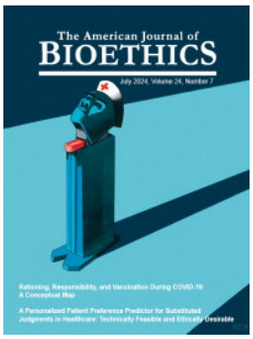

The American Journal of Bioethics

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uajb20

Potentially Perilous Preference Parrots: Why Digital Twins Do Not Respect Patient Autonomy

Georg Starke & Ralf J. Jox

To cite this article: Georg Starke & Ralf J. Jox (2024) Potentially Perilous Preference Parrots: Why Digital Twins Do Not Respect Patient Autonomy, The American Journal of Bioethics, 24:7, 43-45, DOI: 10.1080/15265161.2024.2353810

To link to this article: https://doi.org/10.1080/15265161.2024.2353810

Published online: 24 Jun 2024.

THE AMERICAN JOURNAL OF BIOETHICS
2024, VOL. 24, NO. 7, 43–45
https://doi.org/10.1080/15265161.2024.2353810

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

OPEN ACCESS Check for updates

Potentially Perilous Preference Parrots: Why Digital Twins Do Not Respect Patient Autonomy

Georg Starke[a,b] and Ralf J. Jox[c]

[a]College of Humanities, EPFL; [b]Institute for History and Ethics of Medicine, Technical University of Munich; [c]Lausanne University Hospital and University of Lausanne

The debate about the chances and dangers of a patient preference predictor (PPP) has been lively ever since Annette Rid and David Wendler proposed this fascinating idea ten years ago. Given the technological developments since then, in particular the rise of generative artificial intelligence (AI) and large language models (LLMs), it seems high time that these discussions received an update. In their paper, Earp et al. (2024) meet this need and make a compelling case for a refined, personalized patient preference predictor (called P4), taking "the form of a fine-tuned LLM trained on text produced by, or describing, an individual" (16). Such a system, the authors argue, could result in a "kind of 'digital psychological twin' of the person" (15), ensuring respect for patient autonomy better than currently available alternatives.

The paper provides much food for thought and reminds us once again that the current state of clinical surrogate decision making is far from ideal. While more and more clinical decisions, especially when life is at stake, rely on surrogates, research using hypothetical case vignettes has repeatedly shown that less than 70% of surrogate decision makers accurately predict patients' preferences (Shalowitz, Garrett-Mayer, and Wendler 2006). If one takes respect for patients' autonomy seriously and aims to minimize distress for family members called upon as surrogates, we urgently have to look for ways to improve surrogate decision making in healthcare (Lo 2023).

Yet, given the existing empirical evidence concerning the technical abilities of LLMs, it seems overly optimistic that a P4 could in the near future overcome these fundamental challenges in a realistic setting. Current LLMs have been shown, at least sometimes, to provide inconsistent outputs, to take mutually exclusive stances, and to lead to morally problematic judgments based on their sensitivity to framing (Savage 2023, Bonagiri et al. 2024). Representing purely associative, stochastic models, LLMs have famously been compared with parrots, processing vast language-based data on a probabilistic basis without discernible reference to the texts' meaning (Bender et al. 2021). Crucially, initial research suggests that also fine-tuning LLMs does not mitigate such shortfalls, as has been recently shown with a specific view to morally relevant statements (Kiehne et al. 2024).

If these results are corroborated, it has important implications for the proposed P4. First, a preference predictor in the shape of an LLM could, even after fine-tuning, fall back to replies supported by the pre-trained LLM, reflecting the view most frequently present in the training data instead of mirroring a patient's potentially opposite wishes. Far from being a digital psychological twin, the model may then merely have the persuasive appearance of speaking on the incapacitated person's behalf, e.g. by using words or phrases that they commonly used, without necessarily reflecting their treatment preferences. Instead of safeguarding patient autonomy, such a system could thereby further endanger it by providing an allegedly personalized, confident response with view to treatment decisions while obfuscating the individual's actual preferences.

Second, evidence that fine-tuned LLMs may take mutually exclusive stances uncovers a systematic shortcoming of these models: As of yet, they do not seem to be capable of what could be described as reasoning. We share the authors' view that one needs to be careful to avoid double standards when comparing the performance and capabilities of a P4 and a human surrogate when it comes to appreciating reasons for making clinical decisions. Yet, a human surrogate who is at the same time in favor and against prolonging

CONTACT Georg Starke georg.starke@epfl.ch College of Humanities, EPFL, Lausanne, Switzerland.

life-sustaining treatment, depending merely on the way the question is phrased, would reasonably not be relied upon as surrogate and be excluded from the decision-making process.

To be charitable to the possibility of a P4, let us assume though that the many technical issues, from the consistency and biases of LLMs to the availability of suitable training data, were solved. Even then, respect for the autonomy of the incapacitated patient would involve more than just "accurately" predicting patients' preferences. Respecting autonomy means understanding and applying the personal reasoning underlying a patient's preferences, made possible by rich relational and narrative experience of the person (Entwistle et al. 2010). Thus, respecting autonomy not only depends on the outcome of the decision (which could be accurate by pure chance), but on the way the decision is made. And what is more: it also depends on the kind of agent who makes the decision. Autonomy as normative self-governance and self-determination is made possible by human personhood, involving consciousness, subjectivity, and free will. By consequence, autonomy can only fully be respected by another autonomous agent in a context of interpersonal recognition (Pereira 2013).

To date, digital entities are still far from developing personhood that would enable them to fully respect human autonomy. But what if we keep human surrogates "in the loop" and use P4 simply as an assistive device for surrogates? If their superior predictive accuracy is established, however, it seems unrealistic that surrogates will ever be able and allowed to disregard their predictions and choose otherwise. *De facto*, the P4 will decide, not the surrogate.

The crucial differences between personalized AI models and actual human agents, obscured in the misleading image of a digital twin, also entail important implications for their potential role in ethical decision making. As a purely stochastic model, the P4 in its proposed form can in principle never assume the ethical responsibility inherent in a value judgment on medical treatment. Crucially, this is not merely a question of responsibility gaps or responsibility diffusion. As the authors note, such questions have been extensively discussed, usually with view to "wrong" decisions made by clinical decision support systems (CDSS). Yet, there is an important difference in the use of a classical CDSS and a P4. In standard medical cases, e.g. AI supporting the diagnostic process, the correct or incorrect answer can at least be determined *ex post*. Yet, in many cases for which a preference predictor would be helpful and in fact designed, responsibility goes beyond such concerns. In end-of-life decisions no such ex-post accuracy check is feasible, meaning that the decision maker(s) have to bear—and live with—the moral responsibility for their decisions.

As Earp et al. rightly point out, situations like these can be source of considerable distress in surrogate decision makers, and it is ethically desirable to minimize this burden. Yet, it is precisely due to this distress that great caution seems warranted before introducing any system like the P4. In fact, to avoid distress and evade their individual responsibility, surrogates, professionals, and even patients themselves may be tempted to delegate their human judgment to digital twins, both in and beyond the situation of decisional incapacity—and all the more if such twins are depicted in anthropomorphic terms. To put it bluntly: decision making is so much easier if a machine does it for you.

The discussion about preference predictors seriously suffers from the peculiar fact that they have so far never been empirically proven, let alone tested. While this proof of principle should in fact be attempted, for the time being we should turn our attention to an alternative that has amply proven its effectiveness: comprehensive models of advance care planning that not only increase accuracy of substituted judgment but realize the full respect of patient autonomy in interpersonal encounters (Rietjens et al. 2017; Liu et al. 2024). If we put our hopes on fine-tuned LLMs and digital psychological twins as technical solutions to some of the most existential and intimate human decisions, we risk creating a different kind of P4 instead: a potentially perilous preference parrot.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

FUNDING

The author(s) reported there is no funding associated with the work featured in this article.

ORCID

Georg Starke http://orcid.org/0000-0001-7428-2619
Ralf J. Jox http://orcid.org/0000-0002-3040-4714

REFERENCES

Bender, E. M., T. Gebru, A. McMillan-Major, and S. Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the*

2021 ACM Conference on Fairness, Accountability, and Transparency, 610–23.

Bonagiri, V. K., S. Vennam, P. Govil, P. Kumaraguru, and M. Gaur. 2024. SaGE: Evaluating moral consistency in large language models. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024): 14272–14284. Torino (Italy).

Earp, B. D., S. Porsdam Mann, J. Allen, S. Salloch, V. Suren, K. Jongsma, M. Braun, D. Wilkinson, W. Sinnott-Armstrong, A. Rid, et al. 2024. A personalized patient preference predictor for substituted judgments in healthcare: Technically feasible and ethically desirable. *The American Journal of Bioethics* 24 (7):13–26. doi:10.1080/15265161.2023.2296402.

Entwistle, V. A., S. M. Carter, A. Cribb, and K. McCaffery. 2010. Supporting patient autonomy: The importance of clinician-patient relationships. *Journal of General Internal Medicine* 25 (7):741–5. doi:10.1007/s11606-010-1292-2.

Kiehne, N., A. Ljapunov, M. Bätje, and W.-T. Balke. 2024. Analyzing effects of learning downstream tasks on moral bias in large language models. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024): 904–923. Torino (Italy).

Liu, X., M.-H. Ho, T. Wang, D. S. T. Cheung, and C.-C. Lin. 2024. Effectiveness of dyadic advance care planning: A systematic review and meta-analysis. *Journal of Pain and Symptom Management*. doi:10.1016/j.jpainsymman.2024.01.027.

Lo, B. 2023. Deciding for patients who have lost decision-making capacity—finding common ground in medical ethics. *New England Journal of Medicine* 389 (25):2309–12. doi:10.1056/NEJMp2308484.

Pereira, G. 2013. *Elements of a critical theory of justice*. London; New York: Palgrave McMillan Springer.

Rietjens, J. A. C., R. L. Sudore, M. Connolly, J. J. van Delden, M. A. Drickamer, M. Droger, A. van der Heide, D. K. Heyland, D. Houttekier, D. J. A. Janssen, et al. 2017. Definition and recommendations for advance care planning: An international consensus supported by the European Association for Palliative Care. *The Lancet Oncology* 18 (9):e543–e551. doi:10.1016/S1470-2045(17)30582-X.

Savage, N. 2023. How robots can learn to follow a moral code. *Nature*. doi:10.1038/d41586-023-03258-1.

Shalowitz, D. I., E. Garrett-Mayer, and D. Wendler. 2006. The accuracy of surrogate decision makers: A systematic review. *Archives of Internal Medicine* 166 (5):493–7. doi:10.1001/archinte.166.5.493.

THE AMERICAN JOURNAL OF BIOETHICS
2024, VOL. 24, NO. 7, 45–47
https://doi.org/10.1080/15265161.2024.2353831

Taylor & Francis
Taylor & Francis Group

OPEN PEER COMMENTARIES

The Problematic "Existence" of Digital Twins: Human Intention and Moral Decision

Jeffrey P. Bishop

Saint Louis University

Since surrogates are not good at predicting patient preferences, and since these decisions can cause surrogates distress, some have claimed we need an alternative way to make decisions for incapacitated patients. Fortunately, there's an app for that: a patient preference predictor—PPP (Shalowitz, Garrett-Mayer, and Wendler 2007; Rid and Wendler 2014a, 2014b; Wendler et al. 2016; Lamanna and Byrne 2018). Earp et al, concerned that PPP models draw on demographic preferences rather than the *personal* preferences of patients, now call for a *personalized* patient preference predictor—P4 (Earp et al. 2024). P4 could be developed in several different ways, from looking at personal information found in previous decisions, or found in personal communications, notes, emails, etc., or by completing some sort of inventory of questions designed to uncover preferences, or by some other way of gleaning patient preferences. A patient could even complete a moral psychological inventory to inform the P4. A personal preference input would create a digital twin of the patient, predicting what the real twin would have wanted. Earp et al. argue

CONTACT Jeffrey P. Bishop jeffrey.bishop@slu.edu Health Care Ethics and Philosophy, Saint Louis University, St. Louis, MO, USA.