Inference of Intention and Permissibility in Moral Decision Making

Max Kleiman-Weiner[1] (maxkw@mit.edu), Tobias Gerstenberg[1] (tger@mit.edu),
Sydney Levine[2] (levine@ruccs.rutgers.edu) & Joshua B. Tenenbaum[1] (jbt@mit.edu)
[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139
[2]Center for Cognitive Science, Rutgers University, Piscataway, NJ 08854

Abstract

The actions of a rational agent reveal information about its mental states. These inferred mental states, particularly the agent's intentions, play an important role in the evaluation of moral permissibility. While previous computational models have shown that beliefs and desires can be inferred from behavior under the assumption of rational action they have critically lacked a third mental state, intentions. In this work, we develop a novel formalism for intentions and show how they can be inferred as counterfactual contrasts over influence diagrams. This model is used to quantitatively explain judgments about intention and moral permissibility in classic and novel trolley problems.

Keywords: moral judgment; social cognition; intention; theory of mind; influence diagrams; counterfactuals.

Introduction

Our actions often have multiple effects, whether it's creating a small amount of pollution in order to pick up groceries or making trade-offs between civilian deaths and military objectives during a war. Did the general try to achieve the military objective even at the cost of civilian lives or did his plan use civilian deaths in order to demoralize the enemy? The ability to distinguish between the effects an agent intended versus those that were side-effects are critical in general for social cognition and in particular for assigning responsibility and assessing moral permissibility. Our goal here is to understand these processes in computational terms.

Reasoning about the intentions of other agents relies on theory of mind, the capacity to infer an agent's underlying mental states such as beliefs and desires from her actions. Recently, a lot of progress in computational modeling of theory of mind has been made by formalizing lay intuitions that other agents act as rational actors who maximize expected utility subject to their beliefs. A Bayesian observer can then invert the agent's planning process and reason about the likelihood of certain beliefs and desires given the agent's actions (Baker, Saxe, & Tenenbaum, 2009; Jern & Kemp, 2011).

Although most computational accounts of theory of mind have focused on desires and beliefs, intentions are a third mental state thought to be particularly useful. Intentions can be thought of as plans of action that an agent commits to, chosen in order to bring about its desires given its beliefs about the causal structure of the world (Bratman, 1987; Malle & Knobe, 1997). It is hypothesized that the ability to reason about and with the intentions of others is one the key factors that enables the sophistication of human social behavior (Tomasello, 2014). They are also an important input into the evaluation of moral permissibility such as the doctrine of double effect's requirement against intending harm (Mikhail, 2007; Cushman, 2013; Waldmann, Nagel, & Wiegmann, 2012; Crockett, 2013; Greene, 2014). The relationship between intentions and outcomes is complicated by the fact that it is possible to do the right thing for the wrong reasons (Scanlon, 2009).

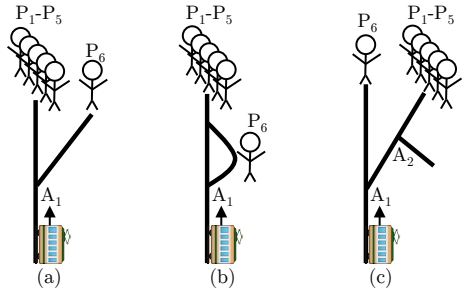

Figure 1: Schematic representation of trolley track geometries: (a) side track, (b) loop track and (c) side-side track.

Here we investigate a novel computational representation for reasoning about other people's intentions based on counterfactual contrasts defined over influence diagrams. This model can distinguish between intended outcomes and unintended side effects as well as represent the future-oriented aspect of intentions as plans (Bratman, 1987). We use this model of intention inference as an input into a computational model of moral permissibility and test how well the model explains both well-studied and novel trolley dilemmas. Before describing our computational model, we motivate the model with some examples.

Side track and loop track The canonical examples for the role of intention in moral permissibility judgments are the *side track* and *loop track* (Thomson, 1985). The *side track*, shown in Figure 1a is a scenario where an out-of-control trolley is heading towards five people. An agent is standing near a switch (A_1) which will turn the trolley from the main track with five people on it (P_1-P_5) to a side track with one person (P_6). The *loop track*, shown in Figure 1b has a loop instead of a split such that the trolley will continue on and hit the five unless it hits the man on the looping track which would cause the train to stop. Consider that in each of the situations, the agent throws the switch.

Empirically, throwing the switch in the *loop track* is judged less morally permissible than the *side track* (Mikhail, 2007). Explanations of this finding usually draw on the agent's intention. In the *side track*, the agent neither intends the hitting nor killing of the man on the side track while in the *loop track* the agent does intend for the trolley to hit the man on the loop but not his death.

Side-side track Following Bratman (1987), future-oriented planning is an important aspect of intention. To probe this aspect of intention in permissibility we developed a novel track geometry which requires inference over the full plan rather

than just a single action. As shown in Figure 1c, the *side-side track* scenario is similar to the *side track* except that the side track has an additional side track with its own switch (A_2). Consider a situation in which there is one person on the main track, five people on the side track and no one on the side-side track. If the trolley is going down the side track, unless the agent throws the second switch directing the trolley down the side-side track, the trolley will continue and hit the people on the side track.

We hypothesize that throwing the first switch is intuitively morally permissible. How can this be explained even though the trolley is now heading towards the five people? Since intentions are forward-directed, they include the agent's intention to throw the second switch, saving all the lives. Only if the agent doesn't intend to throw the second switch does the action become impermissible. This case motivates the central role of planning in our computational model. It is insufficient to consider intentions as merely directed towards the effects of a single action but rather the effects of the entire plan need to be taken into consideration.

Joint inferences: norms, desires and intentions Inferences about the intentions of an agent are often intertwined with inferences about the agent's desires and the social norms to which those desires conform. We contrast a *side track* dilemma that has one anonymous person on the main track and two anonymous people on the side track with a dilemma we call *brother track* where the agent's brother is on the main track and there are two anonymous people on the side track.

In the first dilemma participants rate throwing the switch to be highly impermissible (as shown below) while in *brother track*, participants rate throwing the switch to be permissible. When the agent throws the switch in the first case, participants infer that the agent intended to kill the two people. In *brother track*, participants infer that the agent is following a norm to value loved ones more and doesn't intend to kill the two people on the side track. In a variant of *brother track*, the brother is on the side-track and two anonymous people are on the main track. If the agent throws the switch, we may infer that the agent followed a norm that all lives should be valued equally, or... she might not value all lives equally, she just intended to kill her brother! To infer the intended consequences and judge moral permissibility thus requires jointly inferring the agent's desires and the norms that guided their actions.

Computational Framework

Our computational approach has two parts. The first is a computational account of intention inference and the second uses this account to model permissibility judgments. The model is presented to capture the real-world richness of intentional planning and has greater generality than is needed for our examples.

Our representation of intentions is based on influence diagrams (ID). Influence diagrams are similar to Bayes nets and were used previously to capture reasoning about what other agents know and want during decision-making (Jern & Kemp, 2011). Solving an ID yields an optimal policy (σ^*): the actions the decision-making agent needs to take to maximize her expected utility. We show how the ID and the policy can be used together to compute foreseen outcomes: the most likely outcome of the agent's policy. Using a counterfactual criterion, we refine the foreseen outcomes into a subset of outcomes that are intended.

Overall, we aim to capture that intentions: (1) are partial plans with means-ends correspondence, (2) predict the expected effects of actions, (3) can distinguish between outcomes that the agent is committed to bring about and those that are side-effects, (4) are future-oriented, (5) give reasons for action and are hence inputs to further practical reasoning such as moral permissibility (Bratman, 1987). Indeed, one practical reason for the centrality of intentions in folk psychology is that knowing an agent's intentions allows one to predict how the agent will behave and why.

We then show how an observer with uncertainty about the desires and norms of the agent can rationally update his beliefs about the agent by inverting the planning process using Bayes' rule, and finally, can use these inferences to make judgments about moral permissibility.

Influence diagrams

Our notation follows Koller and Friedman (2009). An influence diagram *ID* is a directed acyclic graph over three types of nodes: state nodes (depicted as circles, $\mathcal{X}$), decision nodes (depicted as rectangles, $\mathcal{D}$), and utility nodes (depicted as diamonds, $\mathcal{U}$). Directed edges between nodes determine causal dependencies. State and utility nodes take values that are a function of the structural equations and depend on the values of their parents, while the value of decision nodes are chosen by the decision making agent such that expected total utility is maximized. Let σ^* be the policy that maximizes the expected total utility of *ID*:

$$\sigma^* = \arg\max_{\sigma} \mathrm{EU}[ID_\sigma]$$

where $\mathrm{EU}[ID_\sigma]$ is the expected total utility of following policy σ in *ID*. Each policy σ, is a setting of the decision nodes to a chosen value. To calculate expected utility for a policy, let ζ be an *outcome*, the setting of each of the state, utility and decision nodes in *ID* to a value. For a node $Z \in ID$, ζ_Z is the value of node Z in outcome ζ. Thus the expected utility of policy σ can be calculated by averaging the total utility of an outcome $U(\zeta)$, weighted by the likelihood of that outcome under the policy $P(\zeta|ID_\sigma)$ for each possible outcome:

$$\mathrm{EU}[ID_\sigma] = \sum_{\zeta} P(\zeta|ID_\sigma)U(\zeta)$$

$$U(\zeta) = \sum_{V \in \mathcal{U}} \zeta_V$$

$$P(\zeta|ID_\sigma) = \prod_{X \in \mathcal{X}} P(X|\mathrm{Pa}_X, \sigma)$$

where Pa_X are the parents of node X. Thus the ID representation concisely factors the agent's decision problem into individual states, decisions, sources of utility and the structural equations that define the dependence relations between them.

Defaults are encoded by requiring any policy that changes the value of a decision node away from its default value to incur a small utility cost (not shown in figures).

See Figure 2a for an influence diagram representation of the *side track* dilemma shown in Figure 1a. The only decision in the policy is the choice to throw the switch A_1. This action determines whether the trolley goes down the left track (T_L) track or right track (T_R) which determines which people are hit and killed affecting the decision maker's utility.

Intention

We now build on the ID representation to first extract the best foreseen outcomes of the optimal policy and then refine the best foreseen outcomes into an intention which excludes outcomes that were unintended.

Definition 1. The best foreseen outcome F is the outcome with the highest expected utility that can be foreseen by the agent acting under the optimal policy:

$$F = \arg\max_{\zeta} \mathrm{U}(\zeta) P(\zeta | ID_{\sigma^*})$$

F captures all the consequences that the agent can optimistically foresee happening as a result of her policy but does not include any backup plans or other types of conditional contingent plans. The decision to choose only a single foreseen state is motivated by efficient planning algorithms which plan only on likely states and replan if necessary rather than directly planning for every contingency (Platt, Tedrake, Kaelbling, & Lozano-Perez, 2010). Specifically, if we assume that number of lives maps to utility, the foreseen effects of throwing the switch in *side track* are that the 5 people on the main track will not be hit by the trolley and live, generating 5 utility while the person on the side track will get hit by the trolley and die generating -1 utility for the decision-maker. This is shown in Figure 2b where each node is assigned to its foreseen value (shown in bold) under the policy of throwing the switch.

While foreseen outcomes optimistically describe the consequences of an action and are brought about "intentionally", not all foreseen consequences are intended by the decision maker (Bratman, 1987). Analogously in causal reasoning, not all of the factors which influence an outcome are judged by human participants to be causes of an observed outcome. This has led to the development of computational models of *actual causation* which try to model the commonsense notion of causality through counterfactual reasoning (Halpern & Pearl, 2005). This formalism has successfully captured aspects of empirical attribution of responsibility (Lagnado, Gerstenberg, & Zultan, 2013; Sloman, Fernbach, & Ewing, 2012). We propose that a similar model can distinguish an agent's intended outcomes from foreseen outcomes. Specifically, intended outcomes are the subset of foreseen outcomes that the choice of the optimal policy (σ^*) counterfactually depends upon. We generalize Halpern and Pearl (2005) to decision problems where outcomes are the policies determined by planning:

Definition 2. An intention I is a subset of nodes and their corresponding values such that the following conditions are satisfied:

1. Nodes in I take on values foreseen under σ^*.
2. Let $ID^{\backslash I}$ be a counterfactual influence diagram that is ID with the nodes in I removed. I are intended if $\sigma^*_{\backslash I} \neq \sigma^*$, i.e., the optimal policy for $ID^{\backslash I}$ is different from the optimal policy for the original influence diagram ID.
3. The sets of nodes in I are a minimal subset, i.e., there are no smaller subsets of intended nodes, which when removed would also satisfy 2.

The intention I for the *side track* is shown in Figure 2c by the nodes and values highlighted in gray. The decision to throw the switch does not depend on the values for hitting and killing the person on the side track (P_6) and the loss of utility that resulted. Even if those nodes were removed from the influence diagram the agent would have still acted the same. Thus those nodes are side effects of the action. In contrast, if the nodes that correspond to the states and utility of the people on the main track were removed, the agent would not have thrown the switch. Since the policy with the nodes removed is not equal to the policy for the full ID, those nodes and their values are treated as intended. We only consider the removal of nodes. However, capturing other aspects of intention may require counterfactual perturbations to the utility values rather than removal (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2015).

Our representation of intentions as counterfactuals over influence diagrams satisfies the five aspects of intentions we aimed to capture: (1) I is a partial plan than contains future expected actions, (2) the outcomes in I are the expected result of the plan, (3) I distinguishes between intended outcomes the agent is committed to bring about and side effects, (4) I contains future-oriented policy information, (5) the nodes and

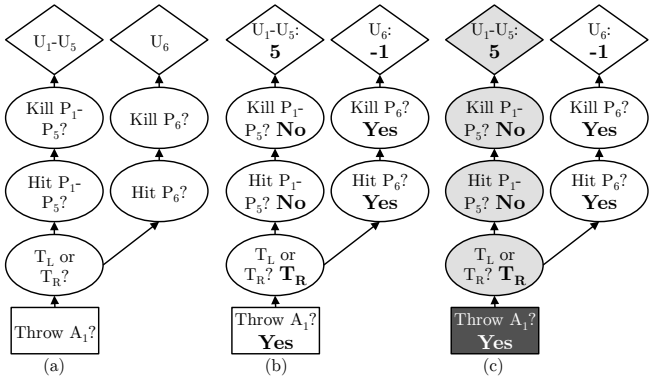

Figure 2: An influence diagram (ID) representation of intention. (a) The ID for the *side track* decision dilemma. (b) The foreseen outcomes F. Each node is set to the best value possible under the policy of throwing the switch (shown in bold). (c) The intention I is shaded in gray. Like the foreseen outcome, each node is set to its most likely value under the policy, however only the nodes shaded in gray and their values are intended by the agent.

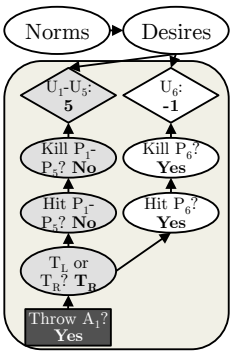

Figure 3: Joint inference of intentions, desires and norms. The nodes in the beige box correspond to the influence diagram the agent is planning over. The nodes outside the box represent the observer's uncertainty over the agent's desires and norms. The observer can use this prior to infer the agent's intention (gray) from the observation of a single action (black).

values in I give the reasons for the action. Thus I is a compressed representation of the actions an agent plans to make and their intended effects.

Joint inference of norms, desires and intentions

Inference of intention through the ID requires knowledge of the agent's desires and beliefs. However, observers often only know these desires and beliefs with uncertainty such as in the *brother track* examples in the introduction. The structure of these priors gives the observer an expressive theory of mind, capable of representing agents with both good and evil desires or adherence to different norms. The observer's beliefs about the agent's desires are modeled by introducing uncertainty over the parameterization of the utility nodes. This uncertainty induces a probability measure over IDs (shown in Figure 3) and since each ID has an intention under rational planning, it also induces a probability measure over intentions. Given observation of an agent's action(s) A, an observer can rationally update his belief about the agent's intentions I, desires D and norms N using Bayes rule:

$$P(I,D,N|A) = P(A|I)P(I|D,N)P(D,N)/P(A)$$

Since $P(A)$ cannot be analytically calculated we used rejection sampling to draw samples from $P(I,D,N|A)$. We first sample from the desire and norm distribution of the observer $P(D,N)$ which defines an influence diagram $ID^{D,N}$. Planning in this ID yields $P(I|D,N)$. If the intended action is the same as the observed action A we keep the sample which is a joint distribution over the intention, desires and norms. If the intended action is not A, the sample is discarded and the processes is repeated.

In order to quantitatively predict observer's judgments of $P(I,D,N|A)$ we must specify the structure of $P(D,N)$, the distribution over how the agent values the lives of the people on the tracks. Let D_T be the utility to the decision maker of the n_T people on track T not being killed. If $k_T = -1$ then the agent wants to kill the people on track T, if not, $k_T = 1$. k_T is negative for all T with probability α_b which means the agent wants to kill as many as possible. Otherwise, $k_T = -1$ for each track independently with probability α_k.

When making decisions about brothers (or other loved ones) we hypothesize that the decision-making agent might apply one of two norms: all lives should be valued equally, or loved ones should be valued more. This norm determines whether a brother is valued to the agent more than an anonymous person. Let *norm* be true when the agent follows the norm that loved ones should matter more which is true with prior probability α_{norm}. If *norm* is true and the person on track T is the agent's brother, then the brother is counted as equal to an α_{bro} number of anonymous people and otherwise treated the same as a single anonymous person. Finally, as is common in discrete choice, we include independent multiplicative exponential noise e_T for each track which captures other unmodeled sources of variation including perceptual and valuation errors. Thus $D_T = n_T k_T e_T$ for anonymous people and brother when the norm is not followed and $D_T = \alpha_{bro} k_T e_T$ when the norm to value loved ones more is true. While only sketched here, these variables specify the structure of the observer's beliefs about the decision-making agent's desires.

Moral Permissibility

Finally, we use the inference of intentions to model moral permissibility. The trolley problem and its variants are well-studied for probing the cognitive processes that generate moral permissibility judgments. However, without a model of graded intention it was not previously possible to quantitatively model these judgments. We consider three models: one based on intentions, one based on differences in the number of people who died and who survived, and a linear combination of the two.

Let *per* be the probability of finding the agent's action morally permissible. The first model only considers the agent's inferred intention to predict permissibility judgments. The more likely the agent is inferred to have an intention to harm the more likely the action is judged to be impermissible:

$$per_{\text{intention}} = \text{logit}^{-1}(1 - P(I_{harm} = \text{Yes}|A))$$

where $P(I_{harm} = \text{Yes}|A)$ is the probability that the agent is inferred to have intended to harm someone given that they took action A. The transform $\text{logit}^{-1}(x) = (1+\exp(-\alpha_1(x-\alpha_2)))^{-1}$ with parameters $\alpha_1 = 7$ and $\alpha_2 = 0.7$ scale the intention judgments onto permissibility. We compare this to a model that predicts moral permissibility judgments based only on the number of lives killed and saved:

$$per_{\text{utility}} = \text{logit}^{-1}(\Delta_{lives})$$

where Δ_{lives} is the expected difference between the number of lives saved and the number of lives lost and the logit parameters $\alpha_1 = 0.3$ and $\alpha_2 = 0$ scale Δ_{lives} onto the unit interval. Finally we consider a full model that weighs both intentions and the difference between number of lives saved and lost:

$$per_{\text{full}} = w * per_{\text{intention}} + (1-w) * per_{\text{utility}}$$

Experiment and Results

We test the predictions of the model for the three examples from the introduction. For the first two we consider only qualitative phenomena from the published literature. For the third we conducted a large scale behavioral study varying the location, number and identity of the people on the tracks.

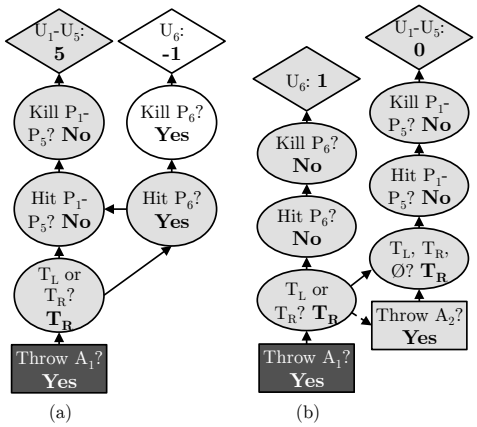

Figure 4: Influence diagram for the (a) *loop track* and (b) *side-side track* with the intention shaded in gray and the action in black.

Side track and loop track As demonstrated before (see Figure 2c), the model correctly predicts that hitting and killing the man on the side track is unintended. In contrast, for the *loop track*, when the man on the loop is hit but not killed, the policy remains unchanged, so the model predicts that the killing of the man is unintended. However, the model predicts that hitting the man on the loop is intended since it is required to stop the trolley from hitting the 5 on the main track. Thus due to this difference in causal structure, the agent in *loop track* intends to hit but not kill the man on the loop. Indeed throwing the switch in *loop track* is found to be less permissible than throwing the switch in *side track*. Given that Δ_{lives} is the same in both conditions suggests that the intention to harm in the loop track case could account for this difference as has been suggested in the literature (Mikhail, 2007).

Side-Side Track In the *side-side track*, the model predicts that if the agent throws the first switch, her intention is to also throw the second switch so that the trolley goes down the side-side track and kills nobody. The model further predicts that both saving the person on the main track and the 5 people on the side track are intended since in both cases they were counterfactually relevant to the policy: if the person on the main track wasn't there, the agent wouldn't throw the first switch. If the people on the side track weren't there, the agent wouldn't have thrown the second switch (since throwing switches has a small action cost associated with it).

The role of intention in evaluating permissibility is clear here even though it plays a different role than in the *loop track*. The Δ_{lives} can only be calculated under the agent's future-oriented plan. Thus the intention captures a key aspect of the permissibility by requiring an inference over future actions rather than through understanding which effects are intended and which are side-effects.

Joint inference: desire and intentions

Experiment To investigate the ability of participants to jointly infer the desires of the agent and her intention, we consider a set of tracks with the same track geometry as in Figure 1a but on each track there were either 1, 2 or 5 anonymous people or the agent's brother. We considered all permutations that had at least one track with a single person on it (this excludes 2v5 and 5v2) yielding a total of 11 track configurations. The tracks are presented in the format XvY where X and Y are the number of people on the main and side track respectively or are a 'B' if it's the agent's brother. 100 participants were recruited via Amazon Mechanical Turk using psiTurk (McDonnell et al., 2012). On each trial, participants read a standard story about the trolley dilemma based on those in Mikhail (2007) but with the identity and number of people varied to reflect the track configuration. After reading the story, participants were asked to answer the following questions with "yes" or "no": "Was it morally permissible for Hank to throw switch?", "Did Hank throw the switch in order to kill *his brother/the man/the two men/the five men* on the side track", "Did Hank throw the switch in order to not kill *his brother/the man/the two men/the five men* on the main track?" and then used a slider to answer "Hank most likely believes:" where the edges of the slider were "all lives should be valued equally" and "only loved ones should be valued". This data was collected as part of a larger experiment where the agent either throws or does not throw the switch. Here we only present the data for when the agent threw the switch.

Results Figure 5 shows the averaged participant responses for moral permissibility. The following trends are apparent: (1) the more lives saved and less lives killed the more permissible the action; (2) killing the brother was seen as less permissible compared to killing an anonymous person for a given number of lives saved; (3) saving the brother became less permissible as the number of lives sacrificed grew. Figure 6 shows the averaged participant data for the intention to kill those on the side track and intention to not kill those on the main track. In all configurations, participants were more likely to infer that the participant acted in order to save rather than kill even though the action had both effects. This reflects the low prior probability on the agent desiring any of the people's deaths. The intention to kill was inferred to be greatest when $\Delta_{lives} \leq 0$ and when the agent switched the trolley onto the track with the brother. The intention to not kill those on the main track (right plot) followed a similar trend corresponding roughly to the inverse of the intent to kill judgments.

Figure 7 shows the averaged participant responses for the agent's relative belief between the two norms: "All lives

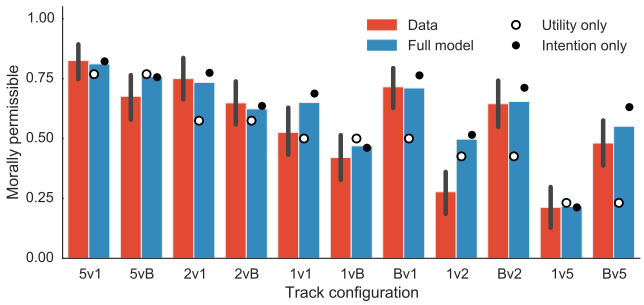

Figure 5: Averaged participant responses for whether throwing the switch was morally permissible. 5vB means 5 people on the main track and the brother on the side track. Error bars in all figures show a bootstrapped 95% confidence interval.

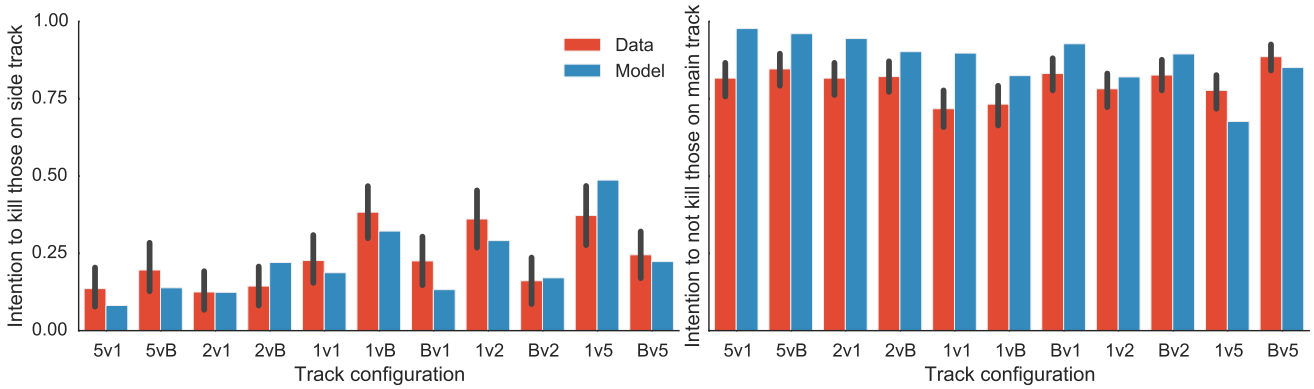

Figure 6: Averaged participant responses for whether the agent threw the switch in order to kill the people on the side track (left) or not kill the people on the main track (right). 5vB means 5 people on the main track and the brother on the side track.

should be valued equally" and "Only loved ones should be valued". When the brother is saved, participants inferred that the agent is following the loved ones norm. When the brother is killed, participants infer that the agent is following the all lives equal norm. The more anonymous people killed, the stronger the inference for the norm to treat loved ones specially.

Model Predictions The predictions of the computational model with parameters $\alpha_k = 0.05, \alpha_b = 0.1, \alpha_{norm} = 0.55, \alpha_{bro} = 30, w = 0.8$ are shown in the above plots. Overall, the model fits the data with $R = 0.97$ and captured the main trends described in the previous section. Several of these free parameters do not significantly affect the quantitative model fit although we include them since they are interpretable and intuitive: there was a low prior probability that the agent wanted to kill people and a high prior probability of endorsing the loved ones norm. A reduced model that only includes 3 of the above 5 parameters ($\alpha_b = 0.15, \alpha_{norm} = 0.55, \alpha_{bro} = 30$) fit with $R = 0.95$.

Discussion

We developed a novel model for intention inference based on counterfactual contrasts over influence diagrams. While we are not the first to give a computational account of intention (see e.g., Cohen and Levesque (1990)), our model is the first probabilistic model based on inverse rational planning that can distinguish between outcomes an agent intended and side effects that were merely foreseen. Our model makes quantitative predictions about both the intentions underlying actions and moral permissibility judgments which correspond well to human judgments. The model also applies to many more situations than we could address in this paper. In future work, we intend to test other aspects of the model through behavioral experiments on trolley dilemmas like the *side-side track* and others where beliefs might be uncertain. We will also apply our model of intention inference to both non-trolley moral dilemmas and non-moral domains such as games and other social interactions (Tomasello, 2014).

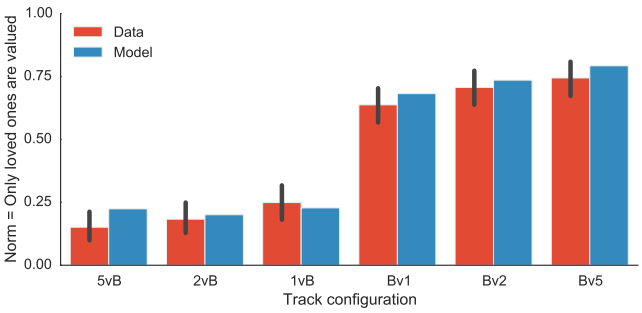

Figure 7: Averaged participant responses for whether the agent most likely believes "all lives should be valued equally" (coded 0) or "only loved ones should be valued" (coded 1).

Acknowledgments MKW was supported by a Hertz Foundation Fellowship and NSF-GRFP. TG and JBT were supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333. SL was supported by a DOD NDSEG.

References

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329–349.
Bratman, M. (1987). *Intention, plans, and practical reason*.
Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial intelligence*, *42*(2), 213–261.
Crockett, M. J. (2013). Models of morality. *Trends in cognitive sciences*, *17*(8), 363–366.
Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and social psychology review*, *17*(3), 273–292.
Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In *Proceedings of the 37th annual conference of the cognitive science society.*
Greene, J. (2014). *Moral tribes: emotion, reason and the gap between us and them*. Atlantic Books Ltd.
Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. part i: Causes. *The British journal for the philosophy of science*, *56*(4), 843–887.
Jern, A., & Kemp, C. (2011). Capturing mental state reasoning with influence diagrams. In *Proceedings of the thirty-third annual conference of the cognitive science society.*
Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT press.
Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive science*, *37*(6), 1036–1073.
Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*(2), 101–121.
McDonnell, J., Martin, J., Markant, D., Coenen, A., Rich, A., & Gureckis, T. (2012). psiturk.
Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, *11*(4), 143–152.
Platt, R., Tedrake, R., Kaelbling, L., & Lozano-Perez, T. (2010, June). Belief space planning assuming maximum likelihood observations. In *Proceedings of robotics: Science and systems*. Zaragoza, Spain.
Scanlon, T. M. (2009). *Moral dimensions*. Harvard University Press.
Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language*, *27*(2), 154–180.
Thomson, J. J. (1985). The trolley problem. *Yale Law Journal*, *94*, 1395–1415.
Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.
Waldmann, M. R., Nagel, J., & Wiegmann, A. (2012). Moral judgment. *The Oxford handbook of thinking and reasoning*, 364–389.