CHAPTER 29

# Causation in Legal and Moral Reasoning

David A. Lagnado and Tobias Gerstenberg

**Abstract**

Causation looms large in legal and moral reasoning. People construct causal models of the social and physical world to understand what has happened, how and why, and to allocate responsibility and blame. This chapter explores people's common-sense notion of causation, and shows how it underpins moral and legal judgments. As a guiding framework it uses the causal model framework (Pearl, 2000) rooted in structural models and counterfactuals, and shows how it can resolve many of the problems that beset standard *but-for* analyses. It argues that legal concepts of causation are closely related to everyday causal reasoning, and both are tailored to the practical concerns of responsibility attribution. Causal models are also critical when people evaluate evidence, both in terms of the stories they tell to make sense of evidence, and the methods they use to assess its credibility and reliability.

**Key Words:** causation, counterfactual, legal reasoning, attribution, moral, causal model

## Introduction

> What or who caused a certain event to occur is essentially a practical question of fact which can best be answered by ordinary common sense rather than by abstract metaphysical theory.
>
> *Lord Salmon, Alphacell Ltd v. Woodward [1972] A.C. 824, 847*

A young child is admitted to a hospital suffering from croup. He is placed under the care of a doctor. The child is settled at first, but then has two episodes of breathing difficulties. The nurse calls the doctor, but she is at a clinic and does not attend. The child recovers from these episodes, but soon after his breathing is severely blocked, and he suffers a cardiac arrest. The child subsequently dies. The child's mother brings a case against the doctor. Medical experts claim that had the child been intubated during these episodes, his death could have been avoided.[1]

We assign responsibility to someone rapidly, often in a matter of seconds, in a process that is internal to us and largely automatic. It seems natural to blame the doctor for the child's death—if she had attended to the child, she could have saved him. Assigning legal responsibility takes longer, often many months. It is an external process with several explicit stages: charging someone, taking her to court, an investigation, and subsequent trial. Crucial to both processes is the construction of causal models of what happened: who did what, how, and why. These models include assumptions about people's actions, beliefs, and motivations, about what actually happened and also what *would have* happened had people acted differently. They also encapsulate assumptions about what *should have* happened: what actions a reasonable person would have taken.

As given, the story about the child's death is underspecified. We do not know exactly what would have happened if the doctor had intubated the child; perhaps the child would have died anyway. Nevertheless, we still blame the doctor for not attending when called, and indeed the legal finding was that she breached her duty of care to the child. But the question of whether the doctor *caused* the child's death is still unclear. We do not yet know what would have happened if the doctor had attended—would she have intubated? In court the doctor claimed that even if she had attended, she would *not* have intubated. That was her practice, and indeed some medical experts supported this decision (although they were in a minority). Now we might be less sure that the doctor caused the death. After all, had she attended and *not* intubated the child, she would have been following an acceptable line of medical practice, but the child would still have died. Thus her failure to attend to the child made no difference to what actually happened.

This scenario illustrates the dependence of both legal and moral reasoning on causal understanding and a relatively sophisticated use of counterfactual reasoning. Indeed, the notion of causation is embedded in legal doctrine, and also implicit in our moral reasoning. But it is not clear exactly what this notion of causation amounts to, how it relates to scientific or everyday conceptions of causality, and how it underpins our legal and moral decisions. Our opening quote from a judge, which is representative of legal opinion, states that the notion of causation used in the law corresponds to our common-sense notion of causality. Moreover, it is also argued (Moore, 2009, 2012) that this is how it should be, because ultimately legal decisions should fit with our moral judgments, and the latter are themselves based on common-sense principles.

But what exactly is our common-sense notion of causation, and how does it underpin our everyday moral and legal judgments? This chapter will explore these questions, building on recent work in philosophy, psychology, and cognitive science that develops a rich picture of how people construct and reason with causal models (Halpern & Pearl, 2005; Pearl, 2000; Sloman, 2005; Sloman & Lagnado, 2015; see also other chapters in this volume).

### *Philosophical Theories of Causation*

Before exploring the role of causation in legal and moral reasoning, it is useful to highlight some key distinctions from the philosophical literature (for more details see Danks, Chapter 12 in this volume).

First, we must distinguish between general and singular causal claims. The former involve claims about causal laws or propensities: that exposure to asbestos causes lung cancer; that reckless driving causes accidents; that poisoning causes death. The latter involve claims about a specific event or state of affairs: that Jim contracted lung cancer due to asbestos exposure; that Joe's speeding on this occasion caused the accident; that Jane died from arsenic in her tea. In legal or moral contexts, the focus is often on singular claims: we want to know the specifics of what happened and blame or praise people accordingly. But general claims also play a crucial role, encapsulating our knowledge of what makes people (and things) tick, and helping us to infer what happened on particular occasions.

Second, philosophical theories of causation divide into two main camps (Paul & Hall, 2013):

1. Dependency accounts: where causation is defined in terms of whether causes "make a difference" to their effects. For singular events, this is often cashed out in terms of counterfactual dependency between events (Lewis, 1986). For example, the arsenic in Jane's tea caused her death because it made the crucial difference; without the arsenic she would not have died.
2. Process or production accounts: where causation is defined in terms of a physical process that transfers energy or momentum from causes to effects (Dowe, 2000). For example, there is a complex physical process from the drinking of arsenic to the Jane's eventual death from poisoning.

While these two approaches are not exclusive—indeed, both apply in the case of Jane's death—they differ on some critical cases, such as their treatment of omissions. An omission, for instance the doctor's failure to intubate, is a perfectly legitimate cause of the child's death according to the difference-making account, whereas on the process view omissions are ruled out because they do not involve a physical process from the omission to the putative effect. What is the physical process that connects the doctor's failure to intubate to the child's death? On the other hand, difference-making accounts notoriously struggle with a different class of cases, such as overdetermination or pre-emption (for examples, see later discussion in this chapter), which seem readily captured in terms of process accounts. Recent research in philosophy and cognitive psychology focuses mainly on counterfactual theories and difference making, but we shall see that legal causation seems to draw on aspects of both dependency and process.

Third, with regard to singular causal claims, another important distinction is between questions of causal *connection* versus *selection* (see Hilton, Chapter 32 in this volume). The former involves determining whether or not an event is "a" cause of some event, where there might be various causes or contributing causes for any specific effect. Thus, the discarded cigarette, the dry wood, and the presence of oxygen in the air are all causes of the subsequent fire in the shed. In contrast, causal selection involves picking out one (or several) of these as "the" cause, and relegating the others to mere background conditions. For example, the dry wood and the presence of oxygen will typically be seen as mere conditions, with the discarded cigarette selected as "the" cause of the fire. But this process of selection depends on contextual or pragmatic issues. In the context of a shed that is usually damp, the dryness of the wood might also be singled out.

This highlights another distinction commonly made in the literature on singular causal judgments, between normal versus abnormal events (Halpern & Hitchcock, 2014; Hart & Honore, 1983; Hilton, Chapter 32 in this volume). Actions or events that are atypical or abnormal relative to the commonplace are often singled out as causes. Whether this reflects a deep feature of causal claims themselves, rather than a pragmatic feature of how we use them, is a controversial question (Blanchard & Schaffer, 2016). But the focus on transgressions from normality is clearly very germane to legal and moral issues, and seems to play a correspondingly large role in human judgment (Kahneman & Miller, 1986; Knobe, 2009).

## Legal Reasoning

### *Causation in the Law*

Causal reasoning is ubiquitous in the law. This holds true in both criminal and civil proceedings. Thus, criminal offenses are analyzed in terms of two key elements: the defendant's action (*actus reus*) and the defendant's mental state at the time of the action (*mens rea*). For example, murder requires both that the defendant *caused* the death of the victim, and *intended* to kill or seriously harm the victim. Both elements invoke questions of causality. This is most straightforward with the *actus reus* element, which amounts to the explicit claim that the defendant's conduct caused the result in question. However, it is also implicit in the *mens rea* condition. The mental element, too, is assumed to play a causal role in the offense. For example, an intention is taken as a causal precursor to the action.[2] Without the assumption that the intention has some causal efficacy, the rationale for establishing *mens rea* for murder would be undermined. Likewise, in civil proceedings, a central concern is to establish that the defendant's action or omission caused the harm suffered by the claimant.

Consequently, proof of causation is often a key issue in court, with the prosecution seeking to prove that the relevant causal link existed, and the defense typically opposing this claim. But even when the defense accepts that the defendant did, in fact, cause the result, they will often argue for countervailing causal factors that justify or excuse the defendant's actions: for example, a claim that the defendant acted in self-defence or under duress. In addition, causal claims infuse the network of hypotheses and evidence that surround the central issues of a crime. Thus, a defendant's motive, opportunity, and means are often causal pre-conditions of a crime, and evidence is offered to support these causal hypotheses. Moreover, evidence is itself a causal consequence of the pattern of events that make up a crime. Fingerprints left at the crime scene are caused by the perpetrator's presence; an eyewitness report is caused by their sighting of the defendant; a defendant's confession is caused by his feelings of guilt (or by the strong arm of the law).

Causality also pervades how we evaluate the process of decision-making and justice. The decisions made by juries and judges are determined by a multitude of causal factors, and these admit of analysis, especially when a trial is appealed. For example, a key question in an appeal is often whether the jury would have reached a different verdict if some aspect of the actual trial had been different—for instance, if a certain piece of evidence had been excluded, or if the judge had given different instructions. These questions often hinge on our causal understanding of how jurors think—and the answers to these questions can determine the result of the appeal. Finally, when a judge decides on an appropriate punishment, the causal impact of these sanctions needs to be considered—what would be the effect of imprisoning or imposing a fine on the guilty party? Here a prospective notion of causation is in play—anticipating the likely consequences of a particular sentence. In short, the law is shot through with causal claims and judgments, and crucial decisions are made based on the various decision-makers' understanding of causality.

### *Common-sense, Legal, and Philosophical Notions of Causation*

Jurists[3] often claim that the law relies on the same basic concepts of causation as those used in everyday thought (Hart & Honore, 1985). Indeed, juries are typically told to rely on their common-sense notion of causation. In complex situations, a judge might instruct the jury on some finer points, but by and large jurists maintain that legal notions of causation appeal to, and elucidate, our ordinary understanding of causation.[4] In contrast, the relation between the legal notion and philosophical analyses of causation is openly debated. Some argue that the law operates with its own notion of causation, and metaphysical analyses are often irrelevant (Green, 2015; Hoffman, 2011; Stapleton, 2008), while others argue for a general theory of causation, suitable for law, science, and metaphysics (Moore, 2009; Schaffer, 2010).

Often missing from this debate is a precise description of everyday causal thought, or any discussion of relevant psychological work, and how this relates to the philosophical theories or the legal accounts that assume it. In this chapter we argue that everyday causal reasoning is indeed closely allied to legal concepts of causation, but also that both are relatively well modeled (but not yet perfectly) by current philosophical theories. This is not to say that we have a single satisfactory theory that encompasses legal and everyday concerns, but recent progress suggests that convergence is possible. And irrespective of whether we attain a unified theory of causation, psychology and law have a lot to learn from each other. One theme that we will develop is that the way in which the law deals with causation, in the face of practical questions and needs, is often mirrored by our everyday causal reasoning.

## *Legal Analyses of Causation*

### THE SHOOT-OUT

The TV series *Breaking Bad* raises many moral issues. Here we use one of its pivotal scenes to illustrate some issues for legal causation. To cut a long story short,[5] Walt is a chemistry teacher with terminal cancer who is cooking and selling crystal meth with his ex-pupil Jesse. Attempting to widen their sales, they get mixed up with Tuco, a psychopathic drug baron. Tuco is on the run from the police, and is holding Walt and Jesse captive at his father's isolated shack. Armed with a machine gun and a handgun, he takes them both outside, and starts beating Jesse. In the struggle, Jesse manages to grab Tuco's handgun, and shoots him in the stomach. Tuco lies dying, and Walt and Jesse leave him for dead. But Tuco gets up and staggers toward Jesse's car, where they have left his machine gun. Suddenly, Hank, a federal drug agent (also Walt's brother-in-law) turns up. He shouts at Tuco to stop and raise his hands. Instead, Tuco seizes the machine gun, and fires repeatedly at Hank. Hank hides behind a car and shoots back. After a battery of firing from both sides, Tuco stops to reload, and Hank shoots Tuco dead. Who caused Tuco's death? The answer seems straightforward: it was Hank.[6] However, even though common sense and the law might agree, it is not trivial to give a principled rationale for this answer. Indeed, as we show next, the standard *but-for* test used in legal contexts cannot deliver this simple answer.

### *Factual and Legal Causation*

Legal analyses of causation operate with two notions—*factual* and *legal* (or proximate) causation. These are supposed to work in two stages: initially, the factual causes in a case are identified; then one (or several) of these is selected as the legal cause. The separability of these steps has been contested, both on theoretical grounds (Green, 2015; Tadros, 2005) and in terms of the actual practice of trial judges (Hoffmann, 2011), but the conceptual distinction is standard in legal texts.

*Factual causation* is assumed to correspond to what actually happened in the case, irrespective of any evaluation or legal judgment.[7] Our knowledge of this depends on the details of the case, the evidence and arguments presented, and our everyday assumptions about how people and things work, sometimes supplemented by expert knowledge (e.g., in medical or scientific contexts). The standard test for causation is the *but-for* test: the defendant's action is a *factual cause* of the result if, *but for* the defendant's action, the result would not have occurred (Herring, 2010, p. 90).

The *but-for* test is appealing in its simplicity, and it has a strong philosophical pedigree in counterfactual theories of causation (Lewis, 1986). In many cases it delivers a clear-cut judgment: when the defendant shoots his victim, it is usually clear that had he not shot, the victim would not have died. However, it suffers from various problems, both as a theoretical and practical principle. These problems will be discussed in the following.

*Legal causation* lacks a crisp definition or test, and is often seen as the juncture where legal and policy issues are introduced. In UK law, a legal cause is defined as "an operating and substantial cause" (e.g., Herring, 2010, p. 91). The question of what counts as a substantial cause is open-ended, but it aims to rule out *but-for* factors that are only remotely linked to the result in question. In many cases it is intuitively obvious whether or not a factor is substantial, but there will be tricky cases where the status of a candidate cause is unclear. Without a precise definition, the decision will rest on the judge's interpretation, and thus allow for non-causal factors to influence the judgment. The notion of an operating cause also lacks a precise definition or rule for application. It is often invoked when the defendant's action is "interrupted" by another person's action or an act of nature. For example, consider the case where the defendant and his two friends beat up the victim, but the defendant then left the scene, after which his friends drowned the victim.[8] The defendant was ruled not to have caused the death, because he was not an operating cause: the subsequent actions of his friends constituted a novel intervention that broke the chain of causation between his actions and the death. Here again, the lack of a definitive test permits leeway in the interpretation of an operating cause, and can allow other non-causal factors to influence the final judgment. Both of these concepts will be illustrated further in the following examples.

So far we have looked mainly at the legal treatment of actions (*actus reus*), but mental states (*mens rea*) can also play a key role in judgments of legal causation. The requirement of an intention in serious offenses like murder is a straightforward example. But a defendant's mental states are also relevant with respect to whether or not they foresaw the adverse result of their actions. For example, consider a defendant who sets light to an adversary's house, intending to frighten them, but a child dies in the fire. Despite not intending to kill the child, the defendant can be convicted of murder because he could have foreseen that someone might be killed.[9] Note how issues of foreseeability are crucial for purpose of legal judgments, but are not clearly tied to causation as normally understood on a scientific view. Why should someone's knowledge or expectations affect the extent to which an action is judged to have caused an outcome? However, the influence of mental factors such as foreseeability does seem to tie in with our everyday conception of causation, as we shall show in the following (e.g., see Knobe, 2009; Lagnado & Channon, 2008).

## *Problems with the But-For Test*

Despite its central role in causal judgments in the law, the *but-for* test suffers from various well-documented problems: it can be imprecise and difficult to prove; it is over-inclusive in what it counts as a cause, but in certain cases it is also too restrictive, ruling out genuine causes. We will discuss these problems in turn, with the ultimate aim of showing how they can be resolved by an extended account of the *but-for* concept (cf. Halpern & Pearl, 2005; Stapleton, 2008).

### THE PROBLEM OF IMPRECISION AND PROOF

One difficulty is that in certain contexts the *but-for* test is imprecise and thus hard to prove (Moore, 2009). The *but-for* is essentially a comparative test: one compares the actual world, in which the defendant acted and the result occurred, with a hypothetical (counterfactual) world in which the defendant did not act, and ask whether or not the result would still have occurred. But this leaves unspecified exactly what takes the place of the defendant's action. Sometimes the contrast case is obvious. For instance, when considering what would have happened if the defendant had not shot the victim, one imagines a world where he did not shoot. One does not consider a world in which he tries to kill the victim in some other way (cf. Schaffer, 2010).

However, sometimes the appropriate counterfactual supposition is less clear. For example, in the preceding medical negligence example, one key question was what would have happened if the doctor had not breached her duty of care, and had attended the child when called. Would the child have survived? The answer to this question depends both on (a) what the doctor would have done if she had attended, and (b) what would have happened as a consequence of this action. Both of these issues are uncertain, and thus require evidence and argument. In this case it was agreed, based on medical opinion, that the child would have had a greater chance of survival if the doctor had intubated. Therefore, if one imagines "doctor intubates" as the contrast case, then the *but-for* test would rule the doctor's negligence as a cause of the death. However, the doctor argued that even if she had attended, she would *not* have intubated. On this counterfactual supposition, the child would still have died, and therefore by the *but-for* test the doctor did not cause the death. The court accepted this argument, and ruled that the doctor's failure to attend, despite being a breach of her duty of care, did not cause the child's death.[10]

The legal question thus shifted to whether the doctor's practice of not intubating was itself reasonable.

Although in this case the two components (a) and (b) of the *but-for* test were resolved to the court's satisfaction, in other situations these issues might be hard to establish. It might be difficult to agree on what the defendant would have done; and, even if this were established, to agree on what would have happened contingent on this action. Thus, although in this case the medical experts agreed that had the doctor intubated, the child would probably have survived, in other medical cases this might not be so clear (and might even go beyond current medical knowledge).

In sum, despite its apparent simplicity, the *but-for* test requires various cognitive operations when applied to real cases. The fact-finder needs to select the appropriate contrast, and to judge how this counterfactual world would have unfolded. These demands are not trivial, and might be hard to prove or provide substantial evidence in support of. Nevertheless, this problem is part and parcel of legal inquiry. Legal cases are often hard to decide, and the *but-for* test, properly analyzed, clarifies what needs to be shown for proof of causation. It seems appropriate that different sides to the dispute might argue for different contrasts, and even make different claims about what would have happened in the relevant counterfactual worlds.

### THE PROBLEM OF PROMISCUITY

Another problem with the *but-for* test is that it generates too many causes. Thus in most cases there will be innumerable *but-for* causes of a specific result. For example, when a defendant shoots his victim, there are all kinds of factors *but-for* which the victim's death would not have occurred: if the defendant's parents hadn't met, if he hadn't been born, if he hadn't moved to the city, if he hadn't been introduced to the victim, and so on. Most of these factors are clearly irrelevant to the legal issue in question. But the *but-for* test by itself is too coarse a tool to demarcate the relevant from the irrelevant factors. This is where the concept of legal causation is supposed to earn its keep, by pruning away those factors that are deemed irrelevant to the legal question at issue.

The notion of legal cause—cashed out in terms of substantial or operating cause—seems to work well in clear-cut cases. It rules out factors that are clearly insignificant, such as distant or coincidental precursors of the defendant's behavior, thus excluding his parents and other factors that were incidental to his behavior on this occasion, and also limits the extent to which the defendant is deemed a cause of the more distant or coincidental consequences of his actions. But there will be hard cases, where the imprecision of the concept of legal cause means that questions of significance or remoteness requires a judgment call by the fact-finder, rather than following explicitly from the causal definition itself.[11]

### PRE-EMPTION

Pre-emption occurs when an action or event brings about an outcome, but if this action had not occurred, an alternative or back-up action would have brought about that same result. The latter action is "pre-empted" by the former. For example, suppose the driver of a rented car fails to brake and injures a pedestrian. Unknown to the driver, the brakes were faulty—the car rental company had not checked or maintained the car's brakes. So even if the driver had applied the brakes, they would not have worked, and the same injuries would have been incurred.[12] Intuitively it is the negligent action (or inaction) of the driver that caused the harm, and not the faulty brakes. But sensible as this claim sounds, the *but-for* delivers the wrong answer, because *but-for* the driver's action the same harm would still have occurred. Such examples are commonplace in legal and everyday settings, and challenge the *but-for* test as an adequate criterion. Once again, legal causation, in particular the notion of operative cause, needs to be invoked. The operative cause of the accident was the driver's failure to use the brakes, not the faulty brakes. The brakes never got the chance to malfunction, and thus were not an operative cause of the accident.

In pre-emption cases, the pre-empted action either fails to occur, or acts but is beaten to the punch by the "actual" cause. Either way there is a clear causal path from the operative cause to the result, but not from the pre-empted action. This asymmetry makes such cases easier to deal with. Certainly our intuitions seem sharper, even if the notion of operative cause is still fuzzy. However, there are related cases in which one action activates a chain of events that would have led to a harmful result, but instead another action intervenes to cause the harm. For example, let us return to the earlier shoot-out example. Jesse shot Tuco and left him to die. The damage from Jesse's shot was slowly killing Tuco, and Tuco would have died in a few hours. However, Hank intervened and shot Tuco dead immediately. Clearly Hank caused Tuco's death. Jesse's causal role in the death is less clear, but given the independent and unforeseeable nature of Hank's intervention, our intuition (and the law?) is that Jesse did not cause Tuco's death.

The *but-for* test, however, appears to rule out both Hank and Jesse. If Hank had not shot Tuco, he would still have died from Jesse's bullet. If Jesse had not shot Tuco, he would still have died from Hank's shot. But it would be crazy to argue that neither man caused Tuco's death. Someone definitely killed him!

The law offers one possible answer to this problem, stipulating that the *but-for* test individuates the result in terms of its timing and manner: "The test for factual causation requires the jury to consider whether, but for the defendant's unlawful actions, the harm would have occurred *at the same time and in the same way that it did*" (Herring, 2010, p. 90). Thus Hank's shot passes the *but-for* test, because the exact timing and manner of Tuco's death would have been different had Hank not shot.[13] What about Jesse's shot? Whether or not it passes the *but-for* test depends on further details about the actual situation (and relevant counterfactuals). The key question is whether Tuco would have died at the same time and in the same way if Jesse had not shot. This is a tricky question. Perhaps Tuco would have died a few minutes later if he hadn't already been wounded by Jesse's shot. More complicated still, if Jesse had not shot Tuco at all, the course of events might have been very different—Tuco would not have been left to die, he might not have been shot by Hank, and so on. Here again our judgments also depend on what contrast case we use—what we substitute for Jesse's shot, and how we play out the counterfactual world subsequent to that change.

The speculation about whether Jesse's shot is a *but-for* cause of Tuco's death can be curtailed if we move to the question of legal causation. Here we can argue that Hank's shot was a substantial and operating cause of Tuco's death, whereas Jesse's shot was not. Hank's action was an intervening cause—independent and voluntary—that broke the chain of causation between Jesse's action and the death. Thus the law has the means to deal with these problem cases, by individuating the outcome at a suitably fine level of grain or by invoking the notion of legal causation (and operative cause). Both approaches have been used in actual legal cases (Herring, 2010, p. 90). Here again, the lack of a precise definition of legal causation allows cases to be dealt with in a flexible manner. This is a practical bonus, but it opens the door for non-causal factors to influence judgment, and can lead to inconsistency and controversy across legal rulings.

#### OVERDETERMINATION

The textbook case of overdetermination is when two people (A and B) independently and simultaneously shoot the victim, and either shot alone was sufficient to kill the victim. On the *but-for* test, neither shooter is a cause of the victim's death, because if A had not shot, the victim would still have died from B's shot, and the same is true for B. But it is counterintuitive to conclude that neither shooter caused the death. What makes this different from pre-emption cases is that each shooter does exactly the same thing and we want both to be judged as causes of the death.

A more complex example of overdetermination is as follows: "A company produced a leather-spray to be used by consumers on their leather clothing. The company discovered that the spray was extremely toxic for certain elderly people and others with respiratory conditions. The relevant group of executives voted unanimously to market the product (the voting rule required only a majority of votes.) Subsequently the product killed a number of consumers" (Stapleton, 2013, p. 43).[14] Each of the executives was prosecuted separately as a cause of the deaths. In their defense, each member argued that his individual vote was not a cause of the deaths, so he should not be held responsible. The court rejected this argument, and each executive was held legally responsible for the deaths incurred. Here again, although everyday judgments and the law converge on the same answer (Gerstenberg, Halpern, & Tenenbaum, 2015), the *but-for* test rules that none of the executives is a cause, because for each member it is true that the motion would still have passed even if he had voted against it. Note that this latter example cannot be dealt with by describing the outcome in more fine-grained detail. The timing and manner of the harmful outcomes of the company's action depend only on whether or not the motion was passed, and not by the exact majority. So a fine-grained *but-for* test still excludes any executive as a cause of the harm.

### *Extensions to the But-For Analysis*

The problems of pre-emption and overdetermination are well known both in philosophy and law, and are often seen as fatal to counterfactual accounts of causation (Paul & Hall, 2013). However, two recent approaches to causation, developed independently in law (Stapleton, 2008, 2009) and in computer science (Chockler & Halpern, 2004; Halpern & Pearl, 2005), offer very similar solutions to these problems.

### INVOLVEMENT AND CONTRIBUTION

Stapleton (2008, 2009) argues against a general-purpose concept of causation on the grounds that causal claims depend on the nature of the interrogation, and that different questions can demand and yield different answers.[15] She identifies a specific notion of causation that fits the wide-ranging purposes of legal inquiry—what she terms "involvement" or "contribution." At the heart of her account is an extension of the *but-for* analysis that allows the counterfactual test to be computed over a wider range of contrasts, thus ruling actions (and omissions) that cause *or contribute* to an outcome as genuine causes, even if their contribution is neither necessary nor sufficient for the outcome.

Stapleton identifies three forms of "involvement" central to legal inquiries. First, a factor is *involved* in an outcome if it satisfies the standard *but-for* test and thus is a necessary condition. One compares the actual world—in which the factor and the outcome both occurred—with a hypothetical world in which the factor is removed. If the outcome would no longer have occurred, then the factor is deemed to be *involved* in the outcome. Second, a factor is also *involved* in an outcome if it satisfies an amended *but-for* test where one compares the actual world with a hypothetical world in which the factor is removed *along with* any other factor that "duplicates" the outcome in question. If the outcome would not have occurred in this hypothetical world, then the factor is judged to be *involved* in the outcome. For example, when two hunters (A & B) independently and simultaneously shoot a hiker (overdetermination), to assess whether hunter A is involved in the outcome, one imagines a world where neither hunter shoots. In this hypothetical world the hiker would not have died, so hunter A is involved in the death: a similar argument rules in hunter B, too. The third form of involvement is the relation of *contribution*. This involves two steps: (1) transform the actual world by removing any factors that are not needed for the result still to occur; (2) compare this world to the hypothetical world where the target factor is also removed; if the result would not have occurred in this latter world, then the target factor *contributes* (3) to the result.

For example, consider a slight variation of the previous voting example (Stapleton, 2008, 2009). Suppose that the vote is 9–0 in favor of marketing the product, and a majority of only 6 is required to pass motion. Take one particular voter (Bob). Did Bob contribute to the harm? First, imagine a world where the motion still passes, but all excess factors are removed, for example by removing three voters, such that the vote is 6–0. Second, establish whether, *but-for* Bob's vote, the motion would have passed. If the answer is no, then Bob contributed to the result. The same argument can be applied to each voter. Therefore, on Stapleton's account, each voter *contributes* to the motion being passed, and thus to the subsequent harm.

Essentially, Stapleton's account generalizes the *but-for* test to allow for comparisons with a broader range of hypothetical worlds, and thus avoids problems of overdetermination. Her account leaves various issues unresolved, for instance, how we specify the exact nature of the hypothetical worlds, how we establish what happens in these worlds, and how we decide which factors to remove to establish contribution. She does refer to the necessary element of a sufficient set (NESS) test as a formal test for involvement (Wright, 1988), but it is unclear that this test can deliver the needed judgments, and is problematic for other reasons (Fumerton & Kress, 2001). Nevertheless, her proposals are a step in the right direction, and have been adopted in some legal rulings.

### STRUCTURAL MODEL APPROACH

Recent work in philosophy and computer science (Chockler & Halpern, 2004; Halpern & Pearl, 2005) has spawned a novel approach to these problems, which bears notable parallels to Stapleton's proposals, albeit couched in a more formal framework. Like Stapleton, the starting point for the structural model is the *but-for* test, where causation depends on a counterfactual relation between putative cause and effect. However, on the structural approach this counterfactual relation is explored in the context of a specific causal model, defined in terms of a set of variables (corresponding to the events of interest) and a set of structural equations (which capture the causal relations between variables). A counterfactual is cashed out in terms of interventions on the causal model, and obeys a specific logic that allows one to update the model and derive the consequences of this intervention (see also Danks, Chapter 12 in this volume). Thus, in a straightforward case where a single hunter (A) shot a hiker (E), A is deemed a cause of E, if an intervention that had stopped A would have undone E. This corresponds to the *but-for* test and (1) in Stapleton's account. Overdetermination cases are dealt with by allowing the counterfactual query to be extended to include additional interventions. For instance, in the overdetermination case where hunter A and hunter B both shoot E, one considers a *but-for* test for each hunter *conditional on* an intervention in which the other hunter is stopped from shooting. On this extended test, both hunters are ruled in as causes of the hiker's death. This corresponds to (2) in Stapleton's account. Note that this also captures the notion of contribution (3), because it allows for multiple counterfactual interventions—for instance, in the company voting example, intervening by removing the votes of three executives would make an individual member's vote a *but-for* cause of the motion being passed. Finally, the structural model approach has been extended to include degrees of causal responsibility (Chockler & Halpern, 2004). This allows us to assess how far a factor is from being a *but-for* cause, by counting the number of changes (interventions) required to make the outcome counterfactually dependent on that factor (see later discussion for more details). This takes us further than Stapleton's account, because we can distinguish situations where the vote is 7–0 rather than 9–0, with each voter receiving a higher degree of casual responsibility in the former case, because each is closer to being a *but-for* cause (see Gerstenberg et al., 2015; Lagnado et al., 2013).

By and large, there is a neat mapping from Stapleton's notion of contribution to the structural model approach. This suggests that a unified framework for causation, applying to both legal and everyday causation, is possible. There are, however, various issues still to be resolved by the structural model approach. One important question is whether the structural account can adequately capture the notion of an active causal process (see Gerstenberg et al., 2015, and Gerstenberg & Tenenbaum, Chapter 27 in this volume), which seems closely related to the legal notion of operative cause. A hint of how this might be achieved can be seen by how the structural model handles difficult pre-emption cases such as the shoot-out example. The problem is that a *but-for* test rules out both pre-empted and non pre-empted causes—for example, neither Jesse nor Hank are *but-for* causes of Tuco's death. To deal with this, the structural approach introduces additional variables into the causal model, effectively capturing the notion of an active causal pathway from one cause that pre-empts other potential causes. For example, in the case of Tuco's death, a fine-grained causal model represents the separate damage caused by Jesse's and Hank's gunshots, and the fact that the damage caused by Jesse's bullet is overridden by the damage caused by Hank's bullet (see Figure 29.1). There is an active causal path (operating cause) from Hank's gunshot to Tuco's death that pre-empts (breaks the chain of causation) from Jesse's shot to Tuco's death.

The extent to which structural accounts (or difference-making approaches more generally) can adequately model such cases is still an open question. Crucial here is whether such approaches can do full justice to our intutive sense that there is a process connecting cause and effect.

Another key set of questions, yet to be fully incorporated into a structural model approach, is the role of mental states such as intentions and foreseeability. Chockler and Halpern (2004) extend the account to include an agent's uncertainty about the outcomes of their actions (formalizing this in terms of a notion of blame), but there is not yet an extension that takes intentionality into account. Modeling an agent's internal mental states, including their beliefs and intentions, should prove a fruitful avenue for further investigation.

## Psychological Research

We have outlined the notion of causation as applied in the law, and argued that despite various problems and complexities, a coherent picture emerges with causal judgments often assessable in terms of

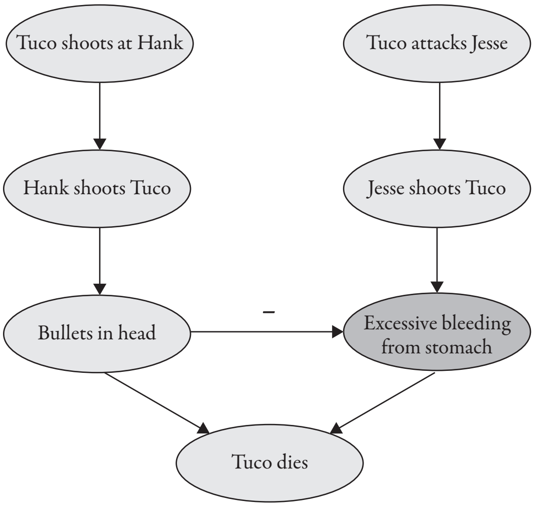

**Figure 29.1** Causal diagram of the shoot-out. Nodes correspond to binary variables, and arrows represent direct causal relations. The arrow with the minus sign represents a preventive relation from bullets in head (from Hank's shots) to the excessive bleeding from stomach. Green variables are those that are true, red variables are false.

counterfactual analyses.[16] We have also seen that a common claim is that the notion of causation in law corresponds to our everyday notion of causation. However, legal theorists usually support this claim only on the basis of intuitions, rather than empirical research. Does it still hold up when we look to empirical studies of causal reasoning in legal contexts?

In the following sections we provide a selective but, we hope, representative look at the psychological research. We believe that this research supports the claimed similarity between legal and psychological conceptions of causation—and also suggests that both legal conceptions and everyday notions of causation serve similar overarching functions and draw on similar abstract conceptions. The details still need to be worked out, but we hold that convergence is not just a claim but also a goal—we should aim for a conception of causation that fits both our everyday understanding and its usage in law (cf. Moore, 2009).

### *Legal Inquiry*

Legal inquiry can be divided into three distinct but interrelated phases:

1. Explanatory: What happened?
2. Evidential: What is the evidential support/proof?
3. Attributive: Who or what is responsible?

All three phases are geared toward the common aim of identifying whether a crime or transgression occurred, apprehending the perpetrators and building an evidential case against them, and deciding guilt or liability on the basis of the evidence. Causal reasoning is involved in all three phases—constructing a causal story or explanation of what happened, using evidence to support this story, and attributing responsibility based on a causal understanding of how and why the guilty parties did what they did.

Given its overarching goals of maintaining justice, legal inquiry has several distinctive features that separate it from a typical scientific inquiry. To start with, the law is concerned with transgressions—disruptions to the normal course of events that violate societal rules and demand correction or punishment. This concern sets the framework for inquiry, and also determines the nature and level of explanation that is sought—predominantly causal explanations about human actions, intentions, beliefs, and desires—explanations that can justify assigning responsibility and blame. The law seeks to identify, punish, and prevent legal violations, and its conception and uses of causal reasoning is geared directly to these aims. This marks a substantial difference from the goals of scientific inquiry, but perhaps not to our everyday concerns and inquiries. Indeed, legal and investigative reasoning seems to provide a more apt metaphor for everyday social reasoning (Fincham & Jaspers, 1980; Tetlock et al., 2007) than the scientific one; for example, Fincham's metaphor of people as "intutive lawyers" (see Alicke et al., 2016, for discussion of various metaphors in causal attribution research). This holds both with respect to the practical role of causal concepts, and the close interrelation between judgments of causality and responsibility.

### *Explanatory and Evidential Reasoning*

A primary goal for legal inquiry is to figure out what happened: Was a crime committed, who was the perpetrator, and why did they do it? To achieve this goal, the fact-finders (judge or jury) must use the evidence presented to them about the specific case, in tandem with their general common-sense knowledge about how the world works, especially their knowledge and assumptions about human behavior. An equally important goal is for the fact-finders to assess how well the evidence and arguments given by the prosecution and defense teams support their respective claims. For example, is there sufficient evidence to uphold a charge of murder, so that you are sure (beyond reasonable doubt)?

Thus the fact-finder has two interlocking tasks—to figure out the best version of what happened, often choosing between competing stories offered by prosecution and defense, and to assess how well the evidence supports either story. Both are required before the fact-finder can decide on guilt. In serious cases, where a case is tried by a jury, the ultimate fact-finders are ordinary members of the public, typically untrained in law or evidential reasoning. They are explicitly asked to use their common-sense understanding of the physical and social world, along with the evidence, to make their decisions. Most psychological research has focused on laypeople (or students) as their subjects of study—how does a layperson reach a decision based on the evidence and arguments presented in court? The dominant answer to this question is given by the story model of juror decision-making (Pennington & Hastie, 1986, 1988, 1992).

### THE STORY MODEL

According to the story model, jurors construct narratives to organize and interpret the mass of evidence presented in court. These stories draw on causal knowledge and assumptions, including scripts about how people typically think and behave. This knowledge is combined with case-specific information to construct causal "situation" models of what happened, typically based around human agency and social interactions. Jurors select the best story—one that explains the evidence, fits with their ideas about stereotypical stories, and satisfies various criteria such as coherence, plausibility, and completeness. This story is then matched against the posible verdict categories to yield the juror's pre-deliberation decision.

### STORY STRUCTURES

One of the key claims of the story model is that people develop rich narrative-based explanations of the evidence. This goes beyond simple evidence-integration accounts (e.g., Hogarth & Einhorn, 1992) in which people compute a weighted sum of the evidence for or against the crime hypothesis. Instead, jurors are assumed to construct a story that makes sense of the evidence and supports a verdict in a more holistic fashion. These narrative structures are usually based around the actions of human protagonists, and are generated from abstract templates known as *episode schemas*. These schemas represent event sequences that occur in real-world contexts as well as fictional stories, and can be used iteratively to produce complex actions and narratives (Bennett & Feldman, 1981; Schank & Abelson, 1977). An archetypal episode schema is depicted in Figure 29.2.

This episode schema is centered on the thoughts and actions of a human protagonist. At the top level are a set of initiating events and background physical states; these events cause specific psychological states in the protagonist (e.g., particular beliefs, desires, and emotions), and lead him or her to formulate goals and intentions, which, in turn, motivate subsequent actions; these actions, in combination with other physical states, generate consequences. This schema can be embedded in a larger episode, and a story structure is often constructed from multiple embedded episodes. We will give a concrete illustration of the schema in the following (see Figure 29.3).

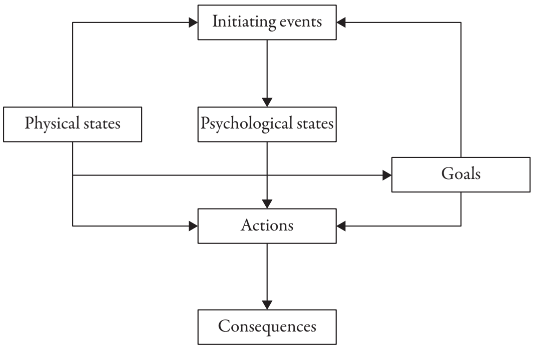

**Figure 29.2** An abstract episode schema.
Adapted from Pennington & Hastie (1986).

Pennington and Hastie used a variety of materials and methods to test the story model (Pennington & Hastie, 1986, 1988, 1992). These included simulated videos of real trials, interviews, and think-aloud protocols for eliciting people's mental representations and reasoning processes. To give a flavor for these studies, and some of their key findings, let us illustrate with the study in Pennington and Hastie (1986).

Participants were sampled from a jury pool and watched a three-hour video of a simulated ~~crirminal~~ trial, based on a real American case: *Commonwealth of Massachusetts v. Johnson.* The defendant, Frank Johnson, was charged with killing Alan Cardwell with "deliberate premeditation and malice forethought." In the trial, both prosecution and defense accepted that Johnson and Cardwell had argued in their local bar on the day of the incident, and that Cardwell had threatened Johnson with a razor. Later that evening, Johnson returned to the bar. He went outside with Cardwell, and they fought, leading to Johnson stabbing Cardwell with a fishing knife. Cardwell died from the wound. The key facts under dispute included the following: whether or not Johnson intentionally returned home to get his knife; whether Johnson returned to the bar specifically to find Cardwell; whether Cardwell drew out his razor during the fight; and whether Johnson actively stabbed Cardwell or held out his knife in self-defense.

After viewing the trial, participants had to decide between four possible verdicts: not guilty, manslaughter, second-degree murder, first-degree murder (these categories were explained in the judge's instructions at the end of the trial). Crucially, participants were asked to think aloud as they considered the case and made their individual decisions. These think-aloud protocols were transcribed and analyzed in terms of content (e.g., story comments versus verdict comments). Story content was

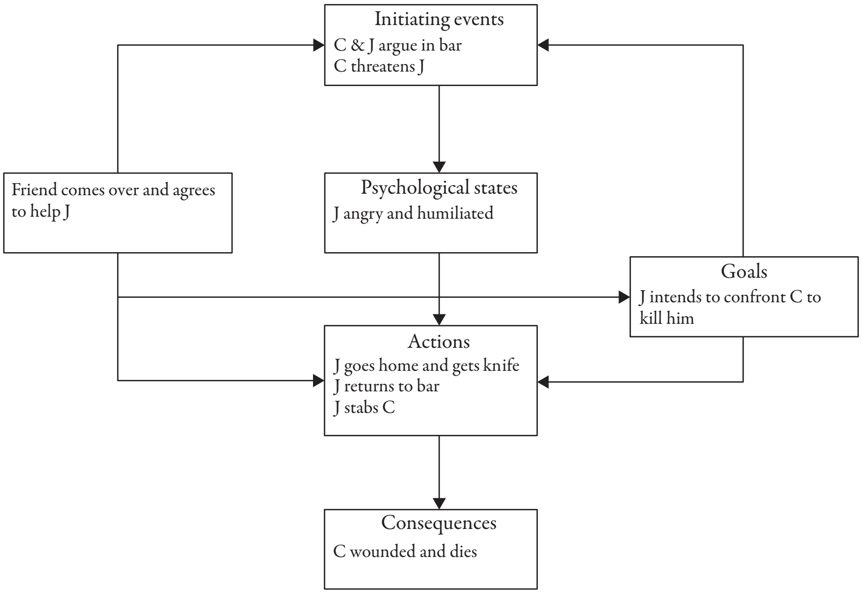

**Figure 29.3** Central story structure for first-degree murder verdict.
Adpated from Pennington & Hastie (1986).

encoded into graphs both at the individual level and at a group level classified by verdict.[17]

Three key empirical findings emerged from these analyses: that people used story structures seeped in causal claims (indeed, 85% of events described in their protocols were causally linked); that people drew numerous inferences beyond the given evidence (only 55% of protocols referred to events in actual testimony; 45% referred to inferred events such as mental states, goals, and actions); that people constructed different stories from the same evidence, and these differences were reflected by correspondingly different verdicts.[18] For example, participants who gave a "first-degree murder" verdict tended to provide a story that elaborated on the events prior to the stabbing, emphasizing Johnson's anger or humilation, and his intention to confront and kill Cardwell (see Figure 29.3). In contrast, those who gave a "not guilty" verdict focused on the altercation itself, spelling out details of how Johnson acted in self-defense. In this story the stabbing was portrayed as a consequence (of Cardwell's behavior) rather than a goal-directed action initiated by Johnson.

Overall, the story model has garnered strong empirical support, and is widely accepted by legal theorists. It encapsulates the core claim that people use causal explanations to draw inferences from evidence. It also highlights the constructive nature of people's explanations, using their causal knowledge to fill in gaps in the evidence and tell a compelling story. The power of a story to summarize and rationalize a mixed body of evidence is also a potential weakness—the most compelling story is not always the one most likely to be true. Nevertheless, there is strong experimental evidence that people use story structures to organize their evidence and make decisions.

### EXTENDING THE STORY MODEL

The story model marks a huge advance in our understanding of juror decision-making. However, several issues remain unresolved. One problem is that the notion of a causal situation model, so central to the account, is not explicitly formalized or defined; this makes it harder to elicit and test people's causal models, or to compare their causal reasoning against normative standards. What makes one situation model better than another? Given that the fact-finder believes a specific causal model, what inferences are licensed? Pennington and Hastie propose several criteria for evaluating stories—such as coherence, plausibility, and completeness—but these notions also lack formal definition and it is unclear how they might trade off against each other.[19] Without a formal framework for causal representation and inference, it is difficult to explain how people construct models based on background knowledge, and unclear how these causal models relate to counterfactual analyses and judgments of factual and legal causation.

We believe that the causal model framework (Pearl, 2000), suitably developed, can address some of these concerns. The framework provides a formal theory of causal representation, learning, and inference, and has been successfully used in numerous areas of causal cognition (Sloman, 2009; see, in this volume, Griffiths, Chapter 7; Oaksford & Chater, Chapter 19; Rehder, Chapter 20). Even though people's actual causal representations and inferences can depart from the formal theory (Sloman & Lagnado, 2015), the causal framework provides a crucial guide to modeling causal cognition. It also suggests how an account of everyday and legal causation can be developed, allowing for appropriate counterfactual reasoning and judgments of causal responsibility (Lagnado et al., 2013).

For instance, the story structures for the Johnson murder case are readily transformed into formal causal networks. We have translated the "first-degree murder" story structure into a formal causal graph (see Figure 29.4). The nodes correspond to events (or propositions), the directed links to causal relations between these events. We have not specified the exact functional relations between events, but it is relatively straightforward to use the formal apparatus to capture the intended combination functions. For example, in this network, Johnson stabbing Cardwell depends on three causes: Johnson is with Cardwell, Johnson is armed with a knife, and he intends to kill Cardwell. If we use an "AND" function, then this states that all three causes are needed for Johnson to stab Cardwell (which seems appropriate to capture this specific story structure).

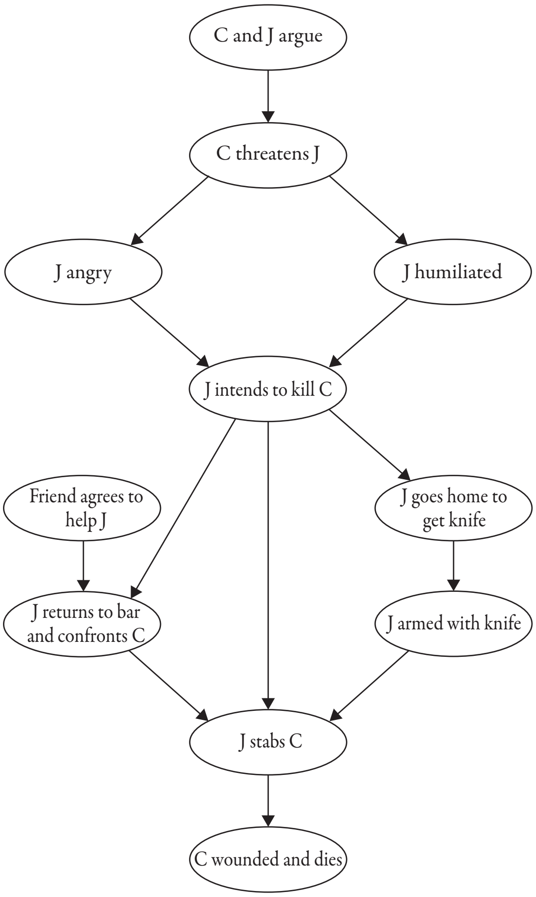

**Figure 29.4** A formal causal network for the first-degree murder story.
Derived from the central story model shown in Figure 29.3.

Furthermore, the story model's claim that people use abstract epsiode schemas to construct specific story structures anticipates recent computational work on how people use intuitive theories of a domain to generate approriate causal models (Griffiths & Tenenbaum, 2009; also see Gerstenberg & Tenenbaum, Chapter 27 in this volume). Applied to the legal domain, the idea would be that people's intutive theories and knowledge (e.g., about criminal behavior and social interactions), combined with case-specific information, allow them to generate specific causal situation models. This is a promising avenue for future research, and would be facilitated by formalizing story structures in terms of causal networks.

### MODELING EVIDENTIAL REASONING

Another underdeveloped area for the story model is the issue of evidence and proof. As well as constructing plausible stories, fact-finders must evaluate how well the evidence supports these stories and assess the strength, credibility, and reliability of the evidence (Schum, 1994). Although the story model supplies some criteria for story evaluation, it does not attempt to model how different items of evidence (e.g., witness testimony, forensic evidence, etc.) relate to different elements in a story, nor does it consider how one captures the credibility or reliability of this evidence.[20] To address this question, Kuhn et al. (1994) proposed that fact-finders lie on a spectrum, from satisficers who maintain a single plausible story, to more adept reasoners who engage in "theory-evidence coordination," assessing how well the evidence supports different hypotheses and stories, as well as the reliability of the evidence. But Kuhn gives no definite explication of this process, nor a normative benchmark for how it should be done.

What we need is a fine-grained analysis of how fact-finders represent and reason about the strength and reliability of the evidence, and how this affects their story evaluation. Just as when they reason about what actually happened, "the story of the crime," fact-finders will draw on causal knowledge and assumptions, here directed to the "story of the trial," where they must reason about people's motivations and beliefs when they give testimony, and use these inferences to update their version of what happened. For example, when a witness gives evidence, the fact-finder must judge whether this testimony is accurate, mistaken, or intentionally deceptive (Schum, 1994). These inferences will modulate the fact-finder's beliefs about what happened, sometimes in complex ways. Thus, when a defendant's alibi is discredited, this can undermine his innocence in two ways—by making it more likely that he was at the crime scene *and* by showing that he is lying (Lagnado, 2011; Lagnado et al., 2013; Lagnado & Harvey, 2008). Similarly, when a victim is shown to be inconsistent in her testimony, even when it concerns a matter peripheral to the crime in question, this can undermine the victim's credibility and thus have the knock-on effect of undermining her testimony about the actual crime (Connor De Sai, Reimers, & Lagnado, 2016). In short, fact-finders draw inferences from the credibility or reliability of the evidence, and these inferences permeate through the fact-finder's network of beliefs about the crime as well. The "story of the crime" and the "story of the trial" are closely intertwined.

Here again, we believe that a careful formal analysis of the relations between hypotheses, evidence, and reliability is crucial to understand how fact-finders actually reason. Again, this is not to claim that fact-finders follow the normative theory; but without a framework to capture reasonable inference, we cannot uncover or appraise how people actually do it. Moreover, a normative framework can also suggest how the evidential reasoning of fact-finders might be improved. One possible approach, closely related to causal model theory, is the Bayesian network framework (Pearl, 1988; Taroni et al., 2006). Fenton et al. (2014) apply Bayesian network analysis to legal arguments, in a framework that allows for the systematic modeling of interrelations between hypotheses, evidence, and reliability. They argue that fact-finders can use *legal idioms*—small-scale causal building blocks tailored to the legal context. These idioms can be combined and reused to represent and reason about large-scale legal cases involving complex and interrelated bodies of evidence. Some recent empirical studies suggest that people follow the qualitative prescripts of this formal account (Lagnado, 2011; Lagnado et al., 2013), even though it is unlikely that they engage in full-fledged Bayesian computations.

To show how the idiom-based approach can be applied, let us return to the Cardwell murder case. One critical issue was whether or not Cardwell pulled out a razor shortly before Johnson stabbed him. Johnson's plea of self-defense would be bolstered if this was true, and mock jurors who found Johnson not guilty tended to include this as an element in their stories. But how do people decide this fact? In the trial they are presented with conflicting testimonies. On the one hand, Johnson claims that Cardwell pulled out the razor, and this is reaffirmed by another eyewitness Clemens (who is a friend of Johnson's!). On the other hand, a policeman and the bar owner both testify that they did not see Cardwell holding a razor. In addition, the reliability of all witnesses is open to question. Johnson has a clear motivation to lie, as does his friend Clemens. And both the policeman and bar owner admit under cross-examination that their views of Cardwell's right hand were partially obscured, so he might have been holding a razor. To complicate matters further, the pathologist who examined Cardwell's body reports that he found a razor in his back pocket. This seems unlikely if Cardwell had indeed pulled out the razor. Could he have put it back into his pocket while dying from a stab wound? If not, might someone else have replaced it? Somehow fact-finders must negotiate these competing claims, and decide how to incorporate the reliabilities of the various witnesses. This requires going beyond story construction.

The idiom-based approach provides a Bayesian network framework to capture the complex interrelations between hypotheses, witness testimony, and reliabilty.[21] In particular, it posits an "evidence-reliability" idiom that explicitly captures how a witness's report is modulated by his reliability, for example, whether the witness is mistaken or intentionally biased. We model the key testimonial evidence about the presence (or absence) of the razor using the Bayesian network shown in Figure 29.5.[22] White nodes represent hypotheses (events that are unknown and need to be inferred), such as whether or not Cardwell pulled out the razor; whether or not the razor was found in his back pocket; whether the policeman is accurate in his testimony, and so on. Gray nodes represent the testimony reports that are heard in court. The key idea is that given the evidence (e.g., the witness reports), the network allows us to revise our beliefs in the unknown hypotheses using Bayesian updating. A fuller model could include evidence about the reliability of the witnesses (e.g., information obtained under

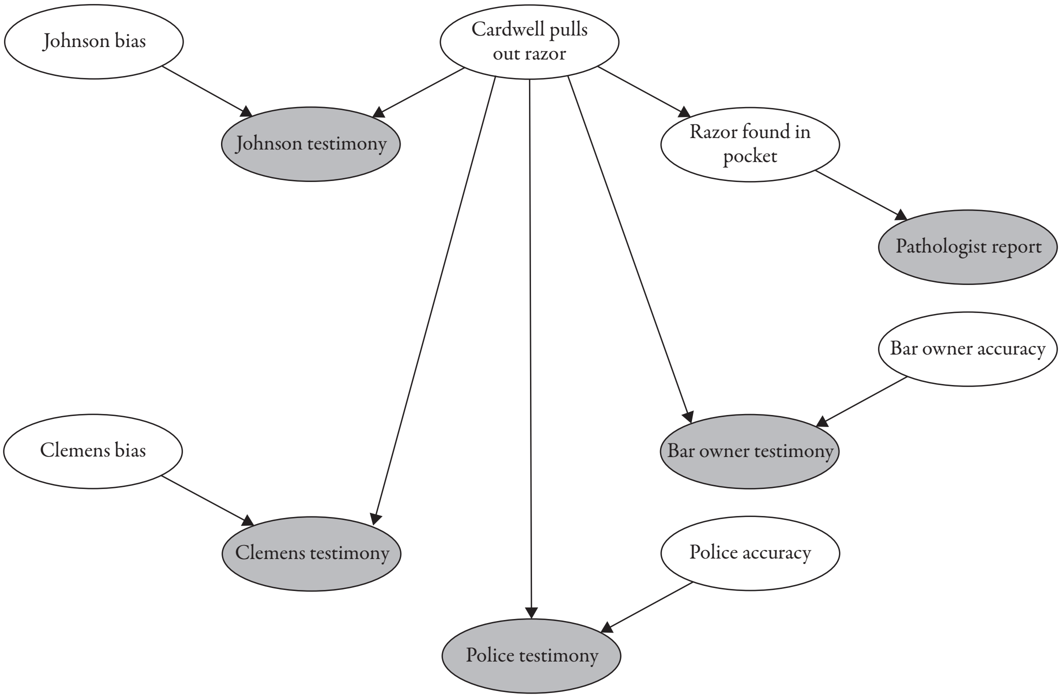

**Figure 29.5** Idiom-based Bayesian network for witness testimonies about whether or not Cardwell pulled out a razor in the fight. Gray nodes represent testimonies presented in court, white nodes represent hypotheses.

cross-examination), and connect up with networks that model other components of the evidence (see Fenton et al., 2014, for details).

The idiom-based approach represents a formal development of ideas from Schum (1994) and Wigmore (1913), capturing the interrelations between hypotheses and evidence in a systematic and probabilistically coherent fashion (see also Dawid & Evett, 1997; Fenton & Neil, 2012; Hepler, Dawid, & Leucari, 2007). The extent to which people actually produce such representations (and computations) is still an open question. While there is some evidence that people's judgments can be captured by network models at a qualitive level (i.e., participants' posterior judgments correlate with the outputs of Bayesian networks constructed from their causal beliefs, priors, and conditonal probabilty judgments; see Connor De Sai, Reimers, & Lagnado, 2016) it seems unlikely that they perform exact Bayesian computations (given the computational demands with mutiple interrelated variables).

### Coherence-Based Reasoning

As noted earlier, although the story model captures many aspects of juror decision-making, it does not provide a formal or computational framework to underpin people's representations or inferential processes. Coherence-based approaches to reasoning and decision-making (Simon & Holyoak, 2002; Simon, Snow, & Read, 2004; Thagard, 2000) aim to provide such a framework, and give a more general account of the kind of complex decision-making faced in legal contexts. Such models were inspired by earlier cognitive consistency theories (Heider, 1958) and were revitalized by advances in connectionism (McClelland & Rumelhart, 1986). The key idea is that people strive for coherent representations of the world, and the decision-making process is driven by the search for a maximally coherent final state, one that best satisfies the multiple constraints faced by the decision-maker. This approach appears well suited to the legal domain, where decision-makers are faced with complex bodies of probabilistic evidence, often ambiguous or contradictory, and need to reach categorical verdicts.

On this view, people represent evidential and decision variables in terms of units in an associative network. These units are connected with excitatory or inhibitory links, depending on whether they are mutually consistent or inconsistent.[23] Units have an initial level of activation that depends on their prior degree of acceptability, with the receipt of new evidence boosting the activation of the corresponding units in the network. Inference or belief updating then involves the spread of activation through the network. Through an iterative process of parallel constraint satisfaction, the network settles into a state that maximizes coherence between units, with the final decision being determined by the units activating above some threshold (for details, see Thagard, 2000).

A key feature of this interactive process is that it can lead to *bi-directional reasoning*—whereby evidence is distorted to fit with emerging decisions and judgments. The decision maker continually ~~readjusts~~ his assessment of hypotheses and evidence until a coherent position emerges, leading to high confidence in a final decision even in the face of initial ambiguity and uncertainty. Advocates of coherence-based approaches maintain that this bi-directional reasoning distinguishes it from Bayesian accounts, arguing that the latter only allow for unidirectional reasoning from evidence to conclusions.[24]

Thagard (2000) applies coherence-based modeling to legal cases, but does not test these empirically. Subsequent work (e.g., Simon & Holyoak, 2002; Simon et al., 2004; Glockner & Engel, 2013) aims to model actual legal reasoning using coherence models. We will illustrate their approach with one key study.

Simon et al. (2004) use a legal decision-making task to show that people engage in bi-directional reasoning and evidence distortion. They use a two-stage paradigm. In the first stage, participants make judgments about a set of social vignettes, including evaluations of various kinds of evidence. For example, in one scenario a mystery man leaves flowers for a woman in an office, and her colleague states that she recognized the man as Dale Brown—whom she has only seen a couple of times before. Participants answer various questions, including one about the value of this identification evidence: "Does the office worker's identification make it more likely that it was Dale Brown who delivered the flowers?" After this initial task, participants complete a distractor task in which they solve analogies. They then move on to stage two of the experiment.

Their main task is to decide a legal case, which involves an employee Jason Wells, who is accused of stealing a large sum of money from the company safe. They are presented with a mixed body of evidence, with various pieces for and against the suspect—for example, a technician claimed to see Jason rushing from the crime scene, a car like Jason's was caught on camera leaving the parking lot around the time of the crime, and Jason had made several large payments shortly after the crime; however, in his defense, another witness claimed to see Jason far away from the crime scene at that time, and Jason claimed his payments were legitimate family transactions. The key evidential manipulation is whether or not participants are told that Jason's DNA was found on the safe. Unsurprisingly, those told that DNA on the safe matched Jason's DNA tended to convict, and those told it did not match tended to acquit. However, participants also assessed the other pieces of evidence in the case. Crucially, they were asked to evaluate the same kind of evidence claims as had been requested in the prior social vignettes, for example, the value of an eyewitness identification. The major finding was that people distorted the value of evidence to fit with their verdicts. Thus, convictors tended to inflate the value of the eyewitness testimony, whereas acquittors tended to deflate it.

Simon et al. take these findings to show that people distort evidence to cohere with their decisions. And further experiments suggest this distortion takes place during the decision-making process, rather than being a post hoc attempt to maintain consistency with their decision. They also contend that bi-directional reasoning does not fit with Bayesian prescripts, and thus undermines a Bayesian updating model. In particular, they argue that the evaluation of one piece of evidence (e.g., the eyewitness identification) should be treated independently from the DNA evidence, but that people violate this prescription.

While these findings appear to support coherence-based effects, we think they can also be explained within a Bayesian framework, if it is extended to include a richer representation of the evidence and its reliability. Applying the idiom-based approach to the Jason Wells case, a Bayesian network that captures the reliability of the eyewitness using a reliability node (see Figure 29.6) can account for the observed change in evidence evaluation. On this network, the presence of a DNA match raises the probability that Jason Wells is

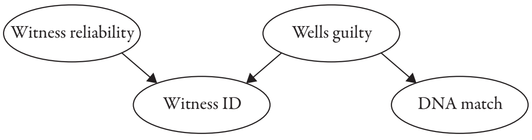

**Figure 29.6** Bayesian network of Jason Wells case using the evidence-reliability idiom to capture the bi-directional reasoning observed in Simon et al. (2004).

guilty, which also (via explaining away) raises the probability that the eyewitness is reliable. In contrast, the same network shows that if the DNA evidence is false, this lowers the probability that Jason Wells is guilty, and in turn (via explaining away) lowers the probability that the eyewitness is reliable. (For more details about explaining away, and the evidence-reliability idiom, see Fenton et al., 2013; Lagnado et al., 2013). Therefore it is perfectly legitimate for participants to modulate their judgments about the reliability of the eyewitness identification according to whether the DNA evidence is positive or negative.

Thus, in this case, bi-directional reasoning need not violate Bayesian updating, given a suitably rich representation of evidence and reliability (see also Jern et al., 2015, for a similar argument about a different legal case used in Simon & Holyoak, 2002). This does not mean that people's reasoning can always be recast in rational terms—especially given the computational demands of even simple Bayesian networks. However, it is useful to see that alleged irrational reasoning is rational relative to a richer representational framework. The exact psychological mechanisms that achieve this are still an open question. Moreover, there are other aspects of human reasoning—such as susceptibility to order effects—where Bayesian prescripts do seem to be violated (Lagnado & Harvey, 2008). But even here, one might argue for heuristic approximations to Bayesian reasoning, rather than throwing out the Bayesian framework altogether (Griffiths, Lieder, & Goodman, 2015).

Another problem with the coherence-based approach is that the cognitive representations that it posits are based solely on associative links. But this lack of directionality means that they cannot fully support causal or counterfactual reasoning (Sloman, 2009; Sloman & Lagnado, 2015; Waldmann et al., 2006). For example, recall the Cardwell murder case from Pennington and Hastie's studies. The critical first-degree murder story claims that Johnson was humiliated by Cardwell, and therefore formulated a plan to kill him. Johnson returned to the bar, and stabbed Cardwell. This is a causal sequence of events, not just an associated set of events. Johnson's plan for revenge is not merely associated with his intimidation by Cardwell, and his stabbing of Cardwell. His plan for revenge is a *consequence* of the intimidation and a *cause* of the stabbing. Capturing this with a causal representation enables counterfactual inference. If someone had stopped Johnson from returning to the bar they might have prevented the stabbing, but they would not have prevented the earlier intimidation. As we argue throughout this chapter, causal representations are critical to legal and moral reasoning. Mere association is not enough.

## *Attributing Causality*

Let us move from the explanatory and evidential phases to the attributive phase. There has been a wealth of empirical research into causal attribution (for reviews, see Alicke et al., 2016; Hilton, Chapter 32 in this volume). We will focus on four key areas: (1) issues of *but-for*, necessity, and sufficiency; (2) intention and foresight; (3) abnormality versus normality; and (4) group attributions.

### *But-For*, NECESSITY, AND SUFFICIENCY

Despite its problems, the *but-for* test occupies a central position in legal causation. It also plays a dominant role in everyday judgments of causation. Numerous studies show that people's causal judgments are sensitive to counterfactual contrasts, and that people are more likely to judge something as a cause of an outcome when they believe that the outcome would not have occurred without the putative cause (Hilton, Chapter 32 in this volume). Moreover, in a set of studies looking specifically at legal reasoning in civil liability cases, Hastie (1999) showed that the majority of mock jurors were concerned with causal aspects such as necessity, sufficiency, and *but-for* reasoning (even though the judge made no explicit mention of causation).

The *but-for* also allows for omissions as causes. This is illustrated by the Bolitho case, where the doctor's failure to attend the child, although a breach of duty of care, was not judged to have caused the child's death because even had she attended she would not have intubated, and thus the child would still have died. Even in a complex case like this, people use *but-for* reasoning and track the legal judgment. Uustalu (2013) asked participants to give causal judgments on the Bolitho case, but varied the counterfactual contrast—whether or not the doctor would have intubated. Particpants (recruited from the general public) judged the doctor significantly more causal (and blameworthy) if they were told that she would have intubated. Moreover, in the absence of any information about what the doctor would have done, participants assumed that she would have intubated, and thus judged her to have caused the death.

Despite the presence of *but-for* reasoning in many studies, it has also been shown that it is neither necessary nor sufficient for judgments of causation. As with legal judgments, there are contexts in which people still judge something as a cause despite it not being a *but-for* condition, or fail to judge something a cause even though it is a *but-for* condition. Thus, Spellman and Kincannon (2001) compared people's causal judgments in legal scenarios with either multiple sufficient (MS) or multiple necessary (MN) causes. For example, two independent gunmen shoot the victim at the same time, and he dies. In the MS condition the coroner rules that either shot would have been sufficient to kill the victim, in the MN condition he rules that both shots were needed. The MS condition is a classic overdetermination case, and the *but-for* fails for both gunshots. Nevertheless, people judged both gunshots as causes of the victim's death. More surprisingly, when asked to rate the strengths of the causes, people rated each gunshot higher in the MS than MN condition, even though the gunshots are *but-for* causes in the latter and not the former. Similar results were obtained using a scenario in which two inanimate factors, lightning or fierce winds blowing down an electrical pole, led to fires that burned down a building. In the MS condition participants were told that either fire alone was sufficient to burn down the entire building, in the MN condition that each fire alone would only have burned down half the building. Spellman and Kincannon concluded that people do not use *but-for* reasoning to assign causality in these cases.

One possible confound here is that the strength of the causes seem different between the two conditions. Thus, the causes in the MS conditions appear stronger, because either alone would have been sufficient to bring about the effect, in contrast to the MN conditions, where neither would have been sufficient. Thus it is possible that the difference in causal ratings reflect differences in the perceived causal strengths of the causes, rather than being due to the contrast between sufficiency or necessity. Nevertheless, the studies clearly show that people assign causality even when the standard *but-for* test does not apply.

Gerstenberg and colleagues (Gertsenberg & Lagnado, 2010; Lagnado et al., 2013; Zultan et al., 2012) also compared MS and MN causes (sometimes in the same scenario), but in contexts where the causal agents were identical across conditions, and with the combination rule (MN or MS) pre-established by the rules of the game. They too found causal assignments even to non-necessary causes, but they also showed that MN were judged as more causal than MS, in line with the Chockler and Halpern model that allows for graded causal responsibility based on an extended *but-for* rule (for more details, see discussion later in this chapter).

Another challenge to *but-for* reasoning arises in situations where there are sequences of actions and events, as in the pre-emption cases discussed earlier. Greene and Darley (1998) studied people's liability judgments in criminal scenarios involving a chain of events between a perpetrator's actions and a final outcome. All scenarios started with the same core setup: Harold intends to kill his colleague Joe and inserts a poisonous pill in his vitamin bottle. Numerous variants were constructed by adding different sequences of actions and events to this initial segment. This allowed the experimenters to vary both the necessity and sufficiency[25] of the perpetrator's actions for the final outcome. For example, in one scenario Joe dies from ingesting the poisonous pill, and thus Harold's action is both a necessary and sufficient cause of his death. In some other scenarios, Joe ingests the poison but is killed by someone/something else before the poison works, thus Harold's action is not a necessary condition of Joe's death; whereas in other scenarios, the poison alone is not strong enough, so Harold's action is not sufficient (and needs to combine with other drugs to kill him). Participants judged Harold's liability for the final outcome, measured in terms of sentence imposed, as well as explicitly asking participants to judge the necessity, sufficiency, and contribution of Harold's action. Overall, Harold's perceived contribution to the outcome was the best predictor of the liability judgments, but perceived necessity and sufficiency were also good predictors. Greene and Darley draw several conclusions from this research: that people favor a graded notion of causation (contribution) rather than a dichotomous yes/no, and that sufficiency, not just necessity, plays a role in liability judgments. This is nicely illustrated by the fact that people assign greater liability for attempted murder when Harold's actions would have been sufficient to kill Joe (e.g., the level of poison was high enough) than when it was insufficient (e.g., the level was too low). They also found a "proximity effect," such that the closer the perpetrator's actions were to bringing about the harm, the more liable he is judged. Finally, they conclude that "while the theory of causation our respondents seem to use is not easy to specify, it has components of sturdy rationality" (Greene & Darley, 1998, p. 467). Building on this conclusion, future work could investigate whether Greene and Darley's pattern of results can be captured by recent extensions of the counterfactual models (e.g., Gerstenberg et al., 2015) with graded causal responsibility and incorporating notions of sufficiency and robustness. It should be noted that Greene and Darley took liability, not causal judgments, but the research discussed in the next section suggests a close link between these two kinds of judgment in lay attributions.

Some of Greene and Darley's scenarios involved pre-emption cases, in which an intervening action or event breaks the chain of causation from the perpetrator's actions and the final outcome. Such situations, as we discussed earlier, can raise interesting issues because the initial action sets in motion a sequence of events, and then the intervening action of a third party (or inanimate process) interrupts this sequence to determine the final outcome. So the initiating action might still be a necessary condition for the final outcome, even though it is not typically judged as the cause. Complications can arise when the initiator and intervening party are connected in some way—for example, as part of a gang attacking a victim. At one extreme, in UK law if the initiator could have reasonably foreseen the actions of the third party, then he too can be convicted of murder, along with the intervening actor who deals the fatal blow. At the other extreme, the intervening actor is judged to break the chain of causation and create a novel causal path, thus vindicating the initiator of murder. Assessing cases as to which action path is "operative" can be a tough judgment call.

These kinds of scenarios have been explored in psychological studies, and people's judgments do seem to fit with some notion of "operative" cause whereby a suitably independent intervening party does alleviate the initiator of causing death (Mandel, 2011), although no studies (to our knowledge) have looked at borderline cases where the foresight of the initiator (as to what the intervening party might do) is varied. However, various studies have explored the more general issue of how people's judgments of a perpetrator are influenced by his foresight of the harmful consequences of his actions.

### INTENTIONS AND FORESIGHT

In legal contexts the mental elements of a crime, such as intention or foresight, are incorporated in judgments of causation. Empirical studies show that this also holds in laypeople's causal attributions (Alicke et al., 2016; Cushman, 2008; Hilton et al., 2010; Lagnado & Channon, 2008). Thus, Lagnado and Channon (2008) explored how people attribute causality and blame in event chains with multiple agents, varying the agents' intentions and foresight. For example, in one scenario, a wife put poison in her husband's medication (either intentionally or accidentally), and then the ambulance was severely delayed (either it got lost or ignored the call), resulting in the husband's death. In more complex scenarios, the foreseeability of the adverse outcome was also manipulated, both in terms of what the agent actually foresaw and what was foreseeable (probable) from an objective viewpoint. For example, a woman makes a self-assembly chair, and she either thinks that it will not break or that it will (subjective foreseeability). The truth state of the world is also varied: the chair is either made properly and is unlikely to break, or made poorly and likely to break (objective foreseeability). These two factors were crossed in a factorial design.

The findings from these studies were systematic across many different scenarios. People assigned more causality and blame to intentional versus unintentional actions, and for outcomes that were foreseeable versus unforeseeable (this applied to both subjective and objective foreseeability). The blame ratings are relatively straightforward to interpret, because most accounts of blame (Shaver, 1985) agree that agents are more blameworthy for intentional and foreseeable consequences of their actions. The causal ratings are harder to explain, even though they seem to fit with the legal notion of causation.[26] According to a counterfactual notion of causation they are puzzling, because the target action, whether intentional or unintentional, is still a *but-for* condition of the outcome.

Several explanations of these findings are possible. A common response is that in situations where human actions lead to adverse outcomes, people are primarily concerned with attributing blame, even when they are ostensibly judging cause. This mirrors the influence of policy factors on causation judgments in legal contexts. This response can divide into distinct psychological accounts (not mutually exclusive). On one view, people's desire to blame someone for an adverse outcome distorts their causal model, exaggerating the degree of causation in the morally reprehensible cases, where the action is intentional and the outcome foreseeable (cf. Alicke, 2000). On an alternative view, people's causal models of the situation (legitimately) incorporate factors that mediate blame, although these factors are not specific just to morality or blame. For example, as well as judging the necessity of a causal relation, people might also be concerned with the *robustness* of the causal relation. Roughly speaking, a causal relation is robust when it would have held even if there had been perturbations to the background conditons, whereas it is sensitive if it relies on a fragile and improbable set of background conditions (cf. Kominsky et al., 2015; Lombrozo, 2010; Woodward, 2006). Thus intentional actions are typically judged more robust than unintentional ones. For example, the wife poisoning her husband is less sensitive to background conditions than her unintentionally doing the same thing; the latter depends on the wife not having her glasses on, misreading the label, not checking the drink, and so on; similarly for foreseeability, which will usually be inversely related to the predictability of the background conditions.

These empircial findings present a challenge for psychological models that rely purely on counterfactual analyses. However, more recent advances are starting to address these issues. One thing that is needed is a more fine-grained modeling of agents' mental states, to incoporate factors such as intentions and foresight (for formal approaches that include foreseeability, see Chockler & Halpern, 2004; for psychological approaches that include intentions, see Kleiman-Weiner, Gerstenberg, Levine & Tenenbaum, 2015; Sloman 2009).

### NORMS

Another systematic finding in the psychological research is that norms play a role in people's causal attributions. Thus, an action or event that violates a norm is often accorded greater causality, or is preferred as "the" cause, of a subsequent outcome (Hart & Honore, 1959/1985; see Hilton, Chapter 32 in this volume). Here norms are a broad category, including moral prescriptions, social rules or conventions, and statistical norms. One much discussed example is the pen vignette (Knobe & Fraser, 2008): A receptionist in an academic department keeps a stock of pens on her desk. Administrative assistants are allowed to take these pens, but academic staff are not (although they often do). One morning both a professor and an assistant each take a pen, leaving the receptionist with none. Participants are asked who caused the problem (the lack of pens). Overall they assign causality to the professor, not the assistant.

Hitchcock and Knobe (2010) argue that the professor is preferred as the cause because he has violated the prescriptive norm of who is allowed to take pens. They explain this in terms of a more general account of how people select singular causes. When judging causation, people consider what actions (or events) made the difference to the outcome in question; and this involves counterfactual reasoning about what would have happened if certain things had been different. Moreover, this reasoning is slanted toward considering typical rather than atypical possible worlds (Kahneman & Miller, 1986), which means that people will focus on abnormal actions rather than normal ones as the relevant difference-makers. For example, in the pen vignette, the professor is rated as more causal than the assistant because the possible world in which he does not take a pen (the norm-conforming world) is considered more relevant than the world in which the assistant does not take the pen.

Hitchcock and Knobe see this selective preference for abnormal events as an effective strategy for future intervention or prevention (cf. Hitchcock, 2012; Lombrozo, 2010). For instance, it is relatively straightforward to address the pen problem by enforcing more stringent measures on the professor. This account links with typical situations encountered in legal cases, where something has gone wrong (a transgression of the normal course of events) and one seeks to assign causality in order to readdress the balance. However, it is unclear whether this means that the notion of abnormality should be built into a definition of causal judgment (Halpern & Hitchcock, 2014), rather than being seen as part of the pragmatics of how we use these judgments.

The finding that norm-violating actions receive greater causal ratings is robust, but it admits of alternative explanations. Alicke (2000) argues that it is people's desire to blame agents for adverse outcomes that drives this effect, with people distorting their causal claims in order to justify assigning blame. For example, the professor is clearly more blameworthy than the innocent assistant. However, on its own this account cannot explain situations where a positive rather than a negative outcome occurs, and norm-violation effects persist.

Another possible explanation is that the term "cause" is ambiguous in such scenarios, and that people see the causal question as a request to assign blame (or praise) to the responsible agents (cf. Hart & Honore, 1959/1985; Lagnado & Channon, 2008). For instance, it seems clear that the professor is most deserving of blame in the pen vignettes, and a similar analysis applies to other scenarios. In support of this claim, Samland and Waldmann (2015) replicate the standard causal preference for the norm-violating agent, but show that on an indirect measure of causal strength, people do not differentiate between norm-violating or norm-conforming agents. For example, both professor and assistant are assigned the same causal strength ratings. Samland and Waldmann argue that in such scenarios, when people are asked to assign causality to agents, they tend to make judgments of accountability rather than causality. This accountability hypothesis resonates with the legal practice of including non-causal factors when judging "legal cause" as opposed to "factual cause." It is too early to rule between these alternative explanations. Indeed, it is possible that each account has some validity depending on the circumstances, for example, whether the norms are moral or statistical, whether the outcomes are bad or good.

Most discussions of norms focus on the causality assigned to the norm violator. But the norms can also have a less direct influence on causal ratings. Thus, Kominsky et al. (2015) investigate situations in which two agents' individual actions combine to bring about an outcome, but where one of the agent's actions is marked out because it violates a norm. They illustrate the issue with a classic legal case (*Carter v. Town*, 1870) in which a child buys gunpowder from the defendant, the child's mother and aunt hide it from the child, but in a place where they know he will find it. The child retrieves the gunpowder and suffers an injury. The court did not find against the defendant because his action was "superseded" by the negligent action of the mother and aunt. Effectively their negligent action (which constituted a norm violation) reduced the causality attributed to the defendant.

Kominsky et al. (2015) explore this notion of supersession in everyday causal reasoning problems, some involving moral norm violations (e.g., stealing something or breaking a rule), others involving statistical norm violations (e.g., throwing double six with a pair of dice). They show that when one agent breaks a norm, the causality attributed to the other agent is reduced. For example, when Sue buys one bookend, and her husband Bob completes the pair by stealing a matching bookend from a friend, Sue is rated as less of a cause (of the couple having a matching pair) than when Bob buys the bookend from his friend. Moreover, they also show that this "supersession" effect only occurs when both agents are necessary for the outcome, not when either agent is sufficient. Kominsky et al. explain these findings in terms of robustness: someone's action (e.g., Sue buys the left-side bookend) is judged as less of a cause (of the matching pair), if it required that someone else acted atypically (e.g., Bob steals the matching bookend). Sue's action is less robust because it relies on someone else breaking a norm. They link this to the counterfactual availability of norm-conforming versus norm-violating behavior. When one assesses the extent to which an agent caused an outcome, one takes into account how readily he would have achieved it under alternative possible worlds, and those worlds in which the other agent conforms with, rather than violates, a norm come more readily to mind (cf. Kahneman & Miller, 1986).

### RESPONSIBILITY ATTRIBUTIONS IN GROUPS

A common finding is that a person's individual responsibility is reduced when several people contributed to the outcome (Alicke, 2000; Kerr, 1996). For example, individuals have a reduced sense of responsibility in situations where multiple people would be capable of helping another person who finds herself in an emergency (Darley & Latané, 1968; Latané, 1981).

A series of studies (Gerstenberg & Lagnado, 2010, 2012; Lagnado et al., 2013; Zultan et al., 2012) shows that attributions to individuals in a group context are sensitive to the causal structure that dictates how individual contributions combine to bring about the outcome. The authors have developed a model of responsibility attribution, the *criticality-pivotality model* (CP model, hereafter), which predicts that people's responsibility attributions are influenced by two key considerations: (1) *criticality*—how important a person's contribution is expected to be for bringing about a positive group outcome (*ex ante*), and (2) *pivotality*—how close a person's contribution was to actually having made a difference to the outcome (*ex post*).

Let us illustrate the different notions via a simple example. Consider a situation in which a two-person company is voting whether to market a product (cf. Stapleton's voting example, discussed earlier). For the vote to pass, both members must vote in favor of the motion. In such a situation, each member's action is critical for the outcome—the motion will not pass unless both vote in favor. Contrast this with a situation in which the motion is passed if *at least one* of the members votes in favor. Here, the criticality of each member's action is reduced. Thus, we say that a member's action is more critical when all of the members have to succeed than when the success of one of the members is sufficient for the outcome.

The CP model predicts that people's responsibility judgments increase, the more critical a person's action was perceived to be for the outcome.

The second component of the model is concerned with how close a person's action was to having made a difference to the outcome. Consider a slighlty more complicated situation in which five members vote, and a majority rule is used to determine of the motion passes (cf. Goldman, 1999). Three people vote in favor and two vote against the motion. In this situation, each of the three people who voted in favor was pivotal for the outcome. Had any of them changed his or her mind, the motion would have failed. Contrast this with a situation in which the outcome of the vote is four to one in favor of the motion. Here, none of the members who voted in favor was pivotal. If one of them had changed her mind, the group would still have passed. Expressed in the legal terminology that we have introduced earlier, none of the individual voter's actions was a *but-for* cause of the outcome. However, intuitively each of the voters should still receive some responsibility for the outcome. Based on Halpern and Pearl's (2005) model of actual causation discussed earlier, Chockler and Halpern (2004) proposed a structural model of responsibility that captures this intuition. Their model predicts that the further away a person's action was from having made a difference to the outcome, the less responsible that person's action will be viewed.

In the case where the vote is 3–2, each of the voter's responsibility is high because his or her action was pivotal in the actual situation. In the case where it is 4–1, each of the voter's responsibility is reduced because none of their individual actions made a difference to the outcome in the actual situation. The responsibility of a person's action for an outcome is predicted to be equal to 1/(N+1), where *N* is the minimal number of changes that are required to render the person's action pivotal. Let's consider how much responsibility Joe, who voted for the motion, is predicted to receive in a situation in which the outcome of the vote was 4–1 in favor. Joe's vote didn't make a difference in the actual situation. However, if we changed the vote of one of the other people who was in favor of the motion, then Joe's vote would have made a difference. Since one change is required to make Joe pivotal, Joe's responsibility is predicted to be equal to 1/2.

Lagnado et al. (2013) tested the CP model by manipulating both the criticality and pivotality of a person's action in a variety of group situations (including group competitons and public goods games). They showed that people's responsibility attributions were sensitive to both criticality and pivotality, and were well predicted by the CP model (see also Gerstenberg, Halpern, & Tenenbaum, 2015). A natural extension would be to apply this model directly to legal and moral contexts.

### *Causal Simulation*

Our look at the psychological research on attribution has focused mainly on empirical findings rather than well-worked-out psychological theory. This is partly due to the lack of any comprehensive theory that explains how people reach their causal judgments. As noted in the section on explanatory and evidential reasoning, the story model presents an attractive approach to juror decision-making, especially if extended to include more rigorous notion of causal explanation and a framework for reasoning about evidence and its reliability. Another promising feature of the story model is the idea that people use their causal knowledge and situation models to simulate possible ways (narratives) in which people's actions might have led to the crime in question.

The idea that people use mental simulations to make judgments of causality and probability was introduced by Kahneman and Tverksy (1982). Athough a fertile idea, and developed in various areas including legal decision-making (Feigenson, 1996; Heller, 2006), causal simulation has mainly been cast as a mental heuristic that avoids complex computation and can yield biased inferences. We endorse Kahneman and Tverksy's original insight that simulation is a crucial aspect of psychological thinking—especially in causal reasoning—but would argue that it is more sophisticated than a mere heuristic (although there might be heuristic ways of achieving it in complicated situations) and should be mapped onto a richer framework of causal representation and inference (cf. Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Goodman, Tenenbaum, & Gerstenberg, 2015).

In short, we should take mental simulation seriously as a complex feature of causal cognition, a capability that involves richly structured representations of the physical and social world, and engages inferential machinery that can often deliver sound inferences, both about actual and counterfactual eventualities. This thesis can be developed in various ways (Gerstenberg & Tenenbaum, Chapter 27 in this volume; Sloman & Lagnado, 2015), and is a promising area for future research.

### *Summary*

We have seen that laypeople's causal attributions accord well with legal judgments, and operate with a very similar notion of causation. This notion is more sophisticated than a simple *but-for*, but can often be captured by an extended counterfactual analysis. At the heart of everyday and legal reasoning is the reliance on causal models, and the focus on human agency and social interactions. Judgments of causation also take factors such as intentions and foresight into account, and thus overlap with issues of responsibility and blame. This could be taken as a departure from rational scientific inquiry, but could also be cast as a consequence of the goals and aims of causal judgments—which in both everyday and legal contexts often serve to identify wrong-doers or deviant behavior.

## Moral Reasoning

Research in moral psychology has focused on a number of different research topics, such as the distinction between moral norms and conventions (Chakroff et al., 2013; Sripada & Stich, 2006), the role of intuition and emotion, such as disgust or empathy, in moral judgment (Greene, 2001; Greene & Haidt, 2002; Haidt, 2001; Haidt et al., 1997), and whether some things, such as a person's life, have sacred values that make them incommensurable with other things, such as money (Tetlock et al., 2000). Here, we will focus on a topic that we believe best illustrates the role of causal and counterfactual thinking in moral cognition: people's evaluative judgments in moral dilemmas.

Work on moral psychology has drawn heavily from philosophical work on normative ethics. In ethics, there are three dominant approaches of how to analyze moral behavior. First, according to *deontological theories*, the morality of a person's action derives from its accordance with a set of moral norms or duties (Darwall, 2003b; Kant, 2002). An action is good if it adheres to set of moral principles or rules. Second, according to *consequentialist theories*, the morality of a person's action is determined by the consequences it brings about (Darwall, 2003a; Smart & Williams, 1973). An action is good if it leads to good outcomes. A third approach, *virtue theories*, focuses on what the action says about the person's character (Darwall, 2003c). An action is good to the extent that it indicates good or virtuous character.

These normative theories emphasize the three elements that are part of any moral analysis of a situation: persons (virtue theories), actions (deontological theories), and consequences (consequentialist theories; Sloman et al., 2009). We will argue that people's moral evaluations of another person's behavior are best understood if we assume that they consider both the causal role that a person's action plays in bringing about the outcome, as well as what the action says about the person's character. Further, we will argue that both action-focused and character-focused considerations are best captured in terms of counterfactual contrasts over people's intuitive causal theories of the domain (cf. Gerstenberg et al., 2015; Goodman et al., 2015).

This section has two parts: in the first part we look at how representing moral dilemmas in terms of causal models that support counterfactual reasoning helps us understand how people make moral judgments. We will see that in order to analyze the causal status of a person's action, we need to have a causal representation of the situation that dictates how the person's action relates to the outcome under consideration. To draw inferences about a person from her action, we need a theory of mind—a causal model of how people plan and choose their actions. We can then invert that model to reason from an observed action to aspects about the character (Baker et al., 2009; Gopnik & Wellman, 1992; Kleiman-Weiner et al., 201~~6~~; Malle & Knobe, 1997; Wellman & Gelman, 1992; Yoshida et al., 2008).

In the second part, we will argue that in order to arrive at a more complete picture of how people make moral evaluations, we will need to shift focus from merely considering the moral permissibility of an action to considering more fully what the action reveals about the person's character.

### *Causality and Counterfactuals in Moral Dilemmas*

Any moral evaluation has to start with a (rudimentary) causal analysis of the situation (Cushman & Young, 2011; Driver, 2008; Guglielmo, Monroe, & Malle, 2009; Mikhail, 2007, 2009; Sloman et al., 2009). Clearly, we would not blame someone whose action played no causal role whatsoever in how the outcome came about. The counterfactual *but-for* test mentioned earlier provides a first pass for evaluating whether a person's action made a difference to the outcome. As we will see, the *but-for* test can also help us to draw a distinction between indended outcomes of an action, and outcomes that were merely forseen but not ~~inteded~~.

In a typical moral dilemma, the agent faces a decision between several actions, each of which is expected to lead to a different negative outcome.

One of the most-studied moral dilemmas is the trolley problem (Foot, 1967; Thomson, 1976, 1985). In a typical trolley scenario, a trolley is out of control and headed toward five people standing on a railroad track. A person, let's call him Hank, observes this. If Hank doesn't do anything, the trolley will kill the five people on the track. However, Hank is close to the control room and he can throw a switch that will change the course of the train onto a side track. As it turns out, there is one person standing on the side track. If Hank throws the switch, the five people on the main track will survive, but the one person on the side track will die. If Hank doesn't throw the switch, the five people on the main track will die, but the person on the side track will survive. Is it morally permissible for Hank to throw the switch?

When faced with this *side-track scenario*, most participants tend to think that it is permissible for Hank to throw the switch (for a review, see Waldmann et al., 2012). Clearly, the consequentialist is on Hank's side: if Hank throws the switch only one person will die, whereas if he doesn't throw the switch five people will die.

Now let's consider another variant of the trolley scenario in which, again, an out-of-control trolley is threatening to kill five people. This time, Hank finds himself on a bridge that crosses the railroad track and the only option he has for stopping the trolley is to push a large man off the bridge onto the track. This will stop the train but kill the large man. Is it morally permissible for Hank to push the large man off the bridge?

Most participants don't think so. For the consequentialist, this is puzzling: in both the *side-track* and the *push* scenario, the person faces a choice between two outcomes: either five people die, if he doesn't act, or only one person dies, if he does act. If all that mattered was the number of deaths, then participants should clearly consider it permissible for Hank to push the large man off the bridge.

Much of research in moral psychology has been devoted to explaining what factors account for the difference in people's intuitions about the moral status of a person's actions between different moral dilemmas. Even though the side-track and push scenarios are similar on an superficial level—there is the same contingency between acting or not acting and the number of people who die as a consequence—they are also different in important respects. So while the consequentialist is somewhat at a loss, the deontologist can attempt to find a principled rule that distinguishes between these cases.[27]

One such rule is the doctrine of double effect (DDE; Foot, 1967; Kamm, 2007; Quinn, 1989). The DDE draws a distinction between two types of effects that can result from a person's action: first, an effect that is desired and intended, and second, an effect that is undesired but foreseen. For example, in the *side-track scenario*, there are two effects when Hank throws the switch. The five people on the main track are saved, and the one person on the side track is killed. The DDE states that an action that would normally be prohibited, such as homicide, may be morally permissible when (1) the (negative) action itself is not directly intended, (2) the good effect of the action is indended but not the bad effect, (3) the good effect outweighs the bad effect, and (4) the actor had no morally better alternative (see Mikhail, 2009).

Thus, throwing the switch is morally permissible according to the DDE, if (1) Hank didn't just throw the switch because he likes throwing switches, (2) he intended to save the five people but didn't intend to kill the person on the side track (even though he foresaw that outcome), (3) the positive effect of saving the five outweighs the negative effect of killing the one, and (4) there was nothing else that Hank could have done which would have led to a morally better overall outcome.

Now what does the DDE say about the *push scenario*? Again, let's assume that Hank is not the kind of guy who enjoys pushing people off bridges just for kicks. What about Hank's intention? Did he intend to kill the man on the bridge? Note that the causal structure between action and outcome in the push scenario is different from the side-track scenario. Let's assume again that Hank intends to save the five people. However, in order to realize his primary intention, he has to kill the large man first. Pushing the large man is not merely a foreseeable side effect of his action, but it features as a causal means in a chain of events that culminates in bringing about the intended outcome. Using the large man as a means implies that Hank intended for the large man to die. Since killing the man was an essential part of Hank's plan to save the people on the track, the DDE rules that this action is impermissible.

How can we tell apart whether a particular outcome was only a side effect of a person's action, or a means for bringing about another outcome? In many situations, the *but-for* test provides a simple procedure to determine whether a particular effect was a means versus a side effect of an action (see Figure 29.7). In the push scenario (Figure 29.7 b),

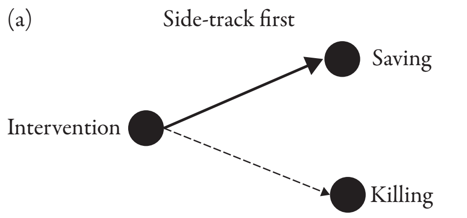

**Figure 29.7** Highlighted causal structures of the side-track and push scenarios as a function of the order of presentation. Highlighted causal paths are ~~shown in bold~~.

Figure adapted from Wiegmann & Waldmann (2014).

if the large man hadn't been shoved off the bridge, then the five people would not have been saved. The survival of the five depends counterfactually on the death of the one. Thus, pushing the large man off the bridge was a means for saving the five. In contrast, in the side-track scenario (Figure 29.7 a), the five on the main track would have been saved even if there had been no person on the side track. Given that Hank threw the switch, the survival of the five was not counterfactually dependent on the death of the one (they would have survived even if the person on the side track had managed to jump off the track). Thus, the death of the person on the side track was a side effect, rather than a means for saving the five.

As discussed earlier, the doctrine of double effect draws a distinction between effects that were intended and effects that were merely foreseen (see also Dupoux & Jacob, 2007; Mikhail, 2007, 2009). Again, counterfactuals can help us to capture this difference. Rather than considering counterfactuals over events that happened in the world, we consider counterfactual contrasts over our model of how the agent made her decision. Assuming that an agent's decisions are determined by her mental states, such as her beliefs, desires, and intentions (Dennett, 1987), we can invert the causal process from decision to action and infer the agent's mental states from her actions (Baker et al., 2009; Yoshida et al., 2008). How can we explain that saving the five on the main track was an intended consequence of throwing the switch, whereas killing the one on the side track was a foreseen but unintended effect? We can do so via considering whether a particular effect influenced the agent's decision (Nanay, 2010; Sloman et al., 2012; Uttich & Lombrozo, 2010) or plan (Kleiman-Weiner et al., 2015).

Let us assume that Hank threw the switch in the side-track scenario. This action is consistent with different intentions that may have driven Hank's action. Hank may have intended to save the five people on the main track. He may have intended to kill that one person on the side track. Or he may have inteded both. Let's first assume that Hank actually intended to save the five people on the main track, and he didn't intend to kill the person on the side track. If that's the case, then he would have also thrown the switch if there had been no person on the side track. In other words, the person on the side track made no difference to Hank's throwing the switch. In contrast, if the five people on the main track hadn't been there, then Hank would *not* have thrown the switch (assuming that people are generally lazy and don't just throw switches for no reason). Thus, the people on the main track, but not the person on the side track, made a difference to Hank's decision-making. By considering a causal model of the decision-maker, we can distinguish intended from foreseen effects: intended effects make a difference to the decision, whereas merely foreseen but not intended effects make no difference (Guglielmo & Malle, 2010).

Of course, Hank's throwing the switch in the side-track scenario is also consistent with the possibility that Hank intended to kill the person on the side track. If that was his intention, then he might not have thrown the switch if the person on the side track hadn't been present (unless he also intended to save the five). As long as we don't have reasons to the contrary, we are generally inclined to assume that a person is more likely to have good intentions (Mikhail, 2007, 2009)—thus, we would consider it more likely that when Hank threw the switch, he intended to save the five, rather than kill the one.

In a recent study, Kleiman-Weiner et al. (2015) showed a close correspondence between the inferences that people make about an actor's intentions and the extent to which they deem the person's action morally permissible. In their experiments, they varied the number of people on the main track and the number of people on the side track. In some of the situations, participants were informed that the person on the track was the decision-maker's brother. In situations in which the decision-maker threw the switch, participants generally judged that the decision-maker didn't intend to kill the people on the side track, and that he had no intention for those on the main track to be killed. However, participants' judgments were sensitive to the number of people on the different tracks and to whether the decision-maker's brother was involved. For example, participants were more inclined to say that the decision-maker actually intended to kill the people on the side track when there was only one anonymous person on the main track but five anonymous people on the side track. In that case, participants were also slightly less willing to say that the decision-maker intended to save the person on the main track. The decision-maker's action is consistent with a desire to kill as many people as possible.

Now consider a situation in which the decision-maker's brother is on the main track, there are five people on the side track, and the decision-maker throws the switch. In this case, participants are less inclined to believe that the decision-maker intended to kill the five people on the main track, and more likely to believe that the decision-maker's intention was to save his brother. While the decision-maker's action is still consistent with a desire to kill as many people as possible, we have a viable alternative for why he acted the way he did. He may simply value his brother more than anonymous strangers.

Kleiman-Weiner et al. (2015) show that a model of moral permissibility that combines inferences about a person's intention with a consideration of how many lives were saved and killed explains people's judgments very accurately.

### *What Counterfactual Contrasts Do People Consider?*

So far, we have used counterfactual contrasts to help us make morally relevant distinctions. By defining counterfactual contrasts over the causal structure of the situation, we were able to tell apart outcomes that were side effects of actions from outcomes that were means for bringing about another outcome. By considering counterfactual contrasts over people's plans, we were able to tease apart outcomes that were intended from outcomes that were merely foreseen but not intended. But what kinds of counterfactual contrasts do people actually consider when they make moral judgments? A host of research on counterfactual thinking has demonstrated how some counterfactuals come to mind more easily than others (e.g., Kahneman & Miller, 1986; Phillips, Luguri, & Knobe, 2015; Roese, 1997) Maybe the the notion of counterfactual availability can also help us make sense of people's moral judgments?

Waldmann and Dieterich (2007) have shown that participants find an intervention on the threat more permissible than an intervention on the victim. They contrasted the side-track scenario (threat intervention) with a scenario in which Hank can intervene by redirecting a bus containing the victim onto the train track and thereby stopping the train (victim intervention). In order to explain the pattern of people's judgments, Waldmann and Dieterich (2007) suggest that, depending on the type of intervention, participants selectively focus on different counterfactual contrasts (cf. Schaffer, 2010). When intervening on the threat, people compare the causal path the trolley would have taken with the path that it actually took. This counterfactual contrast highlights the difference between the five on the main track versus the one on the side track. In contrast, when intervening on the bus with the victim, the counterfactual contrast highlights what would have happened to the victim if Hank hadn't intervened and redirected the bus on the train track. Here, the contrast is between the victim surviving and the victim dying. Waldmann and Dieterich (2007) further argue that the attentional focus triggered by the victim intervention leads to a neglect of other potential victims in the background (i.e., the five on the track)—a phenomenon they call intervention myopia (see also Waldmann & Wiegmann, 2010).

Based on these selective attention effects, Wiegmann and Waldmann (2014) have recently developed an account that explains transfer effects between moral dilemmas. Several studies have shown that the order in which different moral dilemmas are presented affects participants' judgments (e.g., Schwitzgebel & Cushman, 2012). For example, participants judge intervening in the switch scenario less permissible if they were first asked to make a judgment about the push scenario than when the order is reversed.

Wiegmann and Waldmann (2014) explain this order effect by assuming that different scenarios make different causal paths and the associated counterfactual contrasts more salient. Their account focuses on an analysis of the causal structure that underlies the different moral dilemmas as well as people's default evaluations for the different cases (cf. Halpern & Hitchcock, 2015; Hall, 2007). In the switch scenario (Figure 29.7 a), the action has two effects via separate causal paths: saving the five and killing the one. When participants see the switch scenario in isolation, they tend to judge the person's action to be permissible. In line with the good intention prior (Mikhail, 2009), Wiegmann and Waldmann (2014) propose that participants selectively focus on the causal path from intervention to saving rather than the connection between intervening and killing (Figure 29.7 a). In the push scenario (Figure 29.7 b), there is a single causal path from intervening to saving via killing. Here, it is not possible to selectively focus on the relationship between intervening and saving since the causal path is mediated via the killing of the large man. In contrast, the relationship between intervention and killing is salient (Figure 29.7 b).

The key idea is now that participants have a tendency to map salient aspects of the causal structure from one situation to another if such a mapping is possible (cf. Gentner, 1983; Holyoak et al., 2010). Consider a participant who judged the switch scenario before the push scenario (top pair in Figure 29.7). In the switch scenario, the causal path from intervention to saving is highlighted. However, it is not possible to map this path onto the causal structure of the push case since there is no direct causal path between intervention and saving. In contrast, consider a participant who saw the push scenario before the switch scenario (bottom pair in Figure 29.7). The push scenario highlights the link between intervention and killing (what Waldmann & Dieterich, 2007, termed *intervention myopia*). Now, it is possible to map this highlighted causal path onto the causal structure in the switch case. Having judged the push scenario first highlights the relationship between intervention and killing in the switch scenario. Wiegmann and Waldmann (2014) argue that the reliable order effect arises from the asymmetric way in which selectively attended parts of the causal structure can be mapped from one situation to another.

In the trolley problems discussed earlier, the counterfactual analysis was fairly straightforward since the vignettes explicitly stipulated the action–outcome contingency. In the real world, however, we normally cannot be certain about what would have happened in the relevant counterfactual world. We have to rely on our causal understanding of the situation to simulate how the world would have unfolded if the person had acted differently (see Gerstenberg & Tenenbaum, Chapter 27 in this volume). Generally, we cannot be sure that actions always bring about their intended effects (Cushman et al., 2009; Gerstenberg et al., 2010; Schächtele et al., 2011). Imagine that Hank threw the large man off the bridge, but it turned out that this didn't suffice to stop the trolley. In that case, Hank not only failed to save the five people on the track, he also killed an innocent man for no good effect. Studies have shown that people take the uncertainty associated with different actions into account when making moral evaluations (Fleischhut, 2013; Kortenkamp & Moore, 2014).

### *From Action Permissibility to Character Evaluation*

In the previous section, we have seen how different aspects of how a person's action features in the causal structure that ultimately led to the outcome affects people's moral evaluations. We have also seen that people's judgments are solely determined by the causal role that the action played in bringing about the outcome. The same action is evaluated differently based on the context of the situation and the inferences we can draw about the person's intentions from his actions. The majority of work on judgments in moral dilemmas has focused on explaining how people judge the moral permissibility of different actions.

Recently, scholars in moral psychology have argued that this focus on actions as the unit for moral evaluation is misguided. What people mostly care about is what the action reveals about the person (Goodwin, Piazza, & Rozin, 2014; Kelley & Stahelski, 1970; Malle, Guglielmo, & Monroe, 2014; Pizarro & Tannenbaum, 2011; Sripada, 2012;Uhlmann et al., 2013; Uhlmann et al., 2015; Wojciszke, Bazinska, & Jaworksi, 1999; Woolfolk et al., 2006). Rather than putting the action at the center of analysis, the *person-centered* approach sees the person as the key target for moral evaluation (cf. Uhlmann et al., 2015).

People are evaluating creatures—upon meeting someone for the first time, we try to figure out what makes them tick (Alicke et al., 2015). Moral evaluations of a person's character, such as whether he cares for others (Hamlin et al., 2007; Ullman et al., 2009), and whether he can be trusted (Charness & Dufwenberg, 2006; Rezlescu et al., 2012), are particularly important (Goodwin et al., 2014; Todorov et al., 2008). In contrast to the action-centered view, the person-centered view focuses on people's motivation for engaging in moral evaluations and is concerned with explaining what function these evaluations serve (Gintis et al., 2001, 2008). One such proposed function of moral evaluation is relationship regulation (Rai & Fiske, 2011; Scanlon, 2009). As "intuitive prosecutors," our moral judgments serve to shape the world as we would like to see it (Fincham & Jaspars, 1980; Hamilton, 1980; Lloyd-Bostock, 1979; Tetlock et al., 2007). For example, we want that our friends care for us and we want to be able to rely on them when in need. By blaming them for not having helped us when we moved house, we send a signal to change their behavior in the future (Bottom et al., 2002; Scanlon, 2009). However, one might argue, why do we care to engage in moral evaluation of people with whom we have no direct connection? There is a lot of blaming going on in sports bars! When our favorite football team loses because one of the players slacked off, then we expect that player to put in more effort in the future. Even though he is not a close friend, our utility depends on him, and by blaming him we contribute to a public evaluation of the player that may indeed have an influence on his future behavior (McCullough et al., 2013).

There is also a lot of blaming going on when people watch soap operas together (Hagmayer & Osman, 2012). Here, we know that the characters are fictional and our moral evaluations won't influence the plot. However, we can again make sense of this behavior from a functional perspective: moral judgments in these contexts may serve to coordinate one's expectations and norms with one's friends. By blaming John for cheating on his girlfriend, we demonstrate to our friends that we value faithfulness and expect our friends not to follow John's example.

With a focus on the person, rather than the action, the person-centered view needs to answer the question of how we infer a person's character traits from her actions—in particular, her moral traits. Actions differ in the extent to which they are diagnostic about a person's dispositions (Reeder & Spores, 1983; Reeder et al., 2004; Snyder et al., 1979). We learn most about a person's character from behaviors that are different from how we would have expected others to behave in the same situation (Ditto & Jemmott, 1989; Fiske, 1980; Jones & Harris, 1967; McKenzie & Mikkelsen, 2007; Reeder et al., 2004; Reeder & Brewer, 1979). But what shapes this expectation? One factor is the cost or effort it takes for someone to do a certain action. We generally assume that others behave rationally and act in a way to achieve their desired outcomes efficiently (Dennett, 1987). Thus, the more costly or effortful a particular action was for the agent, the more certain we can be that he really valued the outcome (Jara-Ettinger et al., 2014; Ohtsubo & Watanabe, 2009; Ullman et al., 2009). Similarly, we learn that a person values something when he had an attractive alternative (Ben-Porath & Dekel, 1992). We know that a friend really cares for us when he comes over to help console us after a breakup even though he had been invited to the party of the year.

The time it took someone to make a decision is another factor that influences what we can learn about her motivation for acting (Crockett, 2013; Cushman, 2013; Pizarro et al., 2003). Critcher et al. (2012) found that actors who acted quickly were evaluated more positively for good outcomes and more negatively for bad ones, compared to actors who reached the same decision more slowly. Fast decisions signal that the actor was sure about her action and did not need to resolve any conflicting motives. A person who immediately rushes to help someone in need is likely to be more strongly motivated by another person's needs than someone who first considers how much effort it would take to help (cf. Hoffman et al., 2015) and checks if anyone else might be in a better shape to help out.

Additional evidence in favor of the person-centered view to moral judgment comes from research showing that information about a person's general character influences our moral evaluation of a particular action (Alicke, 1992; Kliemann et al., 2008; Nadler, 2012; Nadler & McDonnell, 2011). We find ways of blaming bad people (Alicke, 2000) and excusing people whom we like (Turri & Blouw, 2014). Often, there is considerable uncertainty about the motives behind a person's action. Thus, interpreting the same action differently, depending on who performed it, must not be taken to reflect a biased evaluation. It may be reasonable to use character information to fill in the gaps (Gerstenberg et al., 2014; Uhlmann et al., 2015).

The person-centered view also provides a natural way of handling the expectations that come with exhibiting a certain role (Hamilton, 1978; Schlenker et al., 1994; Trope, 1986; Woolfolk et al., 2006). For example, if a swimmer is about to drown, then it is foremost the lifeguard's responsibility to try to save him. If the swimmer drowned without anyone having helped, then we would blame the lifeguard more than any of the other people who were around and who could have also helped. Part of what it means to be a lifeguard is to have the (prospective) responsibility of making sure that everyone is safe in the water.

Direct empirical support for the person-centered approach comes from work showing dissociations between person and act evaluations (e.g., Tannenbaum et al., 2011; Uhlmann & Zhu, 2013). Such situations arise, for example, when someone takes the "right" action (for example, from a consequentialist perspective), but taking this action indicates a bad moral character (Bartels & Pizarro, 2011; Koenigs et al., 2007; Uhlmann et al., 2013). The act of throwing an injured person overboard in order to save the boat from sinking is evaluated more positively than not doing so. However, the passenger who decided to throw the injured person overboard was evaluated more negatively than a passenger who decided not to do so (Uhlmann et al., 2013). Similarly, the person-centered view helps to shed light on people's moral evaluations of harmless-but-offensive transgressions (Haidt, 2001; Uhlmann et al., 2013). For example, most people consider it morally wrong to eat a dead dog, but often find themselves at a loss when trying to explain why (a phenomenon coined *moral dumbfounding*; Haidt et al., 1993). Uhlmann et al. (2013) showed that while the act of eating a dead dog is not evaluated more negatively than stealing food, the person who ate the dog is judged to be a worse person than the person who stole. Even though eating a dead dog didn't harm anyone (as long as the dog wasn't killed to be eaten), it is plausible that a person who commits such an act is also likely to engage in other dubious behavior that might actually be harmful. Interestingly, while Uhlmann et al. (2013) replicated the moral dumbfounding effect for judgments about actions, participants had much less difficulty justifying the character inferences they had drawn.

What the person-centered view highlights is the need for a rich model of how people make (moral) decisions. We need such a model in order to make inferences about the person's mental states and preferences from his or her actions (Ajzen & Fishbein, 1975; Baker et al., 2009; Bratman, 1987; Malle & Knobe, 1997). For example, when Hank didn't help a drowning swimmer, we need to infer what Hank's beliefs were (maybe Hank thought the person was just pretending; Young & Saxe, 2011), and what Hank would have been capable of doing (maybe Hank wasn't able to swim; Clarke, 1994; Jara-Ettinger et al., 2013; Kant, 2002; Morse, 2003; van Inwagen, 1978). Not only do we need a causal model that explains Hank's actions in terms of his mental states, we also want to be able to simulate how someone else (maybe with the same beliefs and capabilities as Hank, but with different desires) would have acted in the same situation.

Recently, Gerstenberg et al. (2014) have suggested an account that directly links inferences about a person's character to evaluations of his or her behavior (cf. Johnson & Rips, 2015). In their studies, they had participants judge to what extent actors whose action was either expected or surprising were responsible for a positive or negative outcome. For example, in one scenario, participants evaluated goalkeepers in a penalty shoot-out. The goalkeepers knew about the striker's tendency to shoot in either corner of the goal. However, the strikers didn't know that the goalkeepers knew about their tendency. Participants then saw situations in which the striker either shot in the expected corner or in the unexpected corner, and the goalkeeper either jumped in the correct corner and saved the ball, or jumped in the wrong direction.

The results showed that participants blamed the goalkeeper more for not saving the shot when he jumped in the unexpected direction and the striker shot in the expected direction. Participants also considered the goalkeeper more creditworthy overall when he saved a shot that was placed in the unexpected direction.

In another condition of the experiment, the goalkeeper scenario was replaced with a scenario in which the decision-maker had to predict on what color a spinner will land. This scenario was structurally equivalent to the goalkeeper scenario, and the probabilistic information was matched. Again, participants blamed decision-makers more for negative outcomes that resulted from unexpected predictions (e.g., predicting that the spinner will land on blue when the chance for it landing on yellow was greater, and it actually landed on yellow). This time, however, decision-makers were praised more for positive outcomes that resulted from expected rather than unexpected actions (i.e., correctly predicting that the spinner will land on the more probable color, rather than correctly predicting the less probable outcome).

How can we explain this pattern of results? Gerstenberg et al.'s (2014) account assumes that people's responsibility judgments are mediated by an inference about the agent. Accordingly, we start off with some assumptions about how a reasonable person (or goalkeeper) is expected to act. After having observed what happened in this particular situation, we update our belief about the person. Gerstenberg et al. (2014) propose that responsibility judgments are closely related to how we change our expectations about people (cf. Ajzen & Fishbein, 1975). Intuitively, we credit people more if our expectation about a person improved after having observed their action. We blame people more for a negative outcome if our expectation about their future behavior is lowered.

In their model, Gerstenberg et al. (2014) represent people's intuitive theories about the situation as distributions over agents with different character traits or skills. The key idea is then that people have different prior assumptions for the goalkeeper and the spinner scenarios. A skilled goalkeeper may anticipate an unexpected shot and save it. For spinners, howerver, it's less likely that a person can reliably predict that it will land on the less likely color. Hence, if we observe a goalkeeper saving an unexpected ball, we may either think that he acted unreasonably and was just lucky, or that he in fact correctly anticipated the shot. If we deem the chances of skill being present to be reasonably high, then our expectations about this goalkeeper's behavior increase after having seen him save the unexpected ball (and more so compared to a situation in which he saved an expected shot). In contrast, in the spinner scenario, the most likely explanation for a person correctly predicting the spinner's landing on the unexpected outcome is that he was just lucky. Observing such an action actually leads us to lower our expectation about that person's behavior, and thus he is deserving of less credit.

Gerstenberg et al. (2014) applied their model to an achievement domain in which skill is a critical factor. However, their model could be extended and applied more directly to the moral domain. The key idea of linking people's moral evaluations to a difference in expectations is flexible enough to accommodate different aspects of the situation, such as immoral desires or ulterior motives.

One natural way to think about the role of expectations in judgments of responsibility is again in terms of a counterfactual contrast. Rather than thinking about how the outcome would have been different if the person had acted differently, we may think about what would have happened if we had replaced the person in the situation with someone else (cf. Fincham & Jaspars, 1983). In the law, this idea is referred to as the reasonable-man test (Green, 1967). In cases of negligence, for example, we may have expectations about what sort of precautions a reasonable ~~man~~ would have taken that might have prevented the harm from happening. Relatedly, there is a statistic in baseball called *wins above replacement* that captures the difference that a person makes to the number of games that a team wins over the course of the season (Jensen, 2013). It tries to quantify how many more games the team won over the season compared to a counterfactual team in which the player under consideration would have been replaced with another player.

Thinking about moral evaluations in terms of counterfactual replacements provides a rich framework that links up normative expectations, action evaluation, and character inferences. The richness of this framework comes with great theoretical demands. Not only do we need a causal representation of the situation that allows us to reason about the relationship between the person's action and the outcome, but we also require an intuitive theory of the different factors that influence how people make decisions and plans, and how these personal characteristics translate into morally relevant behavior.

### *Discussion*

In this section on moral reasoning, we have seen that people's moral judgments are strongly influenced by their causal representation of the situation, as well as their intuitive theory of how people make decisions.

To evaluate the causal role that the agent's action played, we need a causal model of the situation that supports the consideration of counterfactual contrasts. Our discussion of the literature on trolley problems (Waldmann et al., 2012) showed that we can distinguish means from side effects in terms of counterfactuals on actions, and intended from merely foreseen outcomes in terms of counterfactuals on plans (cf. Kleiman-Weiner et al., 2015). If we additionally assume that some counterfactual contrasts are more salient than others, we can also make sense of why people's moral intuitions differ depending on whether the action targets the threat or the victim (Iliev et al., 2012; Waldmann & Dieterich, 2007), and we can use the idea of salient causal paths to explain transfer effects between trolley problems that differ in their causal structure (Wiegmann & Waldmann, 2014).

To evaluate what we can learn about a person from his or her action, we need a causal model of how people make decisions and plans (Baker et al., 2009; Wellman & Gelman, 1992). Once we have a generative model of how people's mental states determine their actions, we can invert this process and reason about a person's mental states from having observed her actions. Making person inferences comes natural to us, and we have argued that evaluating others along moral character dimensions serves important functions, such as regulating relationships and coordinating normative expectations (Rai & Fiske, 2011; Uhlmann et al., 2015). Finally, we have briefly sketched one way of how we can go from character evaluations to attributions of responsibility via considering counterfactuals over persons—blame and credit vary as a function of our expectations about how a reasonable person should have acted in the given situation (Gerstenberg et al., 2014).

Considering both the person-centered and action-centered views suggests that differences in people's moral judgments can arise from different sources: (a) Two people might disagree about the causal status of the person's action in bringing about the outcome. One person, for example, might believe that the outcome would have happened anyway, whereas the other person might believe that the negative outcome would have been prevented but for the person's action. (b) Two people might disagree about what the action reveals about the person. Whereas the same action might look like an expression of genuine altruism to one person, another person might infer ulterior motives behind the action. We suggest that rather than artificially providing people with all the relevant information, as is often done in psychological studies on legal and moral judgments, it will be fruitful to design experiments that reflect the uncertainty inherent in our everyday lives. New empirical investigations into how people make person inferences and causal judgments in the realm of uncertainty will have to go hand in hand with the development of a coherent formal framework that encompasses both the action-centered and person-centered view.

## Conclusions

Causality is at the core of people's understanding of the physical and social world. In this chapter, we have shown that causation is key to legal and moral reasoning. We have seen that legal scholars and psychologists struggle with very similar issues. Legal scholars seek a principled account of causation that can be applied to complex scenarios and that fits with our common-sense intuitions. Here formal work on causal models and structural equations (e.g., Halpern & Pearl, 2005) has helped to sharpen our theories of causation. The idea of thinking about actual causes as difference-makers under possible contingencies resonates well with how the law has extended the *but-for* test of causation (e.g., Stapleton, 2008). Legal decision-making also requires people to assess the evidence presented to support their conclusions. Causal reasoning is again critical, both in terms of the stories people tell to make sense of the evidence, and in terms of the methods they use to assess its credibility and reliability. Moreover, causal attributions seem to be shaped by pragmatic goals. People's everyday causal judgments often serve to attribute responsibility and blame, mirroring the way in which legal judgments of causation are geared toward ultimate judgments of legal responsibilty.

Psychologists also try to come up with a principled account of how people make moral judgments. We have shown that representing moral dilemmas in terms of causal models and counterfactuals helps us understand people's judgments. Moreover, we argue that a fuller picture of how people make moral evaluations requires a shift of focus from the moral permissibility of actions to the broader issue of what actions reveal about the person's character. Whereas the law has traditionally put less emphasis on the perpetrator's character (Bayles, 1982; Duff, 1993), work in psychology has shown that our moral evaluations are heavily influenced by our inferences about what the person is like (Uhlman et al., 2015). Thus, there is a possible tension between the factors on which people base their intuitive moral judgments and the factors the law deems relevant in determining legal liability.

For both legal and moral reasoning, then, it it is crucial to understand the causal models that people construct, and the rich factual and counterfactual inferences they draw on this basis. Pragmatic and emotive concerns might shape and possibly distort these models, but without them we cannot get started on the route to blame, praise, or indifference.

## Notes

1. Adapted from *Bolitho v. City and Hackney Health Authority* [1997] UKHL 46. This case is also discussed in Schaffer (2010).
2. This holds for issues of foreseeability, too; thus UK law maintains that an action is intentional if the result was almost certain to occur given the defendant's actions and the defendant was aware of this (Herring, 2010).

3. The term applies to both legal theorists and judges.
4. See, e.g., Kennedy 2007 UKHL 38.
5. Readers who have not seen *Breaking Bad* should skip this example until they have!
6. Empricial research backs this up insofar as Hank is the most commonly cited cause; but it also shows that people attribute some causality to Tuco as well (Wallez & Hilton, 2015).
7. This is a simplification, because the legal issue to be decided will dictate the focus of the causal inquiry about what happened. Thus, for a murder charge, the focus will be on the action of the defendant. Nevertheless, legal judgment is not supposed to enter at this stage, and an "objective" notion of causation is usually assumed (Hart & Honore, 1985; Moore, 2009) but see Green (2015) and Tadros (2007) for alternative views.
8. *R v. Rafferty*, 2007, EWCA Crim 1846.
9. *R v. Nedrick* [1986] 3 All ER 1 Court of Appeal.
10. Despite the subtlety of this case, people's intuitive judgments seem to fit with the legal reasoning here (see later discussion).
11. This need not mean it is not principled, but that the principles are difficult to articulate, and might include legal/policy considerations.
12. *Saunders v. Adams*, 117, So 72, (Ala 1928).
13. Note that this response is not always available—sometimes the pre-empted cause would have brought about the effect in exactly the same way (see the voting example later in the chapter). Moreover, the level of description of the outcome required by law is matched to the offense, and is not excessively fine-grained (for discussion, see Stapleton, 2008, and Lewis, 2000).
14. Adapted from German Court case: (37 BGHSt 106, 6 July 1990).
15. Stapleton's point that causal claims are relative to a frame of inquiry is well taken. And legal frames help filter and narrow the focus of inquiry onto a manageable set of putative causes: the defendant's actions or breach of duty, intervening actions of other parties, and so on. Indeed, inquiry-based filtering also holds in our everyday causal inquiries, where we usually only care about a limited pool of candidate causes, and safely ignore a multitude of *but-for* conditions. However, her claim that the inquiry-relative nature of legal causation militates against a general-purpose theory is less convincing (cf. Schaffer, 2010), especially since, as we shall argue later, the legal account she proposes is in many respects an informal version of the definition of actual causation defended in several current philosophical theories.
16. This is not intended as a reductive analysis of causation—because counterfactuals themselves depend on prior causal knowledge, and work as a test of causation, not a constitutive definition. Moreover, this is not to claim that issues of process and production are fully accomodated by counterfactual analyses.
17. Although not classified as such, these graphs bear strong resemblances to formal causal networks.
18. Later studies showed that people's stories were mediators in their decision-making, and not merely post hoc rationalizations of their verdicts (e.g., Pennington & Hastie, 1992).
19. Pennington & Hastie (1988) acknowledge some of these shortcomings and propose a computational model based on connectionist models of explanatory coherence (Thagard, 2000). We discuss coherence-based models later; while such models address how multiple constaints can be satisfied, it is unclear how purely associative representations can capture certain aspects of causal reasoning (Waldmann, Hagmayer, & Blaisdell, 2006).
20. Again, Pennington and Hastie (1988) acknowledge this shortcoming—and indeed in an early paper (Pennington & Hastie, 1981) discuss various aspects of witness credibility and reliability.
21. For introductions to Bayesian networks, see Fenton and Neil (2012); Taroni et al. (2006).
22. Note that Pennington and Hastie (1981) present a Wigmore chart of this issue that maps closely onto our Bayesian network analysis. However, their subsequent work does not develop this approach, but concentrates on the story structures themselves.
23. It's not always clear how the valence of these links is determined—presumably through prior knowledge or learning.
24. This argument is not conclusive: it depends on how rich the Bayesian modeling is. Coherence theorists refer to a simplistic version restricted to Bayes rule, but broader Bayesian approaches can accommodate some form of bi-directional reasoning (see later in this chapter and Jern et al., 2014).
25. The authors state that they are operating with an informal notion of sufficiency—as used in everyday discourse—rather than a logical or technical notion. On this reading, an action (e.g., getting Joe to ingest the poisonous pill) can be sufficient for an outcome (e.g., killing Joe), even if that exact outcome does not occur (e.g., because Joe dies in an accident before the pill poisons him). Although they do not express it in such terms, this is akin to a counterfactual notion of sufficiency (i.e., the poison would have killed Joe if he hadn't died in an accident first).
26. Note that participants made both cause and blame judgments, and were explicitly instructed that these two might dissociate. This was illustated with the example of a child who accidently shoots one of his parents (unfortunately not as far-fetched an example as it seems).
27. Alternatively, one can also try to combine the best of both worlds. There are several different dual-systems frameworks which state that moral judgments are produced by qualitatively different, and potentially conflicting, cognitive systems (e.g., an emotional and a deliberate system, Greene et al., 2001; or systems that assign value directly to actions versus the outcomes that ultimately result from these actions, Crockett, 2013; Cushman, 2013).

## References

Ajzen, I., & Fishbein, M. (1975). A Bayesian analysis of attribution processes. *Psychological Bulletin*, *82*(2), 261–277.

Alicke, M. D. (1992). Culpable causation. *Journal of Personality and Social Psychology*, *63*(3), 368–378.

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin*, *126*(4), 556–574.

Alicke, M. D., Mandel, D. R., Hilton, D., Gerstenberg, T., & Lagnado, D. A. (2015). Causal conceptions in social explanation and moral evaluation: A historical tour. *10*(6), 790–812.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, 121(1):154–161.

Ben-Porath, E., & Dekel, E. (1992). Signaling future actions and the potential for sacrifice. *Journal of Economic Theory*, *57*(1), 36–51.

Bennett, W. L., & Feldman, M. (1981). *Reconstructing reality in the courtroom.* New Brunswick, NJ: Rutgers University Press.

Blanchard, T., & Schaffer, J. (2016). Cause without default. In H. Beebee, C. Hitchcock & H. Price (Eds.), *Making a difference*. Oxford: Oxford University Press.

Bottom, W. P., Gibson, K., Daniels, S. E., & Murnighan, J. K. (2002). When talk is not cheap: Substantive penance and expressions of intent in rebuilding cooperation. *Organization Science*, *13*(5), 497–513.

Bratman, M. (1987). *Intention, plans, and practical reason*. Center for the Study of Language and Information.

Chakroff, A., Dungan, J., & Young, L. (2013). Harming ourselves and defiling others: What determines a moral domain? *PLoS ONE*, *8*(9), e74434.

Charness, G., & Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, *74*(6), 1579–1601.

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Clarke, R. (1994). Ability and responsibility for omissions. *Philosophical Studies*, *73*(2), 195–208.

Connor De Sai, S., Reimers, S. & Lagnado, D. A. (2016). Consistency and credibility in legal reasoning: A Bayesian network approach. In Papafragou, A., Grodner, D., Mirman, D., & Trueswell, J.C. (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 626–631). Austin, TX: Cognitive Science Society

Critcher, C. R., Inbar, Y., & Pizarro, D. A. (2012). How quick decisions illuminate moral character. *Social Psychological and Personality Science*, *4*(3), 308–315.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–366.

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. (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–292.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a "trembling hand" game. *PloS One*, *4*(8), e6699.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Darley, J. M., & Latané, B. (1968). Bystander intervention in emergencies: diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*(4), 377–383.

Darwall, S. L. (Ed.) (2003a). *Consequentialism*. Oxford: Blackwell.

Darwall, S. L. (Ed.) (2003b). *Deontology*. Oxford: Blackwell.

Darwall, S. L. (Ed.) (2003c). *Virtue ethics*. Oxford: Blackwell.

Dawid, A. P., & Evett, I. W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Science*, *42*, 226–31.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Ditto, P. H., & Jemmott, J. B. (1989). From rarity to evaluative extremity: Effects of prevalence information on evaluations of positive and negative characteristics. *Journal of Personality and Social Psychology*, *57*, 16–26.

Dowe, P. (2000). *Physical causation*. Cambridge: Cambridge University Press.

Driver, J. (2008). Attributions of causation and moral responsibility. In Sinnott-Armstrong, W. (Ed.), *Moral psychology: The cognitive science of morality: Intuition and diversity*, Vol. 2. Cambridge, MA: MIT Press.

Dupoux, E., & Jacob, P. (2007). Universal moral grammar: A critical appraisal. *Trends in Cognitive Sciences*, *11*(9), 373–378.

Feigenson, N. R. (1996). The rhetoric of torts: How advocates help jurors think about causation, reasonableness and responsibility. *Hastings Law Journal*, *47*, 61–165.

Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks.* Boca Raton, FL: CRC Press.

AQ: Please add place of publication.

Fenton, N., Neil, M., & Lagnado, D. A. (2013). A general structure for legal arguments about evidence using Bayesian networks. *Cognitive Science*, 37, 61–102.

Fincham, F. D., & Jaspars, J. M. (1980). Attribution of responsibility: From man the scientist to man as lawyer. In Berkowitz, L. (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 81–138). New York: Academic Press.

Fincham, F. D., & Jaspars, J. M. (1983). A subjective probability approach to responsibility attribution. *British Journal of Social Psychology*, *22*(2), 145–161.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology*, *38*(6), 889–906.

Fleischhut, N. (2013). Moral judgment and decision making under uncertainty. Unpublished PhD thesis.

AQ: Please add name of institution.

Foot, P. (1967). The problem of abortion and the doctrine of the double effect. *Oxford Review*, *5*, 4–15.

Fumerton, R., & Kress, K. (2001). Causation and the law: Preemption, lawful sufficency and causal sufficiency. *Law and Contemporary Problems*, *64*, 83–105.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, *7*(2), 155–170.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., Halpern, J. Y., & Tenenbaum, J. B. (2015). Responsibility judgments in voting scenarios. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 788–793). Austin, TX: Cognitive Science Society.

Gerstenberg, T., & Lagnado, D. A. (2012). When contributions make a difference: Explaining order effects in responsibility attributions. *Psychonomic Bulletin & Review*, *19*(4), 729–736.

Gerstenberg, T., Lagnado, D. A., & Kareev, Y. (2010). The dice are cast: The role of intended versus actual contributions in responsibility attribution. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 1697–1702). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Ullman, T. D., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2014). Wins above replacement: Responsibility attributions as counterfactual replacements. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the Cognitive Science Society* (pp. 2263–2268). Austin, TX: Cognitive Science Society.

Gintis, H., Henrich, J., Bowles, S., Boyd, R., & Fehr, E. (2008). Strong reciprocity and the roots of human morality. *Social Justice Research*, *21*(2), 241–253.

Gintis, H., Smith, E. A., & Bowles, S. (2001). Costly signaling and cooperation. *Journal of Theoretical Biology*, *213*(1), 103–119.

Glöckner, A., & Engel C. (2013). Can we trust intuitive jurors? Standards of proof and the probative value of evidence in coherence based reasoning. *Journal of Empirical Legal Studies*, *10*, 230–252.

Goldman, A. I. (1999). Why citizens should vote: A causal responsibility approach. *Social Philosophy and Policy*, *16*(2), 201–217.

Goodman, N. D., Tenenbaum, J. B., & Gerstenberg, T. (2015). Concepts in a probabilistic language of thought. In E. Margolis & S. Lawrence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 623–653). Cambridge, MA: MIT Press.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*(1), 148–168.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1–2), 145–171.

Green, E. (1967). The reasonable man: Legal fiction or psychosocial reality? *Law & Society Review*, *2*, 241–258.

Green, S. (2015). *Causation in negligence*. Oxford: Hart.

Greene, E. J., & Darley, J. M. (1998). Effects of necessary, sufficient, and indirect causation on judgments of criminal liability. *Law and Human Behavior*, *22*(4), 429–451.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences*, *6*(12), 517–523.

Griffiths, T. L., & Tenenbaum, J. B. 2009. Theory-based causal induction. *Psychological Review*, *116*(4), 661–716

Guglielmo, S., & Malle, B. F. (2010). Can unintended side effects be intentional? Resolving a controversy over intentionality and morality. *Personality and Social Psychology Bulletin*, *36*(12), 1635–1647.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Psychology*, *52*(5), 449–466.

AQ: Please add volume number.

Hagmayer, Y., & Osman, M. (2012). From colliding billiard balls to colluding desperate housewives: Causal bayes nets as rational models of everyday causal reasoning. *Synthese*, ~~1–12.~~

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review*, *108*(4), 814–834.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, *65*(4), 613–628.

Haidt, J., Rozin, P., McCauley, C., & Imada, S. (1997). Body, psyche, and culture: The relationship between disgust and morality. *Psychology & Developing Societies*, *9*(1), 107–131.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, *132*, 109–136.

Halpern, J. Y., & Hitchcock, C. (2015). Graded causation and defaults. *British Journal for the Philosophy of Science*, *66*, 413–457.

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Hamilton, V. L. (1978). Who is responsible? Toward a social psychology of responsibility attribution. *Social Psychology*, *41*(4), 316–328.

Hamilton, V. L. (1980). Intuitive psychologist or intuitive lawyer? Alternative models of the attribution process. *Journal of Personality and Social Psychology*, *39*(5), 767–772.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, *450*(7169), 557–559.

Hart, H. L. A., & Honoré, T. (1959/1985). *Causation in the law*. Oxford: Oxford University Press.

Hastie, R. (1999). The role of stories in civil jury judgments. *University of Michigan Journal of Law Reform*, *32*, 227–239.

Heider, F. 1958. *The psychology of interpersonal relations*. New York: John Wiley & Sons.

Heller, K. (2006). The cognitive psychology of circumstantial evidence. *Michigan Law Review*, *105*, 243–305.

Hepler, A. B., Dawid, A. P., & Leucari, V. (2007). Object-oriented graphical representations of complex patterns of evidence. *Law, Probability & Risk*, *6*, 275–293.

Herring, J. (2010). *Criminal law: Texts, cases, and materials* (4th ed.). Oxford: Oxford Univeristy Press.

Hilton, D. J., McClure, J., & Sutton, R. M. (2010). Selecting explanations from causal chains: Do statistical principles explain preferences for voluntary causes. *European Journal of Social Psychology*, *40*(3), 383–400.

Hitchcock, C. (2012). Portable causal dependence: A tale of consilience. *Philosophy of Science*, *79*(5), 942–951.

Hoffman, Rt. Hon Lord (2011). Causation. In R. Goldberg (Ed.), *Perspectives on causation* (pp. 3–9). Oxford; Portland, OR: Hart.

Hoffman, M., Yoeli, E., & Nowak, M. A. (2015). Cooperate without looking: Why we care what people think and not just what they do. *Proceedings of the National Academy of Sciences*, *112*(6), 1727–1732.

Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief adjustment model. *Cognitive Psychology*, *24*, 1–55.

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and category-based inference: A theoretical integration with Bayesian causal models. *Journal of Experimental Psychology: General*, *139*(4), 702–727.

Iliev, R. I., Sachdeva, S., & Medin, D. L. (2012). Moral kinematics: The role of physical factors in moral judgments. *Memory & Cognition*, *40*(8), 1387–1401.

Jara-Ettinger, J., Kim, N., Muentener, P., & Schulz, L. E. (2014). Running to do evil: Costs incurred by perpetrators affect moral judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 684–688). Austin, TX: Cognitive Science Society.

Jara-Ettinger, J., Tenenbaum, J. B., & Schulz, L. E. (2013). Not so innocent: Reasoning about costs, competence, and culpability in very early childhood. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 663–668). Austin, TX: Cognitive Science Society.

Jensen, S. (2013). A statistician reads the sports pages: Salaries and wins in baseball. *CHANCE*, *26*(1), 47–52.

Jern, A., Chang, K. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, *121*(2), 206–224.

Johnson, S. G., & Rips, L. J. (2015). Do the right thing: The assumption of optimality in lay decision theory and causal judgment. *Cognitive Psychology*, *77*, 42–76.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136–153.

Kamm, F. M. (2007). *Intricate ethics*. Oxford: Oxford University Press.

Kant, I. (1796/2002). *Groundworks for the metaphysics of morals*. New Haven, CT; London: Yale University Press.

Kelley, H. H., & Stahelski, A. J. (1970). The inference of intentions from moves in the prisoner's dilemma game. *Journal of Experimental Social Psychology*, *6*(4), 401–419.

Kerr, N. L. (1996). "Does my contribution really matter?": Efficacy in social dilemmas. *European Review of Social Psychology*, *7*(1), 209–240.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle, Dale, R. and Warlaumont, A. S. and Yoshimi, J. and Matlock, T., Jennings and C. D. and Maglio, P. P.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957.

Knobe, J. (2009). Folk judgments of causation. *Studies in the History and Philosophy of Science*, *40*(2), 238–242.

Knobe, J., & Fraser, B. (2008). Causal judgment and moral judgment: Two experiments. In W. Sinnott-Armstrong (Ed.), *Moral psychology* (pp. 441–447). Cambridge, MA: MIT Press.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature*, *446*(7138), 908–911.

Kortenkamp, K. V., & Moore, C. F. (2014). Ethics under uncertainty: The morality and appropriateness of utilitarianism when outcomes are uncertain. *The American Journal of Psychology*, *127*(3), 367–382.

Kuhn, D., Weinstock, M., & Flaton, R. (1994). How well do jurors reason? Competence dimensions of individual variation in a juror reasoning task. *Psychological Science*, *5*, 289–296.

Lagnado, D. A. (2011). Causal thinking. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 129–149). Oxford: Oxford University Press.

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Lagnado, D. A., Fenton, N., & Neil, M. (2013). Legal idioms: A framework for evidential reasoning. *Argument and Computation*, *4*, 46–53.

Lagnado, D. A., & Harvey, N. (2008). The impact of discredited evidence. *Psychonomic Bulletin & Review*, *15*(6), 1166–1173.

Latané, B. (1981). The psychology of social impact. *American Psychologist*, *36*(4), 343–356.

Lewis, D. (1986). Causal explanation. *Philosophical Papers*, *2*, 214–240.

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229.

Lloyd-Bostock, S. (1979). The ordinary man, and the psychology of attributing causes and responsibility. *The Modern Law Review*, *42*(2), 143–168.

Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry*, *25*(2), 147–186.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101–121.

Mandel, D. R. (2011). Mental simulation and the nexus of causal and counterfactual explanation. In C. Hoerl, T. McCormack, & S. R. Beck (Eds.), *Understanding counterfactuals, understanding causation: Issues in philosophy and psychology* (pp. 146–170). Oxford: Oxford University Press.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Cognitive systems for revenge and forgiveness. *Behavioral and Brain Sciences*, *36*(1), 1–15.

McKenzie, C. R., & Mikkelsen, L. A. (2007). A Bayesian view of covariation assessment. *Cognitive Psychology*, *54*(1), 33–61.

Mikhail, J. (2007). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152.

Mikhail, J. (2009). Moral grammar and intuitive jurisprudence: A formal model of unconscious moral and legal knowledge. *Psychology of Learning and Motivation*, *50*, 27–100.

Moore, M. S. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Morse, S. J. (2003). Diminished rationality, diminished responsibility. *Ohio State Journal of Criminal Law*, *1*, 289–308.

Nadler, J. (2012). Blaming as a social process: The influence of character and moral emotion on blame. *Law & Contemporary Problems*, *2*, 1–31.

Nadler, J., & McDonnell, M.-H. (2011). Moral character, motive, and the psychology of blame. *Cornell Law Review*, 97.

Nanay, B. (2010). Morality or modality? What does the attribution of intentionality depend on? *Canadian Journal of Philosophy*, *40*(1), 25–39.

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*(2), 114–123.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford: Oxford University Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Francisco: Morgan Kaufmann.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.

Pennington, N., & Hastie, R. (1981). Juror decision making models: The generalization gap. *Psychological Bulletin*, *89*, 246–287.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: The effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 521–533.

Pennington, N. & Hastie, R. (1992). Explaining the evidence: Tests of the story model for juror decision making. *Journal of Personality and Social Psychology*, *62*(2), 189–206.

Phillips, J., Luguri, J., & Knobe, J. (2015). Unifying morality's influence on non-moral judgments: The relevance of alternative possibilities. *Cognition*, *145*, 30–42.

Pizarro, D. A., & Tannenbaum, D. (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. In M. Mikulincer, & P. R. Shaver

(Eds.), *The social psychology of morality: Exploring the causes of good and evil* (pp. 91–108). Washington, DC: APA Press.

Pizarro, D. A., Uhlmann, E., & Salovey, P. (2003). Asymmetry in judgments of moral blame and praise: The role of perceived metadesires. *Psychological Science*, *14*(3), 267–72.

AQ: Please add volume number.

Quinn, W. S. (1989). Actions, intentions, and consequences: The doctrine of double effect. *Philosophy & Public Affairs*, 334–351.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review*, *118*(1), 57–75.

Reeder, G. D., & Brewer, M. B. (1979). A schematic model of dispositional attribution in interpersonal perception. *Psychological Review*, *86*(1), 61.

Reeder, G. D., & Spores, J. M. (1983). The attribution of morality. *Journal of Personality and Social Psychology*, *44*(4), 736–745.

Reeder, G. D., Vonk, R., Ronk, M. J., Ham, J., & Lawrence, M. (2004). Dispositional attribution: Multiple inferences about motive-related traits. *Journal of Personality and Social Psychology*, *86*(4), 530–544.

Rezlescu, C., Duchaine, B., Olivola, C. Y., & Chater, N. (2012). Unfakeable facial configurations affect strategic choices in trust games with or without information about past behavior. *PLoS ONE*, *7*(3), e34293.

Roese, N. J. (1997). Counterfactual thinking. *Psychological Bulletin*, *121*(1), 133–148.

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1: *Foundations*. Cambridge, MA: MIT Press.

Samland, J., & Waldmann, M. R. (2015). Highlighting the causal meaning of causal test questions in contexts of norm violations. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 2092–2097). Austin, TX: Cognitive Science Society.

Scanlon, T. M. (2009). *Moral dimensions*. Cambridge, MA: Harvard University Press.

Schächtele, S., Gerstenberg, T., & Lagnado, D. A. (2011). Beyond outcomes: The influence of intentions and deception. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd annual conference of the Cognitive Science Society* (pp. 1860–1865). Austin, TX: Cognitive Science Society.

Schaffer, J. (2010). Contrastive causation in the law. *Legal Theory*, *16*(4), 259–297.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Schlenker, B. R., Britt, T. W., Pennington, J., Murphy, R., & Doherty, K. (1994). The triangle model of responsibility. *Psychological Review*, *101*(4), 632–652.

Schum, D. A. (1994). *The evidential foundations of probabilistic reasoning*. Evanston, IL: Northwestern University Press.

Schwitzgebel, E., & Cushman, F. (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, *27*(2), 135–153.

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

Simon, D., Snow, C. & Read, S. J. (2004). The redux of cognitive consistency theories: evidence judgments by constraint satisfaction. *Journal of Personality and Social Psychology*, 86, 814–837.

Simon, D., & Holyoak, K. J. (2002). Structural dynamics of cognition: from consistency theories to constraint satisfaction. *Personality and Social Psychology Review*, 6, 283–94.

Sloman, S. A. (2009). *Causal models: How people think about the world and its alternatives*. New York: Oxford University Press.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2009). Causal models: The representational infrastructure for moral judgment. In D. M. Bartels, C. Bauman, L. J. Skitka, & D. L. Medin (Eds.), *Moral judgment and decision making. The psychology of learning and motivation: Advances in research and theory* (pp. 1–26). Amsterdam: Elsevier.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind and Language*, *27*(2), 154–180.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247.

Smart, J. J. C., & Williams, B. (1973). *Utilitarianism: For and against*. Cambridge: Cambridge University Press.

Snyder, M. L., Kleck, R. E., Strenta, A., & Mentzer, S. J. (1979). Avoidance of the handicapped: An attributional ambiguity analysis. *Journal of Personality and Social Psychology*, *37*(12), 2297–2306.

Spellman, B. A., & Kincannon, A. (2001). The relation between counterfactual ("but for") and causal reasoning: Experimental findings and implications for jurors' decisions. *Law and Contemporary Problems*, *64*(4), 241–264.

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology*, *48*(1), 232–238.

Sripada, C. S., & Stich, S. (2006). A framework fro the psychology of norms. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (pp. 280–301). New York: Oxford University Press.

Stapleton, J. (2008). Choosing what we mean by "causation" in the law. *Missouri Law Review*, *73*(2), 433–480.

Stapleton, J. (2009). Causation in the law. In H. Beebee, P. Menzies, C. Hitchcock (Eds.), *The Oxford handbook of causation* (pp. 744–769). Oxford: Oxford University Press.

Tadros, V. (2007). *Criminal responsibility*. Oxford: Oxford University Press.

Tannenbaum, D., Uhlmann, E. L., & Diermeier, D. (2011). Moral signals, public outrage, and immaterial harms. *Journal of Experimental Social Psychology*, *47*(6), 1249–1254.

Taroni, F., Aitken, C., Garbolino, P., & Biedermann, A. (2006). *Bayesian networks and probabilistic inference in forensic science*. Chichester, UK: John Wiley & Sons.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853–870.

Tetlock, P. E., Visser, P. S., Singh, R., Polifroni, M., Scott, A., Elson, S. B., Mazzocco, P., & Rescober, P. (2007). People as intuitive prosecutors: The impact of social-control goals on attributions of responsibility. *Journal of Experimental Social Psychology*, *43*(2), 195–209.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist*, *59*(2), 204–217.

Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal*, *94*(6), 1395.

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*(12), 455–460.

Trope, Y. (1986). Identification and inferential processes in dispositional attribution. *Psychological Review*, *93*(3), 239–257.

Turri, J., & Blouw, P. (2014). Excuse validation: A study in rule-breaking. *Philosophical Studies*, *172*(3), 615–634.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science*, *10*(1), 72–81.

Uhlmann, E. L., & Zhu, L. L. (2013). Acts, persons, and intuitions: Person-centered cues and gut reactions to harmless transgressions. *Social Psychological and Personality Science*, *5*(3), 279–285.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition*, *126*(2), 326–334.

AQ: Please add publication details.

Ullman, T. D., Tenenbaum, J. B., Baker, C. L., Macindoe, O., Evans, O. R., & Goodman, N. D. (2009). Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, Vol. 22.

Uttich, K. & Lombrozo, T. (2010). Norms inform mental state ascriptions: A rational explanation for the side-effect effect. *Cognition*, *116*(1), 87–100.

Uustalu, O. (2013). The role of spontaneous evaluations, counterfactual thinking, and expert testimony in causal and blame Judgments. Unpublished MSc thesis, University College London.

van Inwagen, P. (1978). Ability and responsibility. *The Philosophical Review*, *87*(2), 201–224.

Waldmann, M. R., Hagmayer, Y, & Blaisdell, A. P. (2006). Beyond the information given: Causal models in learning and reasoning. *Current Directions in Psychological Science*, *15*(6), 307–311.

Waldmann, M. R., & Dieterich, J. H. (2007). Throwing a bomb on a person versus throwing a person on a bomb: Intervention myopia in moral intuitions. *Psychological Science*, *18*(3), 247–253.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In *The Oxford handbook of thinking and reasoning* (pp. 364–389). New York: Oxford University Press.

Waldmann, M. R., & Wiegmann, A. (2010). A double causal contrast theory of moral intuitions in trolley dilemmas. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 2589–2594). Austin, TX: Cognitive Science Society.

Wallez, C., & Hilton, D. (2015). Unpublished data.

Wellman, H. M. & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43(1):337–375.

Wiegmann, A., & Waldmann, M. R. (2014). Transfer effects between moral dilemmas: A causal model theory. *Cognition*, *131*(1), 28–43.

Wigmore, J. H. (1913). The problem of proof. *Illinois Law Journal*, 8(2), 77–103. Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin*, *24*(12), 1251–1263.

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*(2), 283–301.

Yoshida, W., Dolan, R. J., & Friston, K. J. (2008). Game theory of mind. *PLoS Computational Biology*, *4*(12), e1000254.

Young, L., & Saxe, R. (2011). When ignorance is no excuse: Different roles for intent across moral domains. *Cognition*, *120*(2), 202–214.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.