This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Journal of Bioethics* (2022, volume 22, issue no.7, pp.1-3, https://doi.org/10.1080/15265161.2022.2075968 . For the final version and to cite, please refer to the official published version of the article.

# How to use AI ethically for ethical decision-making

Joanna Demaree-Cotton
Brian D. Earp
Julian Savulescu

Oxford Uehiro Centre for Practical Ethics

What counts as a good decision depends on the domain. In diagnostic imaging, for instance, a good decision involves diagnosing cancer if and only if the patient has cancer. In clinical ethics, good decision-making is defined in terms of the extent to which the following two goals are met:

1. **Accuracy**: The decision is the right one, where the "right" decision is that which best aligns with relevant justifying values, principles and their respective weights as they apply to the case at hand.
2. **Transparency**: The patients are provided with an explanation of the decision in terms of relevant values, principles and how they are weighed. In other words, the patients are offered reasons that explain and justify the decision.

For the use of artificial intelligence in clinical ethics to be ethically justified, it should improve the transparency and accuracy of ethical decision-making beyond that which physicians and ethics committees are currently capable of providing.

The kind of AI proposed by Meier and colleagues (2022) has the fascinating potential to improve the transparency of ethical decision-making, at least if it is used as a decision aid rather than a decision replacement (Savulescu & Maslen 2015). While artificial intelligence cannot itself engage in the human communicative process of justifying its decisions to patients, the AI they describe (unlike "black-box" AI) makes explicit which values and principles are involved and how much weight they are given.

By contrast, the moral principles or values underlying human moral intuition are not always consciously, introspectively accessible (Cushman, Young, and Hauser 2006). While humans sometimes have a fuzzy, intuitive sense of some of the factors that are relevant to their moral judgment, we often have strong moral intuitions without being sure of their source, or without being clear on precisely how strongly different factors played a role in generating the intuitions. But if clinicians make use of the AI as a decision aid, this could help them to transparently and precisely communicate the actual reasons behind their decision.

This is so even if the AI's recommendation is ultimately rejected. Suppose, for example, that the AI recommends a course of action, with a certain amount of confidence, and it specifies the exact values or weights it has assigned to autonomy versus beneficence in coming to this

This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Journal of Bioethics* (2022, volume 22, issue no.7, pp.1-3, https://doi.org/10.1080/15265161.2022.2075968 . For the final version and to cite, please refer to the official published version of the article.

conclusion. Evaluating the recommendation made by the AI could help a committee make more explicit the "black box" aspects of their own reasoning. For example, the committee might decide that beneficence should actually be weighted more heavily in this case than the AI suggests. Being able to understand the reason that their decision diverges from that of the AI gives them the opportunity to offer a further justifying reason as to why they think beneficence should be given more weight; and this, in turn, could improve the transparency of their recommendation.

However, the potential for the kind of AI described in the target article to improve the accuracy of moral decision-making may be more limited. This is so for two reasons. Firstly, whether AI can be expected to outperform human decision-making depends in part on the metrics used to train it. In non-ethical domains, superior accuracy can be achieved because the "verdicts" given to the AI in the training phase are not solely the human judgments that the AI is intended to replace or inform. Consider how AI can learn to detect lung cancer from scans at a superior rate to human radiologists after being trained on large datasets and being "told" which scans show cancer and which ones are cancer-free. Importantly, this training includes cases where radiologists did not recognize cancer in the early scans themselves, but where further information verified the correct diagnosis later on (Ardila et al. 2019). Consequently, these AIs are able to detect patterns even in early scans that are not apparent or easily detectable by human radiologists, leading to superior accuracy compared to human performance.

However, Meier and colleagues propose training AI on ethical cases using solely the very verdicts of ethics committees about those cases. At best, it seems that this sort of AI could learn to emulate the decisions of ethics committees, but not exceed them. To be sure, this could still improve ethical accuracy when a proper committee consultation cannot take place due to time or resource constraints. But the AI could not improve on the accuracy of ethics committees themselves. Consequently, it is not clear we would be justified in relying on an AI to make ethical decisions, since we could not reasonably expect it to be more likely to get it right.

One response would be that AI described by Meier and colleagues could be adapted to sum or integrate a vast set of training data from multiple committees, thus abstracting away from the fallibilities of any one committee in order to bring to bear the best expert ethical judgment. This would likely be better than any individual clinician or ethics committee, particularly when urgent decisions are required.

Still, and secondly, for AI to help committees make the right decision we need AI that can shed light on difficult and controversial ethical cases that cannot be easily resolved by human moral judges. But the training data that Meier and colleagues cite involve rather obvious ethical decisions. Whether AI can be expected to perform more accurately than humans on new or difficult cases depends on whether patterns apparent in training can be successfully extrapolated to new, more difficult cases. This is a plausible assumption in the case of lung cancer: patterns indicative of lung cancer are likely to extend similarly to new patients. But it is less clear that an equivalent assumption holds for ethics, especially for an AI that is trained

This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Journal of Bioethics* (2022, volume 22, issue no.7, pp.1-3, https://doi.org/10.1080/15265161.2022.2075968 . For the final version and to cite, please refer to the official published version of the article.

using "textbook" ethical principles and cases about which clinical ethics committees are able to come to a settled verdict. For example, it's relatively clear that long-term prospects ethically outweigh short-term pain and discomfort when deciding whether to subject a temporarily unconscious adult to a difficult emergency surgery that is necessary to save their life and has a high chance of succeeding. But is it worth exposing a child to 3months of painful intensive care for a tiny chance of improving their quality of life to an acceptable level, such as in the case of Charlie Gard (Wilkinson and Savulescu 2018)? This sort of case is difficult in part because it is unclear whether the same tradeoffs that are acceptable in easier cases are correctly applied here. It's thus unclear whether an AI trained on verdicts of ethics committees can be expected to be more accurate than humans when considering difficult and controversial cases.

One response that Meier and colleagues could plausibly give is that it is not clear what the right answer in these difficult cases is. There is a "grey zone" in ethics or, in pediatric decision making, a "zone of parental discretion" (Gillam 2016). What their tool could do is make explicit that if x, y and z factors are given weight, then the decision would be to terminate treatment; but if a, b and c factors are given weight, then treatment should be continued. In this way, AI could promote rational dissensus (Wilkinson and Savulescu 2018).

What if we could implement training that would allow AI to go beyond merely emulating the decisions of ethics committees? Meier and colleagues reject the use of public surveys as a training metric on the grounds that "[s]urveys of moral preferences [ ... ] cannot possibly reflect the complexity that the making of patient-centred clinical decisions requires. The sheer amount of detail that comes with every single case precludes a broad survey-based approach" (12). While we agree that a broad survey-based approach using full-length clinical ethics cases would not be practicable, this overlooks an important additional possibility—one that takes inspiration from cognitive science. Specifically, there is the potential to train AI on very large samples of ordinary moral judgments about more simple, quick-to-digest cases, thus allowing the AI to develop a "moral competence" that encodes the complex patterns of sensitivity that human moral judgment exhibits to various morally relevant factors—such as intentions, harms, benefits, autonomy (Demaree-Cotton & Sommers 2022), distributive fairness, and social-relational context (Earp et al. 2021). This could be seen as a computational extension of methods in moral philosophy (abstracting theoretical principles from intuitions about particular cases) and moral psychology (which examines the principles governing moral intuitions by probing how moral judgment is affected by various factors; see Awad et al. 2022, on "computational ethics").

Alternatively, a more full-blown process of Collective Reflective Equilibrium could be used (Savulescu, Gyngell, and Kahane 2021) This could involve machine learning of human values from psychological or behavioral research, and using these as "settled intuitions" to be brought into coherence with ethical concepts, principles and theories.[1] Indeed, such intuitions could be

[1] However, see Earp, Lewis et al. (2021) on the potential need to first experimentally probe certain intuitions to ensure that they are robust (e.g., not highly susceptible to normatively irrelevant framing effects; see Demaree-Cotton, 2016, for discussion), to clarify what features of cases they are

This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Journal of Bioethics* (2022, volume 22, issue no.7, pp.1-3, https://doi.org/10.1080/15265161.2022.2075968 . For the final version and to cite, please refer to the official published version of the article.

used to modify some of these concepts, principles and theories (Savulescu, Gyngell, and Kahane 2021; Savulescu, Kahane, and Gyngell 2019).

Even though this AI would be trained initially on simpler cases, the variety and complexity of the range of cases means that it has the potential to process lengthy, complex, real-life clinical dilemmas in ways that reflect a highly complex integration of relevant moral values and principles. Paradoxically, the same is not necessarily true of human moral judges presented with equivalently lengthy and complex cases. While human moral agents can respond quickly and efficiently to simplified examples, limited attention and working memory make it difficult to thoroughly process and integrate all of the factors that are relevant to such a case; our moral intuitions, instead, pull us in different directions depending on which aspect of the case we are currently focused on. While we do not believe that such an AI should replace the decisions of human clinical ethics committees, it has the potential to provide outputs that would be meaningfully informative, going beyond what the clinical ethics committee is already able to compute for itself.

## References

Ardila, D., A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, et al. 2019. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nature Medicine* 25 (6):954–61. doi:10.1038/ s41591-019-0447-x.

Awad, E., S. Levine, M. Anderson, S. L. Anderson, V. Conitzer, M. J. Crockett, J. A. C. Everett, T. Evgeniou, A. Gopnik, J. C. Jamison, et al. 2022. Computational ethics. *Trends in Cognitive Sciences* 26 (5):388–405. doi:10.1016/ j.tics.2022.02.009.

Cushman, F., L. Young, and M. Hauser. 2006. The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science* 17 (12):1082–9. doi:10.1111/j.1467-9280.2006.01834.x.

Demaree-Cotton, J. 2016. Do framing effects make moral intuitions unreliable? *Philosophical Psychology* 29 (1):1–22. doi:10.1080/09515089.2014.989967.

Demaree-Cotton, J., and R. Sommers. 2022. Autonomy and the folk concept of consent. *Cognition* 224:105065. doi:10. 1016/j.cognition.2022.105065.

responding to, and so forth, before relying on them as “settled” intuitions to be entered into a process of (collective) reflective equilibrium.

This is an Accepted Manuscript of an article published by Taylor & Francis in *The American Journal of Bioethics* (2022, volume 22, issue no.7, pp.1-3, https://doi.org/10.1080/15265161.2022.2075968 . For the final version and to cite, please refer to the official published version of the article.

Earp, B. D., K. L. McLoughlin, J. T. Monrad, M. S. Clark, and M. J. Crockett. 2021. How social relationships shape moral wrongness judgments. *Nature Communications* 12 (1):1–13. doi:10.1038/s41467-021-26067-4.

Earp, B. D., J. Lewis, V. Dranseika, and I. R. Hannikainen. 2021. Experimental philosophical bioethics and normative inference. *Theoretical Medicine and Bioethics* 42 (3): 91–111. doi:10.1007/s11017-021-09546-z.

Gillam, L. 2016. The zone of parental discretion: An ethical tool for dealing with disagreement between parents and doctors about medical treatment for a child. *Clinical Ethics* 11 (1):1–8. doi:10.1177/1477750915622033.

Giubilini, A., and J. Savulescu. 2018. The artificial moral Advisor. The "ideal observer" meets artificial intelligence. *Philosophy & Technology* 31 (2):169–88. doi:10.1007/ s13347-017-0285-z.

Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2022. Algorithms for ethical decision-making in the clinic: A proof of concept. *The American Journal of Bioethics* 22 (7):4–20. doi:10.1080/15265161.2022.2040647.

Savulescu, J., C. Gyngell, and G. Kahane. 2021. Collective reflective equilibrium in practice (CREP) and controversial novel technologies. *Bioethics* 35 (7):652–63. doi:10. 1111/bioe.12869.

Savulescu, J., G. Kahane, and C. Gyngell. 2019. From public preferences to ethical policy. *Nature Human Behaviour* 3 (12):1241–3. doi:10.1038/s41562-019-0711-6.

Savulescu, J., and H. Maslen. 2015. Moral enhancement and moral artificial intelligence: Moral AI? In *Beyond artificial intelligence: The disappearing human—machine divide*, ed. Jan Romportl, Eva Zackova, and Jozef Kelemen, 79–95. New York: Springer.

Wilkinson, D., and J. Savulescu. 2018. Hard lessons: learning from the Charlie Gard case. *Journal of Medical Ethics* 44:438–442.