Moral Dynamics

Grounding Moral Judgment in Intuitive Physics and Intuitive Psychology

Felix A. Sosa*
Department of Brain and Cognitive Sciences, MIT

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 the protagonist's actions and thoughts are explicitly communicated. Recent studies have begun to employ visual stimuli, and some have postulated a direct link from processing visual features to making moral judgments. We embrace the advent of visual stimuli in moral psychology, but believe that the connection between visual processing and moral judgments is mediated by an inference about what the observed action reveals about the agent's mental states. We formalize moral judgments as computations over an intuitive theory of physics combined with an intuitive theory of mind. Knowing that mental states lead to action (e.g., the belief that someone is in harm's way and the desire to help them stimulates a decision to shove them out of harm's way), and that these actions are constrained by physics (the shove has to be forceful enough, aimed in the right direction, timed appropriately, etc.), allows an observer to make powerful inferences about moral responsibility. Two experiments show that this model captures moral judgments about physical scenes, both qualitatively and quantitatively.

Keywords: moral judgment, effort, intuitive physics, intuitive psychology

*Corresponding author: Felix Sosa (felixanthonysosa@gmail.com).

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? 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 (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), and (at slightly older ages) will act to punish morally bad actors (Hamlin, Wynn, Bloom, & Mahajan, 2011). Recent work in neuroscience has shown selective activation in the posterior superior temporal sulcus when viewing agent animations, with dissociable responses for goal directed action by individual actors and social interactions (Isik, Koldewyn, Beeler, & Kanwisher, 2017; Vander Wyk, Hudac, Carter, Sobel, & Pelphrey, 2009).

In cognitive science, there is a long tradition of attempting to formally link perception and psychological attributions (such as intention) 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 specifically that moral judgments can be explained by the visual processing of kinematic features, such as the velocity of a car hitting a man, or the distance a person traveled to push someone into harm's way (Caruso, Burns, & Converse, 2016; De Freitas & Alvarez, 2018; Iliev, Sachdeva, & Medin, 2012; Nagel & Waldmann, 2012). In line with the fast, automatic, early-developing, and consistent nature of these judgments, these accounts propose a direct mapping from visual features to moral judgments (such that, for example, traveling longer distances to harm maps onto morally worse judgments by others).

However, a great deal of prior work on moral judgment has focused on the deliberative and abstract components that go into a moral calculation. This line of research (which often relies on carefully written vignettes rather than visual stimuli) has demonstrated that both a person's causal role, and the person's inferred mental states are key determinants of moral judgments (Cushman, 2008; Gerstenberg et al., 2018; Lagnado & Gerstenberg, 2017; Lagnado, Gerstenberg, & Zultan, 2013; Malle, Guglielmo, & Monroe, 2014; Patil, Calò, Fornasier, Cushman, & Silani, 2017; Shaver, 1985; Waldmann, Nagel, & Wiegmann, 2012; Weiner, 1995; Young, Cushman, Hauser, & Saxe, 2007; Young & Saxe, 2008): People judge a person more severely when that person caused the bad outcome (Alicke, 1992; Cushman, 2008), and when that person intended to cause a bad outcome (Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum, 2015; Lagnado & Channon, 2008). Moral judgments are not only sensitive to whether someone caused or intended an outcome, but also to the way the outcome was brought about (Jara-Ettinger, Kim, Muentener, & Schulz, 2014; Waldmann & Dieterich, 2007). While moral dilemmas presented as vignettes, such as the trolley problem, are a rich source for empirical exploration (Foot, 1978; Waldmann et al., 2012), they have their limitations. Vignettes may fail to trigger relevant perceptual processing and related downstream processes, similar to asking someone to solve a physics problem involving trajectories with pen and paper, instead of throwing a ball at them and asking them to catch it.

These two approaches to formalizing moral judgments, one focusing on perceptual processing and the other on inferences of cause and intention, can seem incongruous. But a full model will have to incorporate both. Here, we propose such a synthesis. We believe that the route from visual features to moral judgments is mediated by people's intuitive understanding of how the world works, one that encompasses both an intuitive theory of mind (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 and the causal structure of a scenario. It is through this understanding of scenarios that people relate observed physical actions to mental states such as beliefs, desires, and intentions (Baker, Jara-Ettinger, Saxe, & Tenenbaum, 2017; Baker, Saxe, & Tenenbaum, 2009; Battaglia et al., 2013; Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), and evaluate the causal roles that physical actions played in bringing about the outcome (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017). Our computational synthesis is sensitive to visual features and allows for fast and automatic processing, but not because such features are mapped directly to moral judgments. Rather, these features are indicative of the agent's mental state, and it is these mental states that are the input to the moral calculus.

The rest of the paper is organized as follows. We first describe a model of moral judgment operating over an intuitive theory of mind and an intuitive theory of physics. We then examine this model using two empirical studies. In Experiment 1, we replicate an experiment that links visual cues to moral judgment (Iliev et al., 2012), and show how our model accounts for the results by inferring the agent's desire to do harm via the physical effort it exerted. In Experiment 2, we test participants' moral intuitions in a wider range of situations, and elicit graded judgments which provide a stronger test for the model's predictions. We discuss the implications of our findings, limitations of our current model, as well as a roadmap for addressing these limitations.

Moral Dynamics Model

Our model connects perceptual depictions of an agent's actions to reasoning about the underlying mental states of the agent. The model combines ideas from recent formalizations of intuitive psychology (for reasoning about hidden mental states given actions) and intuitive physics (for reasoning about cost as physical effort). Following prior structured generative approaches to intuitive psychology (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. Following recent prior work on intuitive physics (e.g. Battaglia et al., 2013; Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2012; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018), we constrain this generative process to obey noisy Newtonian mechanics: actions correspond to forces applied by an agent to a patient, where the amount of force is monotonically related to the agent's effort.

As an overview, our "Moral Dynamics" model infers an agent's utilities, and predicts that people's negative moral judgments are related to how much an agent desires to harm a patient. We also assume that agents act to achieve desired rewards, and that actions are associated with a cost in the form of physical effort. Given that 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 expends as indicative of the value the agent places on achieving an outcome. If an agent undertakes a great cost 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. We next discuss in more detail the theoretical background and implementation of the framework.

Computational Framework

We model intuitive psychological reasoning using Bayesian Theory of Mind (see e.g. Baker et al., 2017, 2009; Jara-Ettinger et al., 2016). This framework assumes that people think of others as goal-directed agents who choose actions to maximize their expected reward, subject to their beliefs, constraints, and abilities (see also Gershman, Gerstenberg, Baker, & Cushman, 2016). The underlying notion that people use a 'principle of rationality' to reason about the mental states of others has a long history (Dennett, 1987), and the recent avenue of work in Bayesian Theory of Mind has shown how to use this principle to quantitatively capture human reasoning about mental states. For the purposes of our model, we limit ourselves to a version of the framework dubbed 'the Naive Utility Calculus' (Jara-Ettinger et al., 2016), according to which 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, a combination of the reward R derived from world state s, and the cost C of taking action a.

To this underlying framework we add the following three simple assumptions: 1) We limit ourselves to cases in which the cost C for taking an action is proportional to the physical effort necessary to take that action. 2) We assume that an agent's reward can depend on the utility of another agent (cf. Ullman et al., 2009). 3) We assume an observer's moral evaluations are proportional to the inferred reward that the agent derives from helping or hindering the patient (cf. Gerstenberg et al., 2018). We consider each of these assumptions in turn, and then show how their combination leads to an account of moral judgment from visual scenes.

Physical effort. Physical effort features prominently in decision making and moral judgment (Jara-Ettinger et al., 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 (Jara-Ettinger et al., 2014; Liu, Ullman, Tenenbaum, & Spelke, 2017).

The rationale of using cost to infer utility and through that make moral judgments carries through with other types of cost as well (such as risk or mental effort), but here we limit ourselves to inferences about physical effort. We formalize physical effort in terms of Newtonian dynamics, which are broadly consistent with human intuitive physical reasoning (Battaglia et al., 2013; Bramley, Gerstenberg, Tenenbaum, & Gureckis, in press; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016; Sanborn et al., 2013; Ullman et al., 2017, 2018). As detailed in the Appendix, we model physical effort as the amount of force that an agent expended to bring about the outcome. Importantly, when considering how much effort an agent took to harm another, we only count the physical effort used to achieve that goal. For example, if an agent ran around in circles before or after taking intentional action to harm another agent, we would not count that effort associated with running in circles as effort towards accomplishing its goal.

Helping or harming. We label the utility of the patient as U_P, and the reward the agent receives for helping or harming the patient as $R_A(U_P)$. A pro-social attitude (i.e. high reward for helping) between the agent and patient can be captured as a positive relationship between R_A and U_P and an anti-social attitude (i.e. high reward for harming) between the agent and patient can be captured as a negative relationship. If a pro-social relationship exists, whatever states and actions increases the patient's utility will also increase the agent's reward, and the agent will take actions to move the patient into high-reward states or reduce the patient's costs (modulated by the agent's own costs). This simplified model of 'helping and hindering' can quantitatively account for people's reasoning about social goals (Ullman et al., 2009), and the choice patterns of pre-verbal infants (Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013).

From inferred desires to moral judgments. We assume that moral judgments depend on people's beliefs about an agent's desires. That is, people will judge the agent more negatively, in proportion to the inferred reward that the agent derives from harming an innocent other.

Putting these assumptions together, we model people's negative moral judgments about an agent A, $J(A)$, as being proportional to the inferred positive reward A receives for the negative outcome utility of patient P, $R_A(U_P)$, which can be approximated by the amount of effort that A was willing to exert to bring about that outcome. The effort that A exerted in a scenario is defined as the sum of the costs A incurred c_A for taking some action a_t at every time point in the scenario t:

$$J(A) \approx R_A(U_P) \propto \sum_{t=0}^{T} c_A(a_t). \tag{2}$$

We made a number of simplifying assumptions in this calculation. In general, psychological costs encompass more than physical effort, such as time delay or mental effort (Kool & Botvinick, 2018). Also, physical effort is not just an integral of the force generated over time, but subject to biological notions of expendable energy and fatigue (Hills, Mokhtar, & Byrne, 2014). Pro-social and anti-social relationships are more than just a utility-to-reward transformation, and moral evaluations depend on more than the inferred social relationship between agents (Waldmann et al., 2012).

Nonetheless, we believe this framework is flexible enough to capture many naturalistic decision problems, and it provides the core mechanics that future work can build on, adding in more varied notions of cost, effort, and relationships. Because the notions of force and effort play a central role in our model, we refer to it as *Moral Dynamics* model,

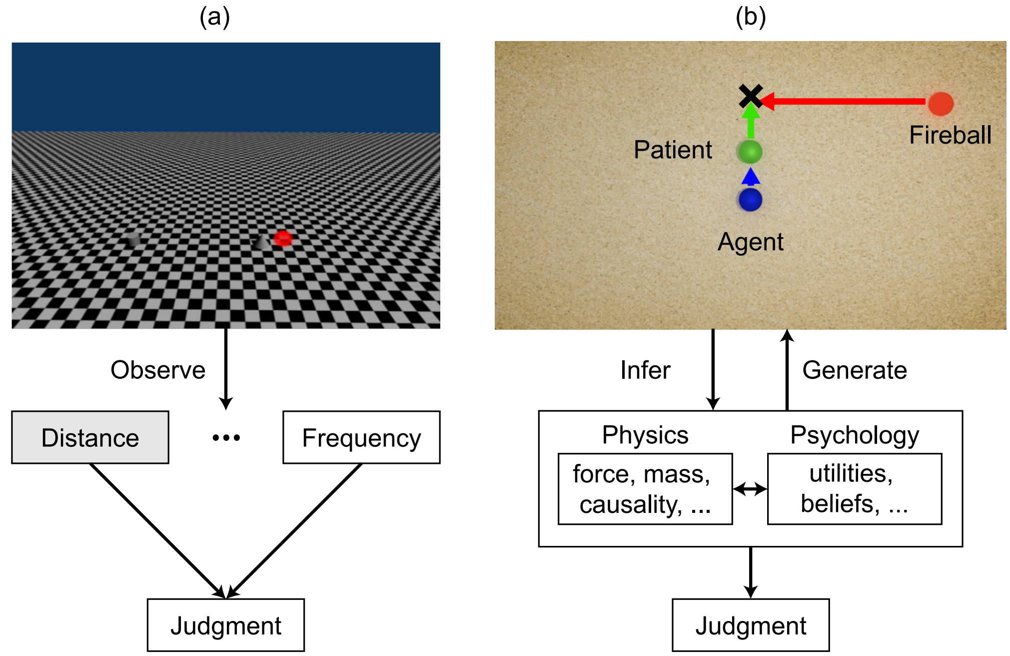


Figure 1. **(a)** Example stimuli from Iliev et al. (2012), and representation of their theory of moral judgment. An observer extracts relevant features of a scene (e.g. distance in the above example) and uses those features directly to form their moral judgment (e.g. greater distance means more negative judgment). **(b)** Example stimuli from our Experiments 1 and 2, and a representation of the *Moral Dynamics* model. An observer infers latent variables related to physics and psychology. Specifically, observers infer the utility an agent attaches to helping or harming by inferring the effort they expended in the scene. These variables inform their moral judgment. Each video shows an agent (blue), patient (green), and fireball (red). Patients cannot see fireballs and are burned by them, agents can see fireballs and are not burned by them. Lines indicate trajectories in the video, 'X' marks the collision of the patient and fireball.

in contrast with a model termed *Moral Kinematics* by Iliev et al. (2012), which predicts moral judgments based on perceptual/kinematic features such as distance, angle, contact, and velocity.

Model Implementation and Domain

We consider a simple domain, based on Iliev et al. (2012), in which different visual scenarios show agents interacting with, and potentially harming, other agents, while exerting physical forces.[1] Figure 1(b) illustrates our experimental setup. In each scenario, a video shows an agent, patient, and fireball. In this domain, the agent can perceive the fireball and is not harmed upon contact with it, while the patient cannot perceive the fireball and is harmed upon contact with it.

[1]We used the 2D physics engine Pymunk (www.pymunk.org) to generate the scenarios in our experiments, and the 3D physics engine Blender (www.blender.org) to render them.

For each agent, the cost for taking a set of actions, C_A, is the sum of the forces that the agent generates on itself (cf. Luo & Baillargeon, 2005). Specifically, at each discrete time step, t, an agent applies an impulse, I_t, to itself. An agent's effort at that time step is proportional to the magnitude of that impulse.

Given this domain, we can use different starting conditions and trajectories to vary the amount of effort an agent is perceived to expend to harm a patient. According to our model, differences in the inferred amounts of effort will translate into different moral evaluations.

Experiment 1

Our first experiment seeks to qualitatively test the computational model developed above, and examine whether participants' judgments can be explained by assuming that they infer the reward the agent has for harming the patient via the amount of effort the agent exerted. Our experiment was closely modeled after Experiment 2 in Iliev et al. (2012). So, an additional goal for this experiment is to verify that our stimuli elicit similar responses to Iliev et al. (2012), so that meaningful comparisons can be drawn when we later expand the stimulus set.

Methods

Participants. 46 participants ($M_{age} = 34.5$, $SD_{age} = 10.4$, 11 female) were recruited via Amazon Mechanical Turk, and compensated for their time. Both Experiment 1 and 2 were run using Psiturk (Gureckis et al., 2016).

Stimuli. In Iliev et al.'s (2012) Experiment 2, participants saw pairs of videos. Each pair tested the effect of a kinematic feature on moral judgment. The videos in each pair differed with respect to at least one of the following factors: The distance the harming agent traveled; whether the agent made contact with the patient; how many times the agent touched the patient; how long the agent made contact with the patient; the force the agent exerted on the patient.

We focused on the video pairs whose physical dynamics could be captured in our 2D, top-view physics engine implementation, and used 9 of the original 15 pairs.[2] The 9 included video pairs were tailored to be similar to the original stimuli used by Iliev et al. (2012), with minor differences beyond the 3D-view vs 2D-top-view (see Figure 1 (a) and (b)): In the original experiment, the agent was a white cylinder, the patient was a white cone, and the floor of the scene was checkerboard. In our videos, the agent was a blue sphere, the patient was a green sphere, and the floor of the scene was visually similar to sand. In both the original experiment and in our stimuli the fireball is a red sphere.

[2]The stimuli that were left out can all be theoretically incorporated into the model, but involve complications that are not relevant for the question at hand. The 6 videos left out were: Three pairs with 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 stimuli uniform in viewpoint); two pairs with agents sliding for long distances after minor collisions (which would require either near-zero friction, very strong agents, or a very low velocity patient after the collision); one pair with an agent that entered the scene from outside the frame (requiring inference over the unseen physics that led to its arrival).

Design and Procedure. As in Iliev et al. (2012), 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 Green collided with the fireball, and that their task was to judge in which video Blue's actions were worse.

Participants then viewed a set of familiarization videos that showed Blues, Greens, and fireballs interacting. The familiarization videos informed participants that Blues and Greens were intelligent, social creatures, and that fireballs were inanimate objects. As in Iliev et al. (2012), participants were told that fireballs 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, there were some reported instances in which Blues harmed Greens. Participants were told that they would see such instances, and would be asked to evaluate what Blue did. Before starting the experiment, participants were required to pass a comprehension check.

The comprehension check ensured participants knew 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.

During the experiment, each participant was shown 9 pairs of videos. The order of the video pairs was randomized. When viewing a pair, 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 counterbalanced across participants. 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.

Results

To best compare our results to those of Iliev et al. (2012), we followed their analysis procedure and binarized participant responses. Responses were coded as 1 if the video in column A of Figure 2 was judged as worse, and 0 otherwise. Iliev et al. (2012) originally found that, based on kinematic features, the agent's actions in videos in column A were predicted and judged as worse than those in column B. Figure 2 shows the percentage of participants that marked the video in column A as worse.

Figure 2 also shows the participants' responses to the equivalent stimuli in Iliev et al. (2012), and the predictions of our model. We used Luce's choice rule as described in Equation 3 to transform the continuous model predictions into a probability of choosing one video over another.

$$P(A) = \frac{\text{Effort}_A}{\text{Effort}_A + \text{Effort}_B} \tag{3}$$

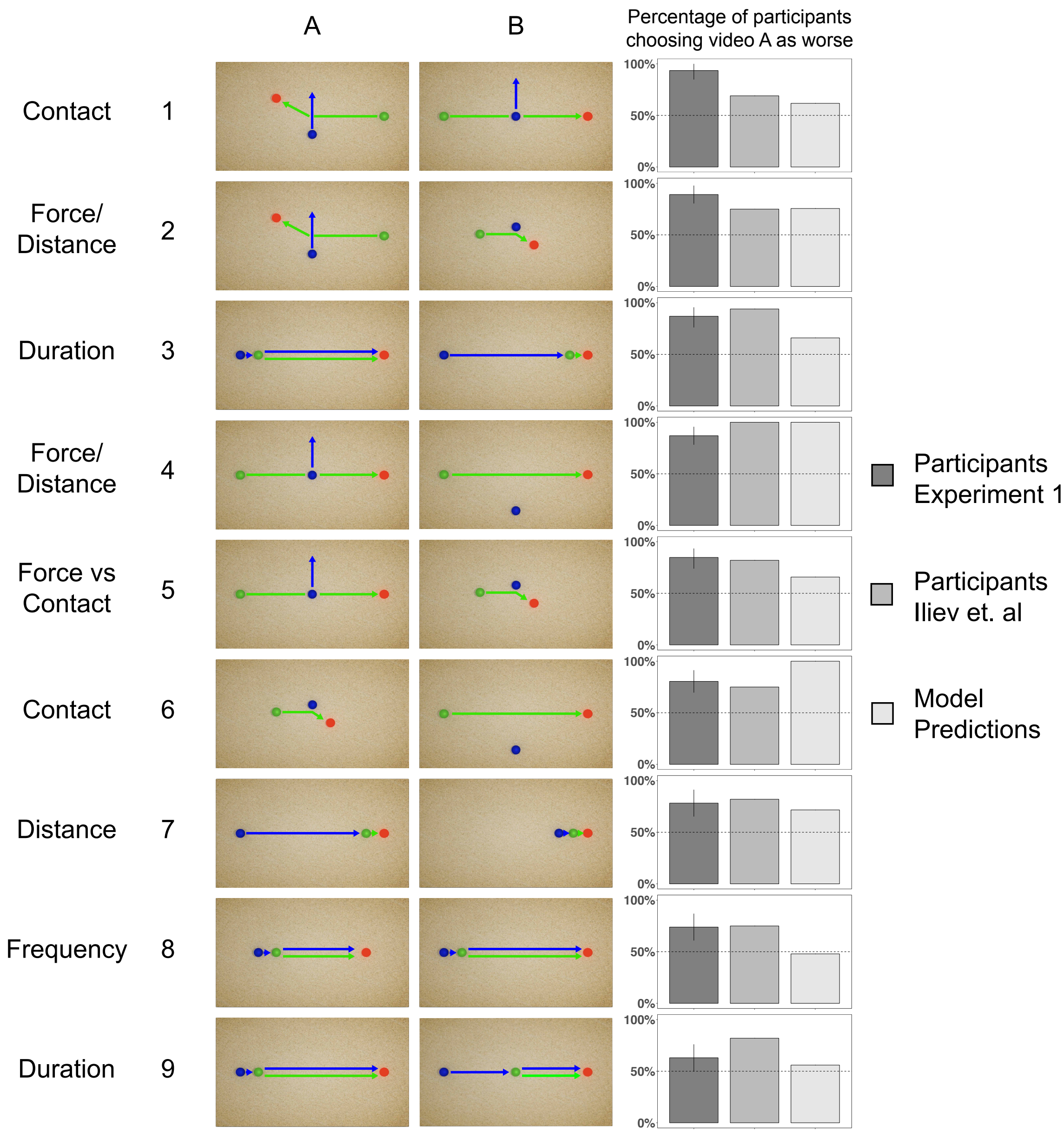


Figure 2. **Experiment 1 stimuli and results**. Each row shows 2 schematics of the videos shown to participants (A and B), as well as the percentage of participants who judged that what Blue did was worse in A compared to B in our replication (Experiment 1), in the original study (Moral Kinematics), and according to the model. Each pair differed with respect to a kinematic feature(s), listed to the left of each pair. The pairs are in descending order of percentage of participants choosing video A as worse. *Note*: The error bars indicate bootstrapped 95% confidence intervals.

Our results replicated the original results reported in Iliev et al. (2012). Across the 9 pairs, a majority of participants judged the agent in video A as being worse. While there are small quantitative discrepancies between our results and what Iliev et al. (2012) found (e.g. we found a stronger preference for A in pair 1 compared to Iliev et al.), we attribute these differences to sampling noise (Iliev et al.'s (2012) Experiment 2 featured only 16 participants). For 8 out of the 9 video pairs, our model predicted the preference found in both our experiment and in Iliev et al. (2012). In pair 8, our model exhibited a (very) slight preference for B over A.

Discussion

The results of our experiment closely replicate what Iliev et al. (2012) found. For each video pair in our study, participants judged the agent's action to have been worse in A compared to B. This successful replication suggests that our stimuli elicit similar moral intuitions, despite being visually somewhat different from the ones used by Iliev et al. (2012).

The *Moral Dynamics* model correctly predicted participants' preference in 8 out of 9 video pairs. Instead of postulating a set of visual and kinematic features that influence people's judgments (see Figure 2 leftmost column), the *Moral Dynamics* model predicted this preference solely based on the effort the agent expended in each video which is diagnostic for how much reward the agent placed on the patient's harm.

Still, the model as realized has several limitations. For example, so far, the model does not try to infer an agent's intention. In video 8A, the agent pushes the patient twice, whereas in video 8B the agent pushes the patient only once, but all the way to the fireball. The two-push scenario provides salient evidence for the agent's intention to harm the patient. A plausible interpretation of what happens in 8A (on the part of a human observer) is that the agent realized that its first push wasn't sufficient to achieve the goal of harming the patient, and then it decided to push again. In 8B, the agent's movement is compatible with a desire to just go in that direction while the patient happens to be in the way. Since the agent's actions are such that it expended almost identical effort in 8B (the long push) and 8A (the double push), our model predicts that participants should have not clear preference in this case.

An additional limitation of the model is that it relies on an estimation of effort that is directly related to the force used in the physics simulation, while participants' own estimations of effort and force may deviate from the underlying dynamics in various ways. For example, in our implementation staying still following a collision requires the active use of an opposing force to cancel out the impact, while people may perceive this as simply the agent staying put (see also stimuli 6A and 6B). In the next experiment, we directly address the use of ground truth effort on the part of the agent by asking participants to judge how much effort the agent exerted. We also expand the number of test cases in order to carry out a more quantitative evaluation of the model.

Experiment 2

For Experiment 2, we turn to a quantitative examination of our model against people's judgments, expanding on the original stimulus set, and again having people judge the relative moral badness of different agents' actions. The expanded stimulus set includes seven additional videos based on the first experiment conducted by Iliev et al. (2012). There, they examined the effect of movement and intervention on moral judgment: whether the agent intervened on the patient or on the fireball to harm the patient and whether the agent, patient, or fireball were moving before the intervention. We added these videos to our set of stimuli and tested whether these additional kinematic features were captured by our model. Since the *Moral Dynamics* model goes from the inferred effort that the agent exerted to how much reward the agent placed on harming the patient, we tested in a separate condition, whether participants' estimate of how much effort the agent exerted was accurately captured by the model.

Methods

Participants. 83 participants ($M_{age} = 35.7$, $SD_{age} = 12.7$, 42 female) were recruited via Amazon Mechanical Turk.

Design and Procedure. Participants were randomly assigned to the *Effort* condition ($N = 42$), or the *Moral* condition ($N = 41$).

The instructions and familiarization videos were largely identical to those of Experiment 1. In both conditions, participants viewed the same videos with slight modifications depending on the condition. This time, instead of pairs of videos in the test phase, participants only viewed a single video at a time. 17 test videos were presented in randomized order.

Participants watched each video twice before being asked to indicate 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 *Moral* condition, the question was "How bad was what Blue did?" and the endpoints were labeled "not bad" (0) and "very bad" (100).

Results

The empirical results of Experiment 2 for both conditions are summarized in Figure 3, showing a schematic of the video stimuli, participants' effort and moral judgments for each individual video, together with the model's predictions. Figure 4(b) and (c) show the fitted effort values from the physics engine against the mean effort and moral judgments, respectively. The model was fitted using separate linear regressions for each condition.

Participants' judgments of effort were closely aligned with the effort values from our model, Spearman's $\rho = .96, p < .001, 95\%$ CI $[.93, .98]$ (see Figure 4b). Further, the mean participant judgments for how much effort the agent exerted in each video positively correlated with participant moral judgments for corresponding videos, $\rho = .68, p = .003, 95\%$ CI $[.55, .80]$ (see Figure 4a). The *Moral Dynamics* model provided a similarly good fit to participants' moral judgments (see Figure 3 light bars as well as Figure 4c) with $\rho = .66, p = .004, 95\%$ CI $[.56, .74]$, 95% CI $[.56, .74]$. As a reminder, the model predicted participants' moral judgments based on how much effort the agent exerted, which is diagnostic for how much reward the agent placed on harming the patient.

Discussion

The results of Experiment 2 support the idea that judgments of physical effort are important for moral judgments in these visual, dynamic scenarios. The *Moral Dynamics* model explains these judgments in terms of an overarching framework rather than postulating a collection of features. Both people's judgments of effort and the effort values from

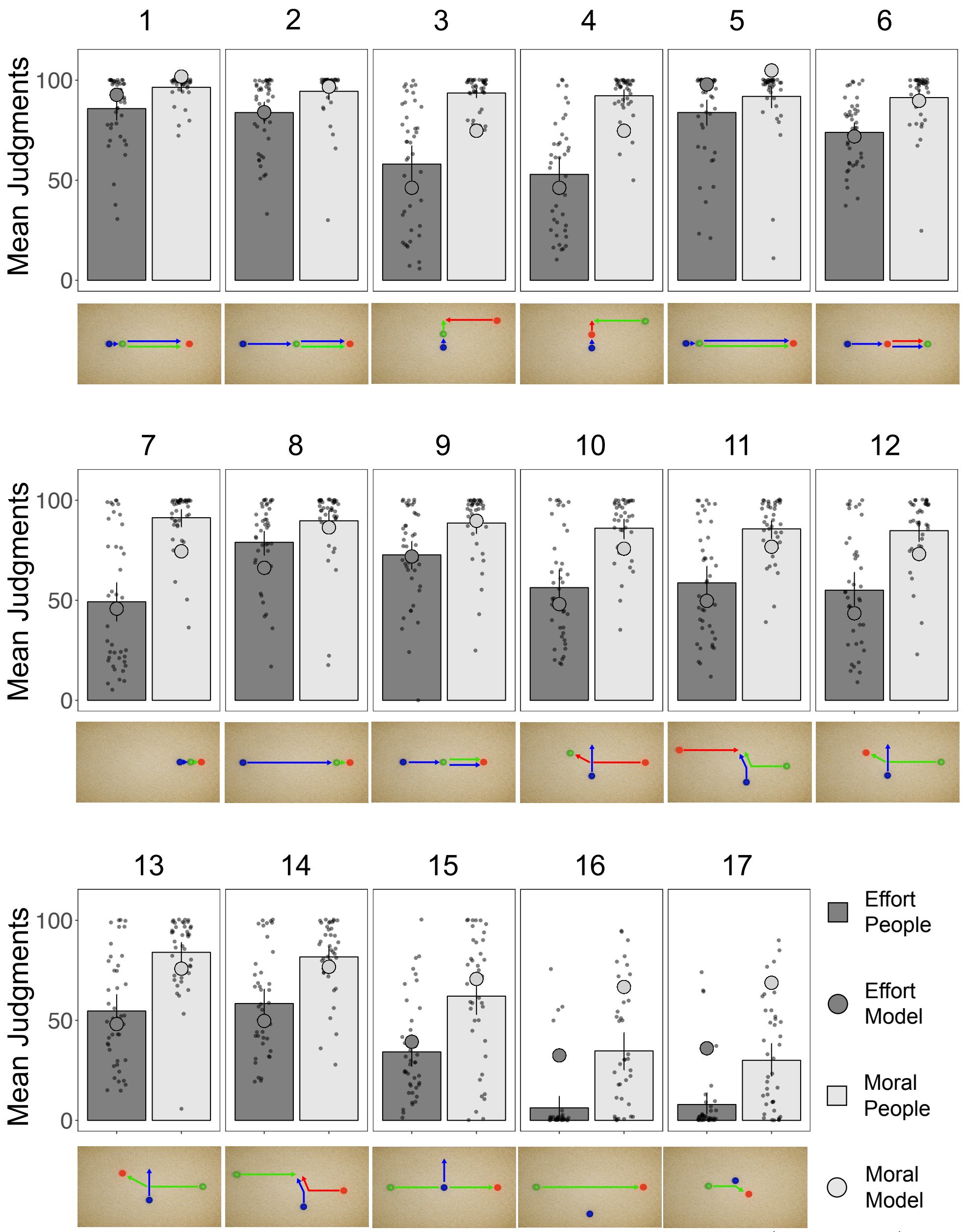
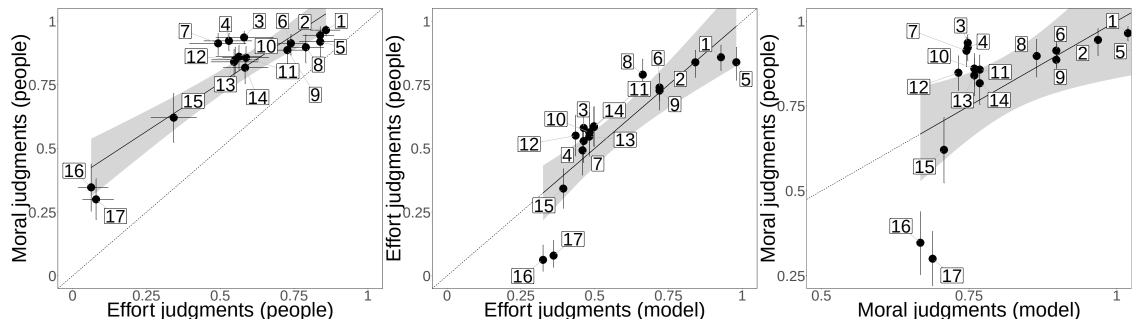


Figure 3. **Experiment 2 results**. Participants' effort judgments (dark gray) and moral judgments (light gray) for each of the 17 scenarios. Bars indicate mean ratings, and small points indicate individual judgments. Error bars indicate bootstrapped 95% confidence intervals. The model predictions are superimposed as circles. Diagrams of what happened in each video are shown below participants' effort and moral judgments. The results are ordered by descending moral judgment from worst (top left) to least bad (bottom right).

the physics engine correlated well with people's moral judgments and, as predicted by the model, the more effort an agent exerted in a scenario, the worse its behavior was perceived to be (4a and c).



(a) (b) (c)

Figure 4. **Experiment 2 results**. Scatter plots of (a) participants' moral judgments against participants' effort judgments, (b) participants' effort judgments against model effort predictions, and (c) participants' moral judgments against model moral predictions. Error bars indicate bootstrapped 95% confidence intervals.

We take these results as supporting the proposal that judgments of physical effort play an important role in a moral calculus over these visual scenarios, as a way of estimating the intention and utility function of an agent. While explaining much of the variance, the correlation between effort (as judged by both people and the model) and people's moral judgment is far from perfect. We attribute the missing variance to our simplifying assumptions. As we stated in Experiment 1, and further elaborate in the General Discussion, our model is likely not capturing salient additional information about the agent's intention to harm the patient. However, such additional information can be incorporated in the future into more sophisticated mental reasoning modules in the overall framework.

We also found that participants' effort judgments in the Effort condition strongly correlate with the effort values from the physics engine (Figure 4b), corroborating the growing body of work that suggests aspects of human reasoning about physics can be captured by physics engines. Given that the model accurately captures participant effort judgments, it is unlikely that deviations between the model's concept of effort and people's perception of effort is responsible for the unpredicted variance in the moral judgments. We also note that while the correlation between the model's effort judgments and people's effort perceptions is high, the linear fit to the model's predictions of effort can deviate noticeably from the mean judgment of effort, due to the intercept term which prevents our model from inferring effort values (and moral values) of zero (for example, in scenarios 16 and 17, Figure 4b).

General Discussion

Our moral evaluations of another person's action depend on our inferences of their mental states, and on the causal role their actions played in bringing about the outcome. Entire research programs have taken this as a given, focusing more on *how* causal and mental state inferences influence moral judgment (e.g. Cushman, 2008; Lagnado & Channon, 2008; Shaver, 1985; Waldmann et al., 2012). At the same time, some moral judgments seem fast and automatic, suggesting a direct route from visual processing to moral judgment (De Freitas & Alvarez, 2018; Iliev et al., 2012; Nagel & Waldmann, 2012). Here, we proposed a framework, the *Moral Dynamics* model, according to which the route from visual processing to moral judgment is mediated by an inference about the agent's mental states. Specifically, we focus on inferences about an agent's desire to harm another based on the effort the agent exerted. We formalized this framework using recent models of intuitive physics and Bayesian theory of mind. In two experiments, we asked participants to evaluate the wrongness of an agent's actions in visual scenes. Experiment 1 replicated earlier work (Iliev et al., 2012) in a new setting, with our model accounting for the qualitative pattern of results with only the underlying parameter of effort. Experiment 2 expanded the range of test cases to allow for finer-grain comparisons, and showed that model was able to quantitatively explain much of the variance in participants' judgments.

We see our *Moral Dynamics* model as a useful framework to build on, not a complete account of moral judgment (cf. Waldmann et al., 2012). We next address several ways in which the model needs to be extended, by expanding on its notion of effort, and incorporating inferences about intentions and causality.

Effort and Cost In this paper, we 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. 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 will be relevant for more complex stimuli where inferring effort becomes more challenging. We also tied effort directly to the use of force by an agent, but effort as a psychological inference may diverge from a simple summation of forces, and intuitive notions of biology and fatigue may enter into the inference (Liu et al., 2017). As a simple 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, our focus in this work was on physical effort, but we expect other types of perceived costs to also be relevant for inferring how much an agent desired a patient's harm (see e.g. Jara-Ettinger et al., 2016). For example, an agent may take risks, forego alternative rewards, or exert great mental effort in realizing their goal. We expect that people take these factors into account and would, for example, judge an agent as morally worse when its action was perceived to be riskier even if the physical effort remained the same.

The role of intention We focused on a simple notion of cost as physical effort, and a simple notion of reward as a direct benefit from harming the patient. But costs and rewards, even if made more sophisticated, will be insufficient to capture the whole range of moral judgments. Specifically, one of the central missing comments in our simple utility calculus is judging the intention of others. Inferring intentions is a non-trivial computational task, but some progress has been made in the past few years (e.g. Kim et al., 2018; Kleiman-Weiner et al., 2015, 2017). These recent models link intentions to plans, and define intended outcomes as those that made a difference to an agent's plan (Bratman, 2009). Incorporating such inferences of intention is an important next step in developing the *Moral Dynamics* framework.

The role of causality Causal inference is critical for moral judgments. However, our current model does not yet feature a full causal analysis of the scene. As a specific proposal for the role of causal reasoning in moral judgment, one could build on the Counterfactual Simulation model of causal judgment (e.g. Gerstenberg et al., 2017). According to this model, causal judgments involve a comparison of what actually happened with what would have happened in a relevant counterfactual world. The more certain an observer is that an outcome would not have happened but for a particular event, the more that particular event is predicted to have caused the outcome. Applied to the domain discussed in this paper, we could determine an agent's causal role by simulating how the dynamics would have unfolded if the harming agent had not been present in a scene. But other counterfactuals may come to mind as well, beyond the simple removal of an agent from the scene. For example, one could consider what would have happened if the agent hadn't exerted any effort, or if the agent had been replaced by a reasonable person (Gerstenberg et al., 2018).

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 judgment are: What happened, who did what, and why did they do that. Moral judgment of a situation follows from how people understand the dynamics of the world that led to that situation, including the minds of other people. Such questions of cause and the mental states of others have been taken as the foundation for a great deal of research into moral reasoning. However, at the same time research has shown that many mental judgments are fast and automatic in nature, suggesting bottom-up reasoning based on visual cues of a scene. Recent work has proposed quantitative models that use visual processing to make moral decisions in particular. 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, for better or for worse.

Code Availability

Code for all models and analyses is available at `https://github.com/flxsosa/MoralDynamicsPaper`

Data Availability

Anonymised participant data and model simulation data are available at `https://github.com/flxsosa/MoralDynamicsPaper`

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).

Author Contributions

F.A.S. and T.G. designed the experiments, collected and analyzed the data, and wrote the paper. T.U., J.B.T., and S.J.G. designed the experiments and wrote the paper.

Competing Financial Interests

The authors declare no competing financial interests.

Appendix

Under the 'Naive Utility Calculus' (NUC) (Jara-Ettinger et al., 2016), people believe other agents act to maximize:

$$U(s, a) = R(s) - C(a), \tag{4}$$

Where U is the agent's utility, a combination of the reward derived from a particular world state s, $R(s)$, and the cost of the action a needed to reach that state, $C(a)$.

For our purposes, we add the following three assumptions to NUC: i) the cost of an action is proportional to the physical effort necessary to take that action, ii) a social agent's reward, R_A can depend on the utility of another agent, U_P, and iii) moral evaluations are based on the inferred reward an agent receives for harming another.

According to our first assumption, the cost of a sequence of actions, $C(a_0, ...)$, is equivalent to the amount of physical effort needed to take those actions:

$$C(a_0, a_1, ..., a_T) \propto \int_{t=0}^{T} F(a_t)dt, \tag{5}$$

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. In practice we consider a discretized time setting in a physics engine, and replace the integral with a sum, and replace F with an impulse I over a short time:

$$C(a_0, a_1, ..., a_T) \propto \sum_{t=0}^{T} I(a_t). \tag{6}$$

According to our second and third assumptions, the moral evaluation J of an agent A is proportional to the inferred reward that the agent derives from harming an innocent patient, P. This relationship can be captured via a simple factor $k < 0$, such that $R_A(U_B) = k \cdot U_P$:

$$J(A) \propto k. \tag{7}$$

Where, a more negative k means a greater reward for agent A if P is harmed, and will lead to a more negative moral evaluation of A.

Bayesian Theory of Mind assumes people perform an inference of the beliefs and utilities of others when observing the actions of others (see Figure 1). Formally, we suppose people jointly infer the reward, R_A that A receives for taking a set of actions and the cost A incurs, C_A, for taking those actions:

$$P(R_A, C_A|\text{Actions}) \propto P(\text{Actions}|R_A, C_A)P(R_A, C_A). \tag{8}$$

We do not compute the full inference of Equation 8.

We assume people judge others actions as rational agents that seek to maximize reward (Dennett, 1987), and use this to support an approximation of R through C. The assumption of rationality leads us to the inequality $R_A(s) > C_A(a)$. That is, we assume people think if A took an action a it must have been because it led to a state of the world s that provided greater reward than the cost of the action.

Using the above inequality and Equation 6 to calculate C for a given agent, we can approximate R as:

$$R_A \propto C = \sum_{t=0}^{T} I(a_t), \tag{9}$$

where a_t is an action taken by an agent at time point t in a given scenario. From this, we approximate U:

$$U_A \propto R_A \propto \sum_{t=0}^{T} I(a_t), \tag{10}$$

Putting everything together, for each scenario containing agent A and patient P, we approximate people's negative moral judgments about A, $J(A)$, as being proportional to the reward A gets for the outcome utility of P, $R_A(U_P)$.

$$J(A) \propto \sum_{t=0}^{T} I_A(a_t), \tag{11}$$

That is, the negative moral evaluation of agent A can be approximated by the amount of physical effort they were willing to put into harming patient P.

Trajectories were deterministic and defined for each of the objects in each video. At each time step, t, in a video, an agent applies an impulse, I_t, to itself and the magnitude of this impulse is recorded. The cost of an action at some time step, $C(a_t)$, is equivalent to the magnitude of the impulse applied at that time step. The amount of effort an agent exerted in a given video is the sum of the recorded impulse magnitudes for that video for that agent.

Two important parameters in our videos are the maximum velocity an agent can reach and friction. Friction was used in each video so that agents had to put in effort to maintain their target velocities at every time step in a simulation. We set the maximum velocity and friction so as to best replicate the dynamics found in the stimuli presented in Iliev et al. (2012).

References

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

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.

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.

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (in press). Intuitive experimentation in the physical world. *Cognitive Psychology*.

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

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

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

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.

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

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. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., 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., & 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.

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.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

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

Hills, A. P., Mokhtar, N., & Byrne, N. M. (2014). Assessment of physical activity and energy expenditure: An overview of objective measures. *Frontiers in Nutrition*, *1*(5), 1–16.

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.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. Retrieved from `https://www.pnas.org/content/114/43/E9145` doi: 10.1073/pnas.1714471114

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.

Kim, R., Kleiman-Weiner, M., Abeliuk, A., Awad, E., Dsouza, S., Tenenbaum, J., & Rahwan, I. (2018). A computational model of commonsense moral decision making. *arXiv preprint arXiv:1801.04346*.

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.

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.

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. Retrieved from `http://science.sciencemag.org/content/358/6366/1038` doi: 10.1126/science.aag2132

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., Guglielmo, S., & Monroe, A. E. (2014, 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

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

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.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological Review*, *120*(2), 411–437.

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*.

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

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

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., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.

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).

Vander Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science*, *20*(6), 771–777.

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.

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.