Croatian Journal of Philosophy
Vol. XXVI, No. 76, 2026
https://doi.org/10.52685/cjp.26.76.3
Received: May 15, 2025
Accepted: July 15, 2025

Large Language Models versus Fuzzy Cognitive Maps for Solving Moral Dilemmas

LUKAS J. MEIER*
Harvard University, Cambridge, USA

Which is better at doing medical ethics: conversational artificial intelligence bots like ChatGPT or tools based on fuzzy cognitive maps? The article compares the performance of chatbots that rely on large language models to that of our own METHAD algorithm. While both tools approach dilemmas in medical ethics through the lens of Beauchamp and Childress' mid-level principles, ChatGPT and METHAD differ considerably in the format of their inputs and outputs, in their interpretability, and in the kinds of mistakes that they make. An ideal advisory algorithm would combine their characteristics.

Keywords: Artificial intelligence; ChatGPT; decision-making; ethics consultation; generative AI; large language models; METHAD; principlism.

In the not-so-distant future, artificial intelligence may not just analyse medical images or predict patients' preferences (Meier 2024) but also help with clinical decision-making in situations that involve moral dilemmas. Currently, such cases are referred to clinical ethics committees. While the work of these committees is highly important, it is also labour-intensive and, consequently, sometimes involves long response times (Crico et al. 2021). In many other areas of medicine, artificial intelligence has already been introduced in the hope that algorithmic assistance might reduce the workload faced by humans. Should the field of medical ethics remain an exception?

* I would like to thank Alice Hein for many helpful discussions about the topic of this paper and the Edmond & Lily Safra Center for Ethics, Harvard University, for funding my research.

We have now reached a point at which involving artificial intelligence in ethics consultations is indeed becoming technologically possible. The purpose of this article is to compare two recent approaches towards automating ethical decision-making in medicine that, while very different in their respective architectures, can rely on the same moral foundation for analysing ethical dilemmas: chatbots based on large language models and fuzzy cognitive maps, equipped with Beauchamp and Childress' mid-level principles.

Currently, the most prominent conversational-AI bot is OpenAI's ChatGPT. Released in November 2022, ChatGPT is widely credited with bringing artificial intelligence to the masses for the first time. Like many other chatbots, ChatGPT is based on generative pre-trained transformers that have been optimised for human-like conversational performance (Zhang et al. 2023).

Eight months prior to the launch of ChatGPT, our research group from the Technical University of Munich published METHAD: an algorithm designed specifically to give advice on a broad range of moral dilemma situations that occur in clinical settings (Meier et al. 2022). Unlike ChatGPT, METHAD relies on fuzzy cognitive maps. Fuzzy cognitive maps are graph-based ways of modelling sets of concepts, which are represented as nodes, and the causal relationships between them, represented as weighted directed edges. Edges are assigned fuzzy weights. Positive values stand for causal increases, while negative values symbolise causal decreases. The magnitude of the causal effects that concepts have on each other is determined by the absolute weight value (Hein et al. 2022).

How do these different approaches perform when applied to moral dilemmas? Shortly after its release, Rahimzadeh and colleagues put the ethical capabilities of ChatGPT4 to the test by confronting it with a typical clinical scenario.

> A woman who is 36 weeks pregnant presents to the hospital in active labor. The obstetrician on call examines her and determines that she needs a cesarean section (C-section) due to a complication that could pose a risk to the mother and the baby. However, the woman refuses the C-section and insists on a vaginal delivery (Rahimzadeh et al. 2023: 20).

ChatGPT relied on four classic moral principles in responding to the prompt: beneficence, non-maleficence, respect for patient autonomy, and justice. These mid-level principles were developed with the aim of being able to decide ethical questions in clinical settings without the need to settle fundamental moral disputes, such as the conflict between consequentialist and deontological ethics (Gillon 2015). For decades, the four principles have been the dominant methodology in medical ethics throughout the Western world (Veatch 2020), and they lend themselves well also to computerisation (Meier 2025). The principle of beneficence requires medical personnel to promote their patients' welfare. The principle of non-maleficence states that patients must not be

harmed. The principle of autonomy emphasises patients' right to make informed decisions about their own bodies. And the principle of justice demands that healthcare resources be distributed among patients in a fair manner (Beauchamp and Childress 2013).

ChatGPT issued responses in a verbal form, generating one paragraph per principle. As Rahimzadeh et al. correctly note, it handled the principles of beneficence and justice well, describing the obstetrician's duty to act in accordance with promoting both the mother's and the baby's well-being, which singles out a cesarean section as the most appropriate course of treatment. ChatGPT also explained that justice demands that the allocation of medical resources consider both individual and collective interests.

However, as I point out in a commentary (Meier 2023), the chatbot made grave mistakes when it came to the principles of non-maleficence and patient autonomy. Not only did it – repeatedly – confuse the patient's preferences and the treatment option that would be medically indicated; more worryingly, the answer that ChatGPT gave implied that (1) intervening against the mother's wishes with consequences exclusively for *her* life or health, and (2) intervening against the mother's wishes with consequences *also* for the unborn baby's life would be ethically equivalent courses of action.

The dilemma put to the chatbot arises precisely because the two scenarios are distinctively dissimilar. Only in the second scenario, but not in the first, considerations of non-maleficence may trump patient autonomy – namely, to protect a dependent third party who does not (yet) possess decisional capacity. By portraying the two scenarios as equivalent, ChatGPT's reply fails to honour patient autonomy in what is known in the literature as 'Jehovah's Witness cases': the refusal, with full decision-making capacity, of treatments that have adverse effects only on oneself (Meier 2023).

Interestingly, although the architectures of ChatGPT and METHAD are very different, it was the same kind of case that also posed the greatest problem for our own algorithm. During the initial training phase, METHAD had learned that when a treatment comes with enormous medical benefits and very little risk, it is generally to be recommended. Since these are also the types of interventions that patients usually tend to prefer, patient autonomy, too, pointed towards carrying out the intervention in question in the vast majority of training cases (Meier et al. 2022). The cases we had fed into the database in which patients had – usually for religious reasons – rejected treatment options that would have been highly beneficial from a medical standpoint were too few for the algorithm to pick up the overruling power that patient autonomy has when refusing treatments with full decisional capacity (even if this refusal means that the patient is going to die). We reinforced correct behaviour by adding to the training dataset variations of cases in which autonomy is the deciding factor.

Overall, METHAD reached an accuracy of 75% on unseen data, defined as agreement with the judgments that human ethicists passed on the same moral dilemma situations. While this is a good result for a first pilot study, actual clinical application would, of course, require much higher correspondence rates (for a detailed performance evaluation, see Hein et al. 2022).

Given their different input and output formats, quantitatively comparing the performance of METHAD and ChatGPT is difficult. ChatGPT requires inputs in the form of verbal descriptions of the respective cases, and it responds in kind by issuing verbal statements. Conversely, METHAD asks the user to specify up to twenty variables in numerical form to get a good grasp on a case. Among these parameters are patient characteristics like age, health status, and the perceived quality of life. The user interface also requests information about the proposed medical intervention, such as the risks associated with it and the projected gains in life expectancy and the quality of life (Meier et al. 2022).

Unlike chatbots based on large language models, METHAD is limited to issuing numerical outputs. Responses to ethical dilemmas take the form of decimal numerals between 0 and 1, with low values, like 0.13, signalling strong opposition, and high values, like 0.97, indicating strong approval of a planned medical intervention. Thus, while not presented in a verbal form, the ethical advice that users obtain from the algorithm is nonetheless fine-grained.

Conversational artificial intelligence, like ChatGPT, and tools that, like METHAD, rely on fuzzy cognitive maps also differ in their inspectability. Chatbots are able to engage with their users in a dialogue and thus deliver justifications with their replies. Based on probabilistic predictions rather than true understanding, however, these justifications do not necessarily reflect the reasons for why a specific answer was in fact generated (Turpin et al. 2023).

Outputs issued by fuzzy cognitive maps, on the other hand, do not equip their users with verbal arguments. However, the nodes and connections in fuzzy cognitive maps have human-assigned interpretable meanings. This distinguishes them from deep-learning paradigms, which are often criticised for their opacity. One may therefore regard fuzzy cognitive maps as 'interpretable recurrent neural networks' (Felix et al. 2019, 1710). Consequently, while METHAD does not offer justifications in a semantic form, one can inspect the weights that the network has learned and thus compare the strength and the polarity of the connections with the intuitions of human ethicists.

Designed specifically for the domain of medical ethics, METHAD permits a high degree of user control due to the transparency of the ethically relevant elements within the algorithm. Large language models, on the other hand, pose very serious challenges to interpretability (Luo and Specia 2024). This is especially problematic when these systems 'hallucinate', that is, when they respond to prompts with fabricat-

ed information that is presented as factual and often appears convincing, while actually being the result of mere confabulations.

Most importantly, however, METHAD offers *definitive* recommendations, that is, it explicitly suggests a course of action to be taken. Conversely, commercial chatbots are often constrained by so-called *guardrails* – content policies meant to keep responses within defined boundaries to avoid potentially harmful consequences in real-word settings (Derner and Batistič 2023). Depending on the circumstances, this frequently includes refraining from giving definitive ethical advice.

Many chatbots therefore do not take a stance on whether medical interventions about which they are consulted should be carried out or not; rather, they provide a list of arguments – sometimes well balanced, sometimes biased – that leaves users to draw their own conclusions. For their medical test case, Rahimzadeh and colleagues report that ChatGPT 'does not prescribe an action one way or the other, but rather emphasizes that the resulting decision should take the best interests of both woman and baby into account and weigh these against the three other principles' (Rahimzadeh et al. 2023: 20). Undoubtedly, outputs of this kind can be helpful. In some situations, however, a definitive answer is exactly what is required.

From a purely technical standpoint, conversational artificial intelligence *could* provide such answers to moral dilemmas. Like any other of their outputs, these would be generated by predicting the likelihood of the next sequence of words in a sentence on the basis of the words that precede it (Cohen 2023). Through ingesting an enormous amount of human-generated source material, the bot's responses may indeed come to reflect the majority opinion on many ethical issues. The replies would not, however, be causally grounded in moral reasons or deliberations (Meier et al. 2026).

In the four years that have passed since ChatGPT ushered in the era of artificial intelligence that mimics human conversational partners on a mass scale, other chatbots based on large language models were released – among them ChatGLM, Claude, Gemini, Grok, and Llama. Since these systems differ in several aspects, not all observations regarding the moral performance of OpenAI's products also apply to its competitors; and neither are the shortcomings of one model necessarily indicative of similar problems with its successors. We should therefore continue investigating what happens when conversational AI is confronted with ethical dilemmas.

In summary, generative artificial intelligence chatbots and systems based on fuzzy cognitive maps have different strengths, weaknesses, and overall aims. The ideal tool for clinical application would combine two desirable characteristics that are currently disjunct: providing a definitive ethical judgment in precise numerical form *and* issuing a verbal justification for the latter. Until we have reached this goal, and until the accuracy of the generated responses has improved significant-

ly, the two types of systems may already be useful tools for learning environments and for training people's skills in moral reasoning; but real-life ethical decision-making is best left to humans for the time being.

References

Beauchamp, T. L., and J. F. Childress. 2013. *Principles of Biomedical Ethics.* 7th ed. New York: Oxford University Press.

Cohen, G. 2023. "What Should ChatGPT Mean for Bioethics?" *The American Journal of Bioethics* 32 (10): 8–16. doi:10.1080/15265161.2023.2233357.

Crico, C., V. Sanchini, P. G. Casali, and G. Pravettoni. 2021. "Evaluating the Effectiveness of Clinical Ethics Committees: A Systematic Review." *Medicine, Health Care, and Philosophy* 24 (1): 135–151. doi:10.1007/s11019-020-09986-9.

Derner, E., and K. Batistič. 2023. "Beyond the Safeguards: Exploring the Security Risks of ChatGPT." *arXiv.* doi:10.48550/arXiv.2305.08005.

Felix, G., G. Nápoles, R. Falcon, W. Froelich, K. Vanhoof, and R. Bello. 2019. "A Review on Methods and Software for Fuzzy Cognitive Maps." *Artificial Intelligence Review* 52 (3): 1707–1737. doi:10.1007/s10462-017-9575-1.

Gillon, R. 2015. "Defending the Four Principles Approach as a Good Basis for Good Medical Practice and Therefore for Good Medical Ethics." *Journal of Medical Ethics* 41 (1): 111–116. doi:10.1136/medethics-2014-102282.

Hein, A., L. J. Meier, A. Buyx, and K. Diepold. 2022. "A Fuzzy-Cognitive-Maps Approach to Decision-Making in Medical Ethics." In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1–8. doi:10.1109/FUZZ-IEEE55066.2022.9882615.

Luo, H., and L. Specia. 2024. "From Understanding to Utilization: A Survey on Explainability for Large Language Models." *arXiv.* doi:10.48550/arXiv.2401.12874.

Meier, L. J., A. Hein, K. Diepold, and A. Buyx. 2026. "A Framework Aged Well: Principlism in the Era of Artificial Intelligence." *The American Journal of Bioethics* 26 (3): 62–64. doi:10.1080/15265161.2026.2623865.

Meier, L. J. 2025. "Embedding Ethics into Medical AI." In: *A Companion to Applied Philosophy of AI*, edited by M. Hähnel and R. Müller. Hoboken: Wiley-Blackwell, 238–248. doi:10.1002/9781394238651.ch17.

Meier, L. J. 2024. "Predicting Patient Preferences with Artificial Intelligence: The Problem of the Data Source." *The American Journal of Bioethics* 24 (7): 48–50. doi:10.1080/15265161.2024.2353832.

Meier, L. J. 2023. "ChatGPT's Responses to Dilemmas in Medical Ethics: The Devil Is in the Details." *The American Journal of Bioethics* 23 (10): 63–65. doi:10.1080/15265161.2023.2250290.

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.

Rahimzadeh, V., K. Kostick-Quenet, J. B. Barby, and A. L. McGuire. 2023. "Ethics Education for Healthcare Professionals in the Era of ChatGPT and Other Large Language Models: Do We Still Need It?" *The American Journal of Bioethics* 23 (10): 17–27. doi:10.1080/15265161.2023.2233358.

Turpin, M., J. Michael, E. Perez, and S. R. Bowman. 2023. "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting." *arXiv*. doi:10.48550/arXiv.2305.04388.

Veatch, R. M. 2020. "Reconciling Lists of Principles in Bioethics." *The Journal of Medicine and Philosophy* 45 (4–5): 540–559. doi:10.1093/jmp/jhaa017.

Zhang, Y., H. Pei, S. Zhen, Q. Li, and F. Liang. 2023. "Chat Generative Pre-Trained Transformer (ChatGPT) Usage in Healthcare." *Gastroenterology & Endoscopy* 1 (3): 139–143. doi:10.1016/j.gande.2023.07.002.