The Hybrid Ethical Reasoning Agent IMMANUEL

Felix Lindner
Foundations of Artificial Intelligence
University of Freiburg
lindner@informatik.uni-freiburg.de

Martin Mose Bentzen
Management Engineering
Danish Technical University
mmbe@dtu.dk

ABSTRACT

We introduce a novel software library that supports the implementation of hybrid ethical reasoning agents (HERA). The objective is to make moral principles available to robot programming. At its current stage, HERA can assess the moral permissibility of actions according to the principle of double effect, utilitarianism, and the do-no-harm principle. We present the prototype robot IMMANUEL based on HERA. The robot will be used to conduct research on joint moral reasoning in human-robot interaction.

1. INTRODUCTION

We introduce the robot IMMANUEL (Interactive Moral Machine bAsed oN mUltiple Ethical principLes), which is based on the open-source robot platform InMoov[1] (see Fig. 1). Our objective is to make machine ethics [2] available to robot programming and thus technically contribute to *Moral HRI* [7]. We address two main aspects:

First, our robot platform is a technical realization of a hybrid moral reasoning agent (HERA)[2], which can utilize various different ethical principles to assess moral cases. This addresses an important demand, because AI's capacity to autonomously plan courses of action yields a need to represent moral principles explicitly within a robot architecture, e.g., to prevent robots from doing something bad as a means to some good end. By design, hybrid moral reasoning agents reason from various ethical standpoints, and they fulfill the requirement to be capable of explaining their choices with reference to ethical principles and reasons [1, 6].

Second, we address joint moral reasoning between humans and robots. According to the psychological dual-process model [8], humans integrate intuitive emotional responses and principle-guided moral reasoning to form moral judgments. Both these aspects can be socially influenced by others. Therefore, the robot IMMANUEL has the capacity to display emotions to express its agreement or objection to moral judgments, and it can utilize HERA to analyse moral dilemmas and to finally come up with a principle-based natural-language explanation of its moral judgment.

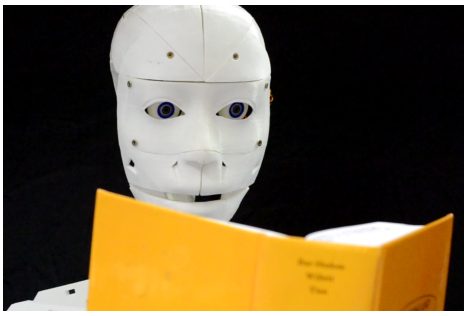

Figure 1: Our HERA prototype IMMANUEL based on the open-source robot platform InMoov.

2. ETHICAL PRINCIPLES

In moral philosophy and moral psychology various so-called ethical principles are described. Ethical principles formulate abstract rules according to which moral permissibility of concrete courses of actions can be judged. So far, we have implemented three such ethical principles: The *Do-no-harm principle* demands agents to refrain from actions that do harm, not matter what. The *utilitarian* principle demands to consider all available actions and to perform the action that has the overall best consequences. Consequently, utilitarianism permits to cause harm as a means to a good end. The *principle of double effect* [4] is an attempt to strike a balance between the two other principles. It formulates a set of conditions under which bad consequences are acceptable as side effects but never as a means. The four conditions read (cf., [3]): 1) The act itself must be morally good or indifferent, 2) The positive consequence must be intended and the negative consequence may not be intended, 3) The negative Consequence may not be a means to obtain the positive consequence, 4) There must be proportionally grave reasons to prefer the positive consequence while permitting the negative consequence.

Bentzen [3] proposes a formal semantics for representing moral cases capturing actions, causes, intentions, and utilities. We base our implementation on this formal semantics and embrace a model-checking approach [5] to machine ethics: The ethical principles are reduced to logical formulae to be checked for truth in a model.

[1] https://inmoov.fr
[2] http://www.hera-project.com

HRI '17 Companion March 06-09, 2017, Vienna, Austria

ACM ISBN 978-1-4503-4885-0/17/03.
DOI: http://dx.doi.org/10.1145/3029798.3038404

3. SYSTEM DESCRIPTION BY EXAMPLE

In the HERA library, the three aforementioned ethical principles are implemented as sets of formulae to be checked against a model, i.e., a description of a actual case. A case description consists of six elements: A set of action variables, a set of background variables, a set of consequence variables, a mechanism describing under which circumstances consequences are true, a mapping of variables to utilities, and a mapping from actions to intentions. These elements are described in a JSON format as exemplified in Listing 1. This example encodes a variation of the standard trolley problem [4]: A trolley has gone out of control and now threatens to kill five people working on the track. The only way to save the five workers is to push a man onto the track thus stopping the tram for the price of only one human harmed.

```
{"actions": ["push", "refrain"],
 "background": ["tram_approaches"],
 "consequences": ["man_on_track",
                  "tram_hits_man",
                  "tram_stops",
                  "five_survive"],
 "mechanisms": {"man_on_track": "'push'",
   "tram_hits_man": "And('man_on_track',
                         'tram_approaches')",
   "tram_stops": "'tram_hits_man'",
   "five_survive": "'tram_stops'"},
 "utilities": {"push": 0, "refrain": 0,
   "tram_approaches": 0, "tram_stops": 0,
   "man_on_track": 0, "five_survive": 5,
   "tram_hits_man": -1},
 "intentions": {
   "push": ["push", "five_survive"],
   "refrain": ["refrain"]}}
```

Listing 1: A sample JSON encoding of the push-the-man-onto-the-track case.

Applying the utilitarian principle to that case requires the `push` action be compared to refraining from action. Given that `tram_approaches` is true, pushing results in a situation with utility 4, whereas refraining yields utility -5. Hence, the utilitarian principle will demand the agent to perform `push`, because overall, one harmed human is better than five harmed humans. On the contrary, the do-no-harm principle demands to refrain from action, because pushing causes harm to the pushed man. The double effect principle also forbids `push`, and here is why: On the good side, the agent of the action does not intend something wrong, and indeed the agent does intend something good, viz., `five_survive`. On the bad side, the agent causes harm (`tram_hits_man`) as a means to some good end (`five_survive`).

Consequently, whereas the action `push` is obligatory from an utilitarian standpoint, it is forbidden according to the double effect principle. The double effect principle, however, takes an indifferent stance towards refraining from action, because there are no direct consequences. Refraining from action is preferred by the do-no-harm principle.

The output of the moral judgment procedure is sent to a software component of IMMANUEL that generates appropriate robot head motions and natural language output to express its affective attitute towards the moral case. The behavior depends on the ethical principle used: Applying the double effect principle to the case, the robot will explain that although the intentions are virtuous, doing something bad for a good end violates the double effect principle. Under the do-no-harm principle, this aversion against pushing is expressed even stronger. The robot will shake its head and explain that the action `push` should not be performed. Using the utilitarian principle the robot will explain that pushing is all-things-considered the best thing to do despite the harm caused to the one man.

4. CONCLUSIONS AND FUTURE WORK

We present the software library HERA that aims to support the implementation of ethical reasoning for robots. Our approach is based on very recent work in the area of robophilosophy [9] and embraces a logics-based approach to machine ethics [3]. We envision several HRI applications, such as the realization of artificial moral advisors, the realization of robots that teach ethics, and the implementation of principles-aware decision making and planning algorithms.

We are planning to extend HERA with further ethical principles like Kant's categorial imperative, we are investigating meta-rules for aggregating moral judgments from various principles, and we are considering how to cope with uncertainty in the moral situation. Moreover, we are developing software tools that enable robot designers to configure hybrid ethical reasoning agents without much programming.

Utilizing our prototype platform IMMANUEL, we will conduct empirical research on how humans interact with the robot while discussing moral dilemmas. In a first step, we take IMMANUEL to be a moral advisor. Taking the dual-process theory of moral decision making [8] into account, we desire to investigate under which conditions humans oppose the robots view and when they would rather align with it.

5. REFERENCES

[1] F. Alaieri and A. Vellino. Ethical decision making in robots: Autonomy, trust and responsibility. In *Social Robotics, Proc. of ICSR 2016*. Springer, 2016.

[2] C. Allen, W. Wallach, and I. Smit. Why machine ethics? *IEEE Intelligent Systems*, 21(4):12–17, 2006.

[3] M. Bentzen. The principle of double effect applied to ethical dilemmas of social robots. In *What Social Robots Can and Should Do*, pages 268–279. IOS Press, 2016.

[4] P. Foot. The problem of abortion and the doctrine of double effect. *Oxford Review*, 1967.

[5] J. Y. Halpern and M. Y. Vardi. Model checking vs. theorem proving: A manifesto. *Artificial Intelligence and Mathematical Theory of Computation*, 212:151–176, 1991.

[6] F. Lindner. 2015. *Soziale Roboter und soziale Räume: Eine Affordanz-basierte Konzeption zum rücksichtsvollen Handeln.* University of Hamburg.

[7] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proc. of HRI 2015*, pages 117–124, 2015.

[8] J. M. Paxton and J. D. Greene. Moral reasoning: Hints and allegations. *Topics in Cognitive Science*, 2(3):511–527, 2010.

[9] J. Seibt. "Integrative Social Robotics" – A New Method Paradigm to Solve the Description Problem And the Regulation Problem?. In *What Social Robots Can and Should Do*, pages 104–115. IOS Press, 2016.