CHAPTER ONE

Causal Models: The Representational Infrastructure for Moral Judgment

Steven A. Sloman, Philip M. Fernbach, *and* Scott Ewing

Contents

Abstract

This chapter has three objectives. First, we formulate a coarse model of the process of moral judgment to locate the role of causal analysis. We propose that causal analysis occurs in the very earliest stages of interpreting an event and that early moral appraisals depend on it as do emotional responses and deliberative reasoning. Second, we argue that causal models offer the best representation for formulating psychological principles of moral appraisal. Causal models directly represent causes, consequences, and the structural relations among them. In other words, they represent mechanisms. Finally, we speculate that moral appraisals reflect the similarity between an idealized causal model of moral behavior and a causal model of the event being judged.

Psychology of Learning and Motivation, Volume 50
ISSN 0079-7421, DOI: 10.1016/S0079-7421(08)00401-5

1. Introduction

Common sense dictates that moral judgment cannot get off the ground until causes are identified and consequences evaluated. Moral condemnation requires first identifying a transgressor as the cause of pain, suffering, or other contemptible consequences. In this chapter, we will not only embrace this commonsense doctrine, but we will argue that causal structure is so central to moral judgment that representations of causal structure, causal models, serve as the representational medium for appraising and reasoning about the morality of events. Our approach stands in contrast to the classical view that people derive their moral conclusions through a process that resembles proof (Kohlberg, 1986; Piaget, 1932), the more recent view that moral conclusions are expressed by a grammar analogous to those found in language (Hauser, 2006; Mikhail, 2000), and the claim that moral judgment is largely unaffected by cognitive operations (Haidt, 2001). We will argue that causal models provide a representation that allows for a direct expression of moral principles. In the course of making the argument, we hope to go some way toward specifying the role of causal analysis in moral judgment.

We distinguish two aspects of moral assessment: a moral appraisal that occurs early in cognitive processing of an event and a moral judgment that reflects a slower more deliberative process, and may also draw on the initial appraisal. Our discussion focuses on the role of causal models in moral appraisal.

Challenges to the commonsense wisdom that moral attribution requires causal attribution have come in the form of hypothetical counterexamples and empirical demonstrations. Here, we simply list those challenges and identify rebuttals to them rather than reviewing the detailed arguments.

Deigh (2008) suggests that there are some situations where people are held morally responsible for an act that they did not cause. For example, when a group of teenagers beats a pizza deliveryman to death, even those who were present but did not participate in the planning and execution of the act may be held criminally responsible. Driver (2008) points out, however, that even the passive participants might have stopped or mitigated the severity of the event. They have at least some causal responsibility by virtue of not preventing the acts. As such, this example does not challenge Driver's claim that moral responsibility entails causal responsibility. Driver also deals (successfully in our view) with a number of other cases that have been offered as counterexamples to this fundamental thesis.

A greater challenge is offered by Knobe (2003) who shows that people are willing to assign someone blame for a negative foreseeable side effect of an action but not to give credit for a positive foreseeable side effect of an identical action. For example, an executive who harms the environment as a foreseen, but unintended side effect of a program instituted to increase

profits is blamed for the harm, but an executive in a parallel situation who helps the environment as a side effect is not given credit. Knobe argues that this implies that moral appraisals affect attributions of intent. As intentions are causes of intentional action, this implies that moral appraisals can determine causal attributions, suggesting that common sense has it backward some of the time: Rather than causal analysis affecting moral judgment, moral judgment can affect causal analysis. In an ongoing debate, Machery (2008) points out that Knobe's study includes a confound. Specifically, when a side effect is negative, there is more reason not to act than when the side effect is positive. A negative side effect is itself a reason not to act. Given the decision maker's selfish motive to act, there is necessarily more conflict when balancing reasons for and against acting when the outcome is negative than when it is positive because the selfish motive to act must be balanced against the side effect only in the negative case. Hence, attributions of blame for a negative side effect may be greater than attributions of credit for a positive side effect not because of a prior moral appraisal but rather because the decision maker acted in the face of greater conflict in the case of blame.

Another empirical challenge is presented by Cushman et al. (2008). They show that people are more likely to construe a morally bad act as actively doing than as passively allowing. A doctor who unplugs the life-support system of a homeless man because the doctor thinks the homeless man is a worthless burden has killed the man. But a doctor who unplugs the life-support system because he believes it could be used more effectively on someone with more promise of survival has enabled the homeless man's death. This indicates that moral appraisal affects how events are evaluated using causal language. Notice though that this evaluation does not necessarily reflect the initial causal construal of the situation; it could well reflect a considered judgment long after initial interpretation of the event and deliberation has occurred.

In sum, we subscribe to Driver's (2008) thesis that an attribution of moral responsibility to an agent for an event presupposes that the agent is causally responsible for the event. But we note that this does not imply that a complete and final causal interpretation and judgment must occur prior to any moral considerations. In the remainder of the chapter, we provide an introduction to the causal models framework and then offer a view of the cognitive architecture of moral judgment, a modal model, the closest we can come to a consensus in the literature. This will allow us to locate the role of causal analysis in moral judgment. We will see that two roles emerge, one in an early moral appraisal and one in deliberative reasoning. Next, we discuss how the canonical principles of moral appraisal depend on causal models and speculate that the principles derive from a comparison between the causal model of an event being judged and an ideal causal model. We end by comparing our view to some others such as the moral grammar idea (Hauser, 2006; Mikhail, 2000).

2. Causal Models

What form does the causal knowledge required by moral judgment take? Here is what must be represented:

- Actors and their causal roles (e.g., perpetrator, bystander, or victim — better, terms that express those causal roles without any moral connotation);
- Physical capacities and their causal roles (e.g., an agent's size or skills with a weapon may have enabled or prevented an outcome);
- Mental states and their causal roles (e.g., intentions, beliefs, and desires are the causes of intentional action; cf. Malle (2007));
- Objects (e.g., weapons, money, etc.) and their causal roles (e.g., are they enablers, disablers, potential disablers, are they on the causal path leading to a consequence or possible consequence?);
- Actions and how they relate actors and objects to one another (e.g., shooting a gun at someone is a mechanism that relates a shooter and a gun as causes to an effect consisting of the end state of the person shot).

Notice that identifying causes and consequences is not nearly enough to make informed decisions and predictions. The structural relations among the causal elements are critical because they link individuals to consequences, specify the requirements for an action to be effective (e.g., that its enablers are present and functioning), and indicate the joint outcome of multiple actions. They also indicate the outcome of counterfactual considerations. That is, inferences about "what might have been" or "what would happen if" can be inferred from knowledge about how events cause other events. Once those causal relations are known, we can use them to determine (for instance) what the effects would be of any assumed set of causes.

The inferences that we are able to make are detailed and specific and therefore require a detailed and specific representation of causal structure. We call such a mental representation a causal model. The causal analyses that people engage in for the sake of moral judgment of a specific situation are likely to involve identifying and making inferences from simple qualitative relations among the elements of an event.

The causal model of a specific event must derive, at least in part, from more abstract, general knowledge. For instance, a model of a specific car accident is constructed using more abstract knowledge about skidding, the effects of contact, brake failure, etc. In this sense, a causal model of a specific event derives from models that describe events of more general types.

A more formal representation of causal structure starts by representing the constituents of the event as random variables that can take different values. A representation of a car accident might include a random variable

for skidding that takes the value 1 in the case that skidding is present and 0 otherwise. Structural relations in the event can be represented by specifying how the values of the constituents of the event change in response to changes in the other constituents. For instance, if the accident representation includes a random variable denoting the presence or absence of ice, this variable should affect the probability of skidding.

One type of model specifies a joint probability distribution over all of the values of all of the constituents of the event. This distribution can be represented economically by a graph with nodes that represent the constituents and links between nodes representing structural relations. The joint distribution can then be expressed by specifying the distributions of the exogenous or root nodes, nodes whose distribution is not determined by any other nodes in the graph, and a set of equations relating the distributions of linked nodes. If the form of the graph obeys certain constraints (e.g., the links are all directional and it has no cycles) it is referred to as a Bayes net (Pearl, 1988).

A Bayes net represents relations of probability, not necessarily causality. Woodward (2003) argues that what distinguishes a causal relation from a merely probabilistic one is that it supports intervention. Roughly speaking, A causes B if a sufficiently strong intervention on A by an external agent would affect the value of B. A causal Bayes net (Pearl, 2000; Spirtes et al., 1993) is a Bayes net in which the links represent causal mechanisms and operations are defined that support the logic of intervention. Pearl (2000) defines an intervention as an action that is external to the graph that sets a variable in the graph to a particular value. An intervention that sets a variable X to a value x is written as $do(X = x)$. The effect of an intervention is to remove the incoming links to the intervened-on variable, rendering it independent of its normal causes. One way to represent an intervention is as a new node in an augmented graph. The intervention $do(X = x)$ is encoded by drawing a link between the new intervention node and the target of intervention X and erasing all other incoming links to X. Figure 1A shows a very simple causal Bayes net representing a traffic accident. Figure 1B shows the same network after an intervention which cuts the brake lines.

The intervention $do(\textit{brake failure} = 1)$ sets the variable ''brake failure'' to 1. This is encoded by drawing a link between the intervention and its target and by erasing the link from the normal cause of brake failure, worn out brake pads. One outcome of severing the link between brake failure and its normal cause is that the variables are no longer informative of one another. Under normal circumstances, the failure of the breaks would increase the probability that the car has worn out brake pads. After the intervention no such diagnostic inference is possible. Predictive inferences are still possible because the outgoing links remain intact. After the brake lines are cut, the probability of an accident is high.

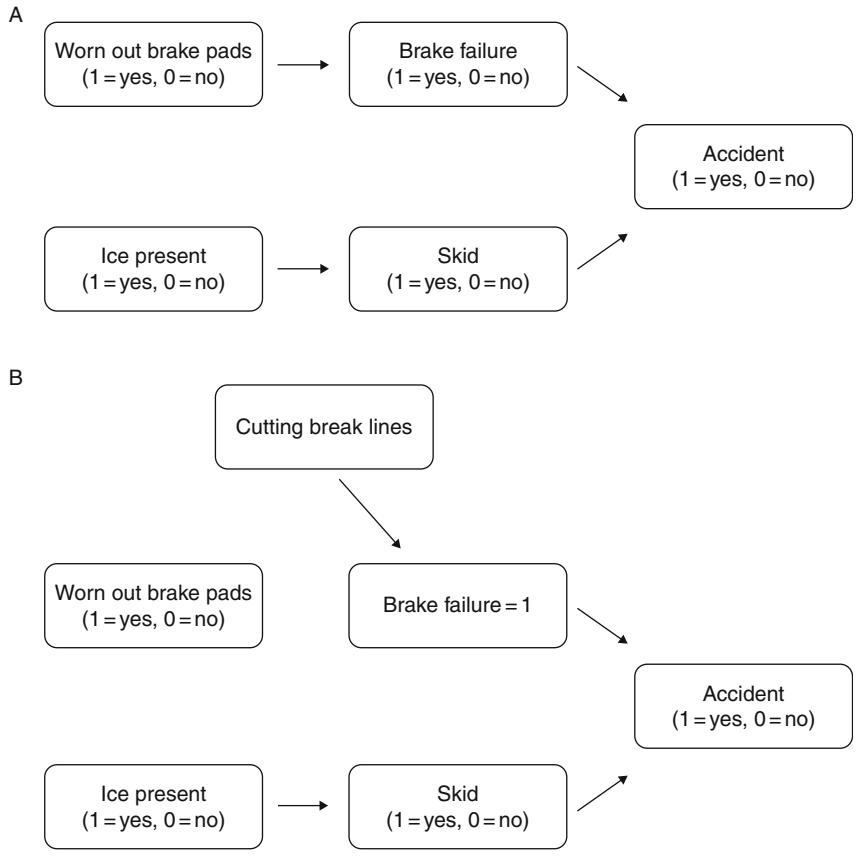

Figure 1 (A) A Causal Bayes net representing a traffic accident. (B) the same network after an intervention which cut the brake lines.

The *do* operation is a device for representing actual physical interventions as well as counterfactual interventions. A causal model can be used to answer counterfactual questions about what would be the case if some condition X had value x simply by using the operator to set the value of X, $do(X = x)$.

Our use of the term ''causal model'' is inspired by causal Bayes nets in that both use sets of links to represent mechanisms, have graph structures that correspond to causal structure, and subscribe to the logic of intervention. However, we do not intend to suggest that judgments are always coherent in the sense of being internally consistent as prescribed by probability theory. Moreover, for reasons spelled out below, we will sometimes draw links between values of variables and interpret them not merely as mechanisms describing potential relations between causes and an effect, but

as an active mechanism that supported the transfer of a conserved quantity along its path.

Sloman (2005) makes the case that causal models describe a basic form of mental representation. Clearly many cognitive functions presuppose assignment of causal structure to the world. Fernbach, Linson-Gentry, and Sloman (2007) and Humphreys and Buehner (2007) show that causal knowledge mediates perception. Causal structure plays a role in language (e.g., Abelson and Kanouse, 1966; Brown and Fish, 1983), category formation (Rehder and Hastie, 2001), and attributions of function (Chaigneau et al., 2004). Causal structure is also central to reasoning and decision making (Sloman and Hagmayer, 2006). Our working hypothesis is that causal models serve as the primary representational medium for moral judgment.

3. Architectural Considerations

Both behavioral and brain imaging data indicate that moral judgments have at least two bases, a deliberative one and a more intuitive one (e.g., Cushman et al., 2006; Greene, 2008; Pizarro and Bloom, 2003). Although there is a debate about the precise nature of the underlying systems (Greene, 2007; Moll and Oliveira, 2007; Moore et al., 2008), wide consensus obtains that two systems support moral judgment in the same way they support reasoning and decision making (Evans, 2003; Kahneman and Frederick, 2002; Sloman, 1996; Stanovich and West, 2000). In the case of moral judgment, the intuitive basis consists of an almost immediate sense of right or wrong when presented with a situation that merits moral appraisal. Figure 2 illustrates our interpretation of the modal model of the processing of moral judgments at the time of writing.

Haidt (2001) offers an intuitionist theory of moral judgment according to which judgments are not the product of deliberative reasoning; rather,

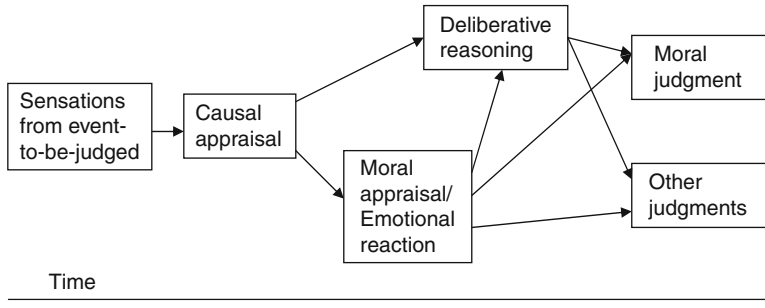

Figure 2 The modal model of the order of mental operations resulting in moral judgment.

reasoning comes after judgment in the form of justification rather than motivation. The modal model accommodates this idea by treating moral appraisal as an input into the deliberative process. Our deliberations take into account the causal structure of the situation along with our immediate moral appraisals and emotional reactions to try to make sense of them. If they can, they are in essence justifications and the final moral judgment will be consistent with them. But if they cannot, or if the initial appraisal is trumped by a conflicting moral rule, then we will change our minds and produce a moral judgment whose motivation really does lie in a deliberative process. Thus, in contrast to suggestions by Nichols and Mallon (2006) and Bartels (2008), we consider strict moral rules (e.g., "Do not have any other gods before me") to be enforced by deliberative reasoning. This explains why emotional reactions (like sexual attraction) can conflict with moral rules.

We refer to the initial determination of causal responsibility as "causal appraisal" and the initial determination of moral responsibility as "moral appraisal" but these are not to be confused with final, observable judgments. Hauser (2006) follows Mikhail (2000) and Rawls (1971) in arguing that moral appraisals are made by moral grammars, analogous to the kind of linguistic grammars proposed by Chomsky (1957). For supporters of the linguistic analogy, moral appraisals are unconsciously formed by principles that are not accessible to introspection.

We agree that moral appraisals emerge early in the flow of processing. Consider:

> A biker gang has moved into a rural community and is intimidating the local residents. One day a father and his son are walking down the street and they cross paths with a couple of the bikers. One of the bikers steps in front of the son and, while staring into the father's eyes, punches the son in the face and says to the father, "What are you going to do about it?"

We surmise that the reader has at least two initial reactions to this story. One is cognitive, absolute contempt for the biker's action. Let us call this immediate appraisal "moral disapproval." The second is emotional, namely anger and perhaps loathing of the biker and his ilk. What causes what? Does the anger cause the disapproval, the disapproval the anger, or are they both caused by a third state? Surely emotions influence moral judgment (Greene et al., 2001), but in this case the anger could not be responsible for the disapproval because there would be no anger if the act were not disapproved of. The anger presupposes something to be angry about which in this case is that the act is morally out of bounds. It may be that anger can arise in the absence of moral appraisal. For instance, serious sudden pain can cause anger (e.g., accidentally closing a car door on your finger). But in the story just told, the reader is merely an observer and thus has no such pathway to anger; the only available source is the heinousness of the action. So either the

disapproval is the cause of the anger or, as Prinz (2006) argues, the disapproval is a moral sentiment that exists in virtue of the anger. The disapproval may be constituted in part by the anger. Either way, a moral appraisal has occurred and it has occurred no later than a very swift emotional reaction. So moral appraisals can occur at the very earliest stages of processing. Indeed, appraisal theories of emotion also stipulate that an emotional response presupposes an appraisal (Clore and Canterbar, 2004; but see Berkowitz and Harmon-Jones, 2004).

The reason to distinguish moral judgment from moral appraisal is that final judgments do not always coincide with initial appraisals or with emotional reactions. In the original trolley problem (Foot, 1978), the initial distaste for killing a single person is overcome after considering the other option of allowing five to die.

4. Roles for Causal Models

4.1. Appraisal

Causal models play several roles in moral judgment. At the early moral appraisal stage, they structure the understanding of events by assigning causal roles to actors, their capacities and mental states, objects, and actions. This involves such inferences as determining the intention of actors (did he want to hurt the victim or was the outcome accidental?) and attributing causal responsibility for the outcome.

Causal models do not offer a process model of moral judgment; rather they describe a form of representation that different judgment processes rely on. They can be used to make different inferences depending on the judge's task. Cushman (2008) shows that people use different criteria for moral inference depending on the question they are asked. Judgments of wrongness and permissibility depend on analysis of mental states whereas judgments of blame and punishment are more sensitive to an analysis of causal responsibility. Cushman attributes these different criteria to the operation of different judgment systems. Our analysis, in contrast, allows that the different criteria reflect the same cognitive operations on the same representation of the event. They merely reflect different directions of causal inference. If asked for a judgment of wrongness or permissibility, people make a diagnostic inference from an outcome upstream to a cause of the outcome, the intention of the action that produced it, a mental state. But when asked to evaluate blame or punishment, the inference tends to go in the opposite direction, from the person's intention to the outcome of their actions. Whichever direction is focused on, a causal model relating intention to action and outcome is necessary.

Moral judgments sometimes require prediction of an outcome or a counterfactual assessment after an outcome occurs of the probability of that or some other outcome given the presence or absence of the action. Prediction is necessary if the outcome has not yet occurred. The moral worth of a government's environmental policy depends on an assessment of the probability that it will lead to serious environmental harm in the future. Counterfactual likelihoods are relevant when an event has already occurred. A moral assessment of the invasion of Iraq in 2003 depends in part on a judgment of the probability that there would have been terrorist strikes on American soil in the immediately subsequent period if the invasion had not occurred. Naïve individuals have little to go on to make these kinds of probability judgments other than their causal models of global warming or terrorist trends, respectively. They can of course appeal to expert judgment, but experts themselves rely heavily on their own causal models to determine the likelihoods of various consequences (Klein, 1999).

4.2. Deliberation

Prediction and counterfactual inference are sometimes deliberative affairs that involve time and reflection. But they nevertheless tend to involve causal structure. Debates about the value of an invasion often involve differences of opinion about the likelihood of various outcomes. Sometimes statistics based on historical precedent enter into consideration, but even then the relevance of the statistics depends on causal beliefs. For instance, whether or not the prevalence of terrorist activity can be generalized from one country to another depends entirely on how the environment in one country differs from the other in terms of its degree of logistical and political support for and responsiveness to terrorist activity. These are the kinds of considerations that are represented in a causal model.

Note that utility analysis depends on prediction and counterfactual judgment. So, if such judgments make use of causal models, it follows that utility analysis does too. In other words, causal models are critical to apply utilitarian principles to derive solutions to moral problems. This point has been made by a number of philosophers (Meek and Glymour, 1994) including some who have proposed causal utility theories (Nozick, 1993; Joyce, 1999; Skyrms, 1982) though there have been dissenting voices (Levi, 2000). And people turn out to reason causally about choices very naturally and effectively (Sloman and Hagmayer, 2006). The probability of the possible consequences of an action refers to the probability of possible effects of a cause.

What about deliberative moral judgments based on deontological principles rather than utilitarian analysis? Do people require causal models to draw conclusions based on universal principles like "one should never push people off bridges to their death" or "killing an innocent human being is

never permitted''? Notice that such principles themselves require an appropriate underlying causal structure. This is most easily seen by the fact that they are expressed using causal verbs. ''Pushing'' and ''killing'' both imply an agent, a recipient, and a change of state. So applying such principles requires instantiating the agent (cause) and change of state of the recipient (effect). Even the golden rule is a causal principle that involves the *sina qua non* of causality, intervention. Doing unto others is an act of intervention that directly affects someone else.

Deliberating about abstract moral principles is largely an exercise in abstract causal reasoning. Once primitive values are specified (life, liberty, pursuit of happiness, etc.), deciding how to formulate a code of ethics is largely a matter of choosing causal laws that have primitive values as consequences (e.g., thou shalt not kill, no detention without due process, etc.). These causal laws are not causal models themselves but rather principles for generating causal models that will ensure justice for different classes of situations (e.g., a court system to try criminals, a medical system to maximize health, etc.). In this sense, Hauser's (2006) moral grammar can be conceived of as a set of causal laws.

5. Moral Principles That Draw on Causal Structure

A guiding puzzle in moral psychology is to determine the aspects of an event that modulate moral judgments about it. Our central claim is that every moral principle that has been seriously considered as a descriptor of the process of moral appraisal depends on a causal model representation of event structure. Cushman et al. (2006) suggest three principles. The intention principle states that bad outcomes that are brought about intentionally are morally worse than unintended outcomes. The action principle, usually referred to as *omission/commission*, states that, *ceteris paribus*, actions are more blameworthy than inactions. The contact principle states that an action that brings about harm by physical contact is morally worse than an analogous action that does not involve contact. We see the contact principle as a special case of *causal proximity* (Alicke, 2000). Actions that are connected to bad outcomes through fewer intermediate causes are more blameworthy. To this list we add *locus of intervention*. Work by Waldman and Dieterich (2007) suggests that an intervention on a victim is more reprehensible than intervention on the agent of harm. Throwing a person on a bomb is worse than throwing a bomb on a person. We also assess the principle of fairness, which is fundamental to many moral judgments (Rawls, 1971). Fairness is usually construed acausally, as an evaluation of the way in which goods ought to be distributed. But fairness is also influenced by causal structure.

5.1. Intention

A critical factor in attributions of moral responsibility is the intention of the actor. In Western law, attributions of intention are generally required to convict for certain crimes like murder. The importance of an actor's intention for attributing blame is manifest in many philosophical principles including the principle of double effect which concerns the ethics of bad side effects of good actions. It is generally accepted that acts with the same consequences should be judged differently depending on their guiding intentions (Foot, 1978). For example, killing civilians intentionally in wartime is not the same as killing civilians, even knowingly, as a side effect of destroying a valuable military target.

Young et al. (2007) offer a demonstration. Grace and her friend are taking a tour of a chemical plant. Grace goes to the coffee machine to pour some coffee, and her friend asks for sugar in hers. The white powder by the coffee is not sugar but a toxic substance left behind by a scientist. In the intentional condition, the substance is in a container marked "toxic," and Grace thinks that it is toxic. In contrast, in the nonintentional condition the substance is in a container mislabeled "sugar," and Grace thinks that it is sugar. In both conditions, participants are told that Grace puts the substance in her friend's coffee. Her friend drinks the coffee and dies. Participants are asked to rate the moral status of Grace's action. The result is an enormous effect of intention. In the intentional condition, participants judge Grace's action as forbidden. In the nonintentional condition they judge it as permissible.

The effect of intention pertains even when the outcome is foreseen. Mikhail (2000) gave participants a scenario that was similar to the standard trolley problem but was varied such that the actor's intention was to kill the single individual, not to save the five. Participants were told that the bystander who had the choice to throw the switch hated the man on the alternate track, wanted to see him dead, and that his decision to divert the train was explicitly intended to kill the man. Throwing the switch was viewed as far worse than in the standard dilemma where the outcome, the death of the single individual, is a foreseen but unintended side effect of saving the five.

In one sense, an intention is a root cause in a causal model. It represents the will of an agent. If one attributes free will to the agent, then the intention is not determined by anything else. Of course, intentions are influenced by other variables. In Malle's (2001) model of intentional action, intentions are influenced by an agent's beliefs and desires. One relevant belief is that the action will produce the intended outcome and a relevant desire is that the outcome will come about. A judge might ask what an agent's beliefs and desires are or even why the agent has such beliefs and desires. If answers are forthcoming, then the causal model is peeled back one more layer by assigning causes to intentions or even two more by assigning

causes to beliefs and desires. But the causal models necessary for everyday moral judgment do not usually require historical elaboration. People rely on the minimal structure necessary by not thinking beyond, or too far beyond, an actor's intention.

In many vignettes in the literature, the intention of the actor is stated outright or strongly implied. In that case, intention can simply be represented as a node in the causal model of the event. Often, however, intention is an unobservable variable that must be inferred prior to making a moral appraisal. Consider a case where a young man pushes an old woman. A moral evaluation of the action is contingent on the young man's intention. It may not be necessary that he intends the outcome, we may be satisfied that he is to blame if he was merely negligent, but the valence of his intention will nevertheless affect our appraisal. If his intention is to push the woman out of the way of a car, it suggests a different judgment than if his intention is to injure her. One feature of causal models is that they support diagnostic reasoning to hidden causes from the status of effects of those causes. Our moral infrastructure hypothesis suggests that people represent the causal structure among relevant variables of an event prior to making a moral appraisal. This includes information that *prima facie* may seem irrelevant to assessing the morality of the outcome but is made relevant by its evidential power to diagnose intention. For instance, if the young man had yelled ''watch out'' to the old woman prior to pushing her, it would support a diagnostic inference to a good intention. Thus causal models not only supply a way to represent intention as a cause of an action to be judged, but also a computational engine for inferring intention on the basis of causal structure.

5.2. Omission/Commission

Acts of commission that lead to bad consequences are usually judged more blameworthy than acts of omission with the same consequences. Spranca et al. (1991) asked people to judge a number of such cases. In one vignette, a tennis pro is eating dinner with an opponent the night before a match. In one condition, the commission case, the tennis pro recommends a dish that he knows will make his opponent ill in order to gain an advantage in the match. In the omission case, the tennis pro fails to warn his opponent against the dish that will make him ill. In both cases the outcome is the same. The opponent orders the dish, becomes ill, and the tennis pro wins the match. People judged the tennis pro to be more blameworthy in the commission case.

It turns out that a parallel law to commission/omission distinction in moral judgment applies to attributions of causality. Counterfactual relations are often insufficient for attributions of actual cause. The fact that an event B would not have occurred if event A had not even in the absence of other

causes of B is not generally sufficient for people to assert that A causes B (Mandel, 2003; Walsh and Sloman, 2005; Wolff, 2007). People sometimes require that a mechanism exist from A to B, what philosophers call causal power. For instance, if Suzy opens a gate that allows a boulder to pass through and knock a man off a cliff to his death, then people tend to assert that Suzy was the cause of the man's death. But if Suzy merely sat in a parked car beside the open gate and failed to close it, then she is not the cause even if she could foresee the outcome (Clare Walsh, personal communication). Omission involves an outcome due to failure to act, which is similar in the sense that no active mechanism links the omitted action to the causal path that leads to the outcome. Commission involves precisely such a mechanism. In that sense, the commission/omission distinction can be reduced to the principles of operation of naïve causal reasoning.

The absence of an active mechanism from the action to the outcome in the case of omission means that there must be some other sufficient cause of the outcome that is independent of the agent. In other words, acts of omission involve a failure to intervene on a causal system that is heading toward a bad outcome; the system's preexisting causes are sufficient for the outcome. Consider the tennis pro's act of omission. The opponent's desire to order the dish that would make him ill is sufficient to bring about the outcome, irrespective of the tennis pro's desire that he get sick. No such independent sufficient cause is required in cases of commission because the action itself is such a cause. Recommending the dish to the opponent is part of the mechanism leading to the opponent's sickness in the commission condition.

In sum, acts of commission and omission differ in two structural ways (presence versus absence of a mechanism and of an alternative sufficient cause). Graphs in standard Bayes nets do not represent mechanisms *per se*, they represent relations of probabilistic dependence. But the notion of mechanism requires more than probabilistic or even counterfactual dependence. One interpretation is that a mechanism involves a conserved quantity (like energy or symbolic value) that travels from cause to effect (Dowe, 2000). Philosophers often talk about this notion of mechanism that entails more than probabilistic dependence in terms of causal power. In our causal models of specific events, we will represent the passing of a conserved quantity as an active mechanism. Using links to represent only active mechanisms, we can then illustrate the difference between commission and omission using causal models as we do in Fig. 3. What the graphs make apparent is that, when the cause is an action, the presence or absence of an active mechanism identifies whether or not the agent intervened. An idle intervention is equivalent to no intervention and occurs only in the absence of an active mechanism and an active mechanism in turn requires that the agent be intervented actively.

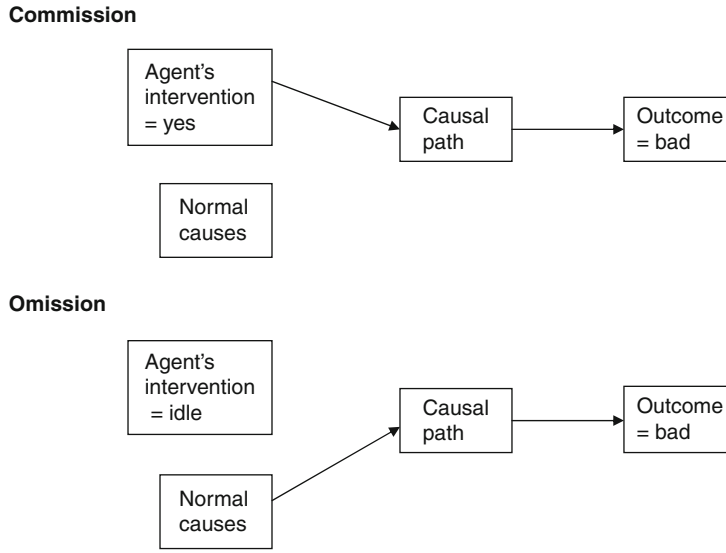

Figure 3 Causal models of abstract cases of commission and omission.

Causal models dictate how causes combine to bring about outcomes. In the case of omission, there exists a sufficient cause that would bring about the outcome in the absence of the action. In the case of commission, there is no such sufficient cause. Commission and omission are distinguished by the active mechanisms that produce the outcomes.

Why should this structural difference have a moral implication? The relevant moral principle, expressed in terms of causal models, is that an agent is more morally responsible for an outcome if a mechanism links them directly to the outcome. In other words, an agent is more morally responsible if their action can be construed as an intervention that led to the outcome. One implication is that an agent is not more morally responsible merely for increasing the probability of the outcome but has to be linked to it by a mechanism.

One effect of the difference in causal structure is that commission and omission suggest different counterfactual possibilities. Causal models support counterfactual reasoning (Pearl, 2000). By hypothetically setting the variables in the model to particular values and observing the values of other variables, one can infer counterfactual possibilities. Like causal attribution (Lewis, 1973; but see Mandel, 2003), moral reasoning can be guided by counterfactual considerations. In the example above, a judge might wonder what would have happened had the tennis pro not been present at all. The sufficiency of the alternative cause in the omission case suggests that the opponent would have gotten sick anyway. Conversely, the lack of an alternative cause in the commission case suggests that the tennis pro really is to blame.

Spranca et al. (1991) report that participants justified their responses to omission/commission vignettes by appealing to a variety of factors including alternative causes, sufficiency, and counterfactual considerations. Our analysis suggests that all of these factors can be understood as properties of an underlying causal model.

5.3. Causal Proximity

The principle of causal proximity is that people who meet other necessary requirements are morally responsible for an outcome to the extent that their actions are direct causes of the outcome. They are reprieved of moral responsibility to the extent that the effect of their actions is indirect. Alicke (2000) posits a model of blame assignment that explicitly includes causal proximity as a factor. Many other theories have included principles that can be construed as variants of causal proximity, such as directness (Cushman et al., 2006), whether an act is personal or impersonal in its relation to a victim (Greene et al, 2001), and whether battery is committed, that is, the victim has been touched without his or her consent by another person (Mikhail, 2000). The most extreme case of causal proximity is when a perpetrator has physical contact with a victim. This is one possible reason for the divergence in responses to the ''trolley'' and ''footbridge'' dilemmas. On this interpretation, the reason that pushing a fat man off the bridge seems so reprehensible is that it requires direct contact with the victim unlike pulling a lever to send a trolley onto a different track. To test this idea Cushman, Young, and Hauser introduce a scenario in which, instead of pushing the fat man off the bridge, one must pull a lever that opens a trap door that drops him onto the tracks. People were more willing to pull the lever than to push the fat man directly.

The impact of causal proximity has also been shown in cases that do not include direct physical contact. One example is derived from a pair of vignettes reported in Hauser et al. (2008). In the proximal case,

> Wes is walking through a crowded park on a cold winter evening. He is nearly home when he sees a homeless man. The man has no winter clothing, and soon he will freeze and die. Wes is wearing a warm coat that he could give to the man, saving his life. If Wes keeps his coat, the homeless man will freeze and die. If Wes gives the homeless man his coat, the homeless man will survive.

In the less proximal case,

> Neil is walking through a crowded park on a cold winter evening. He is nearly home when he sees a collection station for donations to the homeless. A sign explains that most homeless people have no winter clothing, and that dozens will freeze and die every night in the winter. Neil is wearing a warm coat that he could put in the collection station, saving the life of one

> homeless person. If Neil keeps his coat, a homeless person will freeze and die. If Neil puts his coat in the collection station, a homeless person will survive.

Most people think it is more morally permissible for Neil to keep his coat than for Wes. This may because they believe Wes's action is more certain than Neil's to save a homeless person or because of the number or type of alternative possible actions that the scenarios conjure up. We propose that all of these possibilities are a consequence of the greater causal distance between Neil's potential action and saving a homeless person than Wes's.

In a causal model representation, causal proximity depends on the number of mediating causes between the action to be judged and the outcome. For example, pulling a lever is more distant than pushing a man because there is a mechanism, or series of mechanisms between the action on the lever and the outcome of the fat man falling. In the direct case there is no such mechanism.

As in the case of omission/commission, the effect of causal proximity may be to dilute the causal responsibility of the actor. The presence of a causal chain between actor and outcome has at least two implications for assigning causal responsibility. First, in attributing cause, there are salient alternatives to the action being judged, namely the intermediate causes separating the actor from the outcome. If a captain commands a private to shoot a civilian, then the private becomes a cause of the death. Further, there might be supporting additional causes for the private's action. He might like killing and planned to shoot the civilian before receiving the order. In either case, causal attribution of the effect to the captain might be attenuated. More generally, as causal distance increases the number of intervening causes increases, and the greater the possibility for attenuating responsibility to the root cause. Second, in a causal chain with probabilistic links, the probability of the ultimate effect decreases with the length of the chain. This means that the more intermediate causes, the less likely the action will lead to the outcome. For instance, in the example above the private might fail to follow the captain's order. Thus, the ability of the actor to predict the outcome with certainty decreases with causal distance. This could dilute causal attribution by increasing the possibility that the outcome came about due to chance. It could also weaken judgments of intention. According to the folk theory of intentional action (Malle, 2001; Malle and Knobe, 1997), intention attribution is a function of belief that the outcome will happen. The greater the causal distance, the less belief the actor has that the outcome will come about. As discussed above, intention strongly influences moral appraisal. If one effect of causal distance is to weaken judgments of intentionality then it would follow that it should also weaken appraisals of moral responsibility.

5.4. Locus of Intervention

As discussed above, acts that bring about outcomes in a moral situation can be seen as interventions on that situation. Waldmann and Dieterich (2007) argue that moral judgments are influenced by the locus on which the actor intervenes in the underlying causal model. Interventions that influence the path of the agent of harm are more permissible than interventions on the potential victim. This is one possible explanation for the divergence in judgments between the ''trolley'' and ''footbridge'' dilemmas. In the classic trolley problem, the actor redirects the path of the agent of harm, the trolley. In the ''footbridge'' problem, the intervention is on the victim, the fat man who subsequently dies.

Waldmann and Dieterich (2007) compared a number of scenarios where an intervention either influenced the path of the agent or the victim. Interventions on agents were always judged more permissible. The effects cannot all be explained by causal proximity. Here is one example:

> *Agent intervention*: A torpedo threatens a boat with six soldiers. Destroying the torpedo by remote control would sink a nearby submarine with three soldiers.
> *Victim intervention*: A torpedo threatens a boat with six soldiers. Three soldiers could be ordered to move their boat in a way that would divert the torpedo from the original target to their boat.

The scenarios vary with respect to locus of intervention but do not obviously vary in terms of causal proximity. Still, participants judged the agent intervention to be more permissible.

Kant (1785/1998) argued that human beings should never be used as a means to achieve a goal. This suggests the possibility that people's intuitions about these scenarios come not from a causal analysis of the locus of intervention but rather from a strict deontological principle, a prohibition against using human beings as a means under any circumstances. Waldmann and Dieterich (2007) show that people are sometimes willing to use the victim as a means to save others as long as the intervention is on the agent of harm. Participants were given a variant of the trolley problem where in order to stop a train from killing five people the train can be diverted onto a sidetrack where one person is standing. The key manipulation was whether the sidetrack loops back to the main track. In the means condition, the effect of the person on the sidetrack is to stop the train. If the person were not there, the train would continue on its course back to the main track and still kill the five, even if originally diverted. Thus the intervention is on the agent of harm, the train, but the person on the sidetrack is used as a means to stop the train from looping back to the main track. Participants rated diverting the train as permissible. Locus of intervention and the means principle are closely related. Victim interventions tend to violate the

principle. Agent interventions usually do not. Evidently, though, people's moral intuitions are not captured by a strict deontological principle but rather are a function of their causal model.

According to Waldmann and Dieterich (2007), the reason that the locus of intervention is morally relevant is psychological; it shifts attention to the target of the intervention. In the case of the agent intervention, it is natural to consider the two possible causal paths of the agent of harm (e.g., the train continuing on its path or being diverted). In that case, the utilitarian comparison between five dead and one dead comes into focus. Conversely, the victim intervention leads to a focus on the possibilities associated with the victim, the comparison between the victim living versus dying. This backgrounds the utilitarian considerations and makes the intervention harder to justify. This is reminiscent of the difference between commission and omission: The moral principle derives from the counterfactual possibilities that come to mind when considering the effect of the action or inaction. The counterfactual possibilities are brought to mind by the knowledge of the causal structure.

5.5. Fairness

Rawls (1971) proposes that the central principle of justice in a society, the principle from which all others derive, is fairness. Fairness is most naturally thought about in terms of how goods are distributed. All else being equal, available goods should be distributed equally and all deviations from equality need justification. In actual distributions of goods in the world, deviations are commonplace and justifications have a causal rationale.

An idealized illustration of how this works comes from experiments on the ultimatum game, a simple game involving two players. The first player, "the proposer," is given a fixed amount of money to split any way he chooses. The second player, "the responder," decides whether to accept or reject the split. If he accepts, the money is distributed according to the proposal. If he rejects, neither player receives anything. Rational agent models predict that the proposer will make the smallest possible offer and that the responder will accept it. In fact, proposers tend to offer much more than is predicted by these models, and responders often reject even fairly large offers (Oosterbeek et al., 2004). The ultimatum game is thus a good test bed for assessing goods distributions that people deem fair.

Research on the ultimatum game has shown substantial cross-cultural differences in how people play the game. We suggest that at least some of this difference can be explained by the players' causal beliefs about how the proposal is generated. For example, Gypsies in the Vallecas neighborhood in Madrid, Spain often accept an offer of zero, and when asked to justify their behavior they say that the proposer probably needed the money (Branas-Garza et al., 2006). The evaluation of whether the proposal is fair is

contingent on an analysis of the causes of the proposal. This can also be seen in Blount (1995) who found greater willingness to accept small offers when players believed that they were generated by a chance device than by other players. Another way that causal considerations enter into decision in the ultimatum game is that proposers and responders consider the effects of their decisions. For example, an important determinant of behavior is fear that one's reputation will be damaged by appearing to be unfair (Gil-White, 2003). Our causal infrastructure hypothesis makes sense of these effects by assuming that people represent the causal structure. This allows them to make diagnostic inferences about the causes of the proposal and to make predictions about the consequences of their decisions. Causal structure supports a moral appraisal about the fairness of the proposal. It also supports decision making, which is based on moral considerations and other considerations like effects on reputation.

Rawls (1971) assumes that the overriding determinant of fairness is egalitarianism. And we are not suggesting that causal models provide any justification for egalitarianism or indeed for any basic values. Nevertheless, causal models do make a contribution to our sense of fairness by providing a framework for expressing the reasons for deviations from egalitarianism.

5.6. Putting It All Together

We have reviewed how causal models contribute to five principles of moral appraisal. Much of the contribution depends on the specific content of our causal beliefs. In particular, the role of causal models in determining fairness is content specific. However, the other four principles may derive from a more basic cognitive process. We offer a speculation that much of moral appraisal reflects the extent to which the causal model of the event being judged deviates from an idealized causal model (cf. Shaver, 1985). The idealized causal model is exceedingly simple. It states that the most extreme good or bad action consists of an intention for a respectively good or bad outcome with the consequence that the intended outcome occurs (see Figure 4). Normally, of course, an action mediates the causal relation between intention and outcome. But in the ideal case, where it is possible for mere intentions to cause outcomes, this would not be necessary. Our idea is that the moral appraisal of an event is positive or negative in proportion to the degree of similarity of the causal model of the event to the good or bad ideal, respectively. Ideals are not necessarily fixed entities. How good or bad an outcome is may depend on which comparison outcomes come to mind most easily. More generally, ideals may be generated on the spot in response to specific events the same way that surprise is determined by how a contrasting comparison event is constructed at the moment of perception (Kahneman and Miller, 1986). This latter possibility

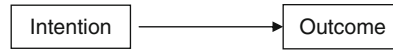

Figure 4 Idealized causal model for evaluating morality of an event.

is consistent with the idea that the repugnance of certain acts depends on what other acts are under consideration (cf. Unger, 1995).

This form of the ideal causal model suggests three dimensions of similarity that modulate appraisals of moral responsibility:

1. Is there an intention to bring about the outcome?
2. Is the intention the cause of the outcome?
3. How good or bad is the outcome?

Our proposal is that moral appraisal varies directly with these three factors and that basic principles of moral appraisal are reflections of this dependence. We consider each of the four principles in turn.

The principle that the actor must intend the action follows immediately from the first dimension. The fact that some actions that are not intended are nevertheless culpable, like acts of negligence, reflects the fact that the causal model of a situation could be similar to the ideal even in the absence of an intention. If an actor did something to make a foreseeable bad outcome very likely through negligence, then the model of the event is similar to the ideal even in the absence of a bad intention.

The difference between omission and commission rests primarily on the second dimension. In the case of omission, the outcome would have occurred anyway so the intention does not have as much causal force as it does in the case of commission. Perhaps more important, attributions of cause are stronger when there is a mechanism that connects a cause to an outcome as in the case of commission.

Causal proximity also reflects differences on the second dimension because the presence of other causes dilutes causal responsibility. The more the outcome depends on other mediating causes, the less power the target cause has to produce the outcome. Causal proximity may also influence attributions of intention.

Locus of intervention depends in part on the same considerations as causal proximity. But it also affects the third dimension because it changes how we view the outcome. An outcome is more acceptable when it is compared to even less desirable alternatives. In the case of the agent intervention, the natural comparison is between the alternative actions of the agent of harm, for example, the death of one and the death of five in the trolley problem. Victim interventions lead to a comparison between actual and counterfactual consequences to the victim. This highlights how bad the outcome is for the victim.

6. Conclusions

We have tried to accomplish three objectives in this chapter. First, we have formulated a coarse model of the process of moral judgment that allowed us to locate the role of causal analysis. We have proposed that causal analysis occurs in the very earliest stages of interpreting an event and that early moral appraisals depend on it. In turn, at least some emotional responses depend on moral appraisals. Deliberative reasoning also relies on causal structure.

Second, we have argued that the causal model formalism is appropriate for formulating psychological principles of moral appraisal. This could be construed as an argument that causal models serve as the underlying representation on which the "moral grammar" (Hauser, 2006; Mikhail, 2000; Rawls, 1971) operates. The primary utility we see for causal models is that they directly represent causes, consequences, and — most importantly — the structural relations among them. In other words, they represent mechanisms.

Mikhail (2000) offers a contrasting formalism that draws on linguistic structure to represent events and formulates moral principles as operations on a graphical tree representing a sentence's semantics. Although the specificity of his proposal is admirable, language does not seem the right place to find the structure relevant to moral appraisal. Moral appraisal concerns events and only indirectly depends on how we talk about them. The structure of an utterance obviously has some correspondence to the event it refers to, but it also manifests additional purely linguistic constraints. These constraints reflect specifics about the particular language used to express the utterance. They also emerge from the fact that language is composed of a linear stream of symbols and that its primary function is communication. These additional linguistic constraints are uninformative about the variables that matter for moral judgment, variables about the event itself like agents' intentions, alternative causes, and the valence of an outcome. Mikhail clearly believes that these nonlinguistic variables are somehow represented; our complaint is that his representation incorporates extra baggage brought along by sentence structure. Causal models in contrast represent events directly via the mechanisms that relate causes to effects, and thus offer a representation much more streamlined to capture facts relevant to moral judgment and only those facts.

Our proposal is that causal models serve as a representation for operations that govern moral appraisals. We emphasize our focus on early operations of appraisal. Moral judgments involve more than such quick and dirty appraisals. For instance, they take into account emotional responses and noncausal moral principles like equity. They also depend on basic values about good and bad outcomes (e.g., charity is good, causing

pain is bad). Such considerations are largely independent of causal knowledge. But even some emotional responses like indignation depend on moral appraisals and thus causal structure. Equity assessments frequently include considerations that require causal analyses like determinations of effort or merit. Causal structure is not the only thing, but no judge would get far without it.

Finally, we have offered a speculation that moral appraisals reflect the similarity between an idealized causal model of moral behavior and a causal model of the event being judged. Admittedly, the evidence for this specific hypothesis about the cognitive operations involved in moral judgment is weak. Support would come from studies demonstrating a gradient of judgment that varies monotonically with the dimensions of similarity that we proposed above. We offer the hypothesis as a relatively concrete alternative to the more standard view that moral judgment involves a reasoning process that derives conclusions in the spirit of logical proof (Kohlberg, 1986; Mikhail, 2000; Piaget, 1932). The evidence for this latter view is no stronger.

If our proposals are correct, then moral agents have a causal model of their environment whenever they are in a position to make moral appraisals. Of course, having a causal model is not sufficient to make one a moral agent. Emotional responses help and moral principles, largely defined in terms of causal structure, are necessary. Moreover, one must have the desire to be moral. Our guess is that all humans who are functional have causal models. Not everyone satisfies the other conditions for moral agency.

REFERENCES

Abelson, R. P. and Kanouse, D. E. (1966). Subjective Acceptance of Verbal Generalizations, in edited by Feldman, S. (Ed.), *Cognitive Consistency: Motivational Antecedents and Behavioral Consequents* (pp. 171–197). Academic Press, New York.

Alicke, M. (2000). Culpable Control and the Psychology of Blame. *Psychological Bulletin*, 126, 556–574.

Bartels, D. M. (2008). Principled Moral Sentiment and the Flexibility of Moral Judgment and Decision Making. *Cognition*, 108, 381–417.

Berkowitz, L. and Harmon-Jones, E. (2004). Toward an Understanding of the Determinants of Anger. *Emotion*, 4, 107–130.

Blount, S. (1995). When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes*, 63, 131–144.

Branas-Garza, P., Cobo-Reyes, R. and Domniguez, A. (2006). "Si él lo necesita": Gypsy Fairness in Vallecas. *Experimental Economics*, 9(3), 253–264.

Brown, R. and Fish, D. (1983). The Psychological Causality Implicit in Language. *Cognition*, 14, 237–273.

Chaigneau, S. E., Barsalou, L. W. and Sloman, S. A. (2004). Assessing Affordance and Intention in the HIPE Theory of Function. *Journal of Experimental Psychology: General*, 133, 601–625.

Chomsky, N. (1957). *Syntactic Structures*. Mouton: The Hague.

Clore, G. L. and Centerbar, D. (2004). Analyzing Anger: How to Make People Mad. *Emotion*, 4, 139–144.
Cushman, F. A. (2008). Crime and Punishment: Distinguishing the Roles of Causal and Intentional Analyses in Moral Judgment. *Cognition*, 108(2), 353–380.
Cushman, F., Knobe, J. and Sinnott-Armstrong, W. (2008). Moral Appraisals Affect Doing/Allowing Judgments. *Cognition*, 108, 281–289.
Cushman, F., Young, L. and Hauser, M. (2006). The Role of Conscious Reasoning and Intuition in Moral Judgments: Testing Three Principles of Harm. *Psychological Science*, 17, 1082–1089.
Deigh, J. (2008). Can You Be Morally Responsible for Someone's Death If Nothing You Did Caused It? in edited by Sinnott-Armstrong, W. (Ed.), *Moral Psychology, Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*, MIT Press, Cambridge.
Dowe, P. (2000). *Physical Causation*. Cambridge University Press, New York.
Driver, J. (2008). Attributions of Causation and Moral Responsibility, in edited by Sinnott-Armstrong, W. (Ed.), *Moral Psychology*, *Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*, MIT Press, Cambridge.
Driver, J. (2008). Kinds of Norms and Legal Causation: Reply to Knobe and Fraser and Deigh, in edited by Sinnott-Armstrong, W. (Ed.), *Moral Psychology*, *Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*. MIT Press, Cambridge.
Evans, J. St. B. T. (2003). In Two Minds: Dual-Process Accounts of Reasoning. *Trends in Cognitive Science*, 7(10), 454–459.
Fernbach, P. M., Linson-Gentry, P. and Sloman, S. A. (2007). Causal Beliefs Influence the Perception of Temporal Order. *Proceedings of the Twenty-Ninth Annual Conference of the Cognitive Science Society*, Earlbaum, Mahwah, NJ.
Foot, P. (1978). *Virtues and Vices and Other Essays in Moral Philosophy*. University of California Press, Berkeley.
Gil-White, F. (2003). Ultimatum Game with an Ethnicity Manipulation: Results from Khovdiin Bulgan Sum, Mongolia, in edited by Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., and Camerer, C. (Eds.), *Foundations of Human Sociality: Ethnography and Experiments in 15 Small-Scale Societies*. Oxford University Press, Oxford.
Greene, J. (2007). Why Are VMPFC Patients More Utilitarian? A Dual Process Theory of Moral Judgment Explains. *Trends in Cognitive Sciences*, 11, 322–323.
Greene, J. D. (2008). The Secret Joke of Kant's Soul, in edited by Sinnott-Armstrong, W. (Ed.), *Moral Psychology, Vol. 3, The Neuroscience of Morality*. MIT Press, Cambridge, MA.
Greene, J., Sommerville, R., Nystrom, L., Darley, J. and Cohen, J. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 293, 2105–2108.
Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, 108, 814–834.
Hauser, M. (2006). *Moral Minds: How Nature Designed our Universal Sense of Right and Wrong*. Ecco, New York.
Hauser, M., Young, L. and Cushman, F. (2008). Revising Rawls' Linguistic Analogy: Operative Principles and the Causal Structure of Moral Actions, in edited by Sinnott-Armstrong, W. (Ed.), *Moral Psychology, Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*, MIT Press, Cambridge.
Humphreys, G. and Buehner, M. J. (2007). In Temporal Judgments in Causal and Non-Causal Event Sequences. *Paper Presented at the 15th Escop Conference*, Marseilles, France, August 2007.
Joyce, J. (1999). *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge.
Kahneman, D. and Frederick, S. (2002). Representativeness Revisited: Attribute Substitution Inintuitive Judgment, in edited by Gilovich, T., Griffin, D. W., and Kahneman, D. (Eds.), *Heuristics and Biases: The Psychology of Intuitive Judgment*, Cambridge University Press, New York.

Kahneman, D. and Miller, D. (1986). Norm Theory: Comparing Reality to Its Alternatives. *Psychological Review*, 93, 136–153.
Kant, I. (1785). Groundwork of the Metaphysics of Morals (*Mary Gregor, Trans.*). Cambridge University Press, Cambridge.
Klein, G. (1999). *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA.
Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 63, 190–193.
Kohlberg, L. (1986). *The Philosophy of Moral Development*. Harper and Row, San Francisco.
Levi, I. (2000). Review Article. James Joyce: The Foundations of Causal Decision Theory. *The Journal of Philosophy*, 97(7), 387–402.
Lewis, D. (1973). *Counterfactuals*. Blackwell, Oxford.
Machery, E. (2008). The Folk Concept of Intentional Action: Philosophical and Experimental Issues. *Mind and Language*, 23, 165–189.
Malle, B. F. (2001). Folk Explanations of Intentional Action, in edited by Malle, B. F., Moses, L. J., and Baldwin, D. A. (Eds.), *Intentions and Intentionality: Foundations of Social Cognition*, MIT Press, Cambridge, MA.
Malle, B.F. (2007). Attributions as Behavior Explanations: Toward a New Theory, in edited by Chadee, D. and Hunter, J. (Eds.). *Current Themes and Perspectives in Social Psychology*. (pp. 3–26). SOCS, The University of the West Indies, St. Augustine, Trinidad.
Malle, B. F. and Knobe, J. (1997). The Folk Concept of Intentionality. *Journal of Experimental Social Psychology*, 2, 101–121.
Mandel, D. R. (2003). Judgment Dissociation Theory: An Analysis of Differences in Causal, Counterfactual, and Covariational Reasoning. *Journal of Experimental Psychology: General*, 137, 419–434.
Meek, C. and Glymour, C. (1994). Conditioning and Intervening. *British Journal for the Philosophy of Science*, 45, 1001–1021.
Mikhail, J. (2000). *Rawls' Linguistic Analogy: A Study of the 'Generative Grammar' Model of Moral Theory Described by John Rawls in 'A Theory of Justice.'* Cornell University PhD dissertation.
Moll, J. and Oliveira, S. (2007). Response to Greene: Moral Sentiments and Reason: Friends or Foes? *Trends in Cognitive Sciences*, 11, 323.
Moore, A., Clark, B. and Kane, M. (2008). Who Shalt Not Kill? Individual Differences in Working Memory Capacity, Executive Control, and Moral Judgment. *Psychological Science*, 19, 549–557.
Nichols, S. and Mallon, R. (2006). Moral Dilemmas and Moral Rules. *Cognition*, 100, 530–542.
Nozick, R. (1993). *The Nature of Rationality*. Princeton University Press, Princeton.
Oosterbeek, H., Sloof, R. and Van de Kuilen, G. (2004). Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Journal of Experimental Economics*, 7 (2), 171–188.
Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers Inc., San Francisco.
Pearl, J. (2000). *Causality*. Cambridge University Press, Cambridge.
Piaget, J. (1932). *The Moral Judgment of the Child*. Kegan, Paul, Trench, Trubner and Co., London.
Pizarro, D. and Bloom, P. (2003). The Intelligence of Moral Intuitions: Comment on Haidt (2001). *Psychological Review*, 110, 193–196.
Prinz, J. (2006). The Emotional Basis of Moral Judgment. *Philosophical Explorations*, 9, 29–43.
Rawls, J. (1971). *A Theory of Justice*. Harvard University Press, Cambridge.
Rehder, B. and Hastie, R. (2001). Causal Knowledge and Categories: The Effects of Causal Beliefs on Categorization, Induction and Similarity. *Journal of Experimental Psychology: General*, 130(3), 323–360.

Shaver, K. G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. Springer-Verlag, New York.
Skyrms, B. (1982). Causal Decision Theory. *Journal of Philosophy*, 79(11), 695–711.
Sloman, S. (2005). *Causal Models: How People Think About the World and Its Alternatives*. Oxford University Press, New York.
Sloman, S. A. (1996). The Empirical Case for Two Systems of Reasoning. *Psychological Bulletin*, 119, 3–22.
Sloman, S. A. and Hagmayer, Y. (2006). The Causal Psycho-Logic of Choice. *Trends in Cognitive Science*, 10, 407–412.
Spirtes, P., Glymour, C. and Scheines, R. (1993). *Causation, Prediction, and Search*. Springer, New York.
Spranca, M., Minsk, E. and Baron, J. (1991). Omission and Commission in Judgment and Choice. *Journal of Experimental Social Psychology*, 27, 76–105.
Stanovich, K. E. and West, R. F. (2000). Individual Differences in Reasoning: Implications for the Rationality Debate. *Behavioral and Brain Sciences*, 23, 645–726.
Unger, P. (1995). Contextual Analysis in Ethics. *Philosophy and Phenomenological Research*, 55, 1–26.
Waldmann, M. and Dieterich, J. (2007). Throwing a Bomb on a Person versus Throwing a Person on a Bomb: Intervention Myopia in Moral Intuitions. *Psychological Science*, 18, 247–253.
Walsh, C. R. and Sloman, S. A. (2005). *In* The Meaning of Cause and Prevent: The Role of Causal Mechanism. *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, Erlbaum, Mahwah, NJ.
Wolff, P. (2007). Representing Causation. *Journal of Experimental Psychology: General*, 136, 82–111.
Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, Oxford.
Young, L., Cushman, F., Hauser, M. and Saxe, R. (2007). The Neural Basis of the Interaction between Theory of Mind and Moral Judgment. *Proceedings of the National Academy of Sciences*, 104, 8235–8240.