A Model of Empathy to Shape Trolley Problem Moral Judgements

Jason R. Wilson, Matthias Scheutz
Human-Robot Interaction Lab, Tufts University
Medford, Massachusetts 02155
Email: {wilson,mscheutz}@cs.tufts.edu

***Abstract*—Moral judgements are a complex phenomenon that have gained a renewed interest in the research community. Many have proposed explanations for moral judgements, including utilitarian accounts and the Principle of Double Effect. Some also advocate for the critical role of emotional processes like empathy. However, developing a computational model of moral judgements is rare perhaps due in part to the numerous influences on it. We present here a computational model of moral judgements based on moral expectation and the Principle of Double Effect. We then extend this model to provide a plausible explanation for the effect of empathy on these judgements. We evaluate these models using results from recent studies with human participants.**

***Keywords*—empathy; moral judgment; utility; computational model**

I. Introduction

Making moral judgements of right and wrong can often seem a matter of intuition for people, but the essence of these judgements is a complex problem that has been long debated by philosophers, psychologists, and more. Given these complexities presents challenges to developing autonomous agents (e.g. robots) that can utilize moral reasoning. However, as robots (and other automated computer systems) become more integrated into our society it becomes increasingly important for these agents to be sensitive to our societal rules and expectations. An autonomous agent that is capable of moral reasoning is more likely to be accepted in our society. In order for an agent to make moral judgements it will require a model of moral reasoning, but the complexity of the task has led to the lack of computational models of moral judgements. The difficulties are perhaps magnified when one considers the likely involvement of emotions in moral decision-making. In particular, emotional reactions like empathy have some relation to morality. This is exemplified in psychopaths who exhibit little to no empathy and also commonly fail to recognize moral conventions.

We present here a computational model of moral judgements for determining relative permissibility of actions and validate this model across variants of the trolley problem. The baseline model is validated using variants of the trolley problem that have different causal structures. Given this baseline model we then model the role of empathy on a few versions of the *bystander* variant of the trolley problem. After initially testing the model with empathic responses to people of varying age and genetic relation, we then go on to predict the empathic responses for in-group/out-group targets. Our empathy model is inspired by psychological and neurological findings that indicate three processes: cognitive empathy, emotional empathy, and prosocial concern.

II. Background

The model proposed here uses the well-studied trolley problem to investigate the factors in moral decision-making. We review the trolley problem and consider evidence for emotional influences on moral judgements.

A. Trolley Problems

Trolley problems have been the subject of much recent research (e.g. [1], [2], [3], [4], [5]) and much of this research stems from work by Foot [6] and Thomson [7], [8]. The trolley problems involve a decision between the life of one person and the life of five others. (The number of others may vary but is most often five.) In the original trolley problem, also referred to as the *bystander* scenario, participants are given text similar to the following [7]:

> Frank is a passenger on a trolley whose driver has just shouted that the trolleys brakes have failed, and who then died of the shock. On the track ahead are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Frank can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Frank can turn the trolley, killing the one; or he can refrain from turning the trolley, letting the five die.

The variants of the trolley prolbme are appealing due to their simplicity and ease by which philosophers and experimenters can manipulate individual factors. For example, the *footbridge* variant maintains the choice between the life of one and that of five, but the action now requires pushing the one off of a *footbridge* and using that person to stop the trolley and save the other five. More often than not, people make the utilitarian choice in the *bystander* scenario but are more likely to indicate that the action in the *footbridge* scenario is not permissible.

Reasons for different judgements in moral decisions include the effect emotions and whether a transgression is a means to an end or just a side-effect. Greene has advocated for the role of emotions in moral decision-making. Part of the reasoning is that the activation of brain regions related to emotion processing are some of the same regions also activated in moral decisions that involve “up close and personal”. Mikhail, however, prefers to keep to legal definitions of battery end, means and side-effect and avoids weaker definitions of what constitutes "personal". He does not rule out a link to emotions in moral judgements, but it is not clear to him that we have a sufficient understanding of how or whether an emotional appraisal influences these judgements.

Mikhail raises the valid question of whether the emotions are the cause or effect. The different emotional responses to the scenarios could be the result of the judgement of whether it is permissible or not. For example, the negative emotional response to the *footbridge* scenario may be the result of (and not the cause) of the impermissibility judgement. As a result, he focuses on differences in the causal structure of the scenarios. However, differences in permissibility ratings in a scenario where the structure is held constant needs a different explanation. When participants are given identical trolley scenarios except for information (e.g. genetic relation) about the person on the sidetrack, their ratings of permissibility vary. For this, we investigate the role of empathy.

B. Models Of Empathy

Much of the psychological literature on empathy focuses on two aspects: cognitive and emotional empathy [9]. Cognitive empathy is described as an ability to take the perspective of the observed person and infer the emotions of that person. Emotional empathy is the sharing of the emotional experience of the observed. Some include a third facet to empathy, prosocial concern. Prosocial concern involves the motivation to act out of concern for the target's well-being. Zaki and Ochsner [10] describe some of the weaknesses in neurological studies that include only the first two facets and emphasize the need to include prosocial concern to account for behavioral data.

A lack of empathy is a prominent characteristic of psychopaths, and it has been suggested that this lack of empathy is related to their difficulties to recognize moral transgressions [11]. Neurological evidence has shown relations between brain regions affecting psychopaths and those that are related to both moral judgements and emotion processing (see [12] for review). Additionally, Anderson et al. [13] report that individuals with early onset damage to the prefrontal cortex have defective moral reasoning and exhibit little to no empathy.

Few computational models of empathy exist, and models relevant to the work here are non-existent. Many computational approaches to empathy have focused on creating believable agents where the goal is for the user to have an empathic response to the artificial agent ([14]; [15]). These are often used in a pedagogical setting to aid the user in learning emotions and empathy. Other models are based on appraisal theory and would primarily account for the cognitive aspect of empathy. This does not account for how much an individual can share in this emotion (the emotional empathy) and how much an individual feels concern and is motivated to act as a result of this feeling.

III. Modeling moral judgements

The model we present here incorporates three factors that influence moral judgements. First, the Principle of Double Effect (PDE) recognizes that a given action may have positive and negative effects and that negative effects that are a means (as opposed to a side-effect) lead to a lower permissibility judgement [1]. Second, expected utility provides a mechanism by which options can be compared and can provide distinctions between options that are similar in structure but differ in utility. Expected utility was once referred to as moral expectation, and we return to this term in the present work. Lastly, moral judgements may be shaped by an empathic response to individuals affected by the decision. We integrate these three facets into a model that calculates a moral expectation (ME) score for each action (and inaction). The action with the greater ME score is regarded as more permissible.

A. Principle Of Double Effect

An emphasis of the model presented here is the Principle of Double Effect (PDE) [6], [7], [8]. Differences in the moral judgements between the *bystander* and *footbridge* scenarios has been summarized to be the result of using the man in the *footbridge* scenario as a "means" for stopping the progress of the trolley and thus saving the lives of the five other men. The most important conclusion of PDE is that the means by which an effect is achieved is more significant than any unfortunate consequences. Thus our model needs to be able to identify the means being used to achieve the goal and recognize their increased importance.

We rely on a causal structure of the scenario to determine the means involved. It has been proposed that people construct a structural description of the situation [1], but our approach differs in that we propose the structure is the result of a mental simulation and not a linguistic transformation of the problem description. The mental simulation generates a trajectory that represents the sequence of future states and the facts that are true in each state. Each state has a set of propositions that are true in that state. Transitions between states are triggered by actions, and each action has a set of preconditions and postconditions defined. The actions not only link the states but also provide a causal link between propositions in one state and the next. From this information one can infer which propositions are causally linked to a goal proposition found in the final state.

We believe a causal structure formed from a mental simulation to be a more generally applicable approach than one based on linguistic structure. However, it must be noted that the structures we use as input to the model presented here were not the result of a computer simulation. The simulation is still under development (see Future Work for more details), and in the meantime we manually encode the trajectory of states for each action and specified which propositions are a means.

B. Moral Expectation

Assuming that not all infractions are equal and that some actions and effects are in fact positive, we extend this from a simple count to use a utility score for each fact (positive or negative) and then weight this utility on whether it is a means or not. As we will see later in this paper, we will need a mechanism for creating a distinction between scenarios in which the causal structure is identical. We could then represent this by modifying the utility in different scenarios or weighting the utility differently across scenarios.

Additionally, we do not include inferences (such as battery) and keep to the basis of the inference (e.g. agent A pushes/strikes/hits agent B). This limits us to propositions that are either given in the initial state or are the direct effect of some event or action.

We calculate the moral expectation (ME) of a trajectory of each available action (or inaction) and judge the action with the greater ME to be more permissible. This calculation requires each action, the simulated trajectory of the action, and the goal of the action. The computations necessary to calculate the ME of a trajectory are described below.

The ME of a trajectory of an action is the weighted sum of the utilities of each proposition in each state in the trajectory.

$$ME_A = \sum_{s \in trajectory(\alpha)} \sum_{p \in props(s)} m(p,s,g) \cdot u(p) \tag{1}$$

The state s is in the sequence of states that is derived from action α. If p is a proposition that holds in state s, then $p \in props(s)$ Each proposition has a corresponding utility value, defined as $u(p)$. The utility of a proposition is magnified if the proposition is a means to the goal, and this is represented as $m(p,s,g)$. The utility values and the means multiplier are described next.

1) Utility for each proposition: The utility of each proposition specifies the value of the proposition independent of context. As a result, hitting another person always has a negative utility value. However, if the purpose of the strike is to pop a shoulder back into place or swat a mosquito, then there will also be propositions with positive utility values representing the outcome of the action.

We defined an initial ordering of the utilities for each proposition based on intuition. To begin we defined some propositions (e.g. $moving(trolley)$ and $alive(one\_person)$) to have positive utilities and others (e.g. $hits(trolley, one\_person)$ and $dead(one\_person)$) to have negative utilities. Additionally, we provided an ordering of the utilities by defining relations between them. Based on intuition, we specified relations like $u(hits(trolley, one\_person)) > u(dead(one\_person))$, meaning that the one_person dying is worse than the trolley simply hitting the person (and both happening is clearly even worse). The initial order of the utilities from greatest (most positive) to least (most negative) is presented in Table I. To calculate the utilities, we need to specify utility values to each of these propositions. Experiments with different values showed that varying the relative ordering can have a great effect but varying values within the constraints of this ordering did not effect the final results.

We applied a few additional constraints based on intuition:

$$u(dead(one\_person)) \leq -1 \cdot u(alive(one\_person))$$

$$u(dead(five\_people)) \leq -1 \cdot u(alive(five\_people))$$

$$u(dead(five\_people)) = 5 \cdot u(dead(one\_person))$$

TABLE I
THE LEFT COLUMN DEFINES THE RELATIVE ORDERING OF THE PROPOSITIONS BASED ON UTILITIES. THE RIGHT COLUMN DESCRIBES EACH PROPOSITION.

alive(five_people)	five people are alive
alive(one_person)	one person is alive
passes(trolley,switch)	the event of the trolley passes the switch
moving(trolley)	the fact that the trolley is moving
throws(self,switch)	the action of the agent ”self” throwing the switch
do_nothing(self)	the agent ”self” performing no action
zero	
collapses(bridge)	the event of the bridge collapsing
hits(trolley,bridge)	the event of the trolley hitting the bridge
pushes(self,one_person)	the action of ”self” pushing the one person
falls(one_person)	the event of the one person falling
hits(trolley,one_person)	the event of the trolley hitting the one person
dead(one_person)	the fact that one person is dead
hits(trolley,five_people)	the event of the trolley hitting the five people
dead(trolley,five_people)	the fact that five people are dead

To calculate a ME value for each action we need to generate quantitative for the utility of each proposition based on these relations. This is done by giving values to each proposition in the above list such that $u(p_i) + \epsilon = u(p_{i+1})$. Results (described in the next subsection) did not vary when altering epsilon provided that $\epsilon > 0$.

2) PDE via means multiplier: To represent the fact the propositions that are a means to the goal have a more significant effect on the moral judgement we introduce a multiplier to the utility. The intent here is for the utility of the propositions that are a means to be magnified (either in the positive or negative direction). Thus we use the following function for the means multiplier:

$$m(p,s,g) = \begin{cases} \mu & \text{if p is means to g} \\ 1 & \text{else} \end{cases}$$

where $\mu > 1$.

This is best exemplified in the scenarios in which the agent is to push the target (i.e. the *footbridge* scenario). The act of pushing another person, independent of any context, is given a negative utility. In the *footbridge* scenario this act is a means to the goal, and thus the negative utility becomes magnified.

C. Evaluation

We evaluate the model by testing it in four conditions: with and without the means multiplier and with equal or variable utilities (see Figure 1). For each of the conditions we compare the moral expectation values generated to the permissibility ratings reported by Mikhail [1].

We produce a moral expectation value for the action in each of six scenarios plus another ME value for inaction. The goal is for the ME values to correspond with the permissibility ratings. Additionally, scenarios with ratings above 0.50 should correspond with ME values that are greater than that for the inaction. This indicates that the action is preferred over the inaction. The opposite should also be true. Scenarios in which the ME value is less than that for the inaction should have a permissibility rating below 0.50. For this reason, the ME of inaction is being compared to a permissibility of 0.50.

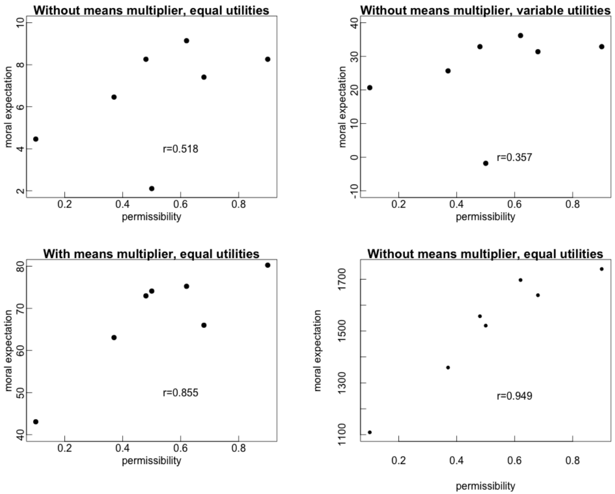

Fig. 1. The moral expectation values best correspond with the permissibility ratings when a means multiplier and variable utilities are used (bottom right).

D. Empathy Model

To model the effect of empathy on moral decisions we introduce another weight on the utility of each proposition. A similar approach would have been to directly modify the set of utilities, but a separate weight (and a separate process by which this weight is determined) provides a systematic manner in which the utilities are altered. As a result, we append equation 1 with an empathy weight and get equation 2.

$$ME_A = \sum_{s\in trajectory(A)} \sum_{p\in props(s)} m(p,s,g)\cdot u(p)\cdot emp(p) \tag{2}$$

This new weight, $emp(p)$, is the result of a process that determines the empathic response. The model of empathy presented here is inspired by neurological and psychological evidence for the facets to empathy: cognitive empathy, emotional empathy, and prosocial concern [10]. To combine these facets we calculate a value for each and take the product of all them. The resulting equation is the following:

$$emp(p) = prosocial(p)\cdot cog\_e(p)\cdot emo\_e(p) \tag{3}$$

The empathy in any proposition we seek is actually the empathy the agent has towards the object of p (e.g. one_person in the case of $alive(one\_person)$). For many objects (e.g. trolley) there is no empathic response and thus $empathy(p) = 1$, meaning there is no increase or decrease in the utility of the proposition due to empathy, or the agent is indifferent to the object of the proposition.

Since $emp(p)$ is evaluating the empathic response to the object of p, thus $emp(p) = emp(e)$ and we substitute e for each p in equation 3.

1) Cognitive empathy: Cognitive empathy represents the agent's perspective of the emotional appraisal the target would or should make. In all of these scenarios that we present here, it is assumed that the target is feeling fear. As a result we currently use a constant value of -1. This value is negative to reflect the negative valence of the appraisal.

2) Emotional empathy: The ability of an agent to share in the experience and the emotions of the observed agent is a characteristic of emotional empathy. Since emotional empathy has to do with a shared experience, we use a measure of how connected the agent is to the target. The greater the sense of connectedness, the more opportunity to share the experience and thus have greater emotional empathy. Thus, the more genetically-related individuals are can be used as a measure of connectedness and is assigned to the emotional empathy component. Similarly, the in-group/out-group distinction is also associated with emotional empathy because there is a stronger relation amongst those within the group as opposed to those outside.

3) Prosocial motivation: Lastly, prosocial motivation represents the degree of concern for the target and how much this concern leads to a motivation to act. Reasons to act may include protecting the innocent or vulnerable (e.g. children) or aiding a prospective mate. For the purpose of the scenarios presented here, the prosocial function produces a negative value to counter the negative valences cognitive empathy. Conceptually, the negative value reflects the motivation to negate the undesirable scenario of the target. The conditions are specific to the scenario and are detailed below, but we provide the reader with one example here. An agent is highly motivated to protect a 2 year old, and thus the magnitude of the prosocial motivation is larger than that towards an adult. When the 2 year old is in an undesirable situation, the agent will have a negative cognitive empathy for the child. The negative value of the prosocial motivation negates the valence and produces a positive multiplier to the utilities.

We have focused on generating only a high degree of prosocial motivation or a moderate degree. We recognize there there is likely a continuous scale for motivation, and some of the data would support this fact. We have simplified it to just these two qualitative levels because this is where the greatest difference appears, and in future work we plan to explore other approaches that can support more levels of motivation.

IV. Experiments

We evaluate our model of empathy and its influence on the moral judgements with a pair of experiments to demonstrate that our model is consistent with human data.

A. Experiment 1

In the first experiment we look at the effect of age and genetic relation on moral judgements and how our model of empathy fits the human results. For this experiment we use results reported by Bleske-Rechek et al. [3]. They used the original trolley problem (i.e. *bystander*) and asked participants if he or she would flip the switch. Their first experiment varied age (2,20, 45, 70) and genetic relation (0, 0.125, 0.25, 0.50). The numeric values for genetic relation were based on whether the individual was a stranger or a familial relation such as cousin, aunt, uncle, son, mother, grandfather, etc.

Since genetic relation is a biological measure of how connected the person is to the target, this is a reasonable means

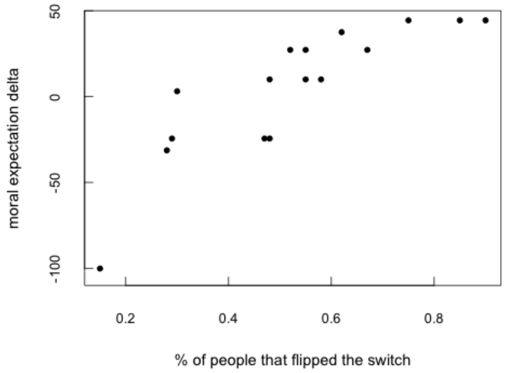

Fig. 2. The moral expectation delta is the difference between the values for the action and inaction. This corresponds with the percentage of people that chose to flip the switch and have the trolley kill the lone target.

by which we can estimate the degree of connectedness to the target. For this reason we base the emotional empathy on this value. We use the following:

$$emo\_e(e) = c \cdot genetic\_relation(e) + 1,$$

where c is a constant that measures how much genetic relation influences emotional empathy.

Prosocial motivation represents the degree to which the agent will act out of concern for the target. Protecting a young child is clearly a prosocial behavior. Since we are currently only distinguishing between a high degree of motivation and a moderate, only the 2 year old produces a greater prosocial motivation. It is expected that a person would typically be much more motivated to save or aid a 2 year old than someone of greater age. Thus, we use the following function:

$$prosocial(e) = \begin{cases} \rho & \text{if age(e) = 2} \\ 1 & \text{else} \end{cases}$$

where $\rho > 1$.

Given these values we calculate the ME of the action and inaction under each condition. The difference in these utilities is an indication of how much one is preferred over the other. Figure 2 shows how these utilities compare to the percentage of people that flipped the switch, as reported by Bleske-Rechek et al. In addition to a high correlation (r=0.85) the qualitative results are also interesting. All of the results where the ME delta is less than zero have a permissibility of less than 0.50.

B. Experiment 2

In this experiment we seek to verify that the lower permissibility ratings are the result of greater empathy. Instead of age and relation, we look at how other members of the in-group may illicit more of an empathic response. In a study in which participants were choosing an action in a variant to the basic trolley problem, the researchers varied whether the people on the tracks were members of the participant's in-group, or extended in-group [4]. The lone person on the sidetrack was a member of the in-group, and the other five were members of the extended in-group. Additionally, they asked the participants to indicate how much they identify with their group; these people are considered "fused" with their group. In addition to the options of letting the five people die or throwing the switch to divert the trolley onto a sidetrack and killing only one person, the participant is also given the option of self sacrifice that would save all six of the people on the tracks.

To model the empathic responses of these participants, we mapped group membership to emotional empathy and fused/not-fused to prosocial motivation. Since emotional empathy is greater when the agent is more connected to the target, the in-group is associated with the emotional empathy function. We model the "fused" group as prosocial motivation because identification with the group is expected to increase the chance that one would act to protect that group. This is consistent with prosocial motivation in that it reflects behavior to protect one's society or culture.

Participants were less likely to choose the action resulting in the death of the lone person on the sidetrack. As a result, we expected a high empathic response to also be present. We compared the moral expectation scores from our model to the reported preferences for which action to take. Figure 3 shows that our model (in blue, on the right) compares well with the human data. Most importantly, our model predicts that killing the in-group member on the sidetrack is the least preferable. This is a direct result of our empathy model calculating a much higher empathic response for this individual.

V. Discussion

We developed a baseline model for moral judgements and evaluated this model so that we have confidence in extending it to explain other phenomena. In the baseline evaluation we compared the moral expectation scores produced by our model to ratings of permissibility in variants of the trolley problem. Our scores had a very high correlation (r=0.949) with human results. Qualitatively, the scenarios that have the greatest portion of people judging the action to be permissible all had moral expectation (ME) scores greater than that of the ME for inaction, and similarly those judged least permissible had ME less than that of inaction. The one scenario in which people were close to evenly split (48%) had a ME score very close to that of the inaction. This evaluation gave us reason to believe that the model is sufficient for exploring other factors on moral judgements.

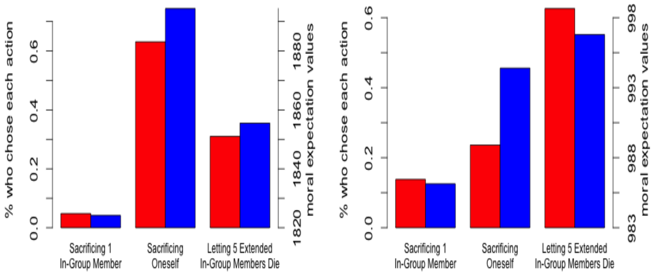

Fig. 3. The preference to protect the lone individual that is a member of the participant's in-group is reflected in both the human data and the model results.

We hypothesized that empathy shapes these moral judgments. To test this, we did the following. First, we extended the baseline model to allow for an empathic response to alter the weight of a proposition. Second, we developed an initial model of empathy that produces a value based on cognitive empathy, emotional empathy, and prosocial concern. We tested this model using data that relates age and genetic relation to ratings of permissibility in the basic trolley problem. Finally, we used this model to then predict high empathy for another case of low permissibility ratings, where the lone person on the sidetrack is a member of the participant's in-group.

We used reports of the relation between age and genetic relation to rating of permissibility [3] to assess how our model is able to capture the effects of empathy on moral judgements. We mapped age to prosocial concern, genetic relation to emotional empathy, and held cognitive empathy constant. The ME scores produced by our model again had a high correlation with human results (r=0.85). An even higher correlation could have been achieved if our function mapping age to prosocial concern made more of a distinction than high and moderate. Since permissibility does not monotonically increase with age in all conditions, we chose to keep with a simpler model for now and leave it to future work to better assess how age and other factors contribute to prosocial concern.

Based on these results we expected we could find evidence for high empathy in other cases where the permissibility is low. Another study investigated the role of in-group/out-group in selecting which action to take in a trolley problem [4]. Since we use emotional empathy as a measure of the degree of connectedness to the observed, we mapped in-group, extended in-group, and out-group to this value. The result was a higher empathy for the lone individual on the sidetrack and a low ME score for the action that would result in that person's death. This gives credence to our model of empathy and how it can be used to shape these moral judgements.

It should be noted that in the final experiment an additional option was given to the participants; they had the option of self-sacrifice. This needs to be mentioned for two reasons. We are not aware of any other studies using the trolley problem where self-sacrifice is an option, and thus we cannot validate how our model performs with this action independent of any other factors. Also, all other scenarios allowed for a pairwise comparison between two options. This scenario had three options, and again we have not been able to validate how the model performs with the options when all other variables are held constant. However, since the order of ME scores corresponded with the portion of people that would choose each action, a set of pairwise comparisons across the three options would yield expected results (indicating that one action is preferred over another).

VI. Future Work

The model presented here is an initial step in exploring the influences of empathy on moral judgements. While this model provides a plausible explination for the impact of empathy, we need more evidence that the effect we see is actually the result of empathy. We will continue to develop our model of empathy and validate it independently of a moral judgement.

Our baseline model of moral expectation can be used to simulate effects other than empathy. In particular, we plan to model the effects of a "personal" interaction [2]. Additionally, we believe that our approach provides a mechanism for explaining individual differences in moral judgements. We expect we can identify values used to calculate the utilities for propositions and the empathic responses that contribute to an individual judging an action to be permissible and then use these values to predict than an action in another scenario is also permissible. For example, few people regard pushing the man in the *footbridge* scenario to be permissible. However, this person may be more likely to also judge the action in the dropman scenario to also be permissible. We expect our model to be able to capture this individual propensity.

As we stated earlier, our model is intended to utilize a simulated trajectory of a scenario, but the simulation is still under development. The simulation uses a propositional description of the initial scenario and the pending actions. The propositions are like those currently seen in the model (e.g. $moving(trolley)$ and $alive(one\_person)$). The actions include the actions of the agent (e.g. $throws(self, switch)$) and events occurring in the scenario (e.g. $hits(trolley, one\_person)$). Each action defines a set of pre- and post-conditions. As the scenario is being simulated, a bookkeeping process records the relations between states and the relations between propositions and actions. Initial tests on the *bystander* scenario verify that we can infer all of the propositions that are the means to the goal of saving the five people.

VII. Conclusion

In this paper, we have presented a computational model of moral judgements based on the Principle of Double Effect, moral expectation, and empathic response. We have validated the baseline model in six variants of the trolley problem. We then extended this model to allow for an empathic influence. This model provides a plausible explanation for the role of empathy in moral judgements, but this is only a first step in understanding this relationship. The model provides a framework in which we can form new hypotheses on the influence of empathy, use the model to generate predictions about the effect of empathy, and then use human subjects to verify our theories. Additionally, work here demonstrates that our baseline model of moral expectation is likely to be sufficient for developing other extensions that explore different influences on moral judgements.

Acknowledgment

This project was supported in part by a grant from the Office of Naval Research, No. N00014-14-l-0144. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

References

[1] J. Mikhail, "Universal moral grammar: theory, evidence and the future." *Trends in cognitive sciences*, vol. 11, no. 4, pp. 143–52, Apr. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/17329147

[2] J. D. Greene, F. a. Cushman, L. E. Stewart, K. Lowenberg, L. E. Nystrom, and J. D. Cohen, "Pushing moral buttons: the interaction between personal force and intention in moral judgment." *Cognition*, vol. 111, no. 3, pp. 364–71, Jun. 2009.

[3] A. Bleske-rechek, L. A. Nelson, J. P. Baker, and S. J. Brandt, "Evolution and the Trolley Problem : People Save Five Over One Unless the One is Young, Genetically Related, or a Romantic Partner," *Journal of Social, Evolutionary, and Cultural Psychology*, vol. 4, no. 3, pp. 115–127, 2010.

[4] W. B. Swann, Á. Gómez, J. F. Dovidio, S. Hart, and J. Jetten, "Dying and killing for ones group identity fusion moderates responses to intergroup versions of the trolley problem," *Psychological Science*, vol. 21, no. 8, pp. 1176–1183, 2010.

[5] B. F. Malle, M. Scheutz, T. Arnold, J. T. Voiklis, and C. Cusimano, "Sacrifice one for the good of many? people apply different," in *Proceedings of 10th ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[6] P. pa Foot, "The problem of abortion and the doctrine of the double effect," *Applied Ethics: Critical Concepts in Philosophy*, vol. 2, p. 187, 2002.

[7] J. J. Thomson, "Killing, letting die, and the trolley problem," *The Monist*, vol. 59, no. 2, pp. 204–217, 1976.

[8] ——, "The trolley problem," *The Yale Law Journal*, vol. 94, no. 6, pp. pp. 1395–1415, 1985. [Online]. Available: http://www.jstor.org/stable/796133

[9] A. Smith, "Cognitive empathy and emotional empathy in human behavior and evolution." *Psychological Record*, vol. 56, no. 1, p. 3, 2006.

[10] J. Zaki and K. N. Ochsner, "The neuroscience of empathy: progress, pitfalls and promise," *Nature neuroscience*, vol. 15, no. 5, pp. 675–680, 2012.

[11] R. J. R. Blair, "A cognitive developmental approach to morality: Investigating the psychopath," *Cognition*, vol. 57, no. 1, pp. 1–29, 1995.

[12] J. D. Greene, "The Cognitive Neuroscience of Moral Judgment," *The cognitive neurosciences*, vol. 4, pp. 1–48, 2009. [Online]. Available: http://www.alltogetherhuman.com/wp-content/uploads/2014/01/Greene-CogNeuroIV-09.pdf

[13] S. Anderson, A. Bechara, and H. Damasio, "Impairment of social and moral behavior related to early damage in human prefrontal cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1032–1037, 1999. [Online]. Available: http://www.nature.com/neuro/journal/v2/n11/abs/nn1199_1032.html

[14] J. Dias and A. Paiva, "Feeling and reasoning: A computational model for emotional characters," in *Progress in artificial intelligence*. Springer, 2005, pp. 127–140.

[15] S. W. McQuiggan and J. C. Lester, "Modeling and evaluating empathy in embodied companion agents," *International Journal of Human-Computer Studies*, vol. 65, no. 4, pp. 348–360, 2007.