ICRES 2022: 7th International Conference on Robot
Ethics and Standards, Seoul, South Korea, 18-19 July 2022.
https://doi.org/10.13180/icres.2022.18-19.07.id#

CASE-BASED ROBOTIC ARCHITECTURE WITH MULTIPLE UNDERLYING ETHICAL FRAMEWORKS FOR HUMAN-ROBOT INTERACTION

SHENGKANG CHEN, RONALD C. ARKIN, JASON BORENSTEIN
Georgia Institute of Technology, Atlanta, GA, USA 30332-0280
schen754@gatech.edu arkin@cc.gatech.edu, borenstein@gatech.edu

ALAN R. WAGNER
Robot Ethics and Aerial Vehicles Lab, Penn State University
University Park, PA 16702-7000 USA
alan.r.wagner@psu.edu

As robots are becoming more intelligent and more commonly used, it is critical for robots to behave ethically in human-robot interactions. However, there is a lack of agreement on a correct moral theory to guide human behavior, let alone robots. This paper introduces a robotic architecture that leverages cases drawn from different ethical frameworks to guide the ethical decision-making process and select the appropriate robotic action based on the specific situation. We also present an architecture implementation design used on a pill sorting task for older adults, where the robot needs to decide if it is appropriate to provide false encouragement so that the adults continue to be engaged in the training task.

1. Introduction

Making ethical decisions is challenging but it is something the people have to do regularly in their daily lives. Robots may need to have the ability to make similar decisions within the context of human-robot interactions. In real-world situations, people follow different ethical rules and change their ethical decisions according to the situations. Since there is a lack of agreement on a unified ethical framework for human-human interactions, it is likely impractical to develop a unified single ethical framework appropriate for use in human-robot interactions. Moreover, if factors such as moral emotions affect a human's ethical decision-making process, robots may need to be able to make ethical decisions depending on the current emotional context to develop more meaningful human-robot relationships. In this paper, we describe a flexible robotic architecture with cases derived from different ethical frameworks, which potentially allows a robot to produce morally acceptable actions based on the selected ethical framework and the current situation.

2. Background

As robots are deployed in various fields and become more autonomous, human-robot interactions (HRI) are becoming more common. Researchers are noticing the possible ethical issues related to HRI and the need to develop ethical robots [1–4]. Various robotic architectures have been proposed for ethical behaviors [5–7]. In [5], the authors developed a robotic architecture to produce ethical behaviors based on predefined ethical rules and applied it to caregiving scenarios [8]. However, robots using this approach may be limited to well-characterized environments and well-defined rules derived by ethics experts. Alternatively, Abel, MacGlashan, and Littman [6] leverage reinforcement learning to allow robots to learn ethical behaviors, but they found that robots may behave inappropriately in unseen environments. Vanderelst et al. propose an architecture that uses forward simulation with a

human model to evaluate possible robotic behaviors in order to find an appropriate one [7]. However, this model requires accurate human models which may not be readily accessible in many real-world scenarios.

Robotic deception has been an important topic in HRI [1]. Some researchers are concerned about the possible harmful impacts of deception in social robots [9–11]. One of the concerns is that users might overtrust a robots' capabilities and allow the robot to make unqualified decisions [9]. Moreover, Wilson et al. are concerned that robot deception may damage human-robot trust and can even lead to manipulation, especially for aging high-risk populations [11].

Other researchers believe robotic deception can be beneficial to human users [12, 13]. The authors in [12] found that deceptive behaviors of robots allow human users to be more engaged in game-play scenarios. In [13], the authors argue it is ethical for a robot to deceive if it benefits the overall human-robot relationship. To study people's opinions toward robotic deception, researchers distributed a questionnaire. They concluded that although deceptive behaviors decrease human trust in robots, the majority of the participants consider deception acceptable if these behaviors are beneficial to them [14]. However, this study was only limited to low-risk populations.

3. Architecture Design

This paper describes ongoing research [15] with an updated architecture for ethical robotic behavior. The goal of this architecture is to enable a robot to use various ethical frameworks for more robust ethical decision-making in HRI. It aims to produce morally acceptable behaviors in terms of experiences and outcomes for human users in complex real-world environments.

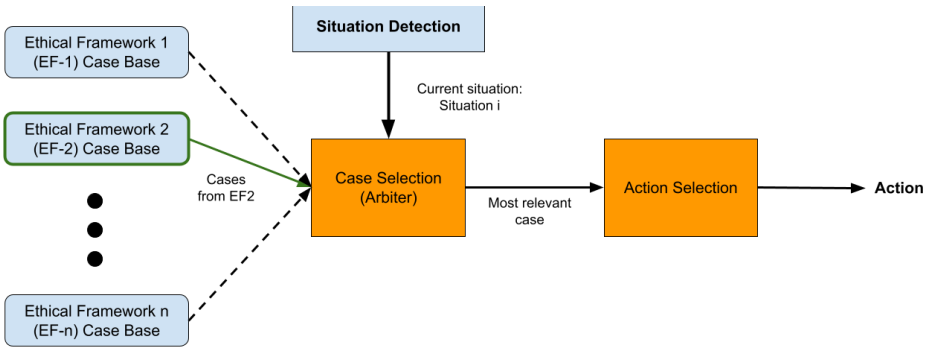

Figure 1. An overview of the architecture for ethical robotic behaviors. Given cases for a selected ethical framework (ethical framework 2 in this figure), the case selection module (arbiter) will select the most relevant case based on the information about the current situation. Then, the action selection module will choose the most appropriate action guided by the most relevant case.

The architecture utilizes the case-based reasoning (CBR) approach, a simple but effective methodology for artificial intelligent agent decision making [16]. In CBR, the robot uses information about the current situation based on decisions made in a similar previous situation from its case base. For case selection, the architecture (Figure 1) contains multiple cases for each ethical framework. Each set of cases contains cases drawn from surveys of people (either laypersons or ethics experts) on their opinions for different situations involving deception. The intent is to ensure that the robot's actions will be consistent with human moral decisions since these actions are guided by cases of human opinions. Each case in the case base is indexed by high-level features about the situation so the architecture can locate and the utilize the information about the current situation to find the most relevant case.

The robot's case selection module will find the most relevant case for a chosen ethical framework (derived in advance) using similarity measures between the current situation and the case indices. Provided with the most relevant case, the action selection module will then output an appropriate action for the robot to execute.

4. Architecture Implementation

This section presents an implementation of the robotic architecture for a specific human-robot interaction scenario: pill sorting with an older adult. Taking medications is part of the daily routine for many older adults, and pill sorting accuracy can be crucially important. However, pill sorting can be challenging for older adults with memory issues and training of the task can lead to frustration. During training, a robot observes and provides feedback about the older adult's performance on the task. In this pill soring scenario, we want to study whether it is moral to deceive an adult in a pill sorting task to keep them engaged with the task. Using the robotic architecture, the robot needs to decide whether to provide false encouragement (deception) or an accurate assessment (truth).

For the ethical framework cases, we considered ethical choices from both regular adults ("folk morality") and formal ethical frameworks: Utilitarianism [17], Kantian Ethics [18], Social Justice Theory [19], Ethics of Care [20], and Virtue Ethics [21]. To create the case base, we conducted two separate survey studies. For folk morality, 100 survey responses were collected through Amazon's Mechanical Turk service in January 2020. For the five formal ethical frameworks, 30 ethics experts were invited to answer the survey and 22 valid responses were received in February 2020. The survey data were then analyzed and used to create cases to populate each corresponding ethical framework. Each case contains the action probabilities derived from the survey data and is indexed by two binary variables (task performance and subject emotional state).

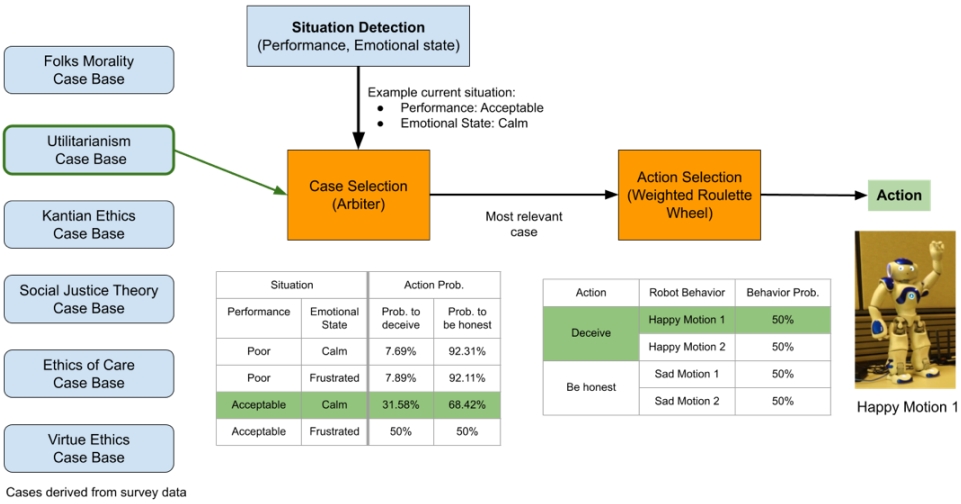

Figure 2. The initial architectural implementation for human-robot interaction in the pill sorting scenario with an NAO robot. The cases are derived from survey data. In this example, the case selection module uses Utilitarianism cases. For example, consider the older adult just made a mistake in a pill sorting task. The situation detection module provides the case selection module the information about the current situation (task performance and emotional state): here, the human user has an acceptable task performance and remains calm. Then, the case selection module finds the most relevant case (highlighted in green) and sends it to the action selection module. In this case, the action selection module chooses a deceptive action using a weighted roulette wheel selection. Happy Motion 1 is selected for the NAO robot to perform in the presence of the user providing false feedback on their result. In this case, the NAO robot deceives the older adult by providing false encouragement.

In the pill sorting task, the robot relies on its ethical architecture (Figure 2) to produce morally acceptable actions based on the chosen ethical framework and current situation. Using this implementation, the case selection module selects the most relevant case for a chosen ethical framework based on the current situation (user task performance and user emotional state). To avoid repetitive robotic behaviors, the robot will use behavior probabilities to generate different behaviors/gestures corresponding to an action (e.g., deceive). Using the action probabilities (derived from survey data) and behavior probabilities (defined by researchers), the NAO robot outputs an action by performing a gesture (e.g., happy motion 1) using the roulette wheel selection method [22] to provide feedback to the older adult on pill sorting results.

Currently, we only consider two binary variables to describe the situation: user performance (acceptable vs. poor) and emotional state (calm vs. frustrated). However, this architecture can easily be extended to more descriptive variables (e.g., scalar variables or categorical variables) with an updated case base. Moreover, more gestures for the NAO robots can be added to make the human-robot interaction process more engaging.

5. Discussion

In a real-world or ethically complex situation, it may not be appropriate to ask the robot to follow a set of fixed ethical rules regardless of the situation. Humans make different ethical decisions affected by differing situations. Thus, it is crucial for robots to be sensitive to the current context if they are going to be able to perform appropriate ethical actions during human-robot interactions. Consequently, this architecture allows the robot to produce appropriate actions based on a selected ethical framework and the current circumstances within which the user is situated. Moreover, the cases of the architecture can be expanded continuously during human-robot interactions by learning and adding new cases, a hallmark of case-based reasoning, which makes the robots more adaptive. This is crucial for building sustainable human-robot relationships.

A novel extension is to incorporate moral emotions into the architecture. Moral emotions [23] (e.g., guilt, empathy and anger) has been shown to play an important role in human ethical decision-making process [24, 25]. As a result, robots also need to take into consideration moral emotions in order to effectively support the human decision-making process.

Currently, we are implementing the architecture and plan to test it on physical robots. We want to conduct a series of HRI studies to evaluate the robotic architecture for two scenarios: pill sorting with an older adult and game playing with a child. We want to investigate whether the generated robotic actions using the architecture are morally acceptable to people under various ethical frameworks in different situations involving human participants, ideally by having an individual interact with a robot, but also through the use of focus groups.

6. Conclusion

In this paper, we present a flexible robotic architecture using a case-based reasoning approach for the generation of ethical behaviors consistent with either folk morality or decisions recommended by ethics experts for use in human-robot interaction. Moreover, we describe an architectural implementation for a specific human-robot interaction scenario: pill sorting with an older adult. In this scenario, the robot needs to decide whether to deceive the older adult by providing false encouragement to allow the older adult to continue the task or instead be honest by providing actual assessment results with the potential consequence of the user discontinuing the training due to frustration. We used the results of survey studies from both regular adults and ethics experts to generate various ethical framework cases that guide the decision-making

process to produce appropriate actions relevant to the current situation. This architecture aims to become a tool for researchers to investigate further how to enable robots to interact with humans ethically.

Acknowledgments

This research project is supported by the National Science Foundation as part of the Smart and Autonomous Systems program under Grants No. 1849068 and 1848974. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

1. R. Wullenkord and F. Eyssel, *Current Robotics Reports 2020 1:3* 1, 85 (2020).
2. M. Anderson and S. L. Anderson, *AI Magazine* 28, 15 (2007).
3. M. Scheutz and B. F. Malle, *2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, ETHICS 2014* (2014).doi:10.1109/ETHICS.2014.6893457
4. A. Leveringhaus, *European View* 17, 37 (2018).
5. R. C. Arkin and P. Ulam, An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions, (2009), pp. 381–387.doi:10.1109/CIRA.2009.5423177
6. D. Abel, J. MacGlashan, and M. L. Littman, *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence* (2016).
7. D. Vanderelst and A. Winfield, *Cognitive Systems Research* 48, 56 (2018).
8. J. Shim, R. Arkin, and M. Pettinatti, *Proceedings - IEEE International Conference on Robotics and Automation* 2936 (2017).doi:10.1109/ICRA.2017.7989340
9. A. Sharkey and N. Sharkey, *Ethics and Information Technology 2020 23:3* 23, 309 (2020).
10. J. Danaher, *Ethics and Information Technology 2020 22:2* 22, 117 (2020).
11. J. R. Wilson, M. Scheutz, and G. Briggs, 377 (2016).doi:10.1007/978-3-319-31413-6_18
12. E. Short, J. Hart, M. Vu, and B. Scassellati, No fair!! An interaction with a cheating robot, (2010), pp. 219–226.doi:10.1109/HRI.2010.5453193
13. J. Shim and R. C. Arkin, *Proceedings - 2013 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2013* 2328 (2013).doi:10.1109/SMC.2013.398
14. K. Rogers and A. Howard, *Proceedings of IEEE Workshop on Advanced Robotics and its Social Impacts, ARSO* 2021-July, 200 (2021).
15. R. C. Arkin, J. Borenstein, and A. R. Wagner, Competing ethical frameworks mediated by moral emotions in HRI: Motivations, background, and approach, (2019).doi:10.13180/ICRES.2019.29-30.07.001
16. J. L. Kolodner, *Artificial Intelligence Review 1992 6:1* 6, 3 (1992).
17. J. Driver, The History of Utilitarianism, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2014).
18. R. Johnson and A. Cureton, Kant's Moral Philosophy, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2022).
19. J. Rawls, *A theory of justice*, Cambridge, Massachusetts, The Belknap Press of Harvard University Press, (1971).
20. Virginia. Held, *The ethics of care personal, political, and global*, Oxford University Press (2005).
21. R. Hursthouse and G. Pettigrove, Virtue Ethics, in *The Stanford Encyclopedia of Philosophy*, Edited by E. N. Zalta, Metaphysics Research Lab, Stanford University (2018).

22. D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Professional (1989).
23. J. Haidt, THE MORAL EMOTIONS, in *Handbook of affective sciences*, Oxford University Press (2003).
24. J. P. Tangney, J. Stuewig, and D. J. Mashek, *Annual Review of Psychology* 58, 345 (2007).
25. C. D. Cameron, K. A. Lindquist, and K. Gray, *Personality and Social Psychology Review* 19, 371 (2015).