MORAL INFERENCES

Edited by
Jean-François Bonnefon and Bastien Trémolière

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

3

CAUSAL MODELS MEDIATE MORAL INFERENCES

Michael R. Waldmann, Alex Wiegmann, and Jonas Nagel

Abstract

Most theories of moral judgments distinguish between acts and outcomes. According to these theories, moral judgments are either primarily based on the evaluation of the acts or the outcomes with multi-system theories allowing for both possibilities. Here we argue that it is not only the acts and outcomes that determine moral evaluations but also the causal relations linking the acts with their outcomes. Causal relations influence moral judgments by shifting attention to aspects of inter-victim relations. We report three projects that demonstrate the usefulness of this framework in tasks that range from moral judgments about trolley problems to basic force-dynamic interpretations of simple perceptual and linguistic scenes.

Introduction

Our central claim in this chapter is that causal model representations play a crucial part in moral judgments. Moral judgments can be very diverse (Haidt & Joseph, 2007); here we will focus on the large class of situations in which potential victims are, without having given their consent, being harmed. A central question of normative moral theories is to specify the boundary conditions that allow agents to harm other people in such situations and when they are prohibited to do so. Why causal representations should mediate moral judgments is not immediately obvious. The goal of the chapter is to empirically demonstrate the role of causal models and offer explanations for their important role in moral judgments.

The distinction between acts and outcomes underlies many theories of moral reasoning both in philosophy and psychology (see Waldmann, Nagel, & Wiegmann, 2012, for an overview). Deontological approaches tend to focus on the permissibility of acts whereas utilitarianism bases judgments on the utility of the outcomes. In psychology, dual-system theories have been proposed that combine

both approaches by postulating separate systems for the two types of judgments. Which of the two systems is activated depends on features of the moral scenarios (e.g., Cushman, 2013; Greene, Sommerville, Nystrom, Darley, & Cohen, 2001).

In our view, these views are too simple to capture moral intuitions. We do not only evaluate acts and outcomes, but are also sensitive to the causal mechanisms that link them. Not only is the fact that an action harmed a victim relevant, but also how exactly the action was causally related to the resulting harm. This has been first demonstrated by philosophers interested in the famous trolley problem (see Foot, 1967; Kamm, 2007, 2015; Thomson, 1985). Trolley scenarios describe a runaway train whose brakes are defect. The train is about to run over, for example, five workmen standing on the track. However, there is an option that allows an agent to change the situation in a way that one worker instead of the five is harmed. Interestingly, in such cases the causal structure of the situation alters intuitions about moral permissibility. In the Switch variant of the trolley problem, the runaway train can be re-directed from the five to a single victim. Most subjects feel that this act is permissible. However, in a different causal setup the only way to stop the train is to push a person off a bridge (Push dilemma). Although here the dilemma also implies a tradeoff between five and one victim, subjects disapprove of the act. We will discuss examples below that show that even when, unlike in Switch and Push, the proximal acts and outcomes are kept constant, moral intuitions are sensitive to the causal mechanisms underlying the dilemma.

Given that the causal structure linking acts and outcomes matters morally, the question is how causal representations trigger moral intuitions. Moreover, it is interesting to reflect on why causal models should be morally relevant. Different possibilities could underlie this effect. Agent-based accounts focus on the person committing the act that triggers a chain of causally connected events. According to this perspective, causal structures provide information about the intentions and beliefs of the agent that are morally relevant. A different possibility focuses on the victims. In moral dilemmas the rights of the victims are about to be breached. For example, in the trolley dilemma one could argue that people have a right not to be harmed. Thus, a possible reason for the importance of causal models may be that they specify variants of precisely how victims are harmed and how the rights of them relative to each other are affected by the action under consideration. We argue that this latter perspective accounts best for the available data on moral judgments.

Moral intuitions in trolley dilemmas

Trolley scenarios are the primary examples for the important role of causal models in explaining moral judgments. Initially, trolley dilemmas have been discussed by philosophers as demonstrations that our moral intuitions do not consistently follow a utilitarian or an absolutist deontological moral theory. Since this was the main goal, the contrasted dilemmas varied in various features. In the meantime, many of these initially confounded features, including physical distance, acts, victims, test questions and many more, turned out to influence moral intuitions (see Waldmann

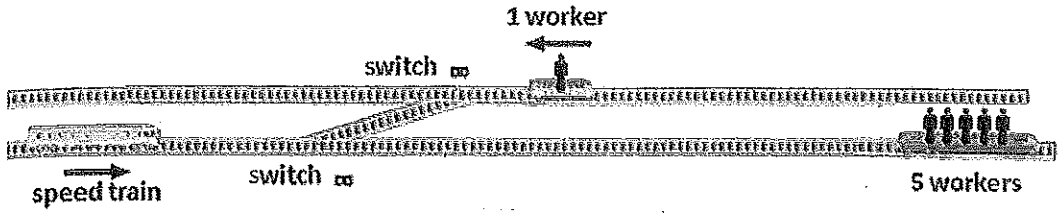

FIGURE 3.1 Threat vs. victim intervention.

et al., 2012, for a review). Our goal here is not to present a complete theory of moral intuitions in the trolley dilemma, but to focus on the role of causal models.

Figure 3.1 provides a graphic depiction of our paradigm. In both variants of the trolley dilemma the trains are steered by employees of the train company by remote control. The trains are cleaning the tracks. In the *threat intervention condition* (a variant of the Switch dilemma), five track workers sit on the large train (located on the bottom track in Figure 3.1) and one on the other train (on the upper track) which is going in the opposite direction (see arrow). The speed train on the left side threatening the lives of the workers is empty. Due to a signaling defect it cannot be stopped. Soon it would hit the train with the five workers. However, an employee in the control room could throw the switch and redirect the speed train on the parallel track where it would hit the train carrying the one worker. In either case the worker(s) hit by the speed train would be killed. Thus, throwing the switch would cause the death of one person, whereas refraining from acting leads to the death of five.

We contrasted this case with the *victim intervention condition* (a variant of the Push dilemma) in which the first part of the story is identical. Here the employee could throw the switch and thereby redirect the train carrying the one worker from the parallel upper track onto the main track. Thereby this train would collide with the speed train and stop it before the speed train reaches the train with the five workers. Again the victims involved in the collisions would be killed.

Figure 3.2 shows the results of the experiment in which we either requested subjects to provide a moral judgment or asked them about the intention of the agent. On the left side, the mean responses to the question whether the employee should do the proposed action using a rating scale ranging from 1 ("certainly no") to 6 ("certainly yes") is shown. One hundred and thirty-nine subjects responded to the moral question. The manipulation yielded a significant difference in the moral ratings, $t(137) = 2.06$, $p < .05$. The effect is smaller than in some other studies because we controlled for the confounds that all contributed to an effect between Switch and Push in previous studies.

In general, the results show that subjects are not solely sensitive to the numeric tradeoff between victims or to the qualities of the concrete proposed proximal action (i.e., throwing a switch), both of which are identical in the two conditions, but also to *how* the (invariant) act of switching is causally related to the (invariant) resulting harm. Thus, it provides *prima facie* evidence for the role of causal models, and contradicts utilitarian theories that simply focus on lives as well as accounts

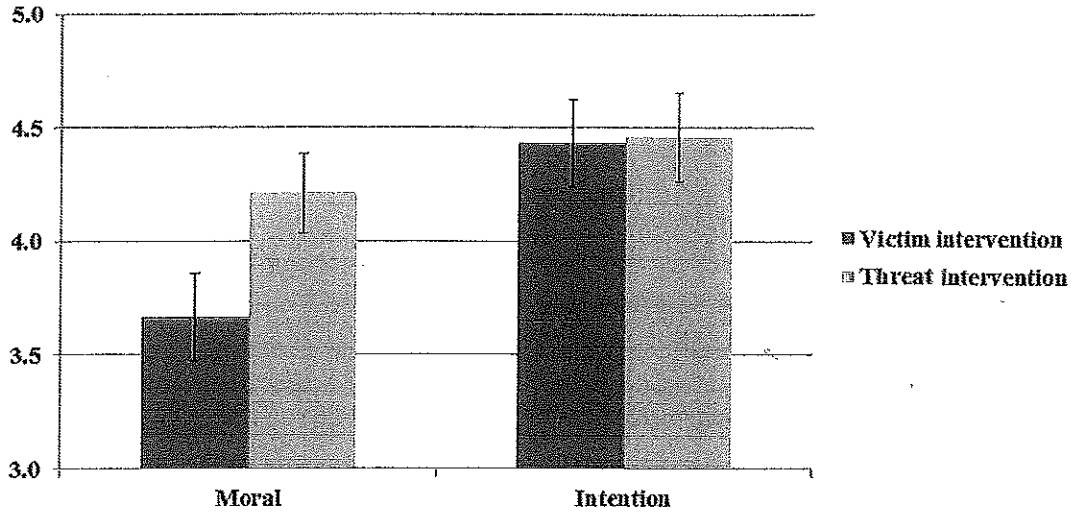

FIGURE 3.2 Results of trolley experiment. Error bars represent standard error of means.

that focus on the quality of the action itself (e.g., throwing a switch vs. pushing a person). The interesting question is to find out how causal relations mediate moral judgments and why.

One possibility, proposed by a number of researchers, attributes differences between Switch and Push to differences in personal proximity between agent and victim ("up close and personal"; see Cushman, 2013; Greene et al., 2001; Greene, Cushman, Stewart, Lowenberg, Nystrom, & Cohen, 2009). This factor certainly makes a difference in scenarios in which it varies. Actually, in recent research focusing on the role of spatial distance in helping scenarios we have shown that it is not pure spatial distance but correlated factors that affect moral judgments (Nagel & Waldmann, 2013, 2016). The present experiment, however, controls for this factor so that this theory does not explain the obtained effect. In both conditions of our experiment the agent throws the switch located in the distant control center.

A second possible causal account explaining the effect uses causal structures to justify different attributions about the agent's intentions. A popular theory is inspired by the *doctrine of double effect* (DDE) that can be traced back to Aquinas (see Mikhail, 2011; Timmons, 2002). According to the DDE, it is wrong to intend harm but permissible to foresee harm if the harmful act serves a greater good and if the action itself is not wrong. The causal structure serves here as a cue to intentionality. According to the DDE, ends (e.g., saving lives) or causal means that bring about the end are *intended*, whereas side effects of the action may just be *foreseen*. Thus, according to the DDE, the threat intervention (or the action in Switch) is permitted because the death of the one victim is merely a foreseen side effect, whereas the action in Push or the victim intervention condition is prohibited because here the single victim is intentionally harmed as a means to stop the threat.

Although there is evidence that intentionality matters when judging whether a person should be punished for an act harming other people (Cushman, 2008), this

factor does not underlie the different moral intuitions in our study. In two separate conditions in which 135 subjects participated, we told subjects they should assume that the employee actually performed the proposed action of throwing the switch, and then asked them whether he "caused the death of the one worker intentionally," again using a rating scale ranging from 1 to 6. Figure 3.2 (right) shows that there was no difference regarding attributions of intentionality across the two conditions. Moreover, the intentionality attributions were between the midline and the ceiling. Apparently subjects saw both scenarios as situations in which the main goal is to save the five with the death of the one employee being subordinated under this goal.

A further variant of the DDE that has been discussed in the philosophical literature focuses on the objective causal relations rather than the agents' intentions (see Kamm, 2007, 2015; Scanlon, 2008). According to the *means principle* that can be traced back to Kant (1785/1959), it is prohibited to use people against their will as means for a greater good. This account could explain the obtained difference between the two scenarios in our experiment if it is assumed that subjects view the single victim as a means in the victim intervention and as a side effect in the threat intervention condition (similarly in Switch vs. Push), and that this causal differentiation is made independently of intentionality attribution.

A different theory that is based on psychological principles of action understanding has been proposed by Waldmann and Wiegmann (2010; see also Waldmann & Dieterich, 2007). According to the *locus of intervention* theory, people tend to focus on the fate of people or relevant objects that their interventions directly target. Although in the trolley dilemmas the actions globally generate a tradeoff between five victims and one victim, the direct targets of intervention differ. In the threat intervention condition, the primary relevant target is the speed train that poses a threat to people. Re-directing this train from five to one is the primary goal of the intervention. Since this intervention highlights the contrast between one versus five dead people, subjects should find the action acceptable. By contrast, in the victim intervention condition the primary relevant target of intervention is the train with its single passenger. According to our theory, this is the primary target because moving the passenger without his consent is the initial morally relevant action and should therefore be evaluated before considering the following causal events. Moving the train with the one passenger highlights the fate of this victim who will be killed by the intervention, but who would stay alive in the absence of this intervention. This contrast places the attentional focus on the act of killing, and should therefore be viewed more negatively. The attentional spotlight does not fully block the more remote goal of saving five from sight, but it shifts the focus to the one victim that is targeted by the intervention.

Both theories, the causal version of the DDE (or means principle) and the locus of intervention theory make similar predictions because they are both sensitive to the causal structure underlying the actions. However, one difference between the theories is that the causal DDE only considers the global causal model of the situation, whereas the locus of intervention theory places a spotlight on the more local causal aspects surrounding the point of intervention. We tested these two theories

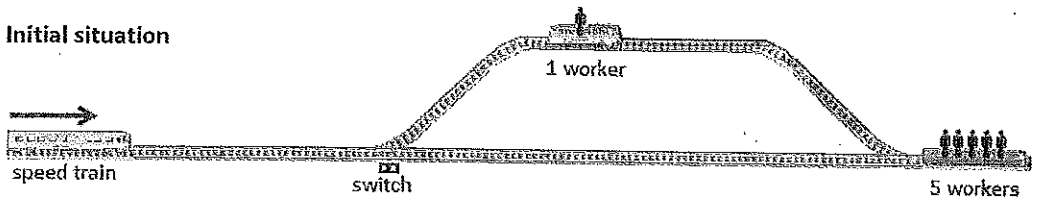

FIGURE 3.3 The loop dilemma.

against each other using Thomson's (1985) famous *loop* version of the trolley dilemma (see Figure 3.3). Here, the side track loops back to the main track which it reaches right before the point where the five victims are located. Thus, the one person being hit on the side track would not only be a side effect of re-directing the speed train in the threat intervention condition (or Switch), it would also play the role of a means to prevent the speed train from coming back to the main track where it would kill the five.

Most published experiments testing the loop case have shown, however, that subjects do not differentiate between the standard Switch (threat intervention) case or related control conditions and the loop condition (e.g., Greene et al., 2009; Liao, Wiegmann, Alexander, & Vong, 2012; Waldmann & Dietrich, 2007; but see Sinnott-Armstrong, Mallon, McCoy, & Hull, 2008).[1] The philosophers Kamm (2015) and Thomson (1985) share the intuitions of the subjects who do not differentiate between the loop case and the Switch condition. The locus of intervention theory predicts the lack of an effect. According to this theory, subjects focus on the direct causal consequences of re-directing the threat to the side track and compare the two alternative outcomes that are expected in the presence versus absence of the act (regardless of whether the side track later loops or not) (see Kamm, 2015, for a different account).

An open question is why moral judgments are sensitive to these causal distinctions. The locus of intervention theory attributes the effect to attentional effects. People tend to focus on the immediate consequences of their actions because they are primarily held accountable for these. Since different interventions highlight the fate of different victims, different moral intuitions arise. However, this explanation may be incomplete because it would apply to all types of victims, including humans, animals, or valuable objects. However, it seems unlikely that subjects would be sensitive to *how* a valuable car would be damaged if this act saved five other cars each having the same value. We actually confirmed this impression in an experiment comparing threat and victim intervention (as in Figure 3.1) in which, instead of using people as victims, we presented tradeoffs between animals (e.g., monkeys, dogs, etc.). Here subjects uniformly favored sacrificing one to save five regardless of how the one was causally harmed.

Therefore, an important boundary condition for the causal effects seems to be that the acts target humans[2] who are considered to be endowed with individual rights which need to be respected, especially the right of not being harmed even

in situations in which greater harm could be prevented. Objects that are treated as valuables do not have such rights unless they are owned by people whose property rights are violated (see Millar, Turri, & Friedman, 2014). Kamm (2015) proposes that in the different trolley dilemmas different inter-victim relations supervene on the different causal relations entailed by the proposed actions. In the Switch scenario, the action of re-directing the threat from five victims to one victim places them in a *substitutive* relation. In the Push case, however, victims are placed in a *subordinative* relation. The right of one person is breached in order to save five others later. This violates our intuitions about human rights despite the fact that inaction prevents a greater good.

In our view, the best psychological account of moral intuitions in trolley dilemmas is provided by a theory that combines assumptions about attentional focus triggered by the locus of intervention combined with consideration of rights of victims that are potentially violated by the actions.

Causal models and transfer of moral intuitions

So far we have shown that attentional focus and, consequently, considerations of inter-victim relations supervening on the attended aspect of the causal model can be shifted by changing the locus of intervention. Another way of manipulating attentional focus is to transfer it to a given dilemma situation from a previous judgment on a structurally analogous judgment problem. In this way, it is possible to demonstrate that putting a different attentional spotlight on one and the same causal structure can lead to differences in moral judgments that are predicted by our account. Wiegmann and Waldmann (2014) worked out in detail under which conditions the attentional focus on a specific causal aspect gets transferred from one moral dilemma to the next, and how this transfer leads to predictable changes in the subsequent judgment.

A particularly interesting case for finding out the relevant boundary conditions are situations in which transfer is asymmetric. A number of researchers have found that trolley dilemmas create such an asymmetry (e.g., Horne, Powell, & Spino, 2013; Lanteri, Chelini, & Rizello, 2008; Lombrozo, 2009; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014; Petrinovich & O'Neill, 1996; Schwitzgebel & Cushman, 2012; Wiegmann, Okan, & Nagel, 2012). Whereas moral intuitions about the Push dilemma stay invariant regardless of its position in a series of moral dilemmas presented to subjects, intuitions about Switch are more malleable. While the act proposed in Switch is considered acceptable by most subjects, its acceptability goes down if it is evaluated subsequent to having responded to the Push dilemma. Figure 3.4 shows the results of an experiment by Wiegmann and Waldmann (2014) demonstrating this asymmetry.

The interesting question is how this asymmetry can be explained. Wiegmann and Waldmann (2014) have proposed that transfer is mediated by selective attention triggered by causal model representations. Figure 3.5 provides a graphic representation of the causal models underlying Switch and Push. Let us first

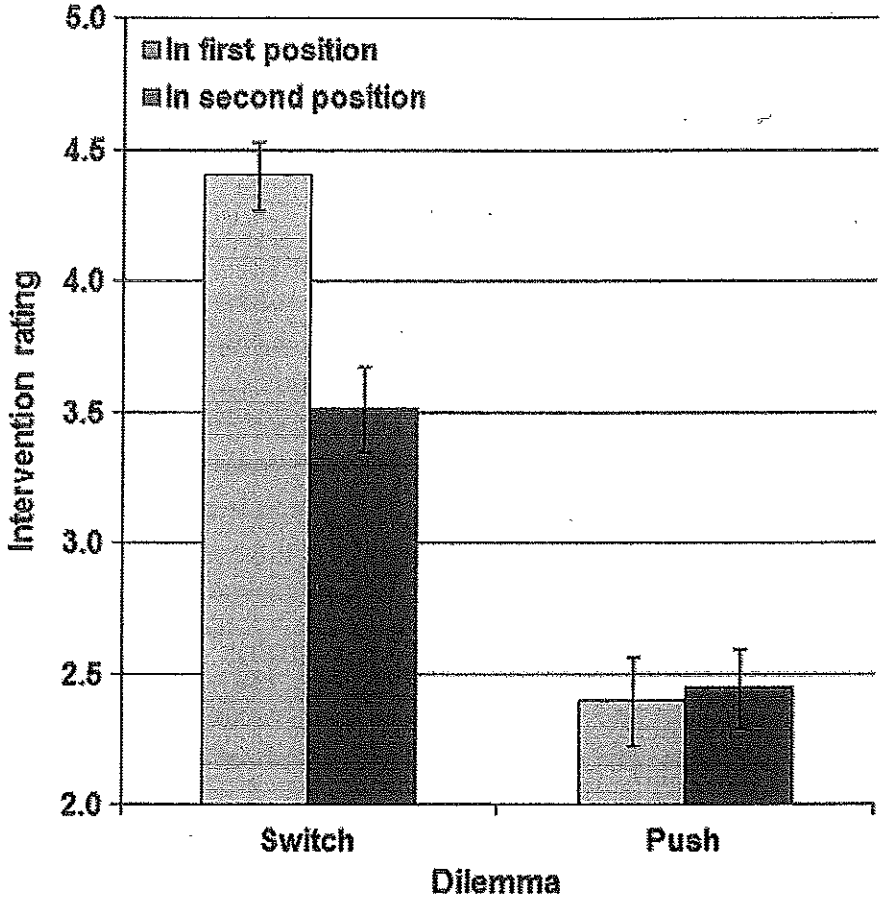

FIGURE 3.4 Ratings for Switch and Push as a function of the order of presentation. Error bars represent standard error of means.[3]

Source: Wiegmann and Waldmann (2014), Experiment 1.

consider the Switch dilemma. As Figure 3.5 (left, top row) shows, there are two causal paths in Switch, one from the intervention to the good outcome (saving five) and one from the intervention to the bad outcome (killing one). Consequently, the bad outcome is not part of the causal path from the intervention to the good outcome. One does not have to intervene on this person in the sense that one does not have to do anything with the one person to save the larger group. One could describe the act of saving the group of people without having to mention the bad outcome ("Three persons were saved by redirecting the trolley"). Moreover, the good outcome is also not part of the causal path from the intervention to the bad outcome. A dilemma with such a causal structure allows us to selectively attend to either the saving or the harming path. Thus, there is some flexibility in how to represent such a dilemma.

In contrast, there are no two independent paths in the causal structure underlying Push (see Figure 3.5, left, bottom row). Here, the bad outcome is part of the causal chain from the intervention to the good outcome. You need to intervene on the one person, that is, push her from the bridge, in order to save the group of five. One cannot describe the intervention of saving the five without

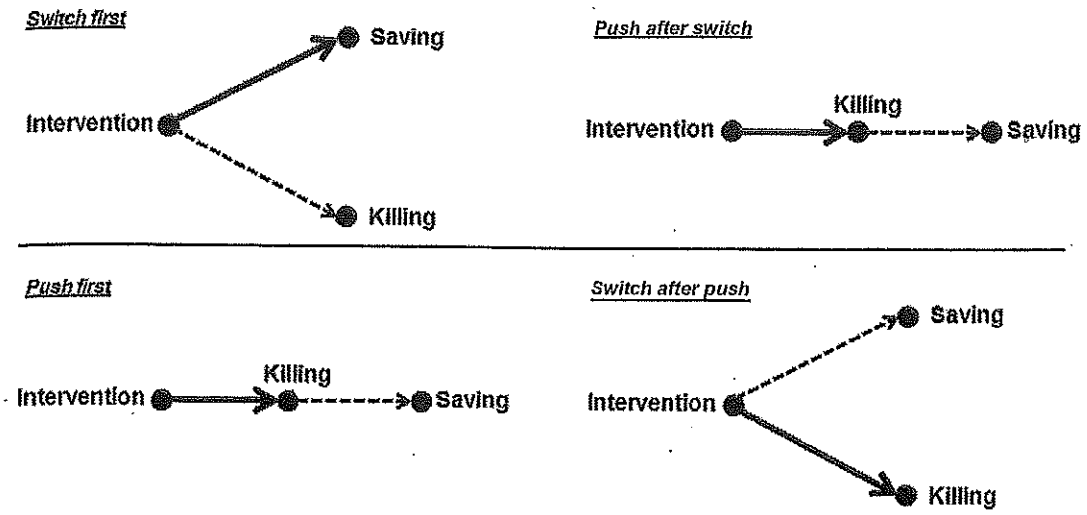

FIGURE 3.5 The highlighted path of the causal structure of Switch and Push in the default evaluation (left side) vs. when preceded by the other dilemma (right side). Bold solid lines represent the highlighted part of the causal structure.[4]

mentioning the one victim on the causal path. The one victim is located in the middle position of a chain where it is used as a means to reach the goal of saving the five. In such chain structures, it is impossible to selectively attend to the causal path from the intervention to the good outcome without simultaneously attending to the bad outcome on the way. One can, however, selectively attend to the causal path from the intervention to the bad outcome. The good outcome is not part of the causal path from the intervention to the bad outcome. One can describe the causal path from the intervention to the bad outcome without having to mention the good outcome ("The heavy person was pushed from a bridge in front of a train and died"). Thus, there is less flexibility regarding how to represent such a dilemma.

One further factor that is needed to explain transfer effects concerns what we called *default evaluation*. As discussed earlier, people have intuitions about moral dilemmas in isolation that can be explained by various theories. Here we do not focus on these theories, but just take it as a given that the action in Switch is considered more acceptable than the action in Push. This empirical finding, which can be explained by the theories presented in the previous section (e.g., locus of intervention theory), is represented here by assuming that by default the causal path from the intervention to the good outcome is highlighted in Switch (bolded arrow in Figure 3.5). If no other dilemma is presented before Switch, the aspect of saving prevails. In contrast, most people disapprove of performing the proposed action in Push. Accordingly, it is assumed that by default the causal path from the intervention to the bad outcome is highlighted (bolded arrow in Figure 3.5), that is, the aspect of killing normally dominates in this dilemma. Again this highlighting of the bad outcome is considered a result of the cognitive processes which we discussed in the previous section as underlying trolley intuitions.

We now have all components needed to explain the asymmetrical transfer effect between Switch and Push. The general assumption is that transfer effects occur when the highlighted causal path from the first dilemma can be mapped onto a similar path of the causal model of the second dilemma. If this is possible, the mapped path in the second dilemma is also highlighted. If highlighting the path in the second dilemma changes the representation of the dilemma, an effect on the moral judgment is to be expected.

For example, if subjects are first presented with Switch, the causal path from the intervention to the good outcome is highlighted by default (see Figure 3.5, left, top row). If subjects then are presented with Push (Figure 3.5, right, top row), the path highlighted by default in the first dilemma (Switch) cannot be analogically mapped onto the causal structure of Push because there is no direct causal path from the intervention to the good outcome in Push that does not include passing the bad outcome. Hence, no transfer is expected in this order of presentation.

Different predictions are entailed for the opposite ordering. When Push is presented first, selective attention highlights the causal path from the intervention to the bad outcome by default (see Figure 3.5, left, bottom row, bolded arrow). If people are then presented with Switch (Figure 3.5, right, bottom row), the highlighted path from the first dilemma can be analogically mapped onto the causal structure of Switch because there is a direct causal path from the intervention to the bad outcome in both dilemmas. Due to this mapping it becomes more likely that the causal path from the intervention to the bad outcome is highlighted in Switch as well. Thus, a lowering of the acceptability of the action in Switch is predicted relative to its evaluation in isolation. This prediction further specifies our theory of the role of inter-victim relations discussed in the previous section. As mentioned there, Switch puts the alternative victims in a substitutive relation, but the relative weights of the aspects of saving and killing can be altered, for example by a preceding dilemma. Thus, acceptability can, within limits, be shifted in Switch by putting more or less weight on the aspect of killing.

In a series of experiments, Wiegmann and Waldmann (2014) have shown that causal model theory can explain a wide range of transfer effects not only in the standard scenarios but also for new cases. Thus, transfer effects cannot only be observed when superficially similar trolley scenarios are paired (ibid., Experiment 1) but also between scenarios coming from other domains. Most interestingly, we also found the predicted asymmetries across different domains which supports the theory that it is the underlying causal structures rather than superficial features that drive transfer (ibid., Experiment 4).

A further test involved pairings of dilemmas in which no transfer effects are predicted. Although Push drags the acceptability of Switch down, no such effect should be expected for a variant of Push (positive Push), which like Push is represented by a chain but in which saving is highlighted by default. In positive Push the intervention saves five by pushing them out of harm's way which, however, has the consequence of killing one victim who is located further downstream on the main track. Pairing these two dilemmas embodying different causal chains

in which the positive and negative outcomes are differently ordered should, according to our theory, not lead to any transfer effects. This prediction was confirmed in Experiment 3 (ibid., 2014).

Experiment 6 (ibid., 2014) provides a more indirect test of our causal model theory. The actions in trolley dilemmas can be generally described as ambiguous. Throwing the switch saves one but kills five people. To measure how subjects represented the actions we asked them with respect to different scenarios whether they consider the action rather a case of killing or of saving. To be able to express the ambiguity, they were presented with a rating scale ranging from 1 ("performing the proposed action is a clear case of killing") to 6 ("performing the proposed action is a clear case of saving"). The effects of selective attention could be replicated with this measure. For example, when Switch or Push are presented in the first position, the action in Switch is more seen as a case of saving and the action in Push as a case of killing, which corresponds to the default evaluations. The assessment of the action in Switch shifted, however, in the direction of killing when preceded by the Push dilemma. No effect on the representation of the action in Push was seen in the opposite ordering.

Force dynamics support patient-centered moral inferences

Causal dependency models (e.g., Figure 3.5) provide one of several formats in which causation can be represented, and we have shown how focusing on different aspects of a causal model can affect moral judgment by highlighting different considerations about the violation of victims' rights. Another way to conceptualize causality is in terms of force interactions (Talmy, 1988; Wolff, 2007). This theory has proven particularly successful in representing causal information extracted from linguistic expression and visual scenarios (see Waldmann & Mayrhofer, 2016, for a discussion of the interaction between different ways of representing causal relations). We will argue that force dynamics also provides a natural link from basic observable causal aspects of behavior constellations to patient-centered considerations about the potential violation of victim rights. In what follows, we first lay out the gist of Talmy's (1988) conceptual framework of force dynamics. The model we present afterwards formalizes how observers can solve the task of inferring morally relevant but unobservable features (e.g., whether or not a negative right of a patient has been violated) from available visual or linguistic information about force-dynamic constellations. We then present evidence that subjects evaluate abstract entities in accordance with the model predictions.

Talmy (1988) conceptualizes force dynamics as an abstract semantic category that structures our language and thought in a variety of domains. When two entities interact with respect to force, our language assigns to them the role of agonist (patient) and antagonist (agent). The patient is the focal entity that has an intrinsic force tendency either towards rest or towards motion. The agent exhibits an opposing force tendency and is mainly relevant with regards to its effects on the patient. Take, for instance, the sentence "The dictator oppressed the people."

The people is the patient in this sentence with an intrinsic tendency to rise. The dictator exerts an opposing force which is stronger than the people's intrinsic tendency. The result is that the people stay down. In Talmy's (1988) terms, this would be an instance of *extended causing of rest*. Note that the people, not the dictator, is the focal entity that this interaction "is about," even though at first it might seem like a passive recipient. An indicator of the patient's primacy is that the gist of the situation could be transported without mentioning the agent ("The people was oppressed"), but not without mentioning the patient.

This example illustrates two important insights. First, force dynamics is an abstract conceptual framework that can be applied not only to physics, but also to mental and social domains. Second, force-dynamic language can be naturally used to describe morally relevant interactions. Agent-patient dyads play an important role in some current theories of moral cognition; some even claim that this dyadic structure makes up the essence of what all morality is about (Gray, Schein, & Ward, 2014; Gray, Young, & Waytz, 2012). Accordingly, it seems plausible that principles of thematic role assignment and force-dynamic pattern interpretation can be exploited to make testable predictions about moral judgments in response to observed or described interactions between two or more entities. In particular, in the absence of mitigating contextual information, agents might be judged negatively for violating (and positively for enhancing) the prima facie right of patients to exhibit their intrinsic tendencies (e.g., the dictator is blamed for preventing the people from manifesting its intrinsic tendency to rise).

Nagel and Waldmann (2012) reported a first attempt to test whether the observation of force-dynamic interactions between abstract shapes leads to systematic evaluative tendencies with regard to the involved entities. They presented subjects with minimal Michotte-type animated events in which two spheres collided on an empty screen. Subsequently subjects were asked to evaluate the movement of the spheres. Nagel and Waldmann hypothesized that people encode from this display the abstract force-dynamic configuration (e.g., *onset causing of motion* in the case of the classic Michotte launching event). When interpreted in the context of morality, this configuration violates a simple non-interference principle (NIP) stating that, by default, patients should be allowed to manifest their intrinsic tendencies. This principle expresses the central Western value of individuals' personal autonomy in the vocabulary of force dynamics. NIP is not implied by the framework of force dynamics itself but constitutes an external normative assumption that lends itself naturally to an operationalization in terms of force-dynamic concepts. In line with this hypothesis, Nagel and Waldmann (2012) observed that subjects evaluated the agent in launching events more negatively than in similar displays in which the underlying force-dynamic pattern did not violate the NIP (e.g., cases in which the patient did not start moving upon collision with the agent).

The entities in the stimuli presented by Nagel and Waldmann (2012) did not display any signs of animacy, such as self-propelled motion or spontaneous change of trajectory. Finding stable evaluative tendencies in response to events devoid of intentionality cues indicates that these judgments are triggered by basic movement cues

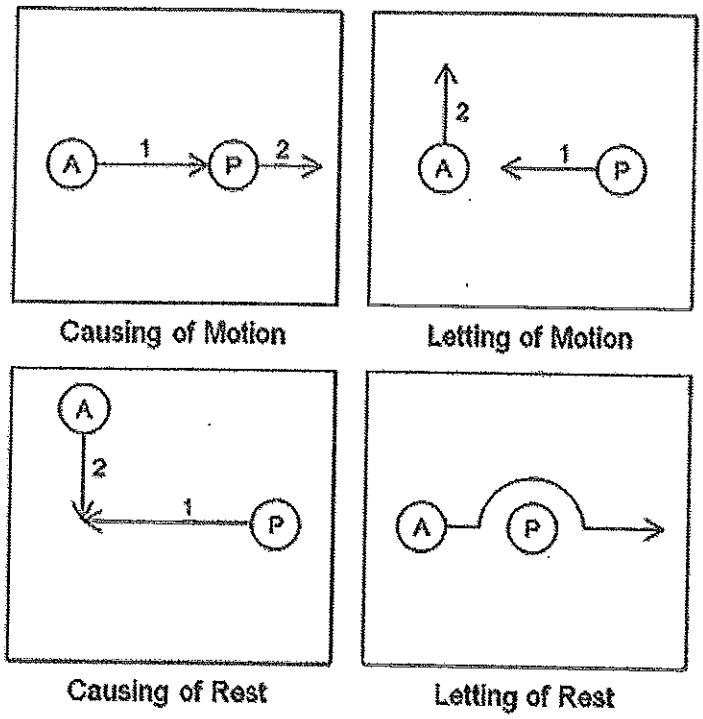

FIGURE 3.6 Illustration of visual displays of different force-dynamic patterns. Numbers at the arrows indicate temporal sequence of movement initiation. Gaps signify that no contact was made. A = agent; P = patient.

and do not require complex cues for animacy or goal-directed action (see e.g., Carey, 2009; Hamlin & Wynn, 2011). The force-dynamic pattern *itself* seems to inform the judgment. We do not claim that the encoding of force-dynamic patterns per se spontaneously elicits moral intuitions in a reflexive, bottom-up fashion, regardless of whether or not the involved entities are appropriate targets of moral evaluation. It seems likely that the mere request to provide moral judgments leads to a top-down interpretation of the entities as being legitimately subject to moral evaluation even in the absence of perceptual animacy cues. What we do claim is that force-dynamic information itself is used to generate a rudimentary evaluative tendency.

Apart from the basic launching case, an interesting other, more complex example of how force dynamics can mediate moral judgments is the *letting* pattern in which the agent could oppose the patient's intrinsic tendency but refrains from doing so. We expect positive agent ratings in these cases. Not only does the agent not violate the NIP, but it proactively assures that the patient can manifest its intrinsic tendency. In one experiment, we presented 52 subjects with animated displays of a blue and a green sphere instantiating either *causing* or *letting* patterns. The patient either had an intrinsic tendency towards rest or towards motion. The agent either moved in a way that caused the patient to change this tendency or to let it continue manifesting it (see Figure 3.6 for an illustration of some of the displays). Subjects were then asked to rate on two 7-point scales "How do you evaluate the movement of the blue/green figure," ranging from –3 ("negative") to

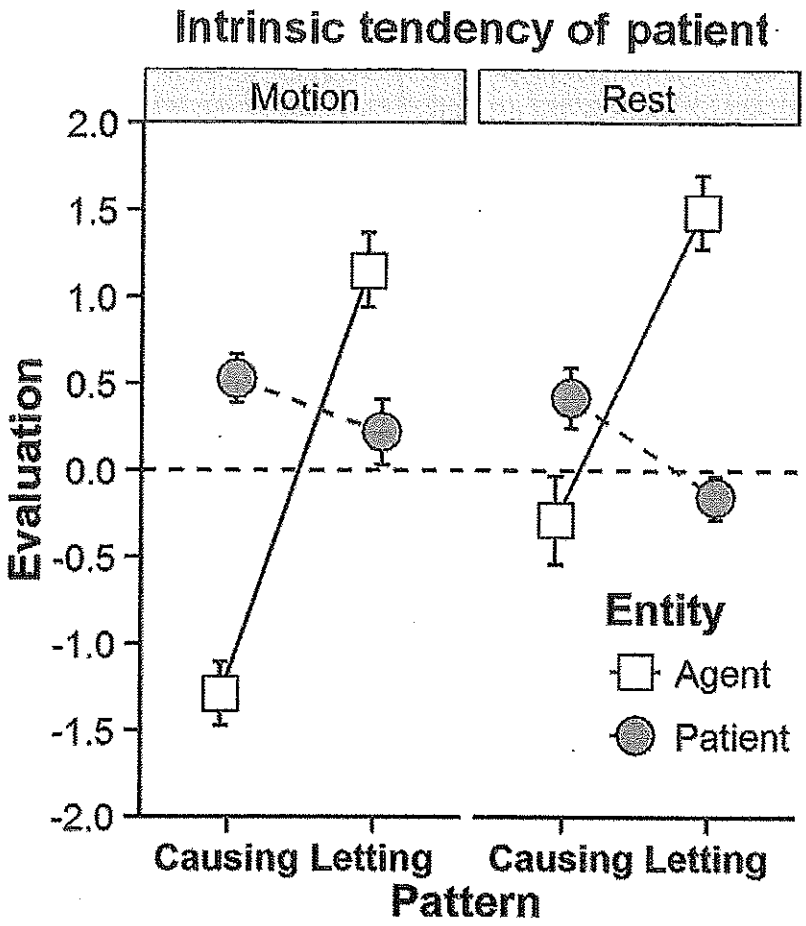

FIGURE 3.7 Results of the force dynamics experiment with visual stimuli. Error bars indicate 95% CIs.

+3 ("positive"). In a second test phase, we assessed the thematic role assignment by asking for each entity in each display whether "the blue/green figure has actively intervened in the situation."

Figure 3.7 summarizes the results of the evaluation task. In cases of *letting*, the agent was evaluated highly positively, while in cases of *causing* a clear tendency to evaluate the agent more negatively than the patient was replicated (see Nagel & Waldmann, 2012). We also found an unpredicted slight decrease in patient evaluations from causing to letting which might be due to a contrast effect. The results are very similar regardless of whether the patient has an intrinsic tendency towards rest or motion, as predicted by our account. The additional agency measure indicated that the agent was more strongly judged to have intervened actively than the patient in all cases. These findings show that our account can be successfully applied to more complex cases.

Another prediction following from the claim that abstract force-dynamic representations inform moral judgments is that the findings should be independent of the concrete input modality. As demonstrated in the introductory example, force-dynamic interactions can be effectively presented not only in visual displays, but also in natural language. Force-dynamic constellations are for instance encoded in

modal verbs (e.g., must, should, can) and in connectors (e.g., despite, because of). This allows for the construction of sentences describing interactions of unknown entities and unknown actions which yet unambiguously describe a force-dynamic interaction between these entities. We presented 80 participants with sentences in which blank entities (a coodle and a doff) engaged in blank actions (broosting and gaking). The assignment of these labels to the patient (P) and its action (X) and to the agent (A) and its action (Y) was counterbalanced across subjects. A and P had to be evaluated as in the previous experiment.

Two baseline sentences merely described the temporal relations of the entities' actions devoid of force-dynamic implications ("P Xed while A Yed"; "A was Ying before P Xed"). We contrasted these baseline scenarios with four descriptions with force-dynamic implications: *extended causing* ("P kept Xing because of A's Ying"), *onset causing* ("A's Ying made P X"), *despite* ("P kept Xing despite A's Ying"), and *letting* ("A's Ying let P X"). We expected the agent to be evaluated negatively relative to the patient in the *causing* patterns (where the NIP is violated) as well as in the *despite* pattern (where the agent attempts to violate the NIP but does not succeed), but not in cases of *letting* (where the agent refrains from interfering with the patient's intrinsic tendency). We furthermore added four more complex descriptions involving richer force-dynamic implications by using the modal verbs *can* and *have to* in combination with *because of* and *despite*. "P could keep Xing because of A's Ying" indicates that P was able to manifest its intrinsic tendency to X only because A opposed the counteracting influence of an (unmentioned) antagonist of P. Our model predicts positive ratings for the agent as it advances the basic value expressed in the NIP. By contrast, in "P could keep Xing despite A's Ying" the agent A itself is P's antagonist attempting (but failing) to keep P from manifesting its intrinsic tendency. We expected A to be evaluated negatively for this attempt to violate the NIP. The pattern should be reversed when *can* is replaced with *have to*. "P had to keep Xing because of A's Ying" indicates that P's intrinsic tendency is not to X, and that the agent A itself is P's antagonist forcing it away from its intrinsic tendency. This pattern should lead to a negative evaluation of A. Finally, in "P had to keep Xing despite A's Ying," it is P's unmentioned antagonist that forces the patient away from manifesting its intrinsic tendency not to X, while A attempts to counteract the unmentioned antagonist but fails. We predicted that A is not evaluated negatively relative to the patient as it attempts to prevent a violation of the NIP.

Figure 3.8 shows the mean agent and patient evaluations in response to the different descriptions. It can be seen that the predictions of our model, by and large, hold for verbally presented blank force-dynamic interactions between unknown entities. The difference in evaluation between agent and patient tended to be larger compared to the baseline in *extended* and *onset-causing*, in *despite*, *can-despite*, and *have-to-because*, but not in the remaining patterns. These results suggest that agents are evaluated negatively relative to the patient whenever they attempt to interfere with the patient's intrinsic tendency. Thus, when subjects are only provided with very basic perceptual information lacking content-based cues about the type of the involved entities and the type of actions, people fall back on abstract force-dynamic

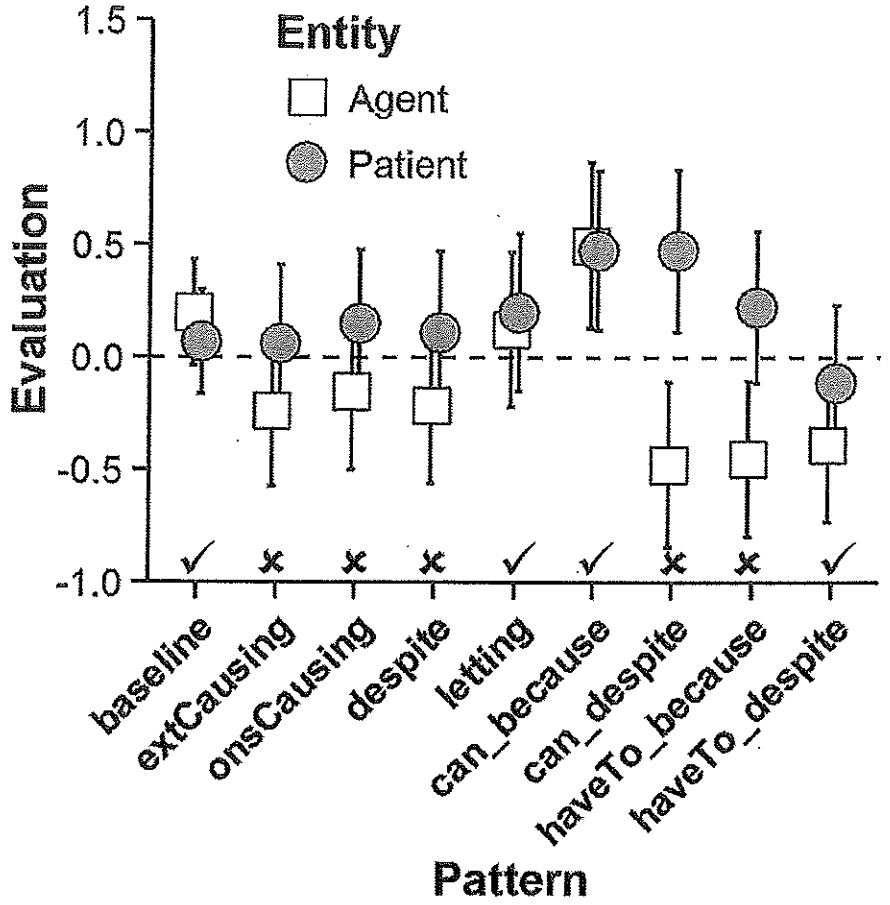

FIGURE 3.8 Results of the force dynamics experiment with verbal stimuli. Crosses indicate patterns in which the agent violates the NIP; check marks indicate patterns in which this is not the case. Error bars indicate 95% CIs.

representations and use these representations to generate a preliminary default evaluation in accordance with the non-interference principle. It can be expected that these default judgments can easily be overridden when rich, contextualized knowledge is available (e.g., the information that P's Xing consists in torturing innocent children for fun; but note that this information would also profoundly alter the force-dynamic structure of the complete event, as the former patient has now become an agent interfering with the children's rights). We do not claim that inferring evaluations from force-dynamic constellations is a reflexive or modularized process. But the seamless abstraction of force-dynamic patterns from a wide range of situations encoded in different modalities make force-dynamic constructs an effective tool in the ubiquitous task of quickly converting observable events into unified, morally interpretable representations.

General summary and discussion

In this chapter we have proposed a novel account of how causal models influence moral judgment. We reported patterns of judgments about a set of trolley

problems that cannot be accounted for by theories that merely focus on acts or outcomes. Instead, we explained the results with the hypothesis that the locus of intervention in the causal model of the situation shifts people's attentional focus to the fate of the targets in the vicinity of the intervention. If the victims of the moral dilemma are recognized to be the bearers of rights (e.g., the right not to be interfered with), subjects consider the inter-victim relations that supervene on the underlying causal model. The two factors, the locus of intervention and the causal model underlying the scenario determine the judged permissibility of the actions under consideration.

In a second project, we have shown that attentional focus cannot only be manipulated by changing the locus of intervention and the causal structure, but also by transferring different attentional foci from analogous cases. Analogical transfer based on causal model representations can lead to systematic changes in moral judgments in response to one and the same moral dilemma.

Finally, we looked at more basic forms of causal interactions to demonstrate the generality of our idea that causal representations entail information about the violation of rights. In our third project we presented virtually content-free visually or linguistically conveyed scenarios which can be represented as simple force-dynamic interactions between abstract entities. We demonstrated that subjects' evaluation of the behavior of abstract entities is sensitive to whether or not the patient of the interaction was prevented from manifesting its intrinsic tendency.

The three projects demonstrate the usefulness of a theory relating causal relations to moral evaluations. However, much more work lies in the future. For one, it would be desirable to more systematically explore the relations between different ways of representing causality and morality (e.g., causal models, force dynamics) (see also Waldmann & Mayrhofer, 2016). Another goal is to explore other factors besides locus of intervention and analogies that may also lead to differential attentional focus on specific aspects of causal models. Moreover, a more general account of the postulated supervenience relation between causal models and inter-victim relations is an important goal for future research. Finally, additional morally relevant factors that have been controlled in our studies need to be let back in and considered in relation to the factors we already specified, including intentionality, types of action, spatial and temporal distance, and type of moral judgments (e.g., blame, permissibility). The three reported projects represent just a first step in these directions.

Notes

1 One reviewer pointed out that in the frequently cited study by Hauser, Cushman, Young, Jin, and Mikhail (2007) the loop condition differed from the conditions in which the death of the single victim was a side effect. We did not list this study above because Waldmann et al. (2012) identified numerous confounds in this experiment. For example, only in the loop condition was the victim first described as a "heavy object" before it was mentioned that he was in fact a human. This alone might have contributed to a higher aversiveness of the act in this condition.

2 Some people might include some animals in this class.
3 Reprinted from *Cognition*, Vol. 131, Wiegmann, A. & Waldmann, M. R., Transfer effects between moral dilemmas: A causal moral theory, 28–43, Copyright (2014), with permission from Elsevier.
4 Reprinted from *Cognition*, Vol. 131, Wiegmann, A. & Waldmann, M. R., Transfer effects between moral dilemmas: A causal moral theory, 28–43, Copyright (2014), with permission from Elsevier.

References

Carey, S. (2009). *The origin of concepts*. New York: Oxford University Press.

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

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

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

Gray, K., Schein, C., & Ward, A. F. (2014). The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. *Journal of Experimental Psychology: General, 143*, 1600–1615.

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

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

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

Haidt, J., & Joseph, C. (2007). The moral mind: How 5 sets of innate moral intuitions guide the development of many culture-specific virtues, and perhaps even modules. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind* (Vol. 3, pp. 367–391). New York: Oxford University Press.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development, 26*, 30–39.

Hauser, M. D., Cushman, F. A., Young, L., Jin, R., & Mikhail, J. M. (2007). A dissociation between moral judgment and justification. *Mind and Language, 22*, 1–21.

Horne, Z., Powell, D., & Spino, J. (2013). Belief updating in moral dilemmas. *Review of Philosophy and Psychology, 4*, 705–714.

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

Kamm, F. M. (2015). *The trolley problem mysteries*. Oxford, UK: Oxford University Press.

Kant, I. (1785/1959). *Foundation of the metaphysics of morals*. Indianapolis, IN: Bobbs-Merill.

Lanteri, A., Chelini, C., & Rizzello, S. (2008). An experimental investigation of emotions and reasoning in the trolley problem. *Journal of Business Ethics, 83*, 789–804.

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology, 25*, 661–671.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*, 273–286.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York: Cambridge University Press.

Millar, J. C., Turri, J., & Friedman, O. (2014). For the greater goods? Ownership rights and utilitarian moral judgment. *Cognition, 133*, 79–84.

Nagel, J., & Waldmann, M. R. (2012). Force dynamics as a basis for moral intuitions. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the thirty-fourth annual conference of the Cognitive Science Society* (pp. 785–790). Austin, TX: Cognitive Science Society.

Nagel, J., & Waldmann, M. R. (2013). Deconfounding distance effects in judgments of moral obligation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 237–252.

Nagel, J., & Waldmann, M. R. (2016). On having very long arms: How the availability of technological means affects moral cognition. *Thinking & Reasoning, 22*, 184–208.

Patil, I., Cogoni, C., Zangrando, N., Chittaro, L., & Silani, G. (2014). Affective basis of judgment-behavior discrepancy in virtual experiences of moral dilemmas. *Social neuroscience, 9*, 94–107.

Petrinovich, L., & O'Neill, P. (1996). Influence of wording and framing effects on moral intuitions. *Ethology and Sociobiology, 17*, 145–171.

Thomson, J. J. (1985). The trolley problem. *Yale Law Journal, 94*, 1395–1415.

Timmons, M. (2002). *Moral theory: An introduction.* Lanham, MD: Rowman & Littlefield.

Scanlon, T. (2008). *Moral dimensions: Permissibility, meaning, blame.* Cambridge, MA: Harvard 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*, 135–153.

Sinnott-Armstrong, W., Mallon, R., McCoy, T., & Hull, J. G. (2008). Intention, temporal order, and moral judgments. *Mind and Language, 23*, 90–106.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 12*, 49–100.

Thomson, J. (1985). The trolley problem. *The Yale Law Journal, 94*, 1395–1415.

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*, 247–253.

Waldmann, M. R., & Mayrhofer, R. (2016). Hybrid causal representations. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 65, pp. 85–127). New York: Academic Press.

Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. In K. J. Holyoak & R. G. Morrison (Eds.), *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.

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology, 25*, 813–836.

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

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General, 136*, 82–111.