	Marissa accidentally took the very last handicapped parking space.
Randy intentionally pulled out some of Eve's hair because he wanted to hurt her.	Randy intentionally pulled out some of Eve's hair because she needed a few hairs for a genetic test.	Randy accidentally pulled out some of Eve's hair. He ran his hand through her hair while playfully pulling on it.	Randy accidentally pulled out some of Eve's hair. He ran his hand through her hair, not knowing she recently started chemo treatments.	Randy intentionally pulled out some of Eve's hair.	Randy accidentally pulled out some of Eve's hair.
Bob intentionally damaged a local wildlife preserve becasue endorsing the new company program would earn him a big promotion with a big salary.	Bob intentionally damaged a local wildlife preserve because he knew it would recover within a year and the program would create 1,000 new jobs.	Bob accidentally damaged a local wildlife preserve. He forgot to read the environmental impact report describing potential damages.	Bob accidentally damaged a local wildlife preserve. He read the environmental impact report and it did not mention such a possibility.	Bob intentionally damaged a local wildlife preserve.	Bob accidentally damaged a local wildlife preserve.
Lisa intentionally shot Tom in the arm because she thought Tom might leave her and couldn't bear to see him with anyone else.	Lisa intentionally shot Tom in the arm because he was breaking into her home and she was trying to defend herself and her children.	Lisa accidentally shot Tom in the arm. She was cleaning the gun, but she forgot to check whether it was loaded.	Lisa accidentally shot Tom in the arm. Her father-in-law had left it loaded and it went off when she was cleaning it.	Lisa intentionally shot Tom in the arm.	Lisa accidentally shot Tom in the arm.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Peter intentionally answered the phone at dinner because he wanted an excuse to not have to talk with his date.	Peter intentionally answered the phone at dinner because he was waiting to hear about his brother's surgery.	Peter accidentally answered the phone at dinner. He meant to silence the phone but pressed the wrong button and the caller started to speak.	Peter accidentally answered the phone at dinner. He tried to silence the phone but it didn't work and the caller started to speak.	Peter intentionally answered the phone at dinner.	Peter accidentally answered the phone at dinner.
Don intentionally accessed Bert's private email because Don wanted to steal Bert's password to his online bank account.	Don intentionally accessed Bert's private email because Bert couldn't get to his computer and asked Don to check something.	Don accidentally accessed Bert's private email. Because they have the same type of computer, Don mistook Bert's computer for his own.	Don accidentally accessed Bert's private email. Bert had checked his email on Don's computer and did not close the browser window.	Don intentionally accessed Bert's private email.	Don accidentally accessed Bert's private email.
John intentionally hit the woman behind him with his elbow because he was tired of her talking on the phone.	John intentionally hit the woman behind him with his elbow because the woman was trying to cut in line.	John accidentally hit the woman behind him with his elbow. He didn't pay attention to those behind him when he put his jacket on.	John accidentally hit the woman behind him with his elbow. He didn't know someone was behind him when he put his jacket on.	John intentionally hit the woman behind him with his elbow.	John accidentally hit the woman behind him with his elbow.
Fred intentionally told his friend the wrong wedding date because he wanted to make it look like the friend was skipping the wedding.	Fred intentionally told his friend the wrong wedding date because the friend was the bride's Ex and wanted to crash the wedding.	Fred accidentally told his friend the wrong wedding date because he was distracted and wrote down the wrong date.	Fred accidentally told his friend the wrong wedding date because the groom had informed Fred of the wrong date.	Fred intentionally told his friend the wrong wedding date.	Fred accidentally told his friend the wrong wedding date.

Intentional + Morally bad reason	Intentional + Morally good reason	Unintentional + Preventable	Unintentional + Not preventable	Intentional	Unintentional
Emma intentionally kicked the dog because she wanted to vent her frustration.	Emma intentionally kicked the dog to prevent it from attacking her little sister.	Emma accidentally kicked the dog. She was dancing and not paying attention to her surroundings.	Emma accidentally kicked the dog. She was carrying a stack of boxes and couldn't see the dog step in front of her.	Emma intentionally kicked the dog.	Emma accidentally kicked the dog.
Ava intentionally went sunbathing nude on a public beach because she enjoyed the attention she received from all the other people.	Ava intentionally went sunbathing nude on a public beach because she was doing a photo-shoot for a top-tier fashion magazine.	Ava accidentally went sunbathing nude on a public beach. She thought it was a nude beach because she missed the signs prohibiting nude sunbathing.	Ava accidentally went sunbathing nude on a public beach. She thought it was a nude beach because the signs prohibiting nude sunbathing were missing.	Ava intentionally went sunbathing nude on a public beach.	Ava accidentally went sunbathing nude on a public beach
Liz intentionally took a t-shirt out of the store without paying because she wanted the shirt but thought it was overpriced.	Liz intentionally took a t-shirt out of the store without paying because it was part of a store promotion.	Liz accidentally took a t-shirt out of the store without paying. She forgot she put the shirt in her bag with other clothes that she already paid for.	Liz accidentally took a t-shirt out of the store without paying. Someone slipped the shirt in her bag with other clothes that she already paid for.	Liz intentionally took a t-shirt out of the store without paying.	Liz accidentally took a t-shirt out of the store without paying.
Jacob intentionally overlooked his old girlfriend at a party because he wanted to make her feel left out and alone.	Jacob intentionally overlooked his old girlfriend at a party. They had a bad break up and he didn't want to upset her.	Jacob accidentally overlooked his old girlfriend at a party. He forgot that she was coming to the party.	Jacob accidentally overlooked his old girlfriend at a party. He didn't know she was at the party.	Jacob intentionally overlooked his old girlfriend at a party.	Jacob accidentally overlooked his old girlfriend at a party.