Moral Attribution in Moral Turing Test

Jolly Thomas and Mubarak Hussain, Indian Institute of Technology, Dharwad, (India)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Abstract: This paper argues Moral Turing Test (MTT) developed by Allen et al. for evaluating morality in AI systems is designed inaptly. Different versions of the MTT focus on the conversational ability of an agent but not the performance of morally significant actions. Arnold and Scheutz also argue against the MTT and state that without focusing on the performance of morally significant actions, the MTT is insufficient. Morality is mainly about morally relevant actions because it does not matter how good a person is at conversing about morally relevant actions. When discussing morality, we consider an agent's ability to perform specific actions in a morally given situation. We show that Allen et al. do not take into account the distinction between the performance of the moral attribution and the performance of the morally relevant action. This distinction gives a robust account of assessing the morality of an AI system in the MTT.

Keywords: Ethics of AI, Moral Turing Test (MTT), Artificial Moral Agent (AMA)

1. Introduction

Allen et al. introduce the Moral Turing Test (MTT), similar to the famous Turing Test. The MTT is based on morality concerning the AI system. Both tests are similar because MTT's arrangements remain similar to Turing's original test. Moreover, in both tests, the roles of participants are also not different. The only difference in the case of MTT is that only morality-related questions and answers are permitted in the MTT. Similar to the Turing Test, if the interrogator cannot distinguish the responses provided by the AI system from the human, then AI system passes the MTT. If the AI system responds in such a way that the interrogator cannot distinguish whether it is the human or the AI system, then the AI system passes the MTT.

a) **The First version of MTT:** In the first version of MTT, the responses from the participants purely focus on the ability to converse about morality. The ability to converse about morality is not enough to consider an AI system or even a human as a moral agent. Because when we think of an AI system or even a human as a moral agent, we think not only that the agent can converse about morality but also that the agent can perform specific actions in a morally given situation.

b) **The Second version of MTT:** Allen et al. redesigned the MTT to overcome the issues in the first form of MTT and tried to introduce the performative aspect of a moral agent. In the revised version of MTT, descriptions of the actions are the focus. A pair of descriptions of morally significant actions are provided to the interrogator: one performed by the human and the other performed by the AI system. The AI system will pass the test if the interrogator cannot identify the AI system from the pairs of descriptions of morally significant actions. In the revised version also, instead of focusing on the morally relevant action, Allen et al. provided descriptions of actions and tried to check whether the AI system would be moral or immoral. Focusing on descriptions of actions cannot guarantee that an agent is actually performing the action in a morally relevant scenario. It is true that descriptions of actions may have some moral value in some situations, but the performance of actions is significantly different from the descriptions of actions. Arnold and Scheutz argue that the primary test for moral attribution should be based on morally significant actions. The agents perform morally significant actions in an actual situation rather than on the agent's ability to converse about morality. Allen et al. overlooked the distinction between the performance of moral attribution and the performance of the morally relevant action. This distinction will give a better account of the possibility of MTT. Through such distinction, we can clearly say that we are testing the moral ability of a machine.

2. The Ascription versus Performance Distinction

Performance of the action is to be distinguished from moral predication or moral attribution. Here one who engages in moral predication actually performs the moral predication. Those whose actions are attributed to moral predicates become the performer of the action. In both version of the MTT, Allen et al. talks about the conversational ability of an agent. The revised version of MTT was designed in such a way that it focuses on the descriptions of the actions. The entire process needs some distinctions, which they have not clearly mentioned. The distinction between the moral predication and the performance of morally significant actions.

We argue that though a particular distinction is assumed in the revised version of the MTT, it is not clearly explained and brought into the analysis. This distinction is important because the distinction makes the following points clear: the act of attributing morality and the act of or performance of the actions or morally relevant actions. Therefore, we bring the moral predication versus moral performance distinctions for accessing morality in the AI system. The utility of this distinction is that we can see whether the proposed kind of Turing test focus on the attribution of morality or does it focus on the performance of morally relevant actions.

The first version of MTT focuses on the ability to converse about morality. The participants of MTT do not see the actions and decide the morality behind the actions. The MTT actually assesses the machine's ability to attribute morality. Suppose you describe (or attribute) to the machines that Mr. X lies in a, b, and c situations. Based on the attribution, the machine will respond to whether it is a morally right or morally wrong action. Let us say the machine says that it is a morally wrong action. In this case, we are basically evaluating a machine's ability to attribute morality. We are not analyzing the ability of a machine to perform morally relevant actions. There is a difference between the machine attributing morality and the machine performing morally relevant actions. Suppose, currently, I am typing this article on my laptop without performing a morally relevant action. Still, I can really be good at performing the attribution. I can say that Mr. X helps Mr. Y, which is morally good. In reality, I might be someone who does not help at all. I can be good at performing this particular moral attribution. However, I might not be a person who would be performing the morally right action. Based on the attribution of morality, I can say that person X helps person Y, which is a morally good action, or helping people in their need is morally right. If I help someone in his/her needs in a morally relevant situation, then only my action would be morally right. This distinction is missing in the entire MTT discussion, which is crucial in evaluating AI systems' morality. The human as attributing moral predicate versus the human as a performer of morally relevant action. If we accept this distinction, we can clearly say that actually, MTT has not come well because it is designed in such a manner that it simply evaluates whether the machine is performing the attribution of moral predicates to the actions or the machine is performing morally relevant actions. The MTT does not assess whether the machine is a performer or actually performing morally relevant actions in a specific situation. Arnold and Scheutz also argue against the MTT (2016, 103-115) but do not consider the abovementioned distinction. The performance of attributing moral predicate or moral attributes is to be disconnected from the performance of the action to which the moral predicates are predicated. According to

Arnold and Scheutz, unless MTT focuses on morally relevant actions rather than the ability to converse or the ability to attribute morality, the MTT actually fails. In the MTT, the interrogator gives descriptions of the actions to the participants and asks the machine to attribute moral attributions.

Acknowledgment: This work was supported by the Technology Innovation Hub on Autonomous Navigation and Data Acquisition Systems (TiHAN) of the Indian Institute of Technology-Hyderabad a project under the Department of Science and Technology's National Mission on Interdisciplinary Cyber-Physical Systems.

References

1. Allen, C., Varner, G. and Zinser, J. 2000. 'Prolegomena to Any Future Artificial Moral Agent.' Journal of Experimental and Theoretical Artificial Intelligence 12, 251-261.

2. Arnold, T., Scheutz, M. 2016. 'Against the Moral Turing Test: Accountable Design and the Moral Reasoning of Autonomous Systems.' Ethics and Information Technology 18, 103–115.
3. Turing, A. M. 1950. Computing Machinery and Intelligence. Mind 59, 433–460.