Towards an Affective Robot Capable of Being a Long-Term Companion

Jason R. Wilson
Human-Robot Interaction Lab
Tufts University
Medford, MA 02155
Email: wilson@cs.tufts.edu

***Abstract*—While it has been well established that affect influences judgments and decision-making, few computational models of the phenomenon exist. The work I have done and propose focuses on the role of affect in complex decision-making or judgment tasks where a pure utilitarian approach does not reflect human behavior. It is especially challenging to develop models that can predict choices made by an individual. However, as we approach having companion robots that have long-term relationships with a user, it becomes increasingly vital for the robot to be sensitive to the affect of the human and to behave in a manner that is consistent with the expectations of the user and society. The models and underlying robot architecture I describe here bring us closer to robots being accepted in our homes.**

Keywords*—*affect; human-robot interaction; decision-making; computational models

I. Introduction

Robots in the home is no longer the subject of science fiction. For years people have had robotic vacuums in their homes. It will not be long before robots are being used in homes to assist in caring for the elderly, the disabled, or children. Unlike a vacuum cleaner, these assistive robots will need to interact with humans, be sensitive to emotions, and may be required to make complex decisions autonomously. Additionally, the level of interaction with these robots goes far beyond that which we can have with a vacuum. Humans will have conversations with the robot and develop an emotional bond with them.

People have emotional bonds and conversations with other people, their pets, and even inanimate objects like their computer. Virtual agents have been shown to be able of developing and maintaining a relationship. In comparing an agent that was strictly focused on the task with one that also developed a relationship with the user, participants rated the relational agent as more liked, more respected, and more trusted [1]. I aim to extend this work to robots because there are numerous challenges introduced by the agent being a robot. For example, the robot may have more capabilities due to its ability to physically interact with the world. Also, some people develop tight emotional bonds with a robot, treat it like a pet, and even mourn its demise [2]. Additionally, physical embodiment of the agent can affect how people interact with it, such as giving fewer instructions to a simulated robot than an embodied one [3].

If robots are to enter our homes and be long-term companions, then we need to be sure that they can handle the complexities of the daily interactions with the human companion. Let us look at a scenario in which a robot is designed to assist an elderly person in daily activities and communications. The robot is not intended to replace the in-home care from a nurse or other human care-taker, and instead can supplement the care in the gaps between the visits and can provide additional information to the care-taker. In the following scenario, A nurse visits Patty, and they begin by having a casual conversation about how Patty is doing. The dialogue might start like the following:

> Nurse: How was your week, Patty?
> Patty: Good, thank you.
> Nurse: And how are things with you daughter?
> Patty: Fine. Why do you ask?
> Nurse: Well, you've had some disagreements with her lately.
> Patty: Oh, that. No, everything is fine.

While Patty is attempting to convey that she has had a good week, she actually had a confrontation with her daughter in which Patty became very angry and later depressed. Patty is ashamed of her behavior and does not want to let the nurse know about it. However, the robot witnessed the confrontation and is also able to detect Patty's current discomfort. The robot is in a difficult position in which it can either relay this relevant information to the nurse and whereby hurt the trust Patty has in the robot, or it can sympathize with Patty's shame and assist in her cover up.

This scenario imposes numerous requirements on the robot, amongst them it must be able to:

- recognize the emotions of Patty,
- recall previous experiences and associated emotions,
- adapt to Patty's emotions and behaviors,
- predict the impact (emotional and otherwise) of the robot's actions,
- consider the current emotions of Patty (and the nurse) in making a decision on how to act, and
- explain why it chose the action and why it did not choose the other available actions.

For robots to enter the home, researchers must address these needs. In general, vital to a long-term companion robot are the

abilities to be:

- sensitive to affect,
- simulate and evaluate many actions and outcomes,
- justify its decisions, and
- adapt to the individual.

My work addresses these issues along two fronts. First, I am developing computational models of decision-making that are influenced by affect and can be tuned to an individual. It is my hypothesis that the influence of affect must be incorporated for reliable decision-making models to predict individual behavior in non-utilitarian scenarios, such as moral dilemmas. Potential influences include empathic responses biasing one to protect a family member [4] or the emotional saliency of personal harm deterring one from choosing an action [5]. Second, in order for a robot to use these models of decision-making, the cognitive architecture of the robot must be able to do mental simulations, learn and adapt how it emotionally appraises a situation, and be able to justify all of its actions. The explanation a robot gives includes the predicted impact of each action available to the robot, how it came to conclude each effect, and why the chosen action was determined to be better than the other options.

II. Background

I am focusing on scenarios in which an agent, a human or a robot, needs to make a choice and there is not necessarily a right answer. Approaches that are purely utilitarian are not always sufficient in these cases. An example of this is a set of scenarios referred to as the Trolley Problems [6], [7], [8]. This is a set of moral dilemmas in which the utilitarian option is not always preferred. The base scenario, called the Bystander, is the following [7]:

> Frank is a passenger on a trolley whose driver has just shouted that the trolley's 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.

More than 90% of the time, people chose to throw the switch, sacrificing the one person while saving five people [9]. This is the utilitarian option. Now let us consider the Footbridge scenario:

> George is on a footbridge over the trolley tracks. He knows trolleys, and can see that the one approaching the bridge is out of control. On the track back of the bridge there are five people; the banks are so steep that they will not be able to get off the track in time. George knows that the only way to stop an out-of-control trolley is to drop a very heavy weight into its path. But the only available, sufficiently heavy weight is a fat man, also watching the trolley from the footbridge. George can shove the fat man onto the track in the path of the trolley, killing the fat man; or he can refrain from doing this, letting the five die.

Now the majority of people would not choose the utilitarian option and prefer not to push the person even though it would save five people [9].

One theory for the differences in judgments between these scenarios is that in the latter one, the man is being used as the "means" for stopping the progress of the trolley and thus saving the lives of the other men. The Principle of Double Effect (PDE) distinguishes between side-effects and means [6], [7], [8]. If the effects of an action are the means to achieving a goal, then that action is less permissible than an action where the same effect is simply a side-effect. This theory has been used to describe human judgments of permissibility on a variety of trolley problems [9].

However, PDE cannot explain differences in judgments for the same trolley problem when the identities of the people on the track are manipulated but all other details of the scenario are held constant. For this we will look at the role of empathy and conditions under which it biases decisions. There are many definitions for empathy, such as "an affective state that stems from the apprehension of another's emotional state or condition, and that is congruent with it" [10]. Observing people being trapped on a track with a runaway trolley coming towards them can illicit an empathic fear response for those trapped people. It is even more likely to illicit a response and alter behavior if the person on the rack is a family member [11]. It is possible that recognizing the pending doom of the situation generates emotions that are congruent with the fear and despair the people trapped on the tracks would feel. I have modeled these variants to the trolly problems and demonstrated that empathy could account for the different judgments one makes in these scenarios [4]. More details of this model and how it fits into my overall research plan is described in the next section.

Trolley problems are convenient scenarios to investigate effects on moral judgment, but it is also important to investigate social scenarios. The effect of emotional expressions of a virtual agent have been shown to effect the cooperativeness of humans in a prisoner's dilemma game [12]. While trolley problems and the prisoner's dilemma are carefully crafted examples that may not bare strong resemblance to real-life scenarios, a robot cannot be expected to make more complicated decisions in scenarios that have many more dimensions if it cannot first handle these more simple ones. That being said, one of my efforts is to develop more ecologically valid scenarios to begin to bridge the gap between these simplistic ones and highly-complex real-life scenarios.

An assistive robot that interacts with a human though natural language and is able to reason about affect and morals requires a comprehensive cognitive architecture. The DIARC architecture has been used to develop natural human-like human-robot interaction capabilities [13]. This provides a mechanism for representing and reasoning about goal and

actions and has a deeply integrated processing for affect to bias goal prioritization. The goal and action representations and reasoning capabilities are being extended to handle the scenarios described here.

III. Computational Models

The foundational work necessary for developing a robot that is designed to be a long-term companion and is sensitive to the social and affective information in daily interactions is the development of the computational models that the robot will utilize. I describe in this section the ongoing work on an appraisal model, an empathy model, and a model of moral decision-making.

The computational models I am developing have two purposes. One is to inform the design of further experiments with human subjects. The second function is to enable a robot to make decisions with respect to emotions by integrating the models into a cognitive architecture running on a robot. My work has included models of appraisal, empathy, and decision-making. My hypothesis is that the computational model of decision-making that incorporates the models of appraisal and empathy will be better predictors of individual results in scenarios in which a pure utilitarian approach fails. In particular, I have been evaluating the model on decisions in moral dilemmas and plan to extend this to social games (e.g. prisoner's dilemma) and more ecologically valid scenarios.

A. Models Of Appraisal

There exists many computational models of appraisal. Some, like EMA, use production rules to specify the conditions for an emotion [14]. For example, *joy* is the appraised emotion when

$$Desirability(self, p) > 0 \,\&\, Likelihood(self, p) = 1.0.$$

The WASABI architecture takes a different approach and models emotions in a three-dimensional space of pleasure, arousal, and dominance [15]. Important features of WASABI are that it distinguishes between non-conscious and conscious appraisals and that it has a memory of previous appraisals that influence future appraisals. I have proposed a multi-phase model of emotion appraisal that is similar to EMA in that it uses nearly identical rules for appraising, but it is also similar to WASABI in that the deliberative appraisal is preceded by determining an affective reaction (an automatic emotional response to the situation without conscious deliberation) and previous appraisals are held in memory and influence future appraisals [16].

One theory of emotion is that it is a feedback system that can be used to shape future behavior [17]. The feedback system is initiated by a retrospective appraisal of an action. This appraisal can be held in memory or used to adjust the mechanism used to do the initial appraisal. Future behavior can then be influenced by the knowledge it has learned. The multi-phase model of emotions I proposed approaches this problem by storing the retrospective appraisal in long-term memory and when a later scenario is encountered that is analogous to the previous scenario, the past appraisals get projected onto the new scenario [16]. This creates a basic prediction of the emotional response the agent should have to the scenario. Afterward, the scenario gets reappraised, and the cycle continues. The feedback loop allows the agent to continuously adapt to its environment and learn new (perhaps more accurate or more appropriate) appraisals.

Future work will include investigating mechanisms to improve the learning of appraisals based on the retrospective analysis. The intent is for the agent to be able to make better predictions of the emotion that will occur by updating the conditions for the emotion.

In addition to making better predictions in general, an approach that is based on learned experiences is able to adapt to an individual's tendencies. This will be necessary when this appraisal model is used to predict the emotional response of another agent instead of just the self. I will discuss this more in relation to cognitive empathy in the next subsection.

B. Models Of Empathy

In many social scenarios, the decision of an agent is not only influenced by the emotion of the decision-maker but also the emotions of other agents involved. In particular, an agent may have an empathic response to the emotions of another. A prototypical example of this would be a mother vicariously feeling the pain or sadness of her child when the child is hurt or fails. In much of the psychological literature on empathy, the focus is on cognitive and emotional empathy (e.g. [18]). Cognitive empathy involves taking the perspective of another agent and making an inference of that agent's emotional appraisal that is distinct from one's own appraisal. Emotional empathy allows the agent to vicariously experience the emotion of the observed. An example of emotional empathy is the greater empathic response for members of an in-group as compared to those from the out-group. Many have also related prosocial behavior to empathy [10], [19], and others have represented prosocial motivation as a distinct facet (in addition to cognitive empathy and emotional empathy) [20]. Prosocial motivation relates to the agent's willingness, or impulse, to engage in prosocial behavior out of concern for the other agent.

While there are numerous computational models of appraisal, there are very few computational models of empathy. Most of these models are designed to visually present empathic responses and do not need to have any impact on the behavior of the agent [21], [22], [23]. These models do not have representations that distinguish between cognitive empathy, emotional empathy, and prosocial motivation. The model I have recently proposed [4] individually represents each of these facets. I have evaluated this model in the context of making decisions in a moral dilemma, trolley problems. I briefly describe here how each dimension of empathy is represented in these scenarios.

Cognitive empathy is the predicted emotional appraisal of another agent. In the trolley problems, where the agents are faced with the pending doom of a runaway trolley, all agents

are likely to feel some sense of hopelessness or despair. For the purposes of the evaluation, I have made the assumption that all the agents involved have the same negative appraisal of the situation. In the future, the agent will make an appraisal using the model described above. It is important to note that since cognitive empathy requires the agent to take the perspective of another, any memories of previous appraisals it utilizes to make the appraisal must be based on the agent's observed emotional response to previous scenarios. This will require architectural mechanisms that can enforce the separation of memories of the agent's one appraisals from the agent's observations of another agent's emotional responses.

Emotional empathy allows the agent to share in the emotional experience of another agent. The transfer of emotions from the observed to the observer is facilitated by how close (by some measurement, but not necessarily metric distance) that observer perceives the observed to be. Since there is a closer connection to one's in-group, as opposed to the out-group, the transfer and sharing of the emotional experience is easier. Similarly, family - who share genetic information - can be the source of greater emotional empathy.

Prosocial concern is based on some motivation to act out of concern for the well-being of another. Babies and young children often evoke greater empathic responses than older people. This is exemplified in trolley problems in which participants were more likely to save a two year old than older people [11].

In some cases it is difficult to distinguish between emotional empathy and prosocial concern. If someone protects one's own family, is this due to emotional empathy, prosocial concern, or both? Ongoing model and experiment development is focused on solidifying operational definitions of these terms and validating the distinction between them. I have hypothesized that previous examples of decisions in which the prosocial concern were manipulated will be a good predictor of future decisions involving prosocial concern but unable to predict decisions in which the emotional empathy is manipulated. Similar developments are planned to validate emotional empathy as a distinct facet in the model.

C. Models of moral decision-making

The model of decision-making I have been developing is based on conducting a mental simulation of the available options and selecting the best option [4]. A single trajectory consists of the path of events and actions that lead from the current state to some projected future state. The *moral expectation* of the trajectory determines which action is preferred. One way to calculate the moral expectation of a trajectory is to take the sum of the utilities of elements in the trajectory. For example, the utility of five dying is less than that of one dying. This approach gives the utilitarian choice. In moral dilemmas, I have implemented influences on these utilities based on the Principle of Double Effect. This increases the significance of utilities that are a means to an end, and consequently diminishing the relative significance of elements that are side-effects. The results is that saving five people is preferred in the Bystander scenario but not in the Footbridge scenario.

In addition to influencing the calculation of the moral expectation based on the Principle of Double Effect, I have added the influence of empathy. The model of empathy described above produces another modifier to the utilities. Empathy for a two year old that is a stranger is enough that the model gives a greater moral expectation to the choice of saving the child. Similarly, a greater value is calculated for the choice of saving a parent.

It is important to validate this model in other domains to verify that it generalizes. To show that my model of mental simulation, moral expectation calculation, and biasing from empathy is robust, I am evaluating it on a broader range of scenarios. There are numerous examples of moral dilemmas with trolleys [24], speedboats [24], food supply trucks [25], and more. An interesting challenge to my model is one in which the available actions have equal utility, such as a story in which a medical officer must choose between one person definitely dying from an injection or doing nothing and allowing each of a 100 people a 1% chance of dying [26]. While this problem was investigated from the perspective of risk aversion, given that the story variants have the decision-maker either an outside medical officer or one of the people affected, a model that incorporates the in-group/out-group effects of empathy must be considered. I have modeled group membership as an effect on emotional empathy, and plan to demonstrate how the empathy model in conjunction with my decision-making model can account for the reported results from human participants.

Lastly, in addition to being able to evaluate the options consistent with how a human would, the model also must be capable of producing judgments that are inline with what a human would expect a robot to make. Just because it is acceptable for a human to sacrifice another human for a greater good does not mean that a robot should make the same decision. Humans may put different blame on a robot than on a human for making the same decision in the same scenario [27]. How the model of decision-making I am developing can account for these recent results is beginning to be investigated. We also continue to explore the nature of the differences between the decisions that a human would make and one a human would expect a robot to make. It is paramount that we understand the human expectations and are able to produce decisions that are consistent with it. Without this capability, robots will be rejected by our society.

IV. Robot Architecture

We are currently lacking a comprehensive robot architecture that can represent and reason about affect in moral dilemmas. The models described above are the necessary foundational work, but they require an architecture that is capable of providing the mechanisms the models rely upon. For example, if the moral expectation of an action is based on the trajectory of what will happen if the action is performed, the architecture must be able to produce this trajectory. As a result, a mental simulation mechanism is required. Given the current state of

the world and the actions available, the simulator projects forward into time the sequence of states that will occur as a result of the action. Actions may have physical effects on the state of the world, and they may also have effects on the emotional state of an agent. The representation of the actions must be able to explicitly represent these effects. Current efforts are completing this simulator, integrating it into the robot architecture, and expanding the action representation.

The simulator and the moral expectation calculation act as a mechanism for predicting which action is best. After the action is performed, a retrospective analysis of the action and its results need to be used to update the simulation and recalibrate the calculations. This feedback mechanism will allow it to continuously adapt to the human user of the robot and the rest of the robot's environment. While work is in the early design phase, the plan is to develop scenarios in which the robot can learn the tendencies of a user. One example of this is predicting the response to a trolley problem in which the prosocial motivation is manipulated. Responses by an individual provide feedback to the system with the goal of learning characteristics of that individual so that the prediction on the next trolley problem is more likely to be accurate.

It is unreasonable to expect the robot to always make a correct prediction. As a result, the robot may chose an action that is not consistent with the expectations of the user. As in human-human interactions, this disconnect can be the source of discontent. In human-robot interaction, the discontent can lead to diminished acceptance of the robot, or worse, fear or hatred of robots. Communication about the reasoning for the robot's decision can help alleviate these issues. As a result, the architecture must support the ability to preserve the reasoning it used and supply it as justification for its actions. Returning to the scenario in which the robot assisting Patty does not share with the nurse the information about Patty's confrontation, a reasonable explanation the robot could give to the nurse might be the following.

"Patty was embarrassed and ashamed of her behavior. She asked me not to share it. If I had shared that information with you, Patty would have felt betrayed and would have lost trust in me. This would have made it more difficult for me to assist her. Not telling you about the argument did violate my obligation to inform you of relevant information but it also increased Patty's trust in me. I was not able to find any other secondary consequences of not telling you. As a result, not sharing the information was the better option."

In addition to maintaining the relationship between the human and the robot, the justification the robot provides is also a means for improving its performance and can be incorporated into the feedback mechanism. If the robot provides an explanation that does not agree with the human, the human can specify which step in the reasoning failed, and the relevant models can be the focus of the update process.

V. Conclusion

I have presented here work leading to a robot that can be a long-term companion in the home of a human. This is enabled by computational models of affective appraisal, empathy, and moral decision-making. I expect my model of affective appraisal that can adapt to the user and its environment by utilizing a feedback loop to be an important step forward in computational models of appraisal and a necessary step for a long-term companion. Also, a computational model of empathy that explicitly represents cognitive empathy, emotional empathy, and prosocial motivation is the first of its kind. The model of moral decision-making I have developed is not only able to closely match the behavior of humans, but I have also demonstrated that it can be integrated with a model of empathy to simulate the decisions made by humans in a variety of moral dilemmas.

These models require a unique decision-making architecture capable of simulating the physical and nonphysical (namely, emotional) effects of actions, updating its models used to simulate and evaluate these actions through a feedback mechanism, and preserving the reasoning conducted so a justification for its actions (and non-actions) can be provided. This combination of capabilities is vital for creating a robot that will be accepted into the homes of users. The simulation is necessary for choosing the appropriate action predicted by the models. These models must be able to adapt to the user to improve performance and to consistently perform within the expectations of the user. Without the ability to generate an explanation for its behavior, there is a significant risk that the robots will not be trusted and not accepted.

The models and architecture I have presented will get us closer to having assistive robots in our homes that we can trust and will adapt to human needs. There are simply not enough human care-takers, and assistive robots can help supplement the services they provide. Many people could potentially benefit from these robots, but more research is necessary to ensure that these robots are capable of being sensitive to human emotions and can perform within the expectations of the user. It is my intent that my work not only make an impact on affective computing but that it would lead to making a meaningful and positive impact on peoples' daily lives.

Acknowledgment

This work and future planned work has been and will be 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] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 2, pp. 293–327, 2005.

[2] N. Gilani, "Soldiers in mourning for robot that defused 19 bombs after it is destroyed in blast," *Daily Mail*, Jan. 4, 2012.

[3] P. Schermerhorn and M. Scheutz, "Disentangling the effects of robot affect, embodiment, and autonomy on human team members in a mixed-initiative task," in *Proceedings of the 2011 International Conference on Advances in Computer-Human Interactions*, Gosier, Guadeloupe, France, February 2011, pp. 236–241.

[4] J. R. Wilson and M. Scheutz, "A model of empathy to shape trolley problem moral judgements," in *Proceedings of the Sixth International Conference on Affective Computing and Intelligent Interaction*, in press.

[5] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen, "An fMRI investigation of emotional engagement in moral judgment." *Science (New York, N.Y.)*, vol. 293, no. 5537, pp. 2105–8, Sep. 2001.

[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.

[9] J. Mikhail, "Universal moral grammar: theory, evidence and the future." *Trends in cognitive sciences*, vol. 11, no. 4, pp. 143–52, Apr. 2007.

[10] N. Eisenberg and P. A. Miller, "The relation of empathy to prosocial and related behaviors." *Psychological bulletin*, vol. 101, no. 1, p. 91, 1987.

[11] 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.

[12] C. M. de Melo, P. Carnevale, and J. Gratch, "The influence of emotions in embodied agents on human decision-making," *Intelligent virtual agents*, pp. 357–370, 2010.

[13] M. Scheutz, G. Briggs, R. Cantrell, E. Krause, T. Williams, and R. Veale, "Novel mechanisms for natural human-robot interactions in the diarc architecture," in *Proceedings of AAAI Workshop on Intelligent Robotic Systems*, 2013.

[14] S. C. Marsella and J. Gratch, "EMA : A process model of appraisal dynamics," *Journal of Cognitive Systems Research*, vol. 10, no. 2000, pp. 70–90, 2009.

[15] C. Becker-Asano and I. Wachsmuth, "Affective computing with primary and secondary emotions in a virtual human," *Autonomous Agents and Multi-Agent Systems*, vol. 20, no. 1, pp. 32–49, May 2009.

[16] J. R. Wilson, K. D. Forbus, and M. D. McLure, "Am i really scared? a multi-phase computational model of emotions," in *Proceedings of the Second Annual Conference on Advances in Cognitive Systems ACS*, vol. 289, 2013, p. 304.

[17] R. F. Baumeister, K. D. Vohs, C. N. DeWall, and L. Zhang, "How emotion shapes behavior: feedback, anticipation, and reflection, rather than direct causation." *Personality and Social Psychology Review*, vol. 11, no. 2, pp. 167–203, May 2007.

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

[19] M. L. Hoffman, *Empathy and moral development: Implications for caring and justice*. Cambridge University Press, 2001.

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

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

[22] 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.

[23] H. Boukricha, I. Wachsmuth, M. N. Carminati, and P. Knoeferle, "A computational model of empathy: Empirical evaluation," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 1–6.

[24] 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.

[25] I. Ritov and J. Baron, "Protected Values and Omission Bias." *Organizational behavior and human decision processes*, vol. 79, no. 2, pp. 79–94, 1999.

[26] W. A. Wagenaar, G. Keren, and S. Lichtenstein, "Islanders and hostages: Deep and surface structures of decision problems," *Acta Psychologica*, vol. 67, no. 2, pp. 175–189, 1988.

[27] 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.