‘Well, at least it tried’ The Role of Intentions and Outcomes in Ethically Evaluating Robot Actions[1]

Daphne Lenders[a] and Willem F.G. Haselager[a]

[a] Dpt. of Artificial Intelligence, Radboud University, Nijmegen, The Netherlands
d.lenders@student.ru.nl

Abstract. In order to make robots more trustworthy, it is important to find out which factors influence a human’s ethical evaluation of a robot. We found that intentions of a robot have a larger positive effect than the outcomes of its actions. Moreover, the influence of outcomes on ethical evaluations is larger when the action was perceived to be based on a good rather than bad intention.

Keywords: Human-robot interaction, Moral judgement, Trust

1 Introduction

Two important factors in morally evaluating the actions of an agent are the intentions behind the actions as well as their outcomes. Current research has shown that humans are especially guided by the intentions of actors, when ethically evaluating humans [1]. As far as we know, no research has however been done on the role of intentions and outcomes in the ethical evaluation of robot actions.

In this study we investigated how the perception of intentions and outcomes related to robot actions influence ethical evaluations of those actions. This was done by showing participants video clips of a robot appearing to have a good or bad intention followed by an action that leads to a good or bad outcome. To each video clip the participants had to give an ethical evaluation of the displayed behaviour of the robot.

Due to the theory of anthropomorphism, which says that humans tend to view inanimate agents like they do humans, it was expected that robots actions would be evaluated similarly to human actions [2]: Perceived good intentions and outcomes were expected to have a positive effect on ethical evaluations, while bad intentions and outcomes were predicted to affect these evaluations negatively. Furthermore intentions were predicted to have a bigger influence on ethical evaluations than outcomes. Finally we expected to find an interaction effect between intentions and outcomes.

Methods

In an online survey, 75 participants were shown videos of 20-50 seconds displaying a Pepper robot that appeared to have either a good or bad intention followed by an action

[1] The full thesis was submitted in fulfilment of the requirements for the degree of Bachelor of Science in Artificial Intelligence at the Radboud University in June 2017

that lead to a good or bad outcome. In a good intention-bad outcome scenario it was e.g. shown how Pepper tried to submit a paper for its owner, but failed because its tablet crashed unexpectedly through which the owner missed an important deadline.

After each video the participants had to answer three questions that together constitute a moral evaluation of the displayed robot.[2] The moral evaluations were used in a two-way repeated measures ANOVA, in order to analyze the influence of ‘intention’ and ‘outcome’ on the moral evaluation of the observed robot behaviour.[3]

2 Results

Just as expected intentions (F(1, 74) = 350.541, p = .000, eta^2 = .826) and outcomes (F(1, 74) = 160.93, p = .000, eta^2 = .685) have large significant effects on the ethical evaluation of actions. Good intentions and outcomes affect ethical evaluations positively, while bad intentions and outcomes affect these evaluations negatively.

Next to that a large interaction effect between intention and outcome was found (F(1, 74) = 49.637, p = .000, eta^2 = .401). The ethical evaluations of actions based on perceived good intentions are more influenced by the outcome of an action than actions based on perceived bad intentions.

3 Conclusion

Good intentions and outcomes affect ethical evaluations positively, but the influence of intentions is stronger than the one of outcomes. This suggests that it may be worthwhile to investigate what makes humans perceive intentions of robots as good or bad. Finding an answer to this may be an essential step towards making robots more trustworthy.

From the results it can also be concluded that humans ethically evaluate robot-actions similar to human-actions, which can be interpreted as further evidence for anthropomorphism: Characteristics that are normally attributed to humans, such as intentions, can also affect the way people view inanimate agents, such as robots [1, 2].

Other studies however found differences in ethical evaluations of human and robot actions [3]. Thus research on ethical evaluation of robots is still in its beginning stage and an overall framework on this topic cannot be established yet.

References

1. Young, L., Saxe, R.: The neural basis of belief encoding and integration in moral judgement. NeuroImage 40(4), 1912-1920 (2007).
2. Fink, J.: Anthropomorphism and human likeness in the design of robots and human-robot interaction. International Conference on Social Robotics, 199-208 (2012).
3. Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., Cusimano, C.: Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction, 117-124 (2015).

[2] Videos and complete survey can be found online: goo.gl/E48acM; goo.gl/P9pP79

[3] In this abstract we chose to focus on our main study only. Additional analyses and their results will not be discussed here