Using Algorithms to Make Ethical Judgements: METHAD vs. the ADC Model

Allen Coin and Veljko Dubljević*

*Corresponding Author: veljko_dubljevic@ncsu.edu

Open Peer Commentary

American Journal of Bioethics

28 April, 2022

In their paper "Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept," Meier and colleagues (2022) present the design and preliminary results of a proof-of-concept clinical ethics algorithm that they claim can use machine learning to make limited recommendations about moral dilemmas that may occur in healthcare, using the ethical framework of principlism as espoused by Beauchamp and Childress. They report some success for the algorithm, with tests producing results agreeing with ethicists in 92% of the training data set and 75% in the test set. A stated limitation of the algorithm is that it is designed based in part on the principlist approach to bioethics and does not account for other ethical frameworks. Additionally, the algorithm described by the authors requires the user to input a number of numerical variables that are subjective and would lead to variable results depending on the human utilizing the algorithm in a clinical setting. Moreover, the lack of a neutral median point in the output leads to bias towards intervention in the interpretation of the results by the authors.

The purpose of this commentary is to relay and contrast some relevant lessons we have learned in our own National Science Foundation-funded work (#2043612), where we combine the virtue theoretic, deontological, and consequentialist approaches for ethical decision-making algorithms within the Agent-Deed-Consequence (ADC) model, along with results of empirical research with human decision makers. The primary difference between METHAD and ADC approaches is that the first explores AI-assisted moral decision making, whereas the latter explores whether AI-empowered moral decision making is possible. The ADC model predicts that moral judgments are positive if all three previously discussed components are positive, and negative if all are evaluated as negative (Dubljević and Racine 2014). Compartmentalizing different aspects of a given situation allows for ease of programming and computation into artificially intelligent systems, as the system would be able to substitute the overall moral

judgment with more accessible information in distinct computations. The capacity for AI to make "decisions" about the health and wellbeing of human beings is especially pertinent when an AI is incorporated into a robot that can autonomously execute the healthcare decision, as has been demonstrated in robots used in in healthcare (Coin and Dubljević 2021). Consider as an example an event like the Surfside building collapse in Florida (see Pflanzer et al. 2022): a building partially collapses, leaving people with crushed limbs and stranded under rubble and debris. The emergency responders must act quickly to get people out of the rubble to ensure that they survive. First responders deploy a small robot that utilizes AI to make decisions based on the in-the-field data. The situation could escalate at any moment, bringing the building down and crushing any potential survivors. The AI finds two people while going through the wreck. One person is conscious but in severe pain due to their leg being stuck under debris. Meanwhile, the other person is not stuck but is unconscious and has a bleeding head wound. Additionally, there are two more victims deeper into the wreckage, but we assume the AI has no way of knowing whether there are more survivors ahead or how deep in the wreckage they may be.

The ADC approach provides AI research with baked-in computations encompassing decisions about Agents (e.g., is a particular person more worthy of saving?), Deeds (e.g., is amputation justifiable?), and Consequences (e.g., could more people have been saved?). To test this approach (*mutatis mutandis* for D and C), we use [A-2] to represent a strong negative designation of a particular agent; [A-1] to represent a weaker negative designation; [A0] when agent information is not available; [A+1] to represent a weaker (low-stakes) positive evaluation; and [A+2] to represent a strong positive evaluation (see Dubljević 2020).

As such decisions are complex and sensitive, they should not be left to a single user to define parameters or bias towards intervention. Multiple possible scenarios need to be rigorously tested in various stakeholder populations by using vignettes such as these:

Scenario 1: The AI system finds the first two people and decides to amputate the legs of the conscious person without consent, as there is no time to waste. It then quickly applies first aid and takes both people out of the rubble. Although there may be more victims yet trapped inside, the AI decides to egress both victims immediately to maximize their odds of survival. The two people recover in due time, though later it is revealed through search that the robot could have saved two more lives if it had continued to search before leaving.

Scenario 2: With the consent of the conscious person, the AI amputates their leg and provides first aid. Then the AI informs the conscious person that there may be more victims further ahead and argues it should attempt to find them. The person orders the AI to take him and the unconscious person out first, as there is no way of knowing if there are more survivors and the building could still collapse on them. The AI obeys and the two people recover, but it is later discovered that there were two other victims who could have been found by the AI if it had proceeded to search further ahead.

Scenario 3: With the consent of the conscious person, AI amputates the leg and provides first aid. Though the person orders the AI to take them and the unconscious person out first, the AI decides it would be better to continue searching for other survivors, even if this action puts the lives of the two people at greater risk. The AI successfully finds two more people and brings them to safety. The AI then returns and successfully rescues the two people it first encountered. In due time, all four victims recover.

These scenarios represent relevant real-world decision-making that AI will have to perform if deployed in situations with outcomes relevant to the health and wellbeing of the humans it is tasked with caring for. As demonstrated by Meier and colleagues, METHAD would not be well suited for making such immediate decisions. However, the ADC approach goes beyond AI-assisted moral decision-making to explore whether and under which conditions AI can be allowed to make decisions affecting humans. The fact that AI systems already adjudicate the wellbeing of humans (Ouchchy et al. 2020) makes this task all the more urgent. While we agree with Meier and colleagues that research into algorithmic ethical decision-making should progress with great care, there are noteworthy differences in our respective approaches. As mentioned above, Meier and colleagues don't utilize a neutral median point, which leads to a bias toward intervention. Our approach is cognizant of uncertainty which needs to be numerically represented (e.g., A0, D0 or C0). Additionally, we assume that AI decisions would first be instantiated only in morally unambiguous (i.e., [A-D-C-] and [A+D+C+]) situations. Additionally, real-world scenarios don't come with neat text-based descriptions. That is why, apart from testing the initial textual input in large-scale surveys (Dubljević et al. 2018) and multiple stakeholder groups in more than one language (see Sattler et al. 2022), we are also creating and testing immersive virtual reality experiences. Establishing cross-cultural human agreement on the evaluation of the specific sub-components of moral decision-making (A,D,C) in audiovisual representation of morally salient situations is necessary before algorithmic solutions can be implemented in fuzzy cognitive maps or neurosymbolic AI (Shah et al. 2019).

Another point of difference is that METHAD seems to limit use cases to desktop applications used by (say) nurse-practitioners, which would save time by automating putative feedback from clinical ethics consultants. Apart from downstream consequences of automation

for clinical ethicists, there are multiple use cases that will inevitably be attempted: so, METHAD may end up being used not only in virtual AI assistive technologies (Bauer & Dubljević 2020), but also in carebot companions and complex humanoid robo-surgeons (Coin & Dubljević 2021). The fact that METHAD over-expresses consequentialist moral deliberations (by essentially duplicating positive and negative evaluations of outcomes) and under-expresses aretaic and deontological aspects of moral decision-making (e.g., by removing fairness considerations inherent in principlism) may be another source of pro-intervention bias.

To conclude, we applaud the valiant efforts of Meier and colleagues to (partially) implement principlism in AI-assisted moral decision-making. At the same time, we urge the authors and the ethics community at large to remain cognizant of the fact that AI-based solutions are and will be implemented in real-world settings and to ensure the algorithmic approaches they are using are vigorously tested in multiple stakeholder groups, languages and modes of input (textual, audiovisual, etc.), and that they are not biased toward any particular moral system (e.g., Utilitarianism) or outcome (e.g., for medical intervention).

References

Bauer, W. A., & Dubljević, V. (2020). AI assistants and the paradox of internal automaticity. *Neuroethics*, *13*(3), 303–310.

Coin, A., & Dubljević, V. (2021). Carebots for eldercare: Technology, ethics, and implications. In *Trust in Human-Robot Interaction* (pp. 553–569). Elsevier.

Dubljević, V. (2020). Toward implementing the ADC model of moral judgment in autonomous vehicles. *Science and Engineering Ethics*, *26*(5), 2461–2472.

Dubljević, V., & Racine, E. (2014). The ADC of moral judgment: Opening the black box of moral intuitions with heuristics about agents, deeds, and consequences. *AJOB Neuroscience*, *5*(4), 3–20.

Dubljević, V., Sattler, S., & Racine, E. (2018). Deciphering moral intuition: How agents, deeds, and consequences influence moral judgment. *PloS One*, *13*(10), e0204631.

Meier, L.J. et al. (2022). Algorithms for Ethical Decision-Making in the Clinic: A Proof of Concept, American Journal of Bioethics

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & Society*, *35*(4), 927–936.

Pflanzer, M., Traylor, Z., Lyons, J.B., Nam, C.S. & Dubljević, V. (2022). Ethics in Human-AI Teaming: Principles and Perspectives. Under review

Sattler, S., Dubljević, V. & Racine, E. (2022). Cooperative Behavior in the Workplace: Empirical Evidence from the Agent-Deed-Consequences Model of Moral Judgment. Under Review.

Shah, N., A. Sheth, S. Bhatt, R. Goswami, V. Shah, R. Kanani, A. Patel, and P. Pathak. (2019). Data processing system and method for computer-assisted coding of natural language medical text, Dec. 17 2019. US Patent 10,509,889.