Personalized Patient Preference Predictors are Neither Technically Feasible nor Ethically Desirable

Forthcoming, American Journal of Bioethics. Please cite published version when available.

Nathaniel Sharadin (natesharadin@gmail.com)
University of Hong Kong

1 Patient Preference Predictors

Except in extraordinary circumstances, patients' clinical care should reflect their preferences. Incapacitated patients cannot report their preferences. This is a problem. Extant solutions to the problem are inadequate: surrogates are unreliable, and advance directives are uncommon. What to do?

One approach to solving this problem would be to increase the proportion of patients with advance directives by very strongly incentivizing their adoption. This approach is easy to implement. For example, we could require advance directives as a condition of receiving (non-emergency) publicly subsidized care, including care that is subsidized through employer-provided health insurance when the employer receives tax-advantages for providing such insurance to their employees. This is an approach I favor.

A very different approach to this problem would be to design an algorithm capable of *p*redicting *p*atients' *p*references -- a "PPP" -- and then use that algorithm to inform care for incapacitated patients. Several authors, including myself, have argued that PPPs face a range of serious problems (N. Sharadin 2024; 2019; N. P. Sharadin 2018; John 2014).

Most importantly, PPPs threaten to undermine patient autonomy in much the same way statistical sentencing threatens to undermine individuals' autonomy (the "autonomy problem"). Intuitively: a choice concerning the criminal sentence for an offender ought to incorporate only considerations that causally bear on a particular individual's disposition to reoffend, rather than also incorporating considerations comprising population-level demographic correlations between (say) race and rate of reoffense. So too with patient care: a choice concerning resuscitation for a patient ought to incorporate only considerations that dispose the particular patient to (dis)prefer resuscitation, rather than population-level demographic correlations between (say) a race and a (dis)preference for resuscitation.

In their recent paper, Earp et al. (2024) propose a new twist on PPPs designed to avoid this (and perhaps other) problems. Earp et al. suggest we *personalize* PPPs using modern machine learning (ML) techniques. According to Earp et al., personalized patient preference predictors (P4s) are an improvement on PPPs because, unlike (mere) PPPs, P4s are responsive to the

grounds of patients' individualized preferences (rather than simply to statistical, population-level demographic data) in making predictions concerning care for incapacitated patients. Here, I'll simply assume that P4s avoid the known problems with PPPs.[1] Nevertheless, contrary to Earp et al.'s claim, personalized patient preference predictors are neither technically feasible nor ethically desirable.

2 Technical Issues

Although Earp et al. offer little in the way of specific technical guidance on the details of the system they have in mind, they suggest implementing a P4 by fine-tuning a large language model (LLM) on (e.g.) a patient's personal data (such as their social media use, blog posts, emails, purchase and browsing history, and so on) and then querying that LLM (presumably with details involving the patient's current physical condition) in order to predict what the patient's would want. While this is clearly a speculative proposal, Earp et al.'s suggestion that we use LLMs to implement the P4 is a crucial component of the view, one they take to distinguish their proposal from nearby alternatives; they say that their proposal "differs from [others] in specifying the use of fine-tuned LLMs" (7) for the purpose of predicting patient preferences. Hence it seems fair to consider in more detail the technical issues facing the use of LLMs for this purpose.[2] For reasons of space, I'll mention only two related issues.[3]

One issue is that Earp et al.'s proposal to use LLMs to implement a P4 is technically undermotivated. They say that a "key assumption for our purposes" is that "[i]n general, the primary function of LLMs is prediction." (9). Therefore (they say), "given data of a sufficient quality and relevance, *prima facie* LLMs should be able to predict medical preferences, too" (ibid.). But it is difficult to understand how they make this conceptual leap. Compare: the primary function of atmospheric general circulation models (AGCMs) is prediction; therefore, given data of a sufficient quality and relevance, *prima facie* AGCMs should be able

[1] Earp et al. discuss the autonomy problem, but they do not say how their proposal aims to avoid what elsewhere I've called the "scope" and the "multiple models" problems (Sharadin 2019). Indeed, P4s appear to exacerbate the scope and the multiple models problems. The scope problem: Earp et al. suggest a P4 might be implemented using a system trained on an individual's "emails, blog posts, or social media posts [...] or even Facebook 'liking' activity" (6). But which emails and posts? Public ones? Which social media posts? Which platforms? The multiple models problem: Earp et al. suggest at least 5 different implementations of the P4 (p. 6), and there are multiple reasonable ways of executing each of these implementations, each of which will vary in their predictions concerning patient care. What principled way can there be for clinicians and other healthcare providers to decide between these models? Worse, the kinds of ML models Earp et al. propose to use (LLMs) are particularly prone to producing widely divergent outputs depending on technical design choices made in deploying the model (such as those involving the inference procedure and other "background conditions"). For discussion, see (Harding and Sharadin 2024). I say more about this issue below, in Section 2.

[2] Earp et al.'s view seems to be that the primary technical challenge to using LLMs as P4s involves acquiring (enough, good) data: "In general, the primary function of LLMs is prediction: given data of a sufficient quality and relevance, prima facie LLMs should be able to predict medical preferences, too" (9). I agree that (enough, good) data will be a barrier to using any LLM as a P4. Here, I'll set this (very big) issue to one side.

[3] For related work on the difficulties associated with evaluating the capabilities of LLMs, see (Harding and Sharadin 2024).

to predict medical preferences, too. That is absurd. Models can be useful for predicting one (kind of) thing but simultaneously useless at predicting another (kind of) thing; and this is something that, sometimes, no amount of data can fix. Worse, LLMs of the sort Earp et al. apparently have in mind are typically trained using a very specific next-token prediction task. Why is this the kind of predictive training task that we would *prima facie* expect to result in a system capable of predicting medical preferences (even after fine-tuning)? Perhaps there is supposed to be some more specific reason to think LLMs could predict medical preferences. Earp et al. note the "large literature on the related notion of aligning AI systems such as LLMs with human values and preferences" (9). But most of that literature -- notably, including the core technical contributions they cite -- emphasizes the fact that LLMs are *not* reliable predictors of human preferences, and that we *don't* have robust, reliable, scalable methods for aligning them with human values. So it's hard to see how the use of LLMs in this context is well-motivated.

But set aside concerns about whether a proposal to use LLMs is undermotivated. Consider the several issues associated with designing and validating a suitable *prompting strategy* designed to elicit accurate predictions of a specific individual patient's preferences. LLM outputs are extraordinarily sensitive to choice of prompting strategy (Liang et al. 2022). Worryingly in the present context, apparently content-irrelevant features of prompts (politeness, sophistication, punctuation, word choice) regularly have substantial effects on model performance. In one striking case, researchers found that models are more likely to provide inaccurate information when a query indicated that the user was uneducated (Perez et al. 2022). One issue therefore involves *designing* a prompting strategy suitable for eliciting a model's actual prediction regarding an individual patient's preferences. A second, related, challenge involves *validating* that prompting strategy. But it's extremely unclear how a particular prompting strategy for a P4 could be validated across all the kinds of use cases that Earp et al. are envisioning. Both these problems are exacerbated by the fact that different strategies work for different models -- there are, as Liang et al. put it, "model-specific incantations" (2022). Worse and worse, research clearly demonstrates that fine-tuning of the sort that Earp et al. propose can itself affect which prompting strategies work (and how well they work) (Stiennon et al. 2020).

3 Moral Issues

Set aside technical reasons for doubting that we should use LLMs as personalized patient preference predictors (P4s). There's a more serious moral challenge with using *any* ML-based P4, independently of whether it's implemented using an LLM. To see this challenge, notice that, in effect, Earp et al.'s P4 is a particular kind of ML-based *content recommendation engine*. The basic idea behind their proposal is that just as content platforms such as TikTok, YouTube, Spotify, etc. might aim to model users' (video, music, etc.) preferences in order to increase metrics such as time spent on the platform and "engagement," we can use ML

techniques to model users' medical preferences in order to increase metrics such as patient satisfaction or perhaps "post-hoc approval of choice." Perhaps as Earp et al. propose this will involve using a fine-tuned LLM, but perhaps (because of technical issues such as those I suggest above) it will involve using some other architecture. The moral problem with ML-based healthcare content recommendation is independent of its specific architectural implementation.

The moral problem is simple. ML-based recommendation systems aim only to maximize their predictive accuracy: they are indifferent as to method. One method for improving the accuracy of a prediction is to make the target of prediction easier to predict. Such systems are therefore indifferent between making predictions that accurately *report* users' preferences and those that effectively *shape* users' preferences. This result shows up in many places, including in studies of users' preferences for (e.g.) extreme content after being exposed to algorithmic recommendations based on milder content (Ribeiro et al. 2020). It's an open technical question why exactly this happens (c.f. Perdomo et al. 2020). But the basic conceptual idea is straightforward: some preferences are easier to predict. Therefore, if a system is incentivized simply to accurately predict preferences, it will be also thereby be incentivized to make predictions that have the result of users having preferences that are easier to predict.

With respect to short-form videos, the moral hazard associated with shaping (rather than merely reporting) users' preferences is relatively minor. Not so with respect to potentially life-altering (or even life-ending) healthcare. The possibility that healthcare content recommendation systems might shape, rather than simply report, a patient's healthcare preferences, is seriously morally concerning. Worse, as I've argued elsewhere (Sharadin 2024), ML researchers presently lack techniques for assuring us that a system is not incentivized to shape its users' preferences in this way. But again, while shaping users' preferences with respect to (e.g.) short-form video consumption might not be (seriously) morally problematic, shaping patients' preferences with respect to medical care clearly is.

References

Earp, Brian D., Sebastian Porsdam Mann, Jemima Allen, Sabine Salloch, Vynn Suren, Karin Jongsma, Matthias Braun, et al. 2024. “A Personalized Patient Preference Predictor for Substituted Judgments in Healthcare: Technically Feasible and Ethically Desirable.” *The American Journal of Bioethics* 0 (0): 1–14. https://doi.org/10.1080/15265161.2023.2296402.

Harding, Jacqueline, and Nathaniel Sharadin. 2024. “What Is It for a Machine Learning System to Have a Capability?”

John, Stephen. 2014. “Patient Preference Predictors, Apt Categorization, and Respect for Autonomy.” *The Journal of Medicine and Philosophy: A Forum for Bioethics and Philosophy of Medicine* 39 (2): 169–77. https://doi.org/10.1093/jmp/jhu008.

Liang, Percy, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, et al. 2022. “Holistic Evaluation of Language Models.” *arXiv Preprint*

arXiv:2211.09110.

Perdomo, Juan, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. 2020. "Performative Prediction." In *Proceedings of the 37th International Conference on Machine Learning*, 7599–7609. PMLR. https://proceedings.mlr.press/v119/perdomo20a.html.

Perez, Ethan, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, et al. 2022. "Discovering Language Model Behaviors with Model-Written Evaluations." arXiv. https://doi.org/10.48550/arXiv.2212.09251.

Ribeiro, Manoel Horta, Raphael Ottoni, Robert West, Virgílio A. F. Almeida, and Wagner Meira. 2020. "Auditing Radicalization Pathways on YouTube." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–41. FAT* '20. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3351095.3372879.

Sharadin, Nathaniel. 2019. "Should Aggregate Patient Preference Data Be Used to Make Decisions on Behalf of Unrepresented Patients?" *AMA Journal of Ethics* 21 (7): 566–74.

———. 2024. "Predicting and Preferring." *Inquiry: An Interdisciplinary Journal of Philosophy*. https://philarchive.org/rec/SHAPAP-28.

Sharadin, Nathaniel Paul. 2018. "Patient Preference Predictors and the Problem of Naked Statistical Evidence." *Journal of Medical Ethics* 44 (12): 857–62. https://doi.org/10.1136/medethics-2017-104509.

Stiennon, Nisan, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. "Learning to Summarize with Human Feedback." In *Advances in Neural Information Processing Systems*, 33:3008–21. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html.