Moral Dynamics: Grounding Moral Judgment in Intuitive Physics and Intuitive Psychology

Felix A. Sosa
Department of Psychology, Harvard University

Tomer Ullman
Department of Psychology, Harvard University

Joshua B. Tenenbaum
Department of Brain and Cognitive Sciences, MIT

Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University

Tobias Gerstenberg*
Department of Psychology, Stanford University

Abstract

When holding others morally responsible, we care about what they did, and what they thought. Traditionally, research in moral psychology has relied on vignette studies, in which a protagonist's actions and thoughts are explicitly communicated. While this research has revealed what variables are important for moral judgment, such as actions and intentions, it is limited in providing a more detailed understanding of exactly how these variables affect moral judgment. Using dynamic visual stimuli that allow for a more fine-grained experimental control, recent studies have proposed a direct mapping from visual features to moral judgments. We embrace the use of visual stimuli in moral psychology, but question the plausibility of a feature-based theory of moral judgment. We propose that the connection from visual features to moral judgments is mediated by an inference about what the observed action reveals about the agent's mental states, and what causal role the agent's action played in bringing about the outcome. We present a computational model that formalizes moral judgments of agents in visual scenes as computations over an intuitive theory of physics combined with an intuitive theory of mind. We test the model's quantitative predictions in three experiments across a wide variety of dynamic interactions between agent and patient.

Keywords: moral judgments, effort, intuitive physics, intuitive psychology, causal inference, counterfactual simulation

*Corresponding author: Tobias Gerstenberg (gerstenberg@stanford.edu).

Introduction

In a popular image, three wise monkeys advise us: see no evil, hear no evil, speak no evil. But do we actually see evil, in the way we see shapes, or colors, or monkeys (Firestone, Scholl, et al., 2016)? When viewing simple shapes moving around a 2D world, people spontaneously and consistently attribute goals and intentions to them (Heider & Simmel, 1944), including social motivations such as helping and hindering (Ullman et al., 2009). Even young children appear to draw consistent conclusions about the goals, intentions, and relations of actors in simple visual vignettes (e.g., Gergely & Csibra, 2003; Gergely, Nádasdy, Csibra, & Bíró, 1995; Hamlin, Wynn, & Bloom, 2007; Luo & Baillargeon, 2005), and at slightly older ages will act to punish morally bad actors (Hamlin, Wynn, Bloom, & Mahajan, 2011).

In cognitive science, there is a long tradition that attempts to formally link perception and psychological attributions, by identifying relevant visual cues in a scene. This line of research can be traced back at least to Michotte (1946/1963), and extends to current work on the visual cues that could underpin perceptions of agency, intention, and various interactions such as courting, chasing, and protecting (e.g., Hubbard, 2005; Scholl & Gao, 2013). Recent work has suggested that even moral judgments may be explained by the visual processing of kinematic features, such as the velocity of a car hitting a person, or the distance a person traveled to push someone into harm's way (De Freitas & Alvarez, 2018; Iliev, Sachdeva, & Medin, 2012; Nagel & Waldmann, 2012). These accounts propose a direct mapping from visual features, such as motion and contact, to moral judgments. For example, De Freitas and Alvarez (2018) argue that our visual system provides all the information we need to make moral judgments. While these views don't deny that inferences about mental states matter as well, they highlight the possibility of a more direct route from perceptual information to moral judgment.

In contrast, a great deal of separate prior work on moral judgment has focused exactly on how mental states enter moral calculations (Cushman & Young, 2011; Mikhail, 2007). This line of research often relies on written vignettes rather than visual stimuli, and it has demonstrated how a person's mental states, such as their beliefs, desires, and intentions, as well as what causal role they played, are key determinants of moral judgments (Cushman, 2008; Gerstenberg et al., 2018; Lagnado & Gerstenberg, 2017; Lagnado, Gerstenberg, & Zultan, 2013; Malle, Guglielmo, & Monroe, 2014b; Patil, Calò, Fornasier, Cushman, & Silani, 2017; Shaver, 1985; Waldmann, Nagel, & Wiegmann, 2012; Weiner, 1995; Young, Cushman, Hauser, & Saxe, 2007; Young & Saxe, 2008): For example, people judge a person more severely when they intended a bad outcome (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Lagnado & Channon, 2008), and when they caused it to come about (Alicke, 1992; Cushman, 2008; Gerstenberg et al., 2018).

Vignette studies are important for mapping out in what ways different factors influence moral judgments (Foot, 1978; Malle, 2021; Waldmann et al., 2012). However, using vignettes to study morality has its limits. Individual participants can only evaluate a small number of scenarios, as reading vignettes quickly gets tiresome. Factors of interest, such as what causal role each agent played, need to be explicitly communicated, which makes it difficult to discern whether people would have spontaneously considered this information in the course of making moral judgments (cf. Malle, Guglielmo, & Monroe, 2014a). It is

also difficult to manipulate the relevant factors in precise quantitative ways in vignettes (cf. Gerstenberg & Icard, 2019). The limitations of this method constrains the kinds of theories that can be tested. The majority of theories of moral judgment make only qualitative predictions (Malle et al., 2014a). For example, they state that having intended for a negative outcome to come about is worse than not having intended it. Instead, we present participants with visual animations of moral interactions that allow us to control relevant factors in precise ways so that different models can be quantitatively compared to one another.

We believe that in order to understand moral judgment, it is critical to bring together quantitative manipulation of morally relevant perceptual information, with qualitative manipulation in the form of written vignettes. Here, we lay out a computational model that combines key insights from both of these approaches. Like much prior vignette-based research, the model emphasizes the role of mental state inferences and causal attribution in moral judgment. However, like recent video-based research, it infers mental states and causal roles based on morally relevant perceptual information.

Rather than assuming a direct mapping from visual features to moral judgments (see, e.g., De Freitas & Alvarez, 2018; Iliev et al., 2012; Nagel & Waldmann, 2012), we propose that the route from visual input to moral judgment is mediated by an intuitive understanding of how the world works, one that encompasses both an intuitive theory of mind (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016; Wellman & Gelman, 1992), and an intuitive theory of physics (Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg & Tenenbaum, 2017; Goodman, Tenenbaum, & Gerstenberg, 2015; Kubricht, Holyoak, & Lu, 2017; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). These intuitive theories support rapid inferences about a person's mental states from their actions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Battaglia et al., 2013), and discern the causal role that a person's action played in bringing about the outcome (Gerstenberg, Goodman, Lagnado, & Tenenbaum, in press; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017).

The proposal that a person's mental states and what causal role they played in bringing about the outcome are critical for others' moral evaluations of their action is not new. There is a rich literature on and what role mental states and causality play in how people judge whether an action was permissible (Mikhail, 2007; Waldmann et al., 2012). For example, according to the doctrine of double effect (Foot, 1967; Thomson, 1976, 1985), an action that causes harm is permissible if that harm was an unintended side-effect of a positive outcome, but impermissible if the harm served as a means for bringing about the outcome (Cushman, Young, & Hauser, 2006; Guglielmo & Malle, 2010; Royzman & Baron, 2002; Waldmann & Dieterich, 2007).

Moral judgments are complex, and the computational model presented here only captures a small proportion of the multitude of factors that go into any moral judgment. Our model lays out what we see as the basic building blocks of a more complete theory of moral judgment. By incorporating these building blocks into a computational model, we derive quantitative predictions and test these predictions against participants' judgments.

The rest of the paper is organized as follows. We first motivate and describe our computational model. We test the model's predictions in three experiments. In Experiment 1, we replicate a prior study which argued for a direct link from visual features to moral judgments (Iliev et al., 2012), and show that the results are consistent with a model that infers

an agent's desire to do harm via the physical effort it exerted. In Experiment 2, we test moral intuitions in a wider range of situations, and elicit graded judgments which provide a stronger test for the model's predictions. In Experiment 3, we further expand the range of test cases and evaluate what role mental state inference and causal attribution plays for different kinds of moral judgments including judgments of responsibility for an outcome, and the moral badness of an action. We discuss the implications of our findings for research in moral psychology, highlight what we see as important limitations of our current model, as well as a road map for these limitations may be addressed by future research.

The Moral Dynamics Model (MDM)

We now introduce the *Moral Dynamics Model* (MDM), a model that produces moral evaluations of agent's actions in dynamic visual scenes. We first motivate the model, describe its two main components – mental state inference and causal attribution – and then show how it predicts moral judgments.

Motivation: From moral kinematics to moral dynamics

Our work was inspired by Iliev et al. (2012). In their paper, Iliev et al. explore the role that physical and perceptual factors play in people's moral judgments about harm. In two experiments, participants watched video clips like the one depicted in Figure 1. The videos include an agent, a patient, and a fireball. Whereas the agent was able to see the fireball and interact with it, the patient couldn't see the fireball and the patient would die upon contact with it. The agent and patient were capable of self-propelled motion, and the fireball was sometimes moved by unobservable winds. All clips ended with the patient colliding with the fireball and dying, though they varied in how the three entities interacted with one another.

Iliev et al. manipulated what causal role the agent played in bringing about the outcome. In Experiment 1, they focused on two factors: motion and contact. Based on prior research on the omission bias in moral judgment – the finding that agents are blamed more for negative outcomes resulting from actions versus non-actions (Royzman & Baron, 2002, see also Gerstenberg & Stephan, 2021) – they hypothesized that the agent's action would be judged as worse when it moved rather than stood still, and when the patient (or fireball) wasn't already in motion. Further, the agent was predicted to be seen as worse when it made direct contact with the patient (see Cushman et al., 2006; Waldmann & Dieterich, 2007). In the experiment, participants saw pairs of clips and were asked to evaluate in which of the two clips the agent's action was worse. The results showed that participants perceived an agent's action as worse when it directly made contact with the patient, and when the intervened-on object was not in motion. This pattern of results was replicated in another experiment that used written vignettes instead of video clips.

In Experiment 2, they expanded on this finding using a new set of video clips that tested whether higher magnitude and frequency of motion and contact would result in harsher moral judgment. Specifically, they predicted that an agent's actions will be judged worse when the force used by the agent was greater, when there was physical contact with the patient, when the agent traveled a longer distance before making contact, when the duration of contact between agent and patient was longer, and when there was greater number of

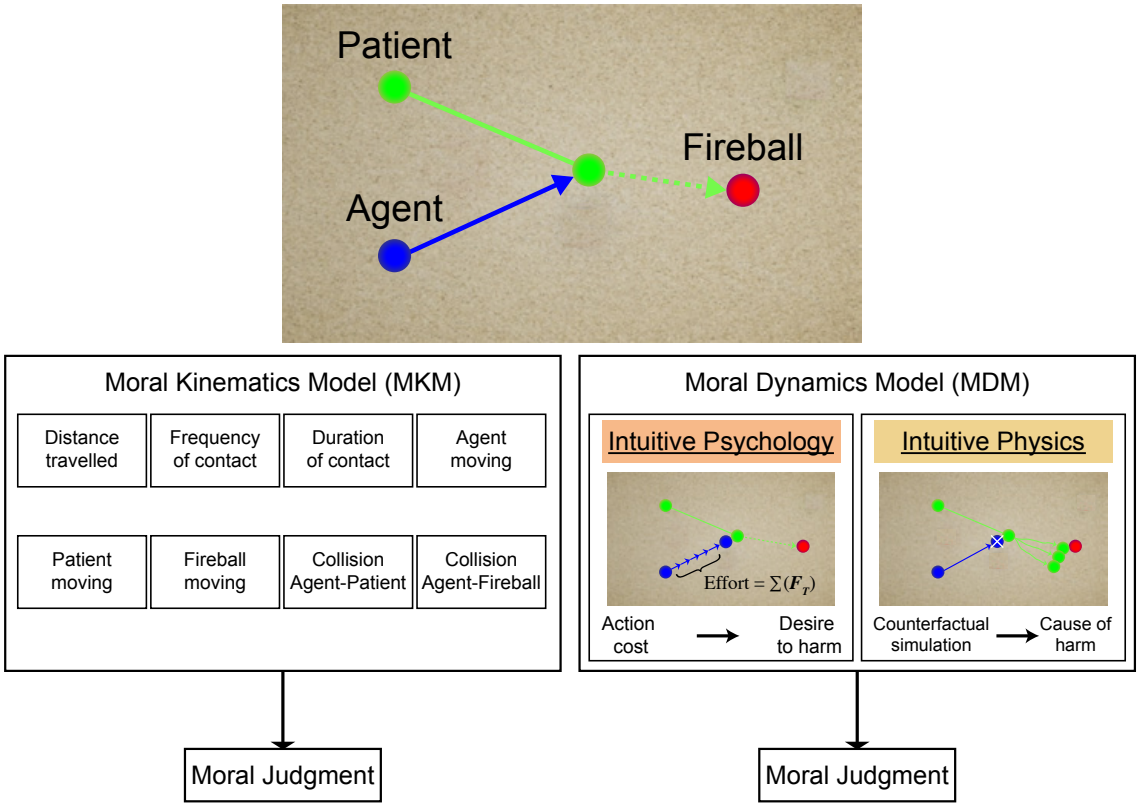

Figure 1. **Top**: A diagrammatic illustration of an interaction between the agent, patient, and fireball. Videos of the stimuli that we used in the different experiments are available here: `https://github.com/cicl-stanford/moral_dynamics`. **Bottom left**: The *Moral Kinematics Model* (MKM) predicts moral judgments via a direct mapping from eight kinematic features to moral judgments (Iliev et al., 2012). These features include the distance that the agent traveled, the frequency with which it contacted the patient, the duration of contact with the patient, whether the agent/patient/fireball were initially moving or at rest, and whether the agent collided with the patient/fireball. **Bottom right**: The *Moral Dynamics Model* (MDM) predicts that the route from visual input to moral judgment is mediated by an intuitive understanding of psychology and of physics. It uses an intuitive psychology module to infer the agent's desire for the harm from the action costs that the agent incurred in the form of physical effort. The MDM computes effort via the sum of force impulses that the agent exerts on itself in order to move. It uses an intuitive physics module to determine what causal role the agent played in bringing about the outcome, by simulating the counterfactual of what would have happened if the agent hadn't been present in the scene. The more certain the model is that the outcome wouldn't have happened without the agent, the more causal the agent is judged. Moral judgments for negative outcomes are predicted to increase the more the agent desired the harm, and the more clear it was that the agent caused the harm.

contacts between agent and patient. Using the same methodology as in Experiment 1, participants viewed 15 pairs of video clips whereby each pair manipulated one of the key dimensions (e.g., the agent traveled a long vs. short distance before it made contact with the patient). Again, the results showed that participants' moral judgments were sensitive to the factors as predicted. Agents' actions were judged worse when they traveled longer distances, used more force, made contact with the patient, made repeated contact, and when the duration of the contact was longer.

Iliev et al. conclude that physical factors that influence what causal role the agent played in bringing about the outcome influence how people make moral judgments. They also acknowledge that in addition to the agent's causal role, patterns of physical motion may also license inferences about the agent's mental and emotional states, as well as their social relations (Barrett, Todd, Miller, & Blythe, 2005; Gao, Newman, & Scholl, 2009; Heider & Simmel, 1944; Scholl & Tremoulet, 2000). They state: "From this perspective, one possibility is that physical factors are used only as cues to infer mental states, which subsequently are used to form moral judgments. A full answer to this question is beyond the scope of this article, yet we believe that inferred differences in mental states will not be enough to explain our main findings." (Iliev et al., 2012, p. 1396)

Here we take on this challenge. The MDM assumes that instead of directly going from kinematic features like information about contact and motion to moral judgments, people instead use this information to infer the agent's mental state (i.e., how much the agent desired the harm), as well as the causal role their action played in bringing about the outcome. While Iliev et al. demonstrated a qualitative link between the different kinematic features and people's moral judgments, we develop a computational model that computes the agent's desire and causal role from the video clip, and that predicts from these factors how bad the agent's actions were, and how responsible the agent was for the outcome. We will now discuss the mental state inference and causal attribution component of the MDM in turn.

Mental state inference: What does the action reveal about the agent's desire?

What mental states drove a person's action is critical for our moral evaluation. In principle, many mental states are morally relevant including a person's beliefs, desires, and intentions. Here, we focus on the role of desires. Following prior approaches to capturing people's intuitive theory of mind (e.g., Baker et al., 2017, 2009; Jara-Ettinger et al., 2016; Kleiman-Weiner, Shaw, & Tenenbaum, 2017; Ullman et al., 2009), we model an observer who reasons about an agent's mental states by inverting the generative process by which mental states give rise to actions. Accordingly, people think of others as goal-directed agents who choose actions that maximize their expected reward, subject to their beliefs, external constraints, and internal abilities (see also Dennett, 1987; Gershman, Gerstenberg, Baker, & Cushman, 2016).

To analyze the role of desires specifically, we draw on a version of the more general Bayesian Theory of Mind framework that is called the 'Naïve Utility Calculus' (Jara-Ettinger et al., 2016). According to this model, people believe that others act to maximize their state-dependent rewards, and to minimize action-dependent costs:

$$U(s, a) = R(s) - C(a), \tag{1}$$

where U is an agent's utility that stems from the reward R derived from world state s, and the cost C of taking action a.

Given this formulation of an agent's utility, we can infer how much reward an agent placed on a particular state, based on the cost that they were willing to incur to bring about that state. More specifically, assuming that agents act rationally to maximize their expected utility, knowing $C(a)$ places a lower bound on $R(s)$. In order for an agent to take on a particular action cost, they must have sufficiently desired the goal state. This leads us to the following inequality which we use to formalize the link between cost, reward, and desire:

$$D_{\text{Agent}}(s) \propto R_{\text{Agent}}(s) > \sum_{t=0}^{T} C_{\text{Agent}}(a_t) \tag{2}$$

which states that an agent's desire to bring about a state of the world, $D_{\text{Agent}}(s)$, is proportional to the reward that agent gets for being in that state, $R_{\text{Agent}}(s)$. Following the principle of rationality, we assume this reward is greater than the sum cost of the series of actions $C_{\text{Agent}}(a_0, a_1, ..., a_T)$, taken by the agent to bring about that state.

In our experiments, we will assume that action costs are related to physical effort, and that one agent's reward can depend on the utility of another agent. We now discuss each of these two assumptions in turn.

Physical effort as action cost. Physical effort features prominently in both decision-making and moral judgment (Bigman & Tamir, 2016; Jara-Ettinger, Kim, Muentener, & Schulz, 2014; Kurniawan et al., 2010). Jara-Ettinger et al. (2014) demonstrated that transgressors are judged more harshly for taking more costly actions to bring about a negative outcome. In those studies, participants were given multiple vignettes involving the same outcome (e.g., stealing someone's wallet), and judged the vignette involving the greatest amount of effort as depicting the worst offender. Even young children are sensitive to the physical effort required by an action, and take it into account when determining the goal of an agent (Liu, Ullman, Tenenbaum, & Spelke, 2017).

Bigman and Tamir (2016) showed that how much effort a person exerted affects others' moral evaluation of their actions. In line with what we propose here, Bigman and Tamir argue that effort serves as an indicator for how much someone desired a particular outcome. In seven studies, they demonstrated that perceived effort influenced participants' judgments of both moral and immoral actions, that this effect was independent of outcome severity, and also had an influence on how much monetary reward participants believed a person should receive. The rationale of using cost to infer reward carries through with other types of cost as well, such as risk or mental effort (Kool & Botvinick, 2018).

Here, we limit ourselves to inferences about physical effort. As detailed below, we record how much physical effort an agent exerted in a given scenario and predict that moral judgments will increase with the perceived amount of effort. Expanding on Equation 2, we can now formalize the cost of a sequence of actions as the amount of physical effort it takes to perform those actions:

$$C(a_0, a_1, ..., a_T) \propto \int_{t=0}^{T} F(a_t)dt, \tag{3}$$

where a_t is the action taken at time t, and $F(a_t)$ is the force an agent generates on itself to take that action. As we discuss later, in practice we consider a discretized time setting in

a physics engine, replace the integral with a sum, and replace F with an impulse I over a short time:

$$C(a_0, a_1, ..., a_T) \propto \sum_{t=0}^{T} I_{\text{Agent}}(a_t) \tag{4}$$

which states our formalization of action cost as the total sum of impulses an agent applies to itself to take an action in a physics engine in order to bring about the end state of a scenario in that physics engine.

Social goals as rewards. Equation 1 states that a person's utility depends on the reward associated with a particular state s. For example, in a foraging task, the reward could be associated with getting food. Here, we assume that agents can have social goals, too. In principle, such social goals can be positive or negative. That is, it could be one agent's goal for another agent to receive positive or negative utility. Ullman et al. (2009) formalized this idea by assuming nested utility functions: the agent reward R_{Agent} depends on patient utility U_{Patient}, such that $R_{\text{Agent}}(s = U_{\text{Patient}})$. A pro-social attitude (i.e., high reward for helping) of the agent towards the patient is expressed as a positive relationship between R_{Agent} and U_{Patient} and an anti-social attitude (i.e., high reward for harming) as a negative relationship. This simplified model of helping and hindering can quantitatively account for adults' reasoning about social goals (Ullman et al., 2009), as well as the choice patterns of pre-verbal infants (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013). Here, we will focus on situations in which the agent's goal is to harm the patient.

Causal attribution: What causal role did the agent play?

When morally evaluating each other's actions, people not only care about mental states; they also care about what causal role the person's action played in bringing about the outcome (e.g., Lagnado et al., 2013; Pizarro, Uhlmann, & Bloom, 2003; Sloman, Fernbach, & Ewing, 2009; Sloman & Lagnado, 2015; Waldmann & Dieterich, 2007). The Iliev et al. experiments were motivated by how causation relates to physical factors. According to *production theories* of causation, causes bring about effects via a spatio-temporally continuous transfer of a quantity such as physical force (Dowe, 2000; Salmon, 1984; Wolff, 2007). In contrast, *dependence theories* of causation construe causal relationship in terms of probabilistic, or counterfactual dependence. For example, for C to qualify as a cause of E according to a counterfactual theory, both C and E must have happened, and E would not have happened if C hadn't happened (Lewis, 1973; Woodward, 2003). A limitation of most existing theoretical frameworks for causation is that they don't make any quantitative predictions about particular causal events, such as whether the agent caused the harm in this instance.

Recently, Gerstenberg et al. (in press) developed the *Counterfactual Simulation Model* of causal judgment (CSM) which yields quantitative predictions for physical scenarios like the ones we explore here. The CSM is inspired both by production and dependence theories of causation. In line with dependence theories (e.g. Lewis, 1973; Woodward, 2003), it computes a counterfactual contrast between what actually happened and what would have happened if the candidate cause had been different. In line with production theories (e.g. Dowe, 2000; Salmon, 1984; Wolff, 2007), it assumes that the fine-grained processes by which causal interactions happen dictate how people's mental simulations unfold. Here, we draw

on this model to quantitatively determine what causal role the agent played in bringing about the outcome.

The CSM predicts that people's judgments about the causal role an agent played in bringing about the outcome are determined by comparing what actually happened with what would have happened if the agent hadn't been present in the scene. For example, consider the scene shown in Figure 1B. Here, the patient and the fireball are initially stationary. The agent then launches the patient toward the fireball with which it collides. In this situation, the agent clearly caused the collision, and the CSM's predictions agree with this intuition. In the counterfactual situation in which the agent had been removed from the scene, the patient wouldn't have collided with the fireball.

Now consider the same situation but assume that the patient was already headed toward the fireball (remember that the patient cannot see the fireball) before the agent collided with it. In this case, the agent didn't cause the collision. The patient would have collided with the fireball even if the agent hadn't been present.

There are also situations in which it is less clear whether or not the agent caused the outcome. For example, imagine a situation in which both patient and fireball were initially in motion, and where the agent again launched the patient into the fireball. Depending on how exactly the patient and fireball were moving, it might not be clear what would have happened in the counterfactual situation in which the agent wasn't present. Would the collision have happened anyhow, or would the patient and fireball have missed each other? In such a situation, the CSM predicts that the agent would receive an intermediate causal rating. Generally, the CSM predicts causal judgments in terms of the observer's subjective degree of belief that the outcome would have been different if the candidate cause had been removed from the scene. The more clear it is that the outcome would have been different, the higher the causal rating is predicted to be.

More specifically, to determine the causal role that the agent A played in bringing about the outcome O (i.e., the collision between patient and fireball), the CSM simulates how a scene S would have unfolded if the agent hadn't been present in the scene. The CSM uses the physics engine that was used to generate the video clip to run the counterfactual simulation. While there is a ground truth answer to the question of whether the collision would have happened if the agent hadn't been present, an observer doesn't have access to that answer. Instead, the observer needs to rely on their intuitive understanding of the physical scene to mentally simulate how the counterfactual would have unfolded (see Battaglia et al., 2013; Ullman et al., 2017). The CSM captures the uncertainty inherent to physical prediction by injecting noise into the simulation process. Concretely, in each counterfactual simulation, it first replays what actually happened, removes the agent from the scene, and introduces movement noise to the object with which the agent collided in the actual situation from the time point on at which that collision happened. This movement noise is implemented by applying a small perturbation to the object's velocity vector and each time step on the simulation.

The CSM then records whether or not the outcome would have happened in that counterfactual simulation. By running several simulations, the CSM estimates the probability that the agent caused the outcome as the proportion of counterfactual simulations in which the outcome would have been different.

More formally, we define the probability that agent caused outcome O as

$$P(\text{Agent} \rightarrow O) = P(O' \neq O | S, \textit{remove}(\text{Agent})) \tag{5}$$

where O' is the outcome in the counterfactual situation, S contains all relevant information about what actually happened,[1] and *remove*(Agent) describes the operation of removing the agent from the scene.

The CSM has been shown to accurately capture people's causal judgments across a large range of dynamic physical scenes (Beller, Bennett, & Gerstenberg, 2020; Gerstenberg et al., in press; Gerstenberg & Stephan, 2021).

Predicting moral judgments

We have now introduced two quantitative components of our model that infer an agent's desire from the physical effort they exerted, and that judge the agent's causal role by simulating what would have happened if the agent hadn't been present in the scene. The MDM predicts that moral judgments depend on people's beliefs about an agent's desires and the causal role that the agent played in bringing about an outcome. That is, people will judge an agent more negatively the more reward that the agent appears to derive from an innocent other being harmed, and the clearer it was that the agent caused that harm.

Specifically, we predict that inferences about both the agent's desire for harm and the extent to which they are the cause of the harm affect people's moral judgments. We model this using a simple linear combination:

$$\text{Moral judgment} = \alpha + \beta_1 \cdot \text{Desire for harm} + \beta_2 \cdot \text{Cause of the harm}, \tag{6}$$

where α is a constant to map between the different scales, and β_1 and β_2 capture how much each component affects the moral judgment. We fit these parameters to participants' judgments.

As discussed above, we infer the lower bound of the agent's desire for the harm via the action costs that they were willing to incur to bring about the outcome (see Equation 2). In our setting, we use the physical effort that the agent exerted as the action cost (see Equation 4). This means that we can rewrite Equation 6 like so

$$\text{Moral judgment} = \alpha + \beta_1 \cdot \text{Effort} + \beta_2 \cdot \text{Causality}, \tag{7}$$

whereby we replaced the 'Desire for harm' with the 'Effort' that the agent exerted.

This model of moral judgment allows us to explore to what extent different kinds of moral judgment are sensitive to these two components (see Cushman, 2008; Malle, 2021). In Experiment 3, we will explore to what extent moral judgments about the badness of actions versus the responsibility for an outcome are sensitive to the agent's inferred desire and their causal role.

To sum up, the MDM infers an agent's desire for harm and what causal role it played in bringing about the outcome. The model assumes that agents act to achieve desired rewards, and that actions are associated with a cost in the form of physical effort. Given that

[1] This includes the objects' motion paths as well as the physical specifications of objects and environment, including, for example, object mass and friction.

a rational agent trades off cost and reward (taking costly actions to receive a greater reward than the cost expended), an observer can use the effort an agent expended as indicative of the desire that the agent had for achieving an outcome (see Bigman & Tamir, 2016). Specifically, the amount of effort exerted by an agent places a lower bound on the reward that the agent expected to receive from its actions. If an agent takes a very costly action to achieve a harmful outcome, that agent likely expected a large reward for causing harm, and thus should be morally blamed to a high degree. Such a unified principle (greater cost means a greater reward) provides an underlying rationale for what may seem like disparate visual features, such as traveling longer distances or taking more actions. To determine what causal role the agent played, the model simulates what would have happened if the agent hadn't been present. The more certain the model is that the negative outcome would not have happened without the agent, the morally worse the agent is judged to be.

The model makes a number of simplifying assumptions. First, psychological costs encompass more than physical effort, such as time delay or mental effort (Kool & Botvinick, 2018). Second, pro-social and anti-social relationships are more than just a utility-to-reward transformation, and moral evaluations depend on more than the inferred desire that the agent places on someone else being harmed (Waldmann et al., 2012). Lastly, exactly how people integrate information about effort and causality is unknown. While we expect it to be more complex than a simple summation, we believe both quantities contribute positively to moral judgment.

Despite these simplifying assumptions, we believe this framework provides a core mechanics and delineation of a moral calculus that further work can build on, adding in more varied notions of cost, effort, and causality. Because the notions of force, effort, and causality play a central role in our model, we refer to it as *Moral Dynamics Model* (MDM), in contrast to the *Moral Kinematics Model* (MKM) by Iliev et al. (2012), which predicts moral judgments based on perceptual/kinematic features such as distance, angle, contact, and velocity.

Stimulus generation and model implementation

Before we test the MDM in a number of experiments, we describe how the video clips in our experiments were generated and detail how the model was implemented.

Stimulus generation. The video clips in our experiments were closely modeled after the clips used in Iliev et al. (2012). Each video clip features an agent (blue), a patient (green), and a fireball (red). The agent can see the fireball and is not harmed upon contact with it, while the patient cannot see the fireball and is harmed upon contact. We generated the physically realistic video clips using the 2D physics engine Pymunk (www.pymunk.org), and we used the graphics engine Blender (www.blender.org) for rendering.

Figure 1 shows an example image of a video clip from Iliev et al. (2012) as well as from our experiments. Iliev et al.'s (2012) video clips were in 3D, featured a checkerboard floor, and different geometric shapes representing the three entities. The patient breaks into pieces when it makes contact with the fireball. Our clips were in 2D shown from a bird's-eye view, featured a floor textured like sand, and used the same geometric shape to represent the different entities. In Experiment 1 and 2, the patient goes up in flames when it contacts the fireball. In Experiment 3, a yellow crash bubble appears when the two make

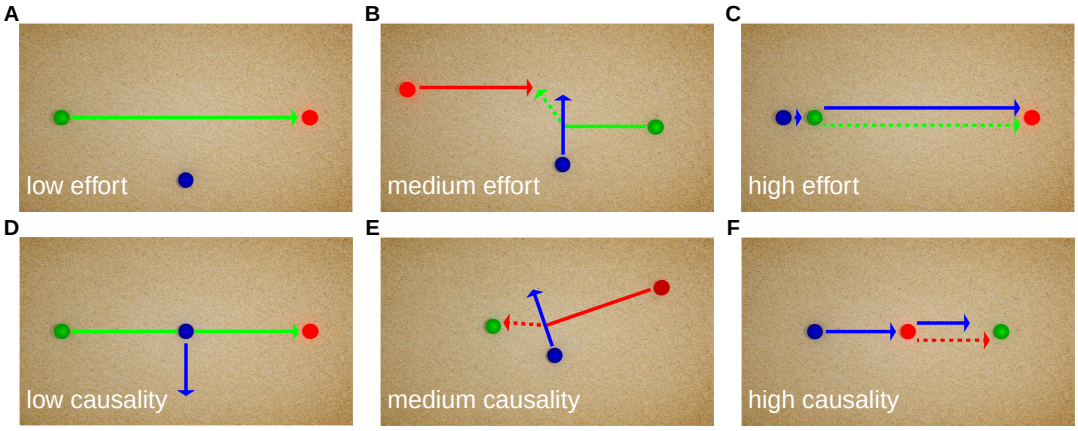

Figure 2. Diagrams for a selection of video clips illustrating interactions between the agent (blue), the patient (green), and the fireball (red). The top row shows situations in which the agent puts in different amounts of effort. The bottom row shows situations in which the agent played a different causal role in bringing about the outcome. The colored arrows show the different entities' movements. Solid arrows indicate that an entity moved by itself. Dashed arrows indicate that an entity was moved/affected by another one. Double arrows show that both entities moved together because one entity pushed the other.

contact (like one would see in a comic book). The video clips are between one and five seconds long.

Model implementation. In order to predict moral judgments, the MDM considers the agent's desire to harm the patient, and the agent's causal role in bringing about the harm. To infer the agent's desire, it takes into account how much effort the agent exerted. To determine the agent's causal role, it simulates what would have happened if the agent hadn't been present in the scene. We discuss how each component was implemented in turn.

Effort. In order for the agent to move, it has to exert a force on itself. If the agent wants to maintain a certain velocity, it has to keep exerting a force on itself at each time point in the physics simulation in order to overcome friction. If an agent were to stop exerting such a force on itself, the friction would slowly bring it to a stop.

The top row of Figure 2 shows diagrams of three video clips in which the amount of effort that the agent exerted differed. In Figure 2a, the agent doesn't move at all so it put in zero effort. In Figure 2b, the agent moves up as the patient get closer. Here the agent put in an a medium amount of effort. In Figure 2c the agent pushes the patient all the way to the fireball, so the agent put in high amount of effort.

We generated the video clips in a way such that the movement of each entity closely resembles the movements in the original clips in Iliev et al. (2012). For each clip, we then calculate the overall amount of effort that the agent exerted. This overall amount of effort is the sum of all the force impulses that the agent applied to itself over the course of the video clip. There are some video clips in which the agent stays put and the patient or fireball bump into it and get deflected off the agent (see Figure 3, clip 5). Even though the agent is not moving here, it still put in some effort because in order for it not to move, it

needs to apply a force impulse that matches the entity's force that bumps into it.

How much effort the agent exerts depends on the parameter settings of the physics engine, which include the masses of each entity and the value of the friction. We incorporate friction in our model by using a damping parameter that determines how much an object's velocity is slowed down at each time step if it doesn't continue to exert effort. A higher damping parameter implies that an agent has to exert more effort to move the same distance. In Experiments 2 and 3, we ask one group of participants to judge how much effort the agent exerted. We fit the damping parameter in our model to participants' effort judgments.

Causality. We determine the causal role that the agent played in bringing about the outcome via the counterfactual simulation model (CSM) of causal judgment. Accordingly, the model compares what actually happened with what would have happened if the agent had been removed from the scene.

The bottom row of Figure 2 shows three situations in which the agent's causal role differs. In Figure 2d, the agent didn't cause the outcome to happen. The patient would have collided with the fireball even if the agent hadn't been present in the scene. In Figure 2e, it's unclear whether or not the agent caused the collision to happen. The patient and fireball might have collided even if the agent hadn't been present. In Figure 2f, the agent clearly caused the collision. Here, both the fireball and the patient were initially at rest and it would have stayed that way if the agent hadn't been present in the scene.

The model yields these causal judgments by probabilistically simulating how the relevant counterfactual situation would have played out. Specifically, the model applies movement noise in the counterfactual simulation to the entity with which the agent collided from the time point on at which this collision happened. In Figure 2d, the agent doesn't feature in a collision so the outcome in the counterfactual situation is clear. The patient would have collided with the fireball even if the agent had been removed from the scene so the probability that the agent caused the outcome O is $P(\text{Agent} \rightarrow O) = 0$ (see Equation 5).

In Figure 2e, the agent collides with the fireball. To determine what causal role the agent played, the CSM simulates how the fireball would have moved if the agent hadn't been present in the scene. To do so, it applies a small perturbation to the direction of the fireball's velocity vector at each time step after the time at which the agent and fireball collided in the actual situation (see Smith & Vul, 2012). We don't apply any noise to the patient in this clip for two reasons: First, because the agent didn't collide with the patient, there is no uncertainty about how the patient would have moved if the agent hadn't been present. Second, because the patient is at rest, perturbing the direction of its velocity vector (which is 0) would have no effect.

The CSM then records whether the patient and fireball would have collided in that noisy simulation, and it repeats this process many times. To compute the probability that the agent caused the outcome, the model looks at the proportion of simulations in which the outcome in the counterfactual situation would have been different from what actually happened. To estimate this probability, we draw 1000 samples for each situation. In this case, the patient and fireball would not have collided in 491 out of 1000 noisy simulations in which the agent was removed from the scene. Thus, the estimated probability that the agent caused the outcome in this clip was $P(\text{Agent} \rightarrow O) = 0.491$.

In Figure 2f, the fireball was initially at rest. Again, because the noisy perturbation is applied to the direction of the entities' velocity vector (rather than its magnitude), the

fireball remains at rest in the counterfactual situation in which the agent had been removed from the scene. So it's clear in this case that the collision would not have happened if the agent hadn't been present, and thus the probability that the agent caused the outcome is $P(\text{Agent} \rightarrow O) = 1$.

In Experiment 3, we asked one group of participants to judge whether the patient and fireball would have collided if the agent hadn't been present in the scene, and we will see that their judgments are highly correlated with the CSM's predictions. The CSM has one free parameter which determines the extent to which an object's velocity vector is perturbed at each time step in the physical simulation (the standard deviation σ of a Gaussian distribution with mean 0). We fit this parameter to participants' judgments in Experiment 3 and assume the same noise parameter for the other experiments.

Experiment 1: A replication of Iliev et al. (2012)

Our first experiment seeks to qualitatively test the *Moral Dynamics Model* (MDM), and examine whether participants' judgments can be explained by assuming that they infer an agent's desire for harming the patient via the amount of effort that the agent exerted. Our experiment was closely modeled after Experiment 2 in Iliev et al. (2012), in which they examined how various kinematic features affected participants' moral judgments. Participants viewed pairs of clips and were asked to evaluate in which clip what the agent did was worse. With this experiment, our goal was to replicate Iliev et al.'s (2012) results, and to make sure that our stimuli yield the same pattern of results as Iliev et al.'s stimuli did (see Figure 1a and b for examples of the different stimuli).

Methods

Participants. 46 participants ($M_{age} = 34.5$, $SD_{age} = 10.4$, 11 female, 34 male, 1 non-binary) participated in the experiment. For all the experiments reported in this paper, we recruited participants via Amazon Mechanical Turk (Crump, McDonnell, & Gureckis, 2013). The experiments were run using PsiTurk (Gureckis et al., 2016). Only participants based in the US with an approval percentage greater than 95% were allowed to participate. All experiments were approved by the MIT Committee on the Use of Experiments with Human as Experiment Subjects (COUHES).

Design. Iliev et al.'s (2012) Experiment 2 included 16 different video clips. The video clips differed in terms of their kinematic features. For example, they varied the distance that the agent traveled, whether or not the agent made contact with the patient, how many times the agent contacted the patient, how long the agent made contact with the patient, and the force the agent exerted on the patient. We included the subset of video clips that could be captured in in our 2D, top-view implementation. We were able to implement 12 out of the 16 video clips from the original experiment.[2] Figure 4 shows diagrams of the

[2]Three out of the four clips that we didn't include featured motion up and down ledges. While such motion can be captured in 2D, it would require a side-view rather than a top-view, and we opted to keep the viewpoint consistent. In the other clip, the agent entered the scene from outside the frame. We excluded this one as it's not clear how much effort the agent had already exerted before entering the scene. For some of the video clips, the patient is initially at rest but then keeps its velocity after it was launched by the patient. For these clips, we assume that the patient exerts effort in order to keep moving. This assumption doesn't affect the predictions of our model as it only considers how much effort the agent exerted.

different video clips that participants saw in each of the 11 trials. In each trial, participants viewed one pair of video clips. The order in which the clips were presented was randomized.

Procedure. Participants were instructed that they would see pairs of videos involving imaginary creatures (Blues and Greens) and a fireball. Participants were further informed that each video shows a situation in which a Green collided with the fireball. Participants' task was to judge in which of the two videos Blue's actions were worse.

Participants then viewed a set of familiarization videos that showed Blues, Greens, and fireballs interacting. As in Iliev et al. (2012), the familiarization videos informed participants that Blues and Greens were intelligent, social creatures, and that fireballs were inanimate objects that were sometimes moved by magnetic winds. Participants were informed that Greens *could not* see fireballs and were burned when they touched them, whereas Blues *could* see fireballs and were not burned when they touched them. Finally, participants learned that while Blues and Greens usually got along, they would see instances in which Blues harmed Greens.

A set of comprehension check questions ensured that participants had read the instructions. These questions assessed that participants knew that only Greens could be harmed by fireballs, only Blues could see fireballs, and that fireballs could sometimes be moved by magnetic winds. Participants were only allowed to move on to the main experiment if they correctly answered all comprehension check questions. If a participant failed the comprehension check, they had to go through the introduction and familiarization videos again, and re-take the comprehension check.

In the main phase of the task, participants watched pairs of video clips. The two clips were presented next to each other, and participants had to watch both videos twice, going from the video presented on the left of the screen to the one on the right, and back again. The left/right placement of videos was randomized. After viewing both videos twice, participants responded to the prompt "The action of Blue was..." with one of six possible responses (presented from left to right): "much worse in the left video", "worse in the left video", "somewhat worse in the the left video", "somewhat worse in the the right video", "worse in the right video", and "much worse in the right video"

At the end of the experiment, participants provided demographic information, and were invited to share any comments. On average, it took participants 10 minutes (SD = 3.1) to complete the experiment.

Results

Figure 3 shows participants' judgments for the 11 trials. On each trial, the two clips are arranged so that the left clip is the one in which the agent exerted more effort. The histograms show what proportion of participants selected the different response options. Bars on the left indicate that participants believed that the agent's action was worse in the left clip, and bars on the right indicate that the action was worse in the right clip. As the figure shows, participants tended to think that the agent's action was worse in the left clip compared to the right clip. If we binarize participants' selections based on whether they thought the agent's action was worse in the left clip (more effort) or in the right clip (less effort), we find that participants chose the left clip 83% of the time. Out of the 46 participants, 37 participants chose the clip in which the agent exerted more effort in at least 9 out of 11 pairs. Participants had a preference for the clip in which the agent exerted more

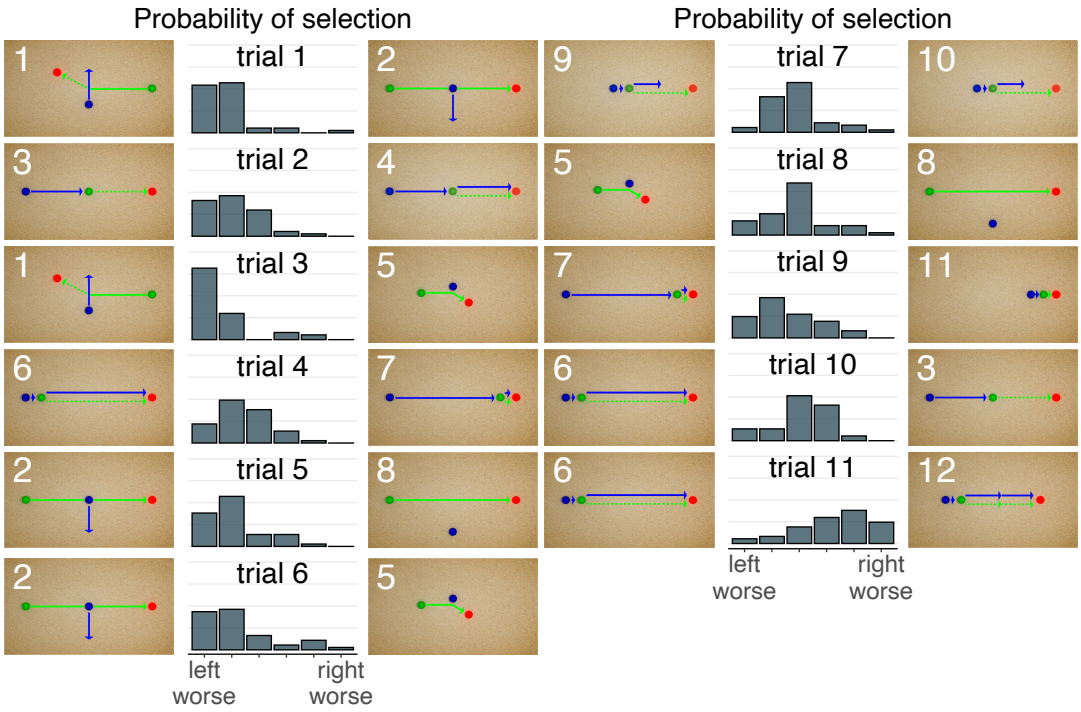

Figure 3. **Experiment 1 results**. Participants' badness judgments for each of the 11 trials. For each pair, participants judged whether blue's action was worse in the left clip or in the right clip. The histograms indicate what proportion of participants selected the different response options. For each trial in this figure, the agent exerted more effort in the left clip (the left/right presentation was randomized in the experiment) so the *Moral Dynamics Model* (MDM) predicts that the agent's action would be judged as worse in the left clip. The trials are ordered from top left to bottom right by participants' overall preference for the left clip in which the agent exerted more effort. Colored arrows in the stimulus pictures show trajectories as in Figure 2. Some clips appeared in multiple trials. For example, clip 1 appeared both in trial 1 and in trial 3.

effort in all trials except for trial 11. On trial 11, the agent pushes the patient for a long duration in clip 6 (left side). In clip 12 (right side), the agent pushes the patient twice, but exerts less effort overall. Despite the fact that the agent exerted more effort in clip 6 compared to clip 12, participants judged the agent's actions as worse in clip 12.

Table 1 shows the proportion of participants who judged that the agent's actions were worse in the clip in which the agent exerted more effort. It shows the data both from our experiment, as well as from Iliev et al.'s (2012) Experiment 2. Our results qualitatively replicate what Iliev et al. found. For all of the trials, our participants preferred the same clip as did participants in Iliev et al. The quantitative differences between our studies are likely due to sampling variation. Iliev et al.'s Experiment 2 featured only 16 participants.

Discussion

The results of Experiment 1 closely replicate what Iliev et al. (2012) found (see Table 1). Whereas Iliev et al. interpreted their results to show that participants' moral judgments are sensitive to different kinematic features, we demonstrate that the results are consistent with the idea that participants glean how much the agent desired to bring about the harm via inferring the amount of effort that the agent exerted. Indeed, participants judged that the agents actions were worse when the agent exerted more effort to bring about the harm. The *Moral Dynamics Model* (MDM) qualitatively predicts this pattern of results. According to the MDM, people judge the agent's actions as worse when the agent exterted more effort, because the cost that the agent incured places a lower bound on the reward that the agent must have received from bringing about the harm. This follows from the principle of rational action according to which agents take into account anticipated costs and benefits when choosing actions.

Instead of postulating a number of different kinematic features, and a mapping from these features to the moral evaluation, the MDM suggests that kinematic features like the distance that the agent traveled, or the duration of contact between the agent and patient, reveal how much the agent desired to bring about the harm. The MDM infers this desire through estimating the amount of effort that the agent exerted in the clip.

The MDM correctly predicts in which of two clips the agent's actions are perceived as worse for 10 out of 11 trials. In trial 11 (see Figure 3), participants judged the agent's actions as worse in clip 12 than clip 6 even though the agent exerted more effort in clip 6. In clip 6, the agent pushes the patient all the way for a long distance. In clip 12, the agent first launches the patient a little bit toward the fireball, and then when the patient slows down, bumps into it again to push it all the way toward the fireball. Given how we implemented the physical model, the agent's effort is greater in clip 6 compared to clip 12. One possibility for the discrepancy between model prediction and people's judgments is that participants' inferences about how much effort the agent exerted differ from the predictions of the physics model. So even though the model states that the agent exerted more effort in clip 6, participants might have inferred that the agent exerted more effort in clip 12. To address this concern, we directly ask participants to judge how much effort the agent exerted in each video clip in Experiment 2.

Another possibility is that the double push scenario (clip 12) makes the agent's intentions particularly clear. One way to describe what happened in this scenario is that the

Table 1
Percentage of participants for each trial (1–11) who judged that the agent's actions were worse in the clip in which the agent exerted more effort (i.e., the left clip for each pair in Figure 3) in our Experiment 1 and in Iliev et al.'s (2012) Experiment 2. Participants' responses were binarized from the original six-point scale.

	1	2	3	4	5	6	7	8	9	10	11
Experiment 1	93	93	89	87	87	85	83	80	78	63	26
Iliev et al.'s (2012) Experiment 2	69	100	75	94	100	82	75	75	82	82	25

agent wanted to launch the patient into the fireball but initially pushed too little. Upon realizing this, the agent took another shot at it and made sure to push the patient all the way. The agent's intention might be less clear in clip 6. Maybe the agent just wanted to walk in this direction and the patient happened to be in the way.

We asked participants at the end of the experiment what factors influenced their badness judgments. While some participants directly referred to the amount of effort that the agent exerted (e.g., "Just whether Blue was active or passive and how much effort Blue seemed to put in to cause harm."), others mentioned that the agent's intentions mattered (e.g., "If the Blue planet purposely attacked, moved out of the way, or pushed the Green planet in a taunting way (such as slightly tapping it and the pushing it the rest of the way), I considered that bad behavior").[3] We will return to the role of intentions for moral judgment in the General Discussion.

Experiment 2: A quantitative test of the model

In this experiment, we will quantitatively compare the model's predictions against people's judgments. Rather than having people evaluate in which of two clips the agent's actions were worse, this time we showed participants a single clip on each trial. By asking participants for a judgment on a continuous sliding scale, we can quantitatively compare their judgments to the predictions of our model. This time, we also test whether our model correctly predicts participants' inferences about how much effort the agent exerted in bringing about the harm.

We will compare how well different models capture participants' moral judgments (see Figure 1). We will test three versions of the MDM that differ in whether they take into account only effort, causality, or both. We will also test the *Moral Kinematics Model* (MKM) which doesn't compute effort or causality, but instead predicts participants' judgments based on kinematic features of the scene (see Figure 1).

Methods

Participants. 83 participants (M_{age} = 35.7, SD_{age} = 12.7, 42 female, 40 male, 1 non-binary) participated in the experiment.

Design

Participants were randomly assigned to the *Effort* condition ($N = 42$), or the *Moral* condition ($N = 41$). The video clips that participants watched included the 12 video clips from Experiment 1 as well as five additional clips that were taken from Iliev et al.'s (2012) Experiment 1 (see Figure 4). These clips focused on the effect of movement and intervention on moral judgment, that is, whether the agent intervened on the patient or on the fireball (e.g., clip 10 vs. clip 11, or clip 12 vs. clip 13), and whether the agent, patient, or fireball were already moving before the intervention (e.g., the patient moves in clip 6 but doesn't move in clip 8).

[3] All of the participants' explanations for what factors influenced their judgments may be viewed in the online analysis document here: `https://cicl-stanford.github.io/moral_dynamics/`

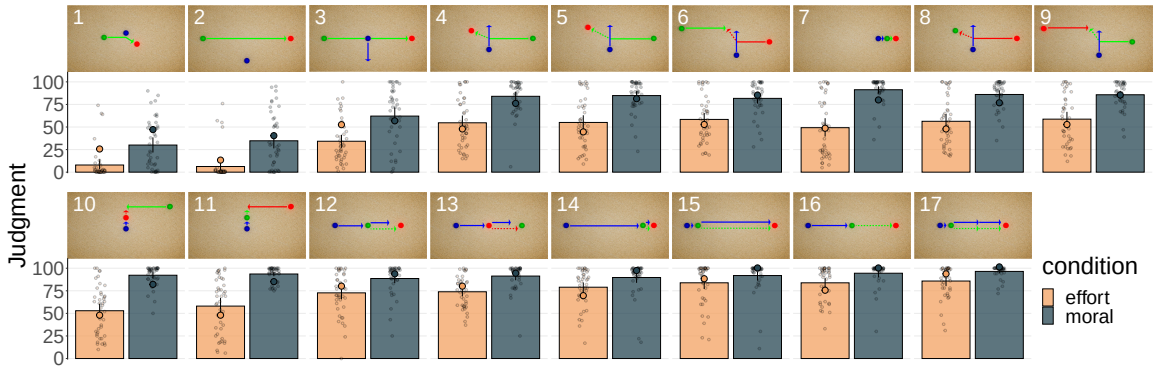

Figure 4. **Experiment 2 results**. Participants' effort judgments (orange, left bar) and moral judgments (gray, right bar) for each of the 17 scenarios. Bars indicate mean ratings and error bars bootstrapped 95% confidence intervals. Small points indicate individual judgments (jittered along the x-axis for visibility). Large circles show model predictions. For the moral condition, the model uses participants' mean effort judgments and the model's computed causality as a predictors (i.e., the effort + causality model in Table 2). The trials are ordered from lowest overall judgments to highest.

Procedure. The instructions and familiarization videos were largely identical to those of Experiment 1. In both conditions, participants viewed the same familiarization videos with slight modifications depending on the condition. This time, instead of pairs of clips in the test phase, participants only viewed a single clip at a time. Participants watched each video twice before indicating their response on a continuous slider. The 17 video clips were presented in randomized order. In the *Effort* condition, participants answered the question, "How much effort did Blue exert in this scenario?" with the endpoints of the slider labeled "very little" (0) and "very much" (100). In the *Moral* condition, the question was "How bad was what Blue did?" and the endpoints were labeled "not bad" (0) and "very bad" (100).

On average, it took participants 8.8 minutes ($SD = 4.4$) to complete the experiment.

Results

Figure 4 shows the results from both the effort and moral condition together with the model predictions. We will report the results from each condition in turn.

Effort condition. The orange bars in Figure 4 show participants' mean effort judgments, and the large circles indicate the model predictions. We fit participants' effort judgments by finding the damping parameter in the physics engine that minimizes the sum of squared errors between model prediction and the mean judgments for each clip. We then ran a Bayesian linear mixed effects model with effort as a fixed effect, as well as random slopes and intercepts for each participant.[4] All Bayesian models reported in this paper were written in Stan (Carpenter et al., 2017) and accessed with the `brms` package (Bürkner,

[4]We rescaled the effort values to range between 0 to 1 by dividing the effort value in each clip by the clip with the highest effort value.

2017) in `R` (R Core Team, 2019). Participants' mean effort judgments were well-accounted for by the model Pearson's $r = 0.91$, rmse (root mean square error) $= 9.25$.

Moral condition. The gray bars in Figure 4 show participants' mean moral judgments, and the large circles indicate the model predictions. We fitted five different Bayesian linear mixed effects models to participants' judgments. Three models are versions of the MDM that either only consider effort, only causality, or both effort and causality. The models also include random intercepts and slopes for each participant. The fourth model is an implementation of the MKM which uses a set of eight kinematic features to predict participants' judgments (see Figure 1). These features are: distance traveled by the agent, duration of contact between agent and patient, whether the agent made contact with the patient, how often the agent contacted the patient, whether the agent made contact with the fireball, as well as one predictor each for whether the agent, patient, or fireball were moving or static at the beginning of the clip. The model includes random intercepts for each participant but no random slopes (because of the larger number of predictors). We also fitted a baseline model that just includes a fixed intercept and random intercepts for each participant.

For the effort predictor in the MDM, we used participants' mean judgments from the effort condition. To compute the causality predictor, we ran the counterfactual simulation model as described above. We fitted the standard deviation of the Gaussian distribution that determines how much an entity's movement direction is perturbed at each time step in the counterfactual simulation by minimizing the squared error between model prediction and participants' mean judgments. Table 2 shows the model predictors for effort and causality, as well as participants' mean effort and moral judgments for the 17 different clips. To predict participants' moral judgments, we used the MDM with 'effort' and '$\text{causality}_{\text{model}}$' as predictors.

Figure 5 shows how well the different models account for participants' moral judgments. A version of the MDM that only considers effort as a predictor does a good job of capturing participants' moral judgments, whereas a model that only considers causality doesn't capture participants' judgments as well. A model that considers both effort and causality as predictors only improves the model fit somewhat compared to the effort-only model. The MKM model achieves the highest fit to participants' judgments albeit with a larger number of free parameters. Whereas the moral kinematics model has nine free parameters (only counting the fixed effects in the linear mixed effects model), the MDM has either three or two free parameters depending on whether both effort and causality or only one of the predictors are included.

Table 2

Experiment 2. *Model predictors for effort and causality, as well as participants' mean effort and moral judgments across the 17 trials. Here all predictors are shown on the same scale from 0 to 100 as the participants' response scale.*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
$\text{effort}_{\text{model}}$	15	0	49	43	39	49	44	43	49	43	43	83	83	70	93	77	100
$\text{causality}_{\text{model}}$	40	0	0	51	86	99	100	49	100	100	100	100	100	100	100	100	100
effort	8	6	34	55	55	58	49	56	59	53	58	73	74	79	84	84	86
moral	30	35	62	84	85	82	91	86	86	92	94	89	91	90	92	94	96

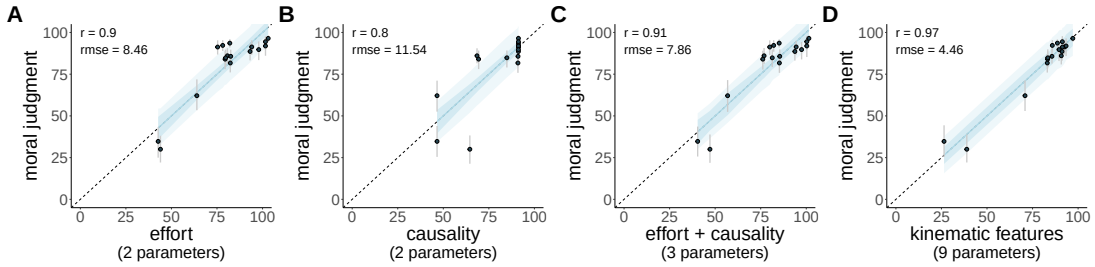

Figure 5. **Experiment 2 Moral Judgments.** Scatter plots showing the relationship between participants' mean moral judgments (y-axis) and versions of the *Moral Dynamics Model* (MDM) that only consider causality (**A**), only effort (**B**), or both causality and effort (**C**), or the *Moral Kinematics Model* (MKM) that uses kinematic features (**D**). Each point indicates one of the 17 trials. The error bars are 95% bootstrapped confidence intervals. The ribbons show the 95% credible interval for each model fit (dark blue), and the 50% prediction interval (light blue).

Table 3 shows a quantitative comparison of the different models. To take into account the varying number of parameters that the different models have, we used approximate leave-one-out cross-validation for model comparison. According to this measure, the 'effort + causality' model accounts best for participants' judgments overall (as indicated in the Δ elpd column; see Vehtari, Gelman, & Gabry, 2017 for details about this measure). While the MKM fits participants' judgments best, it fares less well in the cross-validation (see Figure 6 for the posterior distributions over the different predictors in the MKM). Table 3 also shows how many participants' judgments were best accounted for by the different models (again, using cross-validation). Accordingly, 22 participants judgments' were best explained by one of the versions of the MDM, 18 participants by the kinematic features model, and one participant by the baseline model.

The kinematic features include information about whether the agent collided with the patient and with the fireball (Table A1 shows how well each individual predictor is correlated with participants' badness judgments). In line with prior work (Greene et al., 2009; Waldmann & Dieterich, 2007), Iliev et al. (2012) had found in their Experiment 1 that participants judged the agent's actions as worse when it had collided with the patient rather than the fireball. However, we didn't find such an effect in our experiment. For example, in clip 4 the agent intervenes on the patient and in clip 8 the agent intervenes on the fireball but both clips elicit very similar moral judgments. The same is true for clips 6 and 9, clips 10 and 11, as well as clips 12 and 13.

Discussion

The results of Experiment 2 showed a close correspondence between participants' judgments of how much effort the agent exerted and how bad its actions were judged to be. A model that used the effort judgments from one group of participants accurately predicted the badness judgments of another group of participants. Participants' effort judgments were highly correlated with the predictions from the physics engine that we used to implement the different clips. The MDM predicts that how much effort the agent exerted is important for how its actions are morally evaluated because the amount of effort is indicative for how

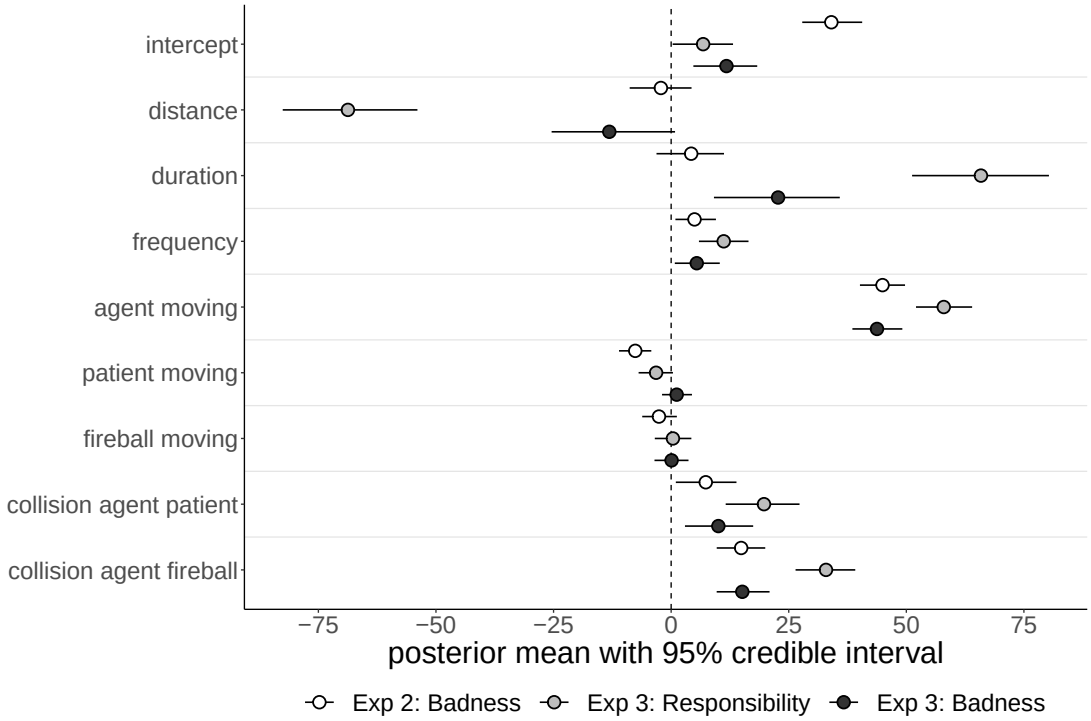

Figure 6. Posterior distributions over the different predictors in the *Moral Kinematics Model* (MKM) for participants' badness and responsibility judgments in Experiment 2 and 3. The points indicate posterior means, and the error bars indicate 95% credible intervals.

much the agent desired to bring about the harm. A version of the MDM that considered both how much effort the agent exerted and what causal role it played in bringing about

Table 3

Experiment 2 Model Comparison. Note*: The 'model' column shows what predictors were included in each model. The baseline model just includes an intercept as a predictor whereas the kinematic features model includes additional predictors (see Figure 6). 'intercept', 'effort', and 'causality' show the posterior means of each predictor together with the 5% and 95% credible interval. 'r' and 'rmse' show the Pearson's correlation and root mean square error. 'Δ elpd' shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models together with the standard error in parenthesis. Lower numbers indicate worse performance. '# best' shows the number of participants whose judgments were best predicted by each model.*

model	intercept	effort	causality$_{\text{model}}$	r	rmse	Δ elpd	# best
effort	37.96 [30.7, 45.35]	75.92 [65.72, 87.13]		0.90	8.46	-20.33 (10.42)	11
causality	46.58 [39.4, 53.39]		44.66 [37.09, 51.87]	0.80	11.54	-103.79 (15.68)	2
effort + causality	36.73 [28.88, 43.8]	58.78 [49.31, 68.87]	14.29 [6.99, 20.79]	0.91	7.86	0 (0)	9
kinematic features	see Figure 6			0.97	4.46	-5.46 (25.39)	18
baseline	81.35 [78.07, 84.63]				19.31	-300.55 (28.35)	1

the outcome didn't do much better than a simpler model that only considered effort. A model that used kinematic features such as the distance that the agent traveled or whether the different entities were initially moving or static fit participants' judgments best. When the different models are compared to one another using cross-validation which takes into account model complexity, the 'effort + causality' model fares best overall. While prior work had shown that an agent's action are judged as morally worse when the agent directly intervened on the patient rather than the harm (Greene et al., 2009; Iliev et al., 2012; Waldmann & Dieterich, 2007), we didn't find such an effect in this experiment.

When looking at how well individual participants' judgments were explained by the different models, we found considerable variation. This variation is also reflected in participants' comments about what factors influenced their judgments, with some participants considering both the agent's intentions and their causal role (e.g., "I made my judgments by looking at whether or not the blue shape influenced the movement of the green shape or the fireball. If it looked like it was intentional or could have been prevented by the blue shape then I considered it bad. If it was not caused by the blue shape or the green shape did it on its own then it was not bad."), and some participants having focused on kinematic features (e.g., "On whether or not the blue orb actively shoved the green one into the fire, stood still, or moved away.").

Why did the agent's causal role make relatively little difference to people's badness judgments overall? One possibility is that participants do in fact care about the agent's causal role but that the stimuli we selected didn't allow us to find support for the role of causality. As Table 2 shows, the model is certain that the agent caused the outcome in 12 out of the 17 clips we used, it gives a medium causal rating for three clips, and a zero rating for the remaining two. In short, there was little variation in what causal role the agent played across the clips. Furthermore, participants' mean effort ratings and the model's causality prediction were relatively highly correlated with $r = .78$ which means that it's difficult to tease apart potential contributions of each factor to people's moral judgments.

Another possibility is that for judgments of badness, what causal role the agent played doesn't really matter. Cushman (2008) found that whereas for judgments of blame and decisions about punishment both the agent's mental state and their causal contribution mattered, judgments about wrongness mostly relied on the inferred agent's mental states.

A limitation of Experiment 2 is that there was relatively little variation in participants' mean badness judgments across the clips. As Figure 4 shows, the agent's action was judged to be very bad for clips 4–17, and received lower badness ratings only in clips 1–3. In both clips 1 and 2 the agent just stays put, and in clip 3 the agent goes out of the way. We address this shortcoming in Experiment 3.

Experiment 3: Badness and Responsibility

The goal of Experiment 3 was to further explore people's moral judgments about dynamic visual scenes, and to address some of the shortcomings of Experiment 2. Whereas for the prior experiments, we re-implemented the video clips based on Iliev et al. (2012), this time we designed a novel set of stimuli that more fully varied the causal role that the agent played in bringing about the outcome, as well as how much effort the agent exerted. In addition to clips in which the agent clearly caused the outcome (e.g., clip 19 in Figure 7), or clearly didn't cause the outcome (e.g., clip 1), we included clips in which the agent's

causal role was less clear (e.g., clip 10). In these clips, the patient might have been harmed even if the agent hadn't been present in the scene.

The results of Experiment 2 showed that participants' badness judgments were close to ceiling for many of the clips. In Experiment 3, we made the outcome less severe. Instead of having the red object be a fireball that sets the patient on fire upon contact (Experiments 1 and 2), a collision between the patient and red object now resulted in a yellow star showing up where the two collided (similar to how collisions are indicated in cartoons). We expected to see more variance in participants' badness judgments now that the negative outcome was less severe. In addition to asking participants to evaluate how much effort the agent exerted and how bad its actions were (Experiment 2), we now also asked participants to evaluate the causal role that the agent played, and to what extent the agent was responsible for the outcome.

In Experiment 2 we found that the agent's causal role had relatively little effect on participants' badness judgments. In line with Cushman (2008) we hypothesized that what causal role the agent played would matter more for responsibility judgments compared to badness judgments (see also Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, submitted; Malle, 2021). To make sure that we'd be able to estimate how much effort and causality contribute to people's judgments, we made sure that the two predictors were only minimally correlated across the stimuli (see Table 4). The pre-registration for Experiment 3 is available here: `https://osf.io/h9cqg`

Methods

Participants. 233 participants (M_{age} = 37.9, SD_{age} = 11.8, 86 female, 144 male, 3 non-binary or preferred not to say) participated in the experiment. 24 participants were excluded because they failed to pass an attention check.

Design. Participants were randomly assigned to the *Effort* condition ($N = 51$), *Causality* condition ($N = 51$), *Responsibility* condition ($N = 55$), or the *Badness* condition ($N = 52$). We pre-registered that we would keep collecting participants until we reached $N = 50$ participants in each condition. The additional participants in each condition are due to the randomized procedure of how we recruited participants. We decided to keep rather then discard the results from the additional participants ($N > 50$) in each condition.

We created the video clips with two goals in mind: 1) sufficient variance in both effort and causality across the clips, and 2) a minimal correlation between effort and causality. To increase the uncertainty about what would have happened if the agent hadn't been present in the scene, we had the entities travel on diagonal rather than horizontal paths. Table 4 shows how correlated the different predictors were with one another (as well as pairwise correlations with participants' mean judgments in the different conditions).

Procedure. The instructions and familiarization videos were largely identical to those of the previous experiments. In all four conditions, participants viewed the same videos with only the query and response variable changing across the conditions. Like in Experiment 2, participants only viewed and rated a single clip at a time. 21 clips were presented in randomized order (including the attention check clip which was excluded from further analysis). In the attention check clip, the agent stayed put and was far away from where the patient collided with the object. Our pre-registered exclusion criterion was to remove any participant who gave a judgment greater than 50 (the midpoint on

the scale) on this clip for the effort, badness, or responsibility question, or less than 50 for the causal question (which was reverse coded, see below). Participants watched each clip twice before indicating their response on a continuous slider. In the *Effort* condition, participants answered the question, "How much effort did Blue exert in this scenario?" with the endpoints of the slider labeled "very little" (0) and "very much" (100). In the *Badness* condition, the question was "How bad was what Blue did?" and the endpoints were labeled "not bad" (0) and "very bad" (100). In the *Responsibility* condition, participants answered the question, "How responsible was Blue for Green colliding with the Red?" with the endpoints of the slider labeled "not at all" (0) and "very much" (100). In the *Causal* condition, participants answered the counterfactual question "Would Green have collided with the Red if Blue had not been present?" and the endpoints were labeled "definitely no" (0) and "definitely yes" (100).

On average, it took participants 10.9 minutes ($SD = 4.7$) to complete the experiment.

Results

Figure 7 shows participants' judgments in the different conditions across the 20 test trials. We will discuss the results from each condition in turn.

Effort condition. Participants' effort judgments were well accounted for by the effort model with $r = .9$ and rmse = 8.47 (see Figure 8A). We used the same damping parameter in the effort model that determines how much an entity slows down if it doesn't exert effort as we did in Experiment 2.

Causality condition. Participants' causality judgments were well accounted for by the causality model with $r = .91$ and rmse = 11.55 (see Figure 8B). The best-fitting parameter for the standard deviation in the Gaussian distribution that determines how much an entity's velocity vector is rotated at each time point in the counterfactual simulations was $\sigma = 1.7$.

There were two situations in which the causality model underpredicted participants' judgments. In clips 3 and 4 (see Figure 7) participants were relatively certain that the collision between the patient and fireball would not have happened if the agent hadn't been there. However, the model only assigns a probability of $P(\text{Agent} \rightarrow O) \approx 50\%$ that the agent caused the outcome in these clips.

Table 4

Experiment 3*. Pairwise correlations between the predictors ($effort_{model}$ and $causality_{model}$) as well as between participants' mean judgments (effort, causality, responsibility, badness) for the 20 trials.*

term	$\text{effort}_{\text{model}}$	$\text{causality}_{\text{model}}$	effort	causality	responsibility
$\text{causality}_{\text{model}}$	.38				
effort	.90	.45			
causality	.22	.91	.22		
responsibility	.65	.80	.83	.62	
badness	.72	.47	.94	.22	.88

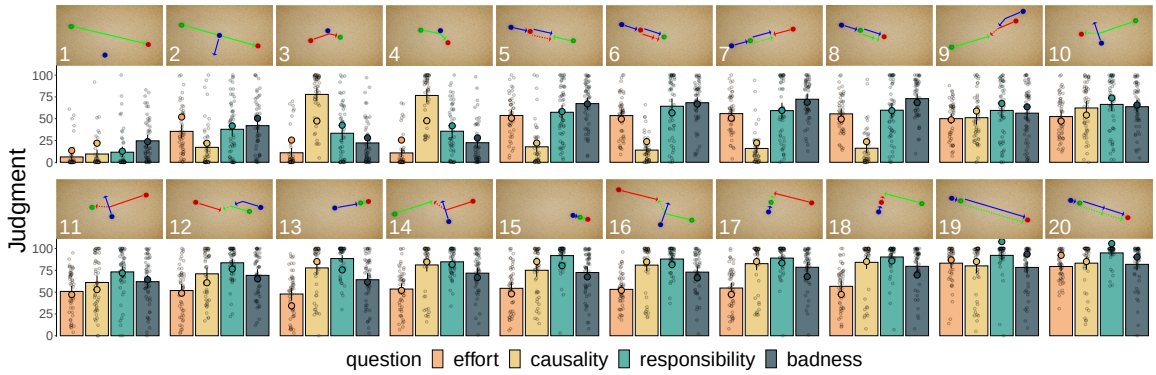

Figure 7. **Experiment 3 results**. Participants' judgments separated by question type (effort, causality, responsibility, badness) for each of the 20 video clips. Bars are mean ratings and error bars are bootstrapped 95% confidence intervals. Small points indicate individual judgments (jittered along the x-axis for visibility). Large circles indicate model predictions. The clips are ordered from lowest to highest overall judgments.

Responsibility condition

We hypothesized that when evaluating the agent's responsibility for the outcome, participants' judgments would be sensitive both to the causal role that the agent played as well as to how much the agent desired the outcome (as evidenced by the effort they exerted to bring it about). To evaluate the putative role that effort and causality played in participants' responsibility judgments, we compared three different models: one model that only considers effort as a predictor, one model that only considers causality, and one

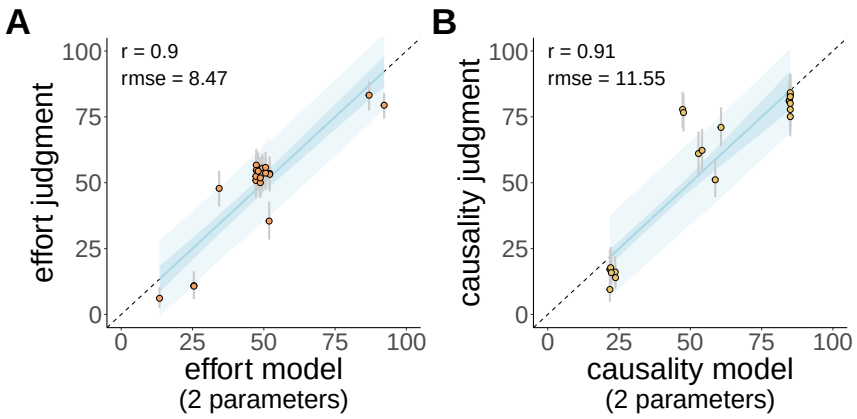

Figure 8. **Experiment 3 Effort and Causality Judgments**. Scatter plots showing the relationship between **A** the predicted effort by the model (x-axis) and participants' mean effort judgments (y-axis), and **B** the predicted causality by the model (x-axis) and participants' mean causality judgments (y-axis). Each point indicates one of the 20 trials. The error bars are 95% bootstrapped confidence intervals. The ribbons show the 95% credible interval for each model fit (dark blue), and the 50% prediction interval (light blue).

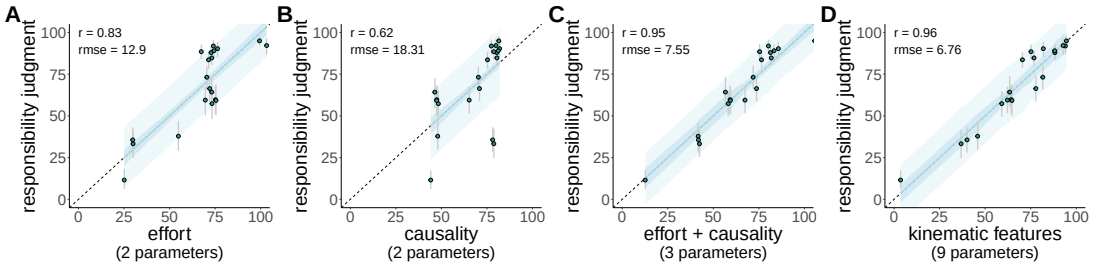

Figure 9. **Experiment 3 Responsibility Judgments.** Scatter plots showing the relationship between participants' mean moral judgments (y-axis) and versions of the *Moral Dynamics Model* (MDM) that only consider causality (**A**), only effort (**B**), or both causality and effort (**C**), or the *Moral Kinematics Model* (MKM) that uses kinematic features (**D**). Each point indicates one of the 17 trials. The error bars are 95% bootstrapped confidence intervals. The ribbons show the 95% credible interval for each model fit (dark blue), and the 50% prediction interval (light blue).

model that considers both effort and causality. Just like in Experiment 2, we also fitted the MKM which uses kinematic features, and a baseline model.

Figure 9 shows how well the different models captured participants' judgments. For the effort and causality predictor in the MDM, we used participants' mean effort and causality judgments from the other two conditions (which were highly correlated with the model effort and causality values, see Figure 8). Like in Experiment 2, we implemented the different models as Bayesian mixed effects models.

Table 5 shows a comparison between the different models. The cross-validation results reveal that the 'effort + causality' model captured participants' overall judgments best. When fitted on the individual participant level, 44 participants were best explained by a version of the MDM, 8 participants by the 'kinematic features' model[5], and 3 participants by the baseline model.

Badness judgments

We hypothesized that participants' badness judgments would be most strongly influenced by how much the agent desired to bring about the harm as indicated by the amount of effort they exerted. We expected that what causal role the agent played would matter less (see Cushman, 2008; Langenhoff et al., submitted).

We fitted the same models as we did for responsibility to participants' badness judgments. Figure 10 shows scatter plots between model predictions and participants' judgments. Table 5 summarizes and compares the model fits. The cross-validation results revealed that the effort-only model accounted for participants' overall judgments best. For the model that considers both effort and causality, the mean of the posterior distribution for the causality predictor was close to 0. On the individual participant level, 40 participants were best explained by a version of the MDM, 6 participants by the kinematic features model, and 6 participants by the baseline model.

[5]See Table A1 for how well each individual feature was correlated with participants' badness judgments.

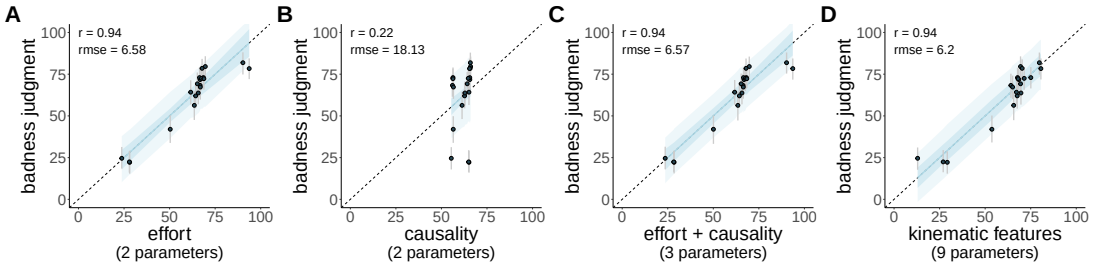

Figure 10. **Experiment 3 Badness Judgments.** Scatter plots showing the relationship between participants' mean moral judgments (y-axis) and versions of the *Moral Dynamics Model* (MDM) that only consider causality (**A**), only effort (**B**), or both causality and effort (**C**), or the *Moral Kinematics Model* (MKM) that uses kinematic features (**D**). Each point indicates one of the 17 trials. The error bars are 95% bootstrapped confidence intervals. The ribbons show the 95% credible interval for each model fit (dark blue), and the 50% prediction interval (light blue).

Discussion

Experiment 3 expanded on the prior experiments in two ways: First, we looked at a larger set of video clips that varied systematically how much effort the agent exerted, and

Table 5

Experiment 3 Responsibility and Badness Judgments. Note*: The table shows the results of fitting different Bayesian mixed effects models to capture participants' responsibility judgments (top part) and badness judgments (bottom part). The 'model' column shows what predictors were included in each model. The baseline model just includes an intercept as a predictor whereas the kinematic features model includes additional predictors (see Figure 6). 'intercept', 'effort', and 'causality' show the posterior means of each predictor together with the 5% and 95% credible interval. 'r' and 'rmse' show the Pearson's correlation and root mean square error. 'Δ elpd' shows the difference in expected log predictive density using approximate leave-one-out cross-validation between the best-fitting model (indicated by 0) and the other models together with the standard error in parenthesis. Lower numbers indicate worse performance. '# best' shows the number of participants whose judgments were best predicted by each model.*

model	intercept	effort	causality	r	rmse	Δ elpd	# best
Responsibility							
effort	18.89 [13.69, 24.26]	101.37 [91.47, 112.25]		0.83	12.90	-130.77 (18.72)	13
causality	39.31 [33.7, 44.88]		50.69 [42.81, 57.86]	0.62	18.31	-251.75 (24.69)	7
effort + causality	3.61 [-2.56, 10.69]	89.36 [76.64, 101.06]	37.38 [29.06, 45.83]	0.95	7.55	0 (0)	24
kinematic features	see Figure 6			0.96	6.76	-45.57 (15.5)	8
baseline	68.1 [65.24, 70.53]				23.33	-371.04 (26.62)	3
Badness							
effort	18.18 [12.04, 24.99]	90.75 [76.1, 105.04]		0.94	6.58	0 (0)	30
causality	54.05 [48.99, 59.63]		14.2 [8.18, 20.01]	0.22	18.13	-328.2 (28.36)	2
effort + causality	18 [10.91, 24.77]	90.09 [76.12, 103.17]	0.82 [-4.06, 5]	0.94	6.57	-0.99 (1.21)	8
kinematic features	see Figure 6			0.94	6.20	-75.7 (13.65)	6
baseline	62.15 [58.16, 66.53]				18.58	-336.86 (28.04)	6

what causal role the agent played in bringing about the outcome. Second, we asked different groups of participants for judgments of effort, causality, responsibility, and badness. The results showed that participants' judgments of effort and causality in this new set of video clips were again well accounted for by our physical simulation model. Participants' effort judgments were highly correlated with the effort that the agent actually exerted as captured by the physics engine that we used to generate the clips. We measured participants' causal impressions by asking whether the collision between Green and Red would have happened if Blue hadn't been present. Participants' counterfactual judgments were highly correlated with the predictions of our model, which runs probabilistic simulations of how the situation would have unfolded if the agent had been removed from the scene.

The results further showed that while effort is important for both judgments of responsibility and badness, causality only mattered for responsibility judgments. How much effort an agent exerted is informative about how much they desired the negative outcome. This result is consistent with prior research that has found that when people evaluate the moral wrongfulness of actions, they care most about the agent's mental states, whereas considerations of how much blame or punishment an agent receives are sensitive to the agent's causal role in addition to their mental states (Cushman, 2008; Langenhoff et al., submitted; Malle, 2021). The *Moral Kinematics Model* (MKM) which uses kinematic features also fits participants' responsibility and badness judgments. However, this model requires a larger number of parameters to account for participants' judgments, and when the models are compared via cross-validation (which takes into account both model fit and complexity), the MDM captures participants' judgments better than the MKM does.

Some of the features that the MKM uses are diagnostic for how much effort the agent exerted, such as whether or not the agent was moving, the distance that the agent traveled, and the duration of contact between the agent and the patient. As Figure 6 shows, whether the agent moved was an important predictor of participants' badness and responsibility judgments. None of the features that the MKM considers allow it to compute the causal role that the agent had in the way that the MDM does via simulating counterfactual simulation. The weights on the different predictors in the MKM for participants' responsibility judgments are less interpretable. Here, the distance that the agent travelled was a negative predictor of responsibility but the duration of contact between the agent and patient was a positive predictor (the correlation between the distance and duration predictor was $r = .75$). So while the MKM is sufficiently flexible to fit participants' judgments, compard to the MDM its predictions are less interpretable, and its parameters shift around in unexpected ways between experimental conditions. While some of the kinematic features are highly correlated with participants' effort judgments ($r(\text{distance}, \text{effort}) = .80$, and $r(\text{agent moving}, \text{effort}) = .86$), none of its features can capture the causal role that the agent played in bringing about the outcome (all r's $\leq .40$).

The differential role that causality plays for judgments of responsibility and badness was also reflected in participants' comments about what factors influenced their judgments. For example, here is a statement from a participant in the 'responsibility condition': "The two main factors I used were: Would the red and green ball hit each other if the blue ball wasn't there at all. And if the blue ball forced either the red or green ball to change direction forcing a collision." Most participants in the 'badness condition' referred to the agent's mental states like this participant here: "The seeming purposefulness of Blue's effort,

rather homicidal. Likely standing by and doing nothing as Green crashes into Red is just as bad though I didn't judge it such. So easy to slip into attributing some emotion to Blue!"

General Discussion

Moral evaluations of other's actions depend on what these actions reveal about the kind of person they are, and about the causal role that their actions played in bringing about the outcome (Gerstenberg et al., 2018). In this paper, we develop the *Moral Dynamics Model* (MDM), a computational model of moral judgment that captures both of these processes (see Figure 1).

First, the MDM infers how much an agent desired a negative outcome by considering the physical effort that the agent exerted to bring about the outcome. The MDM assumes that agents choose actions that maximize their expected utility, which is subject to the reward associated with particular world states, and the costs associated with taking actions (see Baker et al., 2017; Jara-Ettinger et al., 2016; Ullman et al., 2009). The physical effort that an agent exerted to bring about an outcome thus places a lower bound on how much the agent desired for that outcome to happen. Placing a high reward on a negative outcome (such as harming someone else) leads to a negative moral evaluation.

Second, the MDM considers what causal role the agent played in bringing about the outcome. To do so, it simulates what would have happened if the agent hadn't been present in the scene (see Gerstenberg et al., in press, 2017). The more confident the model is that the negative outcome would have been avoided if the agent hadn't been present, the more negative the moral evaluation of the agent's actions.

We tested the MDM in three experiments whose setup was closely inspired by Iliev et al. (2012). In these experiments, participants watched short video clips that show an agent interacting with a patient who comes to harm by colliding with a fireball. In Experiment 1, we successfully replicated one of the experiments from Iliev et al. using our adapted stimuli. Participants viewed pairs of video clips and had to judge in which of the two clips the agent's actions were worse. Our participants' preference matched that of Iliev et al.'s participants for all eleven trials. Participants' selections were consistent with the MDM in all but one trial. That is, participants tended to judge the agent's actions as worse in that video clip in which the agent exerted more effort to bring about the outcome. The trial in which the MDM made the wrong prediction featured one clip in which the agent pushed the patient a long distance (more effort) and another clip in which the agent pushed the patient twice (less effort). We will get back to this discrepancy between model prediction and participants' judgments in our discussion of the role of intentions below.

In Experiment 2, we quantitatively tested the MDM. This time, one group of participants judged for each video clip how much effort the agent exerted, and another group of participants judged how bad it was what the agent did. The results showed that participants' effort judgments were highly correlated with the predicted amount of effort that was derived from the physical simulation of the video clip. More importantly, participants' judgments of how bad the agent's action was were well predicted by how much effort the agent was judged to have exerted. The more effort the agent exerted, the worse its actions were judged to have been. A model that also considers what causal role the agent played in bringing about the outcome didn't improve the model fit much. However, this may have

been due to the fact that the agent's causal role didn't vary much across the clips in this experiment.

In Experiment 3, we expanded the set of video clips. Now the agent's causal role varied more such that there were some clips in which it was clear that the negative outcome would have happened even if the agent hadn't been present, some clips for which the outcome was unclear, and some clips for which it was clear that the negative outcome wouldn't have happened without the agent. In this experiment, we asked four groups of participants to judge how much effort the agent exerted, what its causal role was, how responsible the agent was for the outcome, and how bad it was what the agent did. Participants' effort and causality judgments were well-predicted by the physical simulation model. The results further showed that both effort and causality were important predictors for how responsible an agent was judged. In contrast, for judgments about badness of the agent's action, only effort mattered, not the agent's causal role. The fact that causality mattered more for judgments of responsibility than badness reflects what prior work has found. Cushman (2008) found that while the agent's mental states matter for both judgments of blame and wrongfulness, the agent's causal role matters more for judgments of blame (see also Langenhoff et al., submitted; Malle, 2021).

We compared the MDM with the *Moral Kinematics Model* (MKM) which predicts a direct mapping from the kinematic features of these video clips onto people's moral judgments (Iliev et al., 2012). For example, an agent's actions should be seen as morally worse when the agent travelled a longer distance, made contact with the patient, or when the patient wasn't already moving (see Figure 1). The MKM also fits participants' badness and responsibility judgments in Experiment 2 and 3. However, while it sometimes achieves a better fit than to the MDM in terms of correlation, when model complexity is taking into account by assessing model performance with cross-validation, the MDM provides a better account of participants' judgments than the MKM does. Further, the ways in which the weights on the different predictors in the MKM change as a function of what moral judgment participants were asked to make, is less interpretable.

As of now, the MDM only incorporates a small subset of the factors that are known to influence people's moral judgments about harmful events. Nevertheless, we believe that the work presented here takes an important step toward developing computational models of moral judgment that are grounded in people's intuitive understanding of psychology and physics. People use their intuitive theory of psychology to infer an agent's desires from how much effort it exerted, and they use their intuitive theory of physics to infer that causal role the agent played by considering how the situation would have unfolded without the agent. In the remainder we will discuss some limitations of the model as well as our experimental method and suggest how these limitations may be addressed in future work.

Video clips vs. vignettes

Most research into people's moral judgments has relied on presenting the information in written vignettes, sometimes with images that help to clarify the situation (see, e.g. Christensen, Flexas, Calabrese, Gut, & Gomila, 2014; Waldmann et al., 2012). In this paper, we have used video clips instead of vignettes to elicit moral judgments. We believe that both methodological approaches to studying moral judgment complement each other in their strengths and weaknesses.

Vignettes allow for testing people's moral intuitions in a wide variety of situations. Potentially relevant mental states such as the agent's desires, beliefs, and intentions can directly be communicated and their effect on moral judgments assessed. Arguably, written text (or speech) is also the medium in which we most frequently receive morally relevant information. A potential drawback of vignette studies is that it can be challenging to manipulate information in a fine-grained manner. For example, it can be difficult to precisely manipulate what causal role the agent played in bringing about the outcome. Furthermore, participants can only be asked to read and evaluate a relatively small number of vignettes because reading is exhausting. These constraints limit the usefulness of vignettes for testing computational theories.

Video clips, in contrast, allow for a precise manipulation of relevant factors (see, e.g. De Freitas & Alvarez, 2018; Iliev et al., 2012; Nagel & Waldmann, 2012). It's also possible to manipulate potentially relevant factors, such as the agent's causal role, without the need to explicitly communicate this information. For example, to see whether participants' judgments are sensitive to the agent's causal role, a vignette might have to explicitly stipulate what would have happened if the agent hadn't acted. When participants then take this information into account, it is unclear whether they would have also spontaneously considered the agent's causal role (even without being explicitly told). Video clips allow the researcher to manipulate potentially relevant information without directly having to tell participants. Since watching video clips is less tiring than reading text, it's possible to present participants with a greater number of scenarios which is critical for evaluating and comparing different computational models. Written or oral reports will likely continue to be one of the main forms in which people receive morally relevant information. However, the ubiquity of smartphones with cameras and the ease of sharing video clips online via social media makes it all the more important to study what inferences people draw and what judgments they make based on the video clips they see. Video clips of morally relevant actions not only appear in our social media feeds, they also play a critical role as evidence in legal trials (see Caruso, Burns, & Converse, 2016).

The video clips that we used in our experiments were not realistic. They didn't depict actual people interacting with one another. Some existing work has studied whether and how exposing participants to more realistic moral scenarios affects their judgments (Francis et al., 2016; Patil, Cogoni, Zangrando, Chittaro, & Silani, 2014; Skulmowski, Bunge, Kaspar, & Pipa, 2014; Sütfeld, Gast, König, & Pipa, 2017). This work employs virtual reality (VR) technology to investigate moral behavior in visually immersive environments that sometimes even include haptic feedback as well (see Francis et al., 2017). What's exciting about the VR paradigm is that participants can be immersed into the situation and asked to take action rather than merely making moral judgments. To what extent the general principles that the MDM is built on generalizes to more realistic settings will need to be tested in future research that combines VR technology with computational modeling.

The role of effort

The MDM predicts that people care about other's desires when morally evaluating their actions. In our setting, physical effort played an important role for how desires are inferred. We made the simplifying assumption that an agent's action costs equal the amount of physical effort it exerted. While physical effort has been shown to be an important factor

in people's moral evaluations (Bigman & Tamir, 2016), there are clearly other factors that influence the costs an agent incurred to bring about an outcome. In line with the naïve utilicy calculus (Jara-Ettinger et al., 2016), the MDM predicts that the higher the action cost is, the more the person must have desired the (negative) outcome (see Equation 2). We expect other types of perceived costs in addition to physical effort to also be relevant for inferring how much an agent desired a patient's harm. For example, an agent may take risks (Liu, Pepe, Ullman, Tenenbaum, & Spelke, 2020), forego alternative rewards, or exert great mental effort in realizing their goal (Kool & Botvinick, 2018). We expect that people take these factors into account and would, for example, judge an agent as morally worse when the action it took to harm another was riskier, even if the physical effort remained the same.

We further assumed for simplicity that the observer knows the true amount of effort being exerted by the agent. However, in reality, an observer's perception of effort may deviate from the actual amount of effort an agent exerts (Dik & Aarts, 2007). This is a minor point for the current studies, as our model's estimates of effort correlated highly with people's perceptions of effort for our stimuli, but it will be relevant for more complex stimuli where inferring effort becomes more challenging (Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). We tied effort directly to the use of force by an agent, but effort as a psychological construct may diverge from a simple summation of forces, and intuitive notions of biology and fatigue may enter the equation (Liu et al., 2017; McCoy & Ullman, 2019). As an example of this divergence, consider that a strong agent enacting a large force may be seen as exerting less effort than a weak agent, with downstream repercussions for estimating the reward of the agents.

Finally, it's also important to note that the overall amount of effort an agent exerted in a given scenario doesn't necessarily map onto the costs they incurred to bring about a particular outcome. For example, you could imagine an agent running in circles before (or after) it pushes a patient into the fireball. Would that agent's actions be any worse than those of an agent who didn't run in circles before? Intuitively, the answer is no. Only the action costs that were part of the agent's plan for bringing about the outcome should count. To determine what the agent's plan was is closely linked to the idea of having acted purposefully and intentionally. We will return to the role of intention below.

The role of causality

Moral judgments are sensitive to the causal role that the agent played in bringing about the outcome (Sloman & Lagnado, 2015; Waldmann & Dieterich, 2007). The MDM computes the agent's causal role by probabilistically simulating the counterfactual situation of what would have happened if the agent hadn't been present in the scene (see Gerstenberg et al., in press). The more likely it is that the outcome would not have happened without the agent, the more certain the model is that the agent caused the outcome.

This graded notion of the agent's causal role is appealing and it accurately captured participants' beliefs about what would have happened in the relevant counterfactual situation in Experiment 3. At the same time, it is clear that there is more to causality than this particular form of counterfactual dependence. Gerstenberg et al. (in press) showed that people's causal judgments are sensitive to different aspects of causation. It doesn't only matter to people that the candidate cause made a difference to whether or not the

outcome happened. It also matters *how* the outcome came about (see Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010). For example, launching a stationary ball through a gate is different from knocking an obstacle out of the way of an already moving ball. While in both instances, the outcome would not have happened without the candidate causal event, the way in which the outcome depends on the cause differs (Beller et al., 2020). There is a more fine-grained dependence between cause and effect in the launching case than in the obstacle removal case (see Lewis, 2000). In line with prior work on moral judgment (Greene et al., 2009; Waldmann & Dieterich, 2007), Iliev et al. (2012) had found that participants judged an agent's action to be worse when it directly intervened on the patient rather than on the object.

Currently, the MDM predicts moral judgments by combining an inference about the agent's desire with a counterfactual simulation to capture the agent's causal role. Considering what would have happened if the agent hadn't been present in the scene is only one of many possible counterfactuals. For example, one could consider what would have happened if the agent hadn't exerted any effort, or how another agent may have acted in the same situation (Gerstenberg et al., 2018). The reasonable person test is often employed in the law to assess legal liability – it asks us to evaluate whether the negative outcome would have been avoided if a reasonable person had acted instead of the defendant (Green, 1967; Lagnado & Gerstenberg, 2017; Tobia, 2018). So, a potential path for further unifying the different components in the MDM would be to explain moral judgments in terms of different counterfactuals (see Gerstenberg et al., in press), some of which may operate over physical properties of the scene (such as the presence of absence of objects) whereas others may operate over psychological properties (such as the agent's beliefs and desires).

We have shown that the agent's causal role affects participants' judgments of responsibility. Future work should investigate how action expectations and differences in exactly how the agent brought about the outcome, impact moral evaluations.

Pushing ahead

The MDM explains participants' moral evaluations by taking into account the agent's desire for harm and its causal role in bringing about the harm. What is missing? One of the key missing ingredients is the agent's intention (Malle, 2021; Mikhail, 2007; Reeder, 2009; Rosset, 2008). It not only matters whether an agent had a desire for a negative outcome, it also matters whether the agent acted in a purposeful way to realize that negative desire. We hold others more responsible when they acted intentionally versus accidentally (Lagnado & Channon, 2008), and for outcomes they indended versus ones that were unintended side-effects of their actions (Greene et al., 2009; Kleiman-Weiner et al., 2015). In Experiment 1 we saw that participants judged an agent's action as worse when it pushed the patient multiple times versus a single time for a longer distance (see Figure 3, trial 11). While the agent exerts more effort in the longer push, the double push provides particularly strong cues about the agent's intention. Inferring intentions is a non-trivial computational task (Gao, Baker, Tang, Xu, & Tenenbaum, 2019) – a task that humans are extremely good at (McEllin, Sebanz, & Knoblich, 2018). Recent work has linked intentions to plans (Bratman, 2009; Shu, Kryven, Ullman, & Tenenbaum, 2020), and defined intended outcomes as those that made a difference to an agent's plan (Kleiman-Weiner et al., 2015). Inferring intentions

via inverting an agent's plan while also acknowledging that agents may have uncertainty about what will happen in the future is an important next step in developing the model.

Our experiments only looked at situations in which a patient was harmed. Future research should also investigate situations in which agents may have positive intentions. For example, one could create simulated social dilemmas akin to the trolley dilemma (Awad et al., 2018; Thomson, 1985), in which an agent may have to weigh the costs and benefits of different actions while taking into account the uncertainty of the situation. When a negative outcome happened, this may have been intended, or it may have been the result of an action with a positive intention that failed. For example, imagine that in the setting that we've used in our experiments, a fireball is headed toward the patient, the agent pushes the patient, and the fireball and the patient collide. Maybe the agent tried to push the patient out of the way of the incoming fireball but failed? Or maybe the agent wanted to make sure that the patient is struck by the fireball? In our everyday lives we often resolve potential ambiguities about another person's intentions by drawing on prior knowledge (Kliemann, Young, Scholz, & Saxe, 2008). In our experiments, each video just featured a single interaction between agent and patient. Future work needs to investigate how people learn to infer stable traits from repeated interactions, and how these inferences guide the resolution of potentially ambiguous actions. Ambiguities also arise in situations in which an agent omits to help (Gerstenberg & Stephan, 2021; Henne, Niemi, Pinillos, De Brigard, & Knobe, 2019; Jara-Ettinger et al., 2014). Did the agent not foresee the negative outcome, did they lack the capacity to help, or did they want the negative outcome to happen?

In line with existing work (Cushman, 2008; Langenhoff et al., submitted; Malle, 2021), we showed that different moral judgments, such as how bad an agent's action was or how responsible they are for the outcome, can come apart. Whereas badness judgments were primarily driven by the agent's inferred desire, responsibility judgments were also sensitive to the agent's causal role. More research is needed to better understand the rich nexus of moral judgments, and as well as how and why different types of moral evaluations are differentially sensitive to various aspects of the situation and the mental states of the agents that were involved.

Conclusion

From walking into a messy playroom with two children brawling on the floor, to confronting an elaborate crime scene, the key questions that need answering for assigning moral responsibility are: What happened, who did what, and why did they do it? Moral judgments are based on how people understand the dynamics of the world that led to that situation, including the minds of other people. We proposed a framework for quantitatively formalizing moral judgment as an operation over intuitive theories of the world and others, bringing these two strands of research closer together. We hope this framework pushes the field closer to a comprehensive quantitative account of moral reasoning.

Code Availability

Code for all models and analyses is available at `https://github.com/cicl-stanford/moral_dynamics`

Data Availability

Anonymised participant data and model simulation data are available at `https://github.com/cicl-stanford/moral_dynamics`

Acknowledgments

This work was supported by the Center for Brains, Minds, and Machines (NSF STC award CCF-1231216) and the Office of Naval Research Science of Autonomy program (N00014-17-1-2984). We thank David Rose and Joseph Outa for feedback on the manuscript.

Author Contributions

F.A.S. and T.G. collected and analyzed the data. All authors designed the experiments and wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

Appendix

Table A1

Pearson correlation coefficient between each predictor and participants' mean judgments in Experiment 2 and 3.

	Experiment 2	Experiment 3	
predictor	badness	responsibility	badness
effort	0.90	0.83	0.94
$\text{effort}_{\text{model}}$	0.72	0.65	0.72
causality		0.62	0.22
$\text{causality}_{\text{model}}$	0.80	0.80	0.47
distance	0.54	0.43	0.61
duration	0.33	0.38	0.35
frequency	0.28	0.48	0.39
agent moving	0.92	0.74	0.88
patient moving	-0.57	-0.37	-0.17
fireball moving	0.16	0.06	0.11
collision agent patient	0.25	0.42	0.35
collision agent fireball	0.19	-0.05	-0.04

References

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

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... Rahwan, I. (2018). The Moral Machine experiment. *Nature*. doi: 10.1038/s41586-018-0637-6

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017, mar). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*(4), 0064. Retrieved from `https://doi.org/10.1038%2Fs41562-017-0064` doi: 10.1038/s41562-017-0064

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

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 3133–3139). Cognitive Science Society.

Bigman, Y. E., & Tamir, M. (2016). The road to heaven is paved with effort: Perceived effort amplifies moral judgment. *Journal of Experimental Psychology: General*, *145*(12), 1654.

Bratman, M. E. (2009, sep). Intention rationality. *Philosophical Explorations*, *12*(3), 227–241. Retrieved from `http://dx.doi.org/10.1080/13869790903067717` doi: 10.1080/13869790903067717

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Caruso, E. M., Burns, Z. C., & Converse, B. A. (2016). Slow motion increases perceived intent. *Proceedings of the National Academy of Sciences*, *113*(33), 9250–9255. Retrieved from `http://www.pnas.org/content/113/33/9250` doi: 10.1073/pnas.1603865113

Christensen, J. F., Flexas, A., Calabrese, M., Gut, N. K., & Gomila, A. (2014). Moral judgment reloaded: a moral dilemma validation study. *Frontiers in psychology*, *5*, 607.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013, Mar). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PLoS ONE*, *8*(3), e57410. Retrieved from `http://dx.doi.org/10.1371/journal.pone.0057410` doi: 10.1371/journal.pone.0057410

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral

psychological representations. *Cognitive Science*, *35*(6), 1052–1075.

Cushman, F., Young, L., & Hauser, M. (2006, dec). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science*, *17*(12), 1082–1089. Retrieved from `http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x` doi: 10.1111/j.1467-9280.2006.01834.x

De Freitas, J., & Alvarez, G. A. (2018). Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition*, *178*, 133–146.

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

Dik, G., & Aarts, H. (2007). Behavioral cues to others' motivation and goal pursuits: The perception of effort facilitates goal inference and contagion. *Journal of Experimental Social Psychology*, *43*(5), 727–737.

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

Firestone, C., Scholl, B. J., et al. (2016). Moral perception'reflects neither morality nor perception. *Trends in Cognitive Sciences*, *20*(2), 75–76.

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

Foot, P. (1978). *Virtues and vices and other essays in moral philosophy*. Oxford University Press.

Francis, K. B., Howard, C., Howard, I. S., Gummerum, M., Ganis, G., Anderson, G., & Terbeck, S. (2016). Virtual morality: Transitioning from moral judgment to moral action? *PloS one*, *11*(10), e0164374.

Francis, K. B., Terbeck, S., Briazu, R. A., Haines, A., Gummerum, M., Ganis, G., & Howard, I. S. (2017). Simulating moral actions: An investigation of personal force in virtual moral dilemmas. *Scientific Reports*, *7*(1), 1–11.

Gao, T., Baker, C. L., Tang, N., Xu, H., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive Science*, *43*(8).

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154–179.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gershman, S. J., Gerstenberg, T., Baker, C. L., & Cushman, F. (2016). Plans, habits, and theory of mind. *PLoS ONE*, *11*(9), e0162246.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (in press). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*.

Gerstenberg, T., & Icard, T. F. (2019). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from `https://doi.org/10.1177%2F0956797617713053` doi: 10.1177/0956797617713053

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by

omission. *PsyArXiv*. Retrieved from https://psyarxiv.com/wmh4c/

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Gerstenberg, T., Ullman, T. D., Nagel, J., Kleiman-Weiner, M., Lagnado, D. A., & Tenenbaum, J. B. (2018, August). Lucky or clever? from expectations to responsibility judgments. *Cognition*, *177*, 122-141. doi: 10.1016/j.cognition.2018.03.019

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). MIT Press.

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

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009, jun). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, *111*(3), 364–371. Retrieved from http://dx.doi.org/10.1016/j.cognition.2009.02.001 doi: 10.1016/j.cognition.2009.02.001

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.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., ... Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, *48*(3), 829–842.

Hamlin, J. K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Developmental Science*, *16*(2), 209–226. doi: 10.1111/desc.12017

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

Hamlin, J. K., Wynn, K., Bloom, P., & Mahajan, N. (2011). How infants and toddlers react to antisocial others. *Proceedings of the national academy of sciences*, *108*(50), 19931–19936.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Henne, P., Niemi, L., Pinillos, Á., De Brigard, F., & Knobe, J. (2019). A counterfactual explanation for the action effect in causal judgment. *Cognition*, *190*, 157–164. doi: 10.1016/j.cognition.2019.05.006

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*(5), 822–851.

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., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(10), 785. Retrieved from https://doi.org/10.1016%2Fj.tics.2016.08.007 doi: 10.1016/j.tics.2016.08.007

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.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., & Tenenbaum, J. B. (2015). Inference of intention and permissibility in moral decision making. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 1123–1128). Austin, TX: Cognitive Science Society.

Kleiman-Weiner, M., Shaw, A., & Tenenbaum, J. B. (2017). Constructing social preferences from anticipated judgments: When impartial inequity is fair and why? In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 676–681). Austin, TX: Cognitive Science Society.

Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008, oct). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957. Retrieved from `http://dx.doi.org/10.1016/j.neuropsychologia.2008.06.010` doi: 10.1016/j.neuropsychologia.2008.06.010

Kool, W., & Botvinick, M. (2018, sep). Mental labour. *Nature Human Behaviour*.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017, oct). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. Retrieved from `https://doi.org/10.1016%2Fj.tics.2017.06.002` doi: 10.1016/j.tics.2017.06.002

Kurniawan, I. T., Seymour, B., Talmi, D., Yoshida, W., Chater, N., & Dolan, R. J. (2010). Choosing to make an effort: The role of striatum in signaling physical effort of a chosen action. *Journal of Neurophysiology*, *104*(1), 313-321. Retrieved from `https://doi.org/10.1152/jn.00027.2010` (PMID: 20463204) doi: 10.1152/jn.00027.2010

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

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In M. Waldmann (Ed.), *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

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

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (submitted). Predicting responsibility judgments from dispositional inferences and causal attributions.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Lewis, D. (2000). Causation as influence. *The Journal of Philosophy*, *97*(4), 182–197.

Liu, S., Pepe, W., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2020). Dangerous ground: Thirteen-month-old infants are sensitive to peril in other people's actions. *PsyArXiv*. Retrieved from `https://psyarxiv.com/rvydk/`

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Luo, Y., & Baillargeon, R. (2005, 09). Can a self-propelled box have a goal? psychological reasoning in 5-month-old infants. , *16*, 601-8.

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318.

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

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014b, Apr). A theory of blame. *Psychological Inquiry*, *25*(2), 147-186. Retrieved from http://dx.doi.org/10.1080/1047840x.2014.877340 doi: 10.1080/1047840x.2014.877340

McCoy, J., & Ullman, T. (2019). Judgments of effort for magical violations of intuitive physics. *PloS one*, *14*(5), e0217513.

McEllin, L., Sebanz, N., & Knoblich, G. (2018). Identifying others' informative intentions from movement kinematics. *Cognition*, *180*, 246–258.

Michotte, A. (1946/1963). *The perception of causality*. Basic Books.

Mikhail, J. (2007, apr). Universal moral grammar: theory, evidence and the future. *Trends in Cognitive Sciences*, *11*(4), 143–152. Retrieved from http://dx.doi.org/10.1016/j.tics.2006.12.007 doi: 10.1016/j.tics.2006.12.007

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 34th annual conference of the cognitive science society* (pp. 785–790). Austin, TX: Cognitive Science Society.

Patil, I., Calò, M., Fornasier, F., Cushman, F., & Silani, G. (2017). The behavioral and neural basis of empathic blame. *Scientific reports*, *7*(1), 5200.

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*(1), 94–107.

Pizarro, D. A., Uhlmann, E., & Bloom, P. (2003). Causal deviance and the attribution of moral responsibility. *Journal of Experimental Social Psychology*, *39*(6), 653–660.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

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

Rosset, E. (2008). It's no accident: Our bias for intentional explanations. *Cognition*, *108*(3), 771–780.

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*(2), 165–184. Retrieved from http://dx.doi.org/10.1023/a:1019923923537 doi: 10.1023/a:1019923923537

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press, Princeton NJ.

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment. *Social perception: Detection and interpretation of animacy, agency, and intention*, *4629*.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.

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

Shu, T., Kryven, M., Ullman, T. D., & Tenenbaum, J. B. (2020). Adventures in flatland: Perceiving social interactions under physical dynamics. In *42d proceedings of the annual meeting of the cognitive science society*.

Skulmowski, A., Bunge, A., Kaspar, K., & Pipa, G. (2014). Forced-choice decision-making in modified trolley dilemma situations: a virtual reality and eye tracking study. *Frontiers in behavioral neuroscience*, *8*, 426.

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). Elsevier.

Sloman, S. A., & Lagnado, D. (2015). Causality in thought. *Annual Review of Psychology*, *66*(1), 223–247. Retrieved from `http://dx.doi.org/10.1146/annurev-psych-010814-015135` doi: 10.1146/annurev-psych-010814-015135

Smith, K. A., & Vul, E. (2012). Sources of uncertainty in intuitive physics. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society.*

Sütfeld, L. R., Gast, R., König, P., & Pipa, G. (2017). Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. *Frontiers in behavioral neuroscience*, *11*, 122.

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

Thomson, J. J. (1985, may). The trolley problem. *The Yale Law Journal*, *94*(6), 1395. Retrieved from `http://dx.doi.org/10.2307/796133` doi: 10.2307/796133

Tobia, K. P. (2018). How people judge what is reasonable. *Alabama Law Review*, *70*, 293–359.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. Retrieved from `https://doi.org/10.1016%2Fj.tics.2017.05.012` doi: 10.1016/j.tics.2017.05.012

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, pp. 1874–1882).

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413-1432. doi: 10.1007/s11222-016-9696-4

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.

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

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

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, *104*(20), 8235–8240.

Young, L., & Saxe, R. (2008, 10). An fmri investigation of spontaneous mental state inference for moral judgment. , *21*, 1396-405.