https://doi.org/10.53656/phil2021-04-07

Moral Philosophy
Морална философия

QUESTIONING THE ROLE OF MORAL AI AS AN ADVISER WITHIN THE FRAMEWORK OF TRUSTWORTHINESS ETHICS

Assoc. Prof. Dr. Silviya Serafimova
Bulgarian Academy of Sciences (Bulgaria)

Abstract. The main objective of this article is to demonstrate why despite the growing interest in justifying AI's trustworthiness, one can argue for AI's reliability. By analyzing why trustworthiness ethics in Nickel's sense provides some well-grounded hints for rethinking the rational, affective and normative accounts of trust in respect to AI, I examine some concerns about the trustworthiness of Savulescu and Maslen's model of moral AI as an adviser. Specifically, I tackle one of its exemplifications regarding Klincewicz's hypothetical scenario of John which is refracted through the lens of the HLEG's fifth requirement of trustworthy artificial intelligence (TAI), namely, that of *Diversity, non-discrimination and fairness*.

Keywords: trustworthiness ethics; rational, affective and normative accounts of trust; moral AI as an adviser; HA-AA trust relationships

1. Introduction

The primary objective of this article is to clarify the methodological advantages and concerns about arguing for AI's trustworthiness. For the purposes of mapping the debate, some reasons behind what I call positive projects, viz. projects which encourage the recognition of the concept of trustworthy AI (TAI), with a special focus upon the HLEG project, are explored. In this context, I analyze Nickel's distinction between *justified trust ethics* and *trustworthiness ethics*, since it can contribute to clarifying whether or not one can argue for trustworthiness ethics within scenarios whose operation is underlined by the use of preprogrammed moral algorithms.

Having outlined some concerns about TAI, I also examine the arguments for the recognition of AI's reliability against the background of the human agents (HAs)-artificial agents (AAs) interaction. The latter is considered as involving different types of trustors and trustees. As for an exemplification of the difficulties in grounding the reliability of moral AI, I point out Savulescu and Maslen's model of moral AI as an adviser.

Consequently, the practical outcomes of challenging the reliability of the aforementioned adviser are displayed by Klincewicz's hypothetical scenario of John's racism. In this context, I analyze the role of the HLEG's fifth requirement — that of *Diversity, non-discrimination and fairness* — by demonstrating why it cannot guarantee that the non-discrimination on the input level is a necessary and sufficient condition for eradicating one's discriminative attitudes.

2. What is trustworthiness ethics?

The idea of trustworthiness ethics is introduced as triggering some general concerns about technological trustworthiness which require one to provide a distinction between such ethics and so-called justified trust ethics. By *justified trust ethics* Nickel understands "the application of a plausible minimal rationalist principle" which states that in making important decisions it is best to do it on the basis of adequate reasons (Nickel 2013). As an argument favoring justified trust ethics, he outlines the one that in certain contexts, this ethics generates an obligation on part of the designer, the manufacturer, or the deployer of the technology to "provide evidence of trustworthiness to people in a position to trust that technology" (Ibid.). Consequently, as a major concern about the role of justified trust ethics, Nickel points out the one about most people's ability (such as that of non-experts) to justify adequate reasons for their trust (Ibid.).

In turn, *trustworthiness ethics* is determined against the background of the presupposition that designers, manufacturers and deployers of technology have only two trust-related tasks: 1) "to make the technology as reliable as possible at doing the things it is supposed to do" (then safety is considered as a criterion for being defined as trustworthy) and 2) "to persuade possible users that it is sufficiently reliable and safe" (then trustworthiness is considered as being achievable by the adoption of psychologically effective means) (Ibid.).

Nickel also clarifies that an argument for trustworthiness ethics says that "relying on trustworthy technology makes people better off, whereas having a further justification for that reliance does not add any additional benefit" (Ibid.). This means that "a justification for one's true belief does not make one any better off than simply having a true belief" (Ibid.).

2.1. What are the implications of AI's trustworthiness?

The debates about AI's trustworthiness inherit the outcomes of those about trustworthiness of technical artifacts (Nickel et al. 2010). In both cases, rational accounts of trust provide an opportunity for grounding the idea of trustworthiness by narrowing it to that of reliability, as well as distinguishing between trustworthy technology the and "full-blown notion of trustworthiness associated with interpersonal trust" (Ibid.). Both AI and other technical artifacts cannot reach the level of interpersonal trust due to the fact that they cannot show emotions and held responsibility, as do humans. Why can one call them reliable?

Providing a definition of reliability of technical artifacts which corresponds to that of reliability of humans is a narrow definition in itself. As such a definition, Nickel et al. point out Birolini's definition. According to the latter, "*Reliability* is a *characteristic* of a person, expressed by the *probability* that the person will perform his/her *required function* under *given conditions* for a *stated time interval…*". From a qualitative point of view, reliability is "the *ability of the person to remain functional*", while from a quantitative one, it concerns the *probability* of the lack of operational interruptions during a stated time interval (Birolini in Nickel et al. 2010, 433 – 434).

As Nickel et al. cogently argue, such definitions are acceptable mainly from an engineering point of view because they make perfect sense when the person is the operator of the machine (Ibid., 444). Specifically, the foregoing illustrates that there is "no fundamental difference between the reliability of machines and people" (Ibid.). This conclusion, however, displays a very limited picture of reliability in interpersonal terms.

Regarding the rational-choice account of trust, as extrapolated to the idea of reliability, Nickel et al. provide the following specification: "one trusts a technical artifact to a degree *x* when and only when one is willing to risk the use of that technical artifact, on the basis of a judgement that it will perform (function) with probability *x*" (Ibid., 435). The conclusion is that "the extension of the rational-choice account of trust" in technical artifacts "does not lead to a genuine notion of trust in technical artifacts, different from reliability" (Ibid.). That is why, trust in technical artifacts can be interpreted only as a matter of a derived form of interpersonal trust (Ibid., 436).

In turn, motivation-attributing account of trust in artifacts cannot be straightforwardly applied to technical artifacts, since the latter do not have mental states, whose content is interests and values of the trusting person, nor do they have interests of their own (Ibid., 443).

In this context, what is of crucial importance for the questioning of AI's trustworthiness is the normative account of trust. The main moral concern about defining AI as trustworthy is that AI does not have the ability to care about or be moved by the trust placed in it (Ryan 2020). Therefore, AI's actions are not actions guided in trust, but "acts based on reliance and predictability" (Ibid., 2761).

By extrapolating what Nickel defines as general negative attitudes towards technological failure, one may argue that the moral concern about AI's trustworthiness is that AI's failure can provoke anger and frustration, but not a feeling of blame, as is when interpersonal trust is broken. Furthermore, if we cannot blame AI, we cannot trust it either, but rather rely upon its functional ability to perform properly. The latter performance can be referred to what Ryan calls quasi-trust or misplaced trust whose potential is to deceive the individuals about the AI's capacities, as well as obfuscating responsibility by AI companies (Ibid., 2752).

Consequently, the three dominant accounts on trust – the rational, affective and normative accounts – are summarized by Ryan in the following way. The rational account of AI, which is the only one that meets the requirements of being trustworthy, “is in fact a form of reliance because of its lack of concern about the trustee’s motivation for action” (Ibid.).

The affective account of AI lacks the motivation of the AI to do something as being based on a goodwill towards the trustee (Ibid.). Therefore, AI may be able to act like humans and have intelligence to carry out actions, while “still not possessing the capability of being moved by those actions” (Ibid., 2760).

Going back to the normative account of AI, it lacks its commitment to the relationship with the particular trustee (Ibid., 2753). Comparing and contrasting the aforementioned three accounts, Ryan draws the conclusion that AI “is something we can have a reliance on, but not something that has the capacity to be trusted” (Ibid., p. 2754).

3. What are the practical implications of AI’s trustworthiness (reliability)?

The so-called positive projects of trustworthy AI

The AI4People Scientific Committee aims at proposing a series of recommendations for the development of an AI society in Europe. In this context, the necessity of justifying a trustworthy AI gains new strength, as is demonstrated by Thiebes et al.’s project on trustworthy artificial intelligence (TAI). The building of TAI is underlined by the five principles outlined by Floridi et al., namely, those of beneficence, non-maleficence, autonomy, justice and explicability (Floridi et al. 2018, 696). By relying upon such principles, Thiebes et al. aim at developing “a data-driven research framework for TAI” for the purposes of demonstrating the benefits of its distributed ledger technology-based realization (Thiebes 2020, 447).

The concept of TAI, as is developed by the HLEG project, promotes the idea that the individuals, organizations and societies will be able to achieve the full potential of AI of trust established in its development, deployment and use (Ibid.) (HLEG 2019). What is specific for the HLEG’s project is the way in which it relates TAI to some general ethical principles. AI is perceived as trustworthy by its users (e.g., consumers, organizations, society), when the aforementioned development, deployment and use are grounded in adherence to such general ethical principles, in addition to its compliance with all relevant laws (Thiebes 2020, 451).

For the purposes of showing the limitations regarding the use of moral AI as an adviser, I focus upon the HLEG’s fifth requirement, namely, that of *Diversity, non-discrimination and fairness*. According to this requirement, unfair bias “must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination…” (HLEG 2019).

What is moral AI as an adviser?

In the field of machine ethics, researchers discern between four sub-fields of AI expertise. First, AI could be used as a moral environment monitor (which provides morally-relevant information about one's surroundings), second – AI can work as a moral prompter whose function is to inform the user about morally important situations or events. In turn, the third and fourth types of expertise concern the role of AI as a moral adviser (nudging the user what should do) and a moral observer (observing the behavior of the user) (Savulescu and Maslen 2015, 84) (Klincewicz 2016, 174).

What can be considered as being of particular interest regarding trustworthiness ethics is the function of moral AI as an adviser. In the language of trustworthiness ethics, it means that the moral AI as an adviser should be 1) recognized as morally reliable and 2) accepted as morally reliable by the users so that the AI can be adopted for changing users' moral behavior. However, setting such a task is sufficiently complicated still on the level of clarifying the aspects of interpersonal trust.

Savulescu and Maslen claim that before offering advice in the first instance, the AI should ask the agent to "indicate which of a long list of morally significant values or principles he holds and is guided by" (Ibid., 88) (Ibid.). The agent is also asked to assign a weight (between 0 and 1) to each value including values such as autonomy, benevolence, non-malevolence, justice/fairness, legality, environmental protection, family/significant relationships, fulfilling duties/commitments/promises, and maximizing net utility. According to this model, an agent who strongly cares about not harming others, only a little about legality and not at all about protecting environment "might assign them weights of 1, 0.5 and 0, respectively" (Klincewicz 2016, 174).

Regarding the different scenarios of axiological priorities, "the AI would compute the extent to which the courses of action open to the agent would uphold or compromise these values (fully uphold value=1; fully compromise value=-1), amplifying or diminishing based on the weight indicated by the agent" (Ibid.). Having had the rank of weighed values, the AI would use it to suggest the best course of action (Ibid.) (Serafimova 2020).

Extrapolating the debate to the issue of AI's trustworthiness, one may argue that in so far as Savulescu and Maslen's model addresses the purely computational part of the moral AI as an adviser, it meets the first two requirements of Ryan's rational account of interpersonal trust, namely, those of confidence and competence (Ryan 2020, 2754). If the rational account of trust is classified as "'reliance on another person's qualities' or features of the situation, disregarding the trustor's social right embedded in their relationship of mutual respect to have the trustee's responsiveness…" (Ibid., 2759), one can find some arguments against the assumption that the human agent, who assigns weights to values in the chart, may trust the AI system. This agent may rely upon the AI's confidence (assigning weights to values, the human agent is confident that the AI will do the actions that are entrusted to it) and competence (the AI has the capacity to do the actions it is entrusted with) (Ibid., 2755) in producing an outcome data as a result from the transformation of the input data.

In this context, one may draw the conclusion that the rational account of trust can be applied to this first stage of introducing moral AI as an adviser due to its proven computational reliability. However, one should also keep in mind that the moral AI does not fulfill the requirements of all the trust components of the rational account of trust missing that of vulnerability (Ibid., 2754). Taking into account that the rational account can provide a justification of vulnerability similar to that of interpersonal trust, the human agent becomes vulnerable, when placing their trust in the AI. However, in the process of assigning weights to values, human vulnerability is neglected by default. This is due to Savulescu and Maslen's assumption that the advice, which can be given on the basis of a precise computational process is the best moral advice by default.

3.2.1. Klincewicz's scenario of John's racism

Extrapolating Savulescu and Maslen's model to one scenario, namely, to that of a hypothetical person, a white American named John "who could report that he just saw a police officer verbally abusing a Chinese person" (Klincewicz 2016, 177), M. Klincewicz examines how AI can function as a moral environment monitor, moral prompter, moral adviser and moral observer. He raises the following question. If John is "then hooked up with a moral AI with all of its different subsystems, as presented by Savulescu and Maslen, and faced with an identical situation", can we expect that he will change his behavior and react differently? (Ibid.).

According to Klincewicz, the moral environment monitor in the AI may tell John that the person is Chinese and this is morally relevant. Being a racist, John will use this information to make a decision that he should not report the police officer's abuse because "This is how racism works" (Ibid.). Regarding the function of AI as a moral prompter, it could tell John that the situation with the police officer involves a moral dilemma which assumes reacting "to an observed transgression of laws by someone else" (Ibid.). Thus, John may find it helpful to know that in the given case, the laws are broken by a police officer, but he will "also be compelled by his racist views to not do anything about it" (Ibid.). As Klincewicz relevantly argues, even if the AI functions as a moral adviser, it is highly unlikely that John would act differently (Ibid., 178).

This conclusion raises the issue of examining the implications of the fifth requirement of the HLEG's Ethics Guidelines, namely, that of *Diversity, non-discrimination and fairness* (HLEG 2019). Paraphrasing Klincewicz's concerns, the issue is whether or not fostering diversity with the help of AI, the latter can encourage the process of non-discrimination. As is stated, the issue requires at least two more questions to be raised. First, what can one understand by diversity? Diversity in facts and diversity in opinions do not guarantee a fair treatment of all the agents involved. Second, how can one define what unfair bias looks like, as displayed by the fifth principle, so that one can avoid the diversity of its negative implications?

Judging by the aforementioned questions, one may draw the conclusion that the HLEG's fifth requirement is too general in two senses: 1) it is too general in a moral sense, when evaluated on the level of interpersonal trust and 2) it is too general in the sense that it does not clearly show how AI in particular can contribute to eradicating discrimination which has not been eradicated by humans yet.

In turn, elaborating upon Klincewicz's hypothetical scenario from the perspective of trustworthiness ethics has the following benefit. Even if one may argue for AI's computational reliability when providing non-discrimination data, it does not follow that the moral AI as an adviser can contribute to changing one's discriminative attitudes. One of the reasons is that moral agents' plurality of choices regarding values is driven by some complex factors which are neither fixated nor purely moral.

Following Savulescu and Maslen's rules of the assigning, John should rate only one value as being the highest by ascribing one point to it. I, however, would argue that the main moral concern is that he cannot point out only one value as being the highest due to some of the arguments mentioned above. This specification makes the calculation of the best possible scenario for John highly problematic, since the AI model will display one initially limited picture of John's values (Serafimova 2020, 114).

For instance, in his general value system, John may rate highly both justice and legality and even double check his act of non-interference by assuming (from the perspective of his racist ideology) that, in the case with the Chinese person, justice coincides with legality. Specifically, John can share the assumption that Chinese people are not 'valuable' as people, ergo it is 'fair' to be offended if the offense is committed by the authority figures. Thus, he can encourage the computation of a problematic scenario which will be evaluated as being morally appropriate according to his distorted vision of morality. Certainly, fulfilling such a scenario would have a zero impact upon changing John's immoral attitudes in the process of his interaction with Chinese people. That lack of influence could also be a result of John's conviction that verbal abuse is legal first, because it is exerted by a representative of the authorities and second, because he does not consider verbal abuse as a matter of 'real' (physical) harassment (Ibid.).

In this context, the strongest moral concerns about John's racism arise against the assumption that if the algorithms' correction was a necessary and sufficient condition for correcting one's moral behavior, it would have meant that John's racist attitudes would have been easily changed by asking him to change the weights to the values he considers as being the highest. Thus, the inconsistency between moral and computational weights becomes apparent in one paradoxical manner, namely, in order to change his racist behavior, John should change his priority to justice and look for another value to replace it with.

All the aforementioned concerns derive from the gist of the multi-agent system in which they are set. In so far as the function of the moral AI as an adviser is to nudge the user what to do, both the trustor (the moral AI) and the trustee (HA) need to be recognized as trustworthy. Only then one can confirm that the HA has changed their moral behavior by relying upon the advices given by the AI.

In John's hypothetical scenario, the lack of change can be explained with the lack of internalization of "the moral component" of trust, namely, with what Buechner et al. define as normative expectations regarding trust (Buechner et al. 2014, 69). Referring to their definition that a trustworthy trustor is not only able to differentiate trustworthy from untrustworthy trustees (the epistemic component), but also to identify the normative nature of trust relationships (the moral component) (Ibid.), one may argue that in the best scenario, the moral AI as an adviser can recognize only the first component regarding John's motivation. Being unable to understand the second one, the moral AI cannot understand the complexity of John's normative expectations, since it cannot recognize the impact of John's complex biases which guide his immoral behavior.

In this context, the questionable impact of AI's reliability concerns the fact that it is not a trustor in a human sense. Consequently, the AI cannot meet the general requirements of being a trustee either. In contrast, although John can be defined as untrustworthy due to his immoral behavior at this stage of his life, he has the potential to become a trustor sometime in the future. The reason is that as a human being, John initially has the potential to recognize the role of his normative expectations and change them.

The HA (John)-AA (moral AI as an adviser) multi-agent system is quite complicated because it requires the incorporation of different levels of trust which are mediated by that of reliability. The kinds of stakes involved in John's hypothetical scenario may have the following embodiments. First, the moral AI may be positively biased (by its designers) as an adviser, but due to John's strong negative (racist) biases, it cannot contribute to changing his behavior. Second, the moral AI may be positively biased again and its advices find a fruitful soil in John's own willingness to change his behavior. However, such an assumption does not make clear why if John needs some stable grounds to change his behavior, he looks for a piece of advice provided by an AI instead of another human being.

Third, the moral AI can be negatively biased by its designers as well. Then, the AI is unable to change John's negative biases because its inherited (by its designers) biased moral preferences do not provoke the expected feedback.

Consequently, since the moral AI can be defined as being reliable rather than trustworthy, one cannot say that the failure of John's expectations can be internalized as something more than a disappointment of a technological failure if John is aware of it.

4. Conclusion

In this article, I explore how weighing up pros and cons AI's trustworthiness leads to the conclusion that one can argue for AI's reliability. As one of the main reasons for drawing such a conclusion, I point out the unsuccessful attempts at anthropomorphizing AI technologies, when AI's trustworthiness is recognized by analogy with the interhuman trust.

Specifically, the major methodological concern is that AI's trustworthiness cannot meet the requirements of the affective and normative accounts of trust in Ryan's sense. Revealing the arguments for the justification of TAI, I argue that the experts rely upon one rather ideal scenario which is fulfillable only if the rational account of trust is taken under consideration.

In turn, elaborating upon Nickel's distinction between *justified trust ethics* and *trustworthiness ethics* in respect to AI, I demonstrate that favoring *trustworthiness ethics* contributes to examining the challenges in building AI's reliability. These challenges are considered as being closely tied with the recognition of a HA-AA multi-agent system which provides different types of trustors and trustees.

For the purposes of exemplifying the gist of some of the aforementioned challenges, I examine Savulescu and Maslen's model of moral AI as an adviser. Tackling the role of the rational, affective and normative accounts of trust, I draw the conclusion that the moral AI in question meets only two of the five criteria of Ryan's rational component of interpersonal trust, namely, those of confidence and competence, while explicitly failing to meet that of vulnerability.

The practical consequences of introducing moral AI as an adviser, as well as revealing the associated challenges in grounding its reliability are exemplified by Klincewicz's hypothetical scenario of John's racism. Specifically, AI's reliability is refracted through the lens of the fifth requirement of the HLEG's Ethics guidelines — that of *Diversity, non-discrimination and fairness*. I reach the conclusion that the application of the fifth requirement, as interpreted from the perspective of trustworthiness ethics, has the following benefit. It disenchants the fact that even if one may argue for AI's computational reliability in terms of providing non-discrimination data (on the input level), it does not follow that the moral AI as an adviser, which uses that data, can change one's discriminative attitudes.

In the best scenario, moral AI as an adviser can understand only the epistemic component in Buechner et al.'s sense of John's motivation. Being unable to anticipate the normative component, this AI cannot grasp the complexity of John's normative expectations, since it cannot recognize the impact of John's biases which guide his immoral behavior.

Regarding the future research on AI’s reliability, I would argue that the ‘traditional’ frame of agent-computer trust should be enriched with that of HA–AA trust relationships. In this context, the main concern about the functioning of moral AI as an adviser is one to clarify the moral consequences of the potential harms and associated duties in building *HA’s trust in AI’s reliability*.

REFERENCES

Buechner, J., Simon, J. & Tavani, H. T., 2014. Re-Thinking Trust and Trustworthiness in Digital Environments. In: E. Buchanan et al. (Eds.). *Autonomous Technologies: Philosophical Issues, Practical Solutions, Human Nature: Proceedings of the Tenth International Conference on Computer Ethics - Philosophical Enquiry: CEPE 2013*. Menomonie, WI: INSEIT, 65 – 79.

Floridi, L., Cowls, J., Bertrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Shafer, B., Valcke, P. & Vayena, E., 2018. AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Mind and Machines* [online]. **28**, 689 – 707. [viewed 20 July 2021]. Available from: https//doi.org/10.1007/s11023-018-9482-5.

HLEG, 2019. *Ethics Guidelines for Trustworthy AI*. Available from: https://www.ccdcoe.org/uploads/2019/06/EC-190408-AI-HLEG-Guidelines.pdf.

Klincewicz, M., 2016. Artificial Intelligence as a Means to Moral Enhancement. *Studies in Logic, Grammar and Rhetoric*. **48**(61), 171 – 187.

Nickel, P., J., Franssen, M. & Kroes, P., 2010. Can We Make Sense of the Notion of Trustworthy Technology? *Know Tech Pol*, **23**, 429 – 444.

Nickel, P., J., 2013. Trust in Technological Systems. In: M. J. de Vries, S. O. Hansson & A. W. M. Meijers (Eds.). *Norms in Technology: Philosophy of Engineering and Technology*, vol. 9. Dordrecht: Springer, 223 – 237.

Ryan, M., 2020. In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics* [online]. **26**, 2749 – 2767 [viewed 20 July 2021]. Available from: https// doi.org/10.1007/s11948-020-00228-y.

Savulescu, J. & Maslen, H., 2015. Moral Enhancement and Artificial Intelligence. Moral AI? In: J. Romportl, E. Zackova & J. Kelemen (Eds.). *Beyond Artificial Intelligence. The Disappearing Human—Machine Divide*. Springer, 79 – 95.

Serafimova, S., 2020. How to Assign Weights to Values? Some Challenges in Justifying Artificial Explicit Ethical Agents. *Balkan Journal of*

Philosophy [online]. **12**(2), 111-118 [viewed 20 July 2021]. Available from: https://doi.org/10.5840/bjp202012213.

Thiebes, S., Lins S. & Sunyaev, A., 2020. Trustworthy Artificial Intelligence. *Electronic Market* [online]. **31**, 447 – 464. [viewed 20 July 2021]. Available from: https://doi.org./s12525-020-00411-4.

✉ **Dr. Silviya Serafimova, Assoc. Prof.**
Web of Science Researcher ID: X-8174-2019
Department of Ethical Studies
Institute of Philosophy and Sociology
Bulgarian Academy of Sciences
E-mail: silvija_serafimova@yahoo.com