Inquiry
An Interdisciplinary Journal of Philosophy

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/sinq20

Predicting and preferring

Nathaniel Sharadin

To cite this article: Nathaniel Sharadin (25 Sep 2023): Predicting and preferring, Inquiry, DOI: 10.1080/0020174X.2023.2261493

To link to this article: https://doi.org/10.1080/0020174X.2023.2261493

Published online: 25 Sep 2023.

INQUIRY
https://doi.org/10.1080/0020174X.2023.2261493

Predicting and preferring

Nathaniel Sharadin

Department of Philosophy, University of Hong Kong, Pok Fu Lam, Hong Kong

ABSTRACT
The use of machine learning, or 'artificial intelligence' (AI) in medicine is widespread and growing. In this paper, I focus on a specific proposed clinical application of AI: using models to predict incapacitated patients' treatment preferences. Drawing on results from machine learning, I argue this proposal faces a special moral problem. Machine learning researchers owe us assurance on this front before experimental research can proceed. In my conclusion I connect this concern to broader issues in AI safety.

ARTICLE HISTORY Received 28 April 2023; Accepted 11 September 2023

KEYWORDS PPP; medical ethics; AI; patient preference predictors; preference shaping

1 Introduction: patient preference predictors

There's a convergence of a long-standing problem in clinical medicine and newly maturing capabilities of predictive models trained using machine learning (ML) – what are sometimes called 'artificial intelligences' (AIs).[1] I'll describe the problem first, then I'll describe why it's natural to suggest using ML, or AI, to solve it.

The problem is simple: care ought to reflect patients' preferences, but incapacitated patients cannot indicate their preferences. Clinicians can hope to use *indirect* indicators of patients' preferences, e.g. advance directives and surrogates.[2] But these indirect strategies face serious challenges: most patients do not in fact have an advance directive, and surrogates are systematically epistemically unreliable.[3]

In response, researchers have recently suggested a solution based on ML (Biller-Andorno and Biller 2019; Brock 2014; Ferrario, Gloeckler, and

CONTACT Nathaniel Sharadin natesharadin@gmail.com Department of Philosophy, University of Hong Kong, Pok Fu Lam, Hong Kong

[1]In what follows, I mostly refer to these systems as ML systems, rather than as 'AI', in order to avoid unfortunate and controversial implications about machine 'intelligence'.
[2]For an overview, see (Emanuel et al. 1991; Buchanan and Brock 2019).
[3]For discussion, see (Salmond and David 2005; Shalowitz, Garrett-Mayer, and Wendler 2006; Jezewski et al. 2007).

Biller-Andorno 2023; Rid and Wendler 2014b; Wendler 2021; Wendler et al. 2016). Very roughly, the idea is that we can attempt to accurately model incapacitated patients' preferences; we can then use that model to predict what patients would want under a range of clinical conditions, and then use those predictions as the basis for clinical care. Supposing it is technically feasible to develop the algorithmic part of a patient preference predictor (PPP), such a model would need to be *trained*. Such training is sometimes said to face a serious logistical challenge, viz. somehow (legally, one hopes) acquiring the necessary data.[4] For reasons of space, I'll ignore this challenge in what follows.

Whatever the logistical challenges, there are clear ethical concerns with PPPs.[5] However, extant concerns do not target PPPs *qua* ML models trained using deep learning. Instead, these concerns apply equally well if (say) PPPs are entirely hand-programmed statistical models.[6] In this paper I develop a novel ethical problem for PPPs – one that applies to PPPs specifically in virtue of their nature as ML models trained using modern deep learning techniques. I'll argue this problem is sufficiently morally serious to shift the normative burden of proof: those in favor of developing and deploying PPPs in a clinical setting owe us a solution before patient-involving experimental research can safely begin. Absent a meaningful risk mitigation measure, institutional review boards will and should ban exactly the practical research needed to move the proposal forward. This puts the moral ball in the ML researcher's technical court: show us a way to assure ourselves against the moral hazard, or risk a halt to progress in an area of potentially important impact for AI in medicine. Here is how I proceed.

Section 2 gives a brief overview of why predictive models trained using deep learning are indifferent as to methods for achieving improvements in accuracy. Here, I highlight the fact that improvements in accuracy can potentially be achieved by so-called 'performative prediction.' Section 3 lays out the idea of preference shaping and explains why it is morally illicit as a means to achieving accuracy in the case of PPPs. This yields a normative bar that proposals to use PPPs must clear: they must show that a particular model is not incentivized to improve accuracy by way of preference shaping. Section 4 concludes by connecting the problem identified here to broader concerns about AI safety.

[4]See (Rid and Wendler 2014a) for discussion; for relevant machine learning research, see (O. Evans et al. 2018).

[5]For a selection of moral criticism, see (N. Sharadin 2019; N. P. Sharadin 2018; Ditto and Clark 2014; Kim 2014; John 2014; Dresser 2014; Tretter and Samhammer 2023; Mainz 2022). For a recent reply to autonomy-based criticism, see (Jardas et al. 2022).

[6]Compare (N. P. Sharadin 2018).

2 Reward, accuracy & performative prediction

The performance of predictive models is measured by how close their predictions are to the way the world actually is. For example: if it actually rains 80% of the time a weather forecasting model reports an 80% chance of rain, this is a well-performing model. Call this measure of performance 'accuracy.'[7] Any *in*accuracy can in principle be corrected in one of two ways.[8] The first, familiar method, involves changing the model (and so its predictions) to better match the actual distribution of probabilities. This can be done by retraining the model, fine-tuning it, or by otherwise altering its architecture. Less familiarly, but obviously, inaccuracies in a model's predictions can also be corrected by changing the actual distribution of probabilities, which is to say *changing the world*. For instance, if a climate model predicts that it will be more than 3 degrees warmer by the end of the century, then one way to ensure this prediction is correct is by emitting as much carbon into the atmosphere as humanly possible.

ML models trained using deep learning aim to maximize their expected accuracy (or: to minimize inaccuracy); they are incentivized to be on an entirely accuracy-based metric the best predictor they can be.[9] There are in general no constraints on what counts as the 'right kind' of improvement in accuracy: accuracy is accuracy is accuracy, however it's achieved. One way to think about this feature of deep-learning trained systems is that this kind of training doesn't typically restrict the *permissible means* to maximizing accuracy.[10] By default, then, models are indifferent as to *how* to maximize accuracy. For instance, if it was possible for a weather model predicting rain to make it rain, or for a climate model predicting warming to heat up the planet, then, in principle at least, *ceteris paribus*, the model would be indifferent to improving the accuracy of its predictions by making it rain or heating up the planet as compared to adjusting its predictions.

This might seem worrying: after all, we can expect models to be indifferent between improvements to accuracy arrived at by changing their *representations* of the probabilities and those arrived at by changing

[7]For a technical overview, see (Gneiting and Raftery 2007).
[8]Well, three. We could change our scoring rule, or our performance metric. I ignore this possibility in what follows.
[9]I follow the literature in saying that a learner is *incentivized* to do something just in case doing that thing increases performance (or reward). See (Krueger, Maharaj, and Leike 2020, 2).
[10]If this sounds familiar from the Forever War between consequentialists and Kantians, that's not an accident.

the actual *probabilities* – the *facts themselves*. But these are two very different kinds of thing, and we certainly don't want our predictive models doing the latter! The natural reaction to this worry is that it's misplaced. There aren't (to our knowledge) any (e.g. weather or climate) models in existence or development that can act on the world in the usual kind of way required to change the actual distribution of probabilities (e.g. by making it rain or emitting carbon). Models don't really *do* anything.

But, natural as it is, this reaction is too quick. Models do at least one thing. They make predictions. And predictions can, after all, affect the world. For instance: a hedge fund's model predicts that NYSE:GME will fall, and as a result the fund publicly takes up a short position (on margin). Maybe this causes other investors to lose confidence, and so causes the stock to fall. Maybe not: it might instead cause part of the internet to lose its collective mind, attempt a squeeze, and so cause the stock to (briefly) go to the moon.[11] Either way, the model didn't just make a prediction, it made a (more or less convoluted) difference to the *actual likelihood that that prediction would be correct*, i.e. it made a difference to the facts on the ground.

Machine learning researchers do not agree on a name for this phenomenon, whereby a model's prediction can affect its own accuracy.[12] Here, I begrudgingly agree to call it 'performative prediction.'[13] Performative prediction is the family of phenomenon whereby a model's predictions (somehow) make a difference to the spread of probability distributions it aims to represent. Under what conditions are models of the sort we've described here *incentivized* to make performative predictions? That's a trick question; the answer is: under the same conditions they're incentivized to make *any prediction whatsoever*, viz. that doing so maximizes accuracy. They are indifferent between means for maximizing accuracy.

3 Preference shaping & performative prediction

Preference shaping is when an agent's preferences are induced to change exogenously. Preference shaping *per se* is morally neutral, as in: your preference for cilantro exogenously changes when you move to Mexico City.

[11]See (Good 2021).
[12]Philosophers call a related phenomenon self-fulfilling beliefs (Silva forthcoming; Antill 2019).
[13]Following (Perdomo et al. 2020). *Begrudgingly* because it can make it sound as if the model *itself* is doing something. It isn't: we are doing something with the model.

Of course, sometimes preference shaping is *not* morally neutral, as in: you are beaten daily until you come to enjoy cilantro.[14]

In the clinical context, patients' preferences regarding care are generally regarded as sacrosanct: they ought not be intentionally, wittingly shaped.[15] There are, as everywhere, exceptions. You might try to talk a Christian Scientist into wanting a blood transfusion. The exceptions to the rule are justified by a *trade-off* in values.[16] On the one hand, there are important values (e.g. autonomy) at stake in an agent's preferences being down to her; on the other hand, there are important values (e.g. reducing harm, improving outcomes) at stake in not letting people prefer outcomes that are worse on some objective, non-preference-based measure.

Is it ever morally permissible for a *clinician* to shape a patient's preferences for the sake of improving the accuracy of a predictive model of those preferences? Obviously, no. Consider the following dialogue:

> *Doctor*: Great, I have your results. Our in-house-model, Happy Patient, has predicted that you prefer radiation to surgery. Both are equally effective in your case. So, I'll put you down for radiation.
>
> *Patient*: I actually prefer surgery.
>
> *Doctor*: … <checks notes> … I see … In that case, hmm … how about having a look at these statistics about death during surgery and these gruesome pictures of surgical mishaps.
>
> *Patient*: I'd rather not.
>
> *Doctor*: Look, you're seriously hurting Happy Patient's accuracy score, if you could just …
>
> *Patient*: Can I get a referral?

Not only can *clinicians* shape patient preferences, a *model's performative predictions* can also shape patient preferences. This might seem odd and unfamiliar: how could a model's predictions about a patient's preferences count as performative predictions in this sense – how could they shape a patient's preferences?

In fact this phenomenon isn't odd; at least, it isn't unfamiliar. There are a number of more or less well-studied ways in which people's preferences

[14]Compare (Franklin et al. 2022).
[15]This follows from broader ideas about the importance of informed consent. For an overview, see (Faden and Beauchamp 1986); for critical discussion, see (Manson and O'Neill 2007).
[16]This is not controversial. See (Li and Chapman 2020) for discussion.

over outcomes can be shaped by predictions about those preferences, at least when those predictions in some way causally interact with the patients themselves (e.g. by being presented to them). For example, the literature on 'nudging' is rife with strategies for affecting people's preferences by (e.g.) presenting them information about prospective choices in a particular order, or by carefully curating the list of alternatives (e.g. by removing or including irrelevant ones).[17] In the present context, information presented to patients could be (e.g.) a model's predictions about what a patient will want; those predictions about preferences can shape the preferences themselves in exactly the way other kinds of (irrelevant) information can shape people's preferences.

Not only is it possible to *imagine* models' predictions about people's preferences shaping those preferences, we know that in fact this happens in the real world. Forget about medical preferences for a moment. Content-recommendation models, such as those that determine the next song, movie, news clip, or other piece of 'content' in an algorithmically determined feed, offer a simple illustration of the phenomenon.[18] Here is how it works for content-recommendation models. If you're predicted by a content-recommendation model to like content with feature X, then, *ceteris paribus*, you will be shown more content with feature X. In a widely recognized phenomenon known as the 'mere exposure' effect, 'mere' exposure to content with feature X is extraordinarily likely to increase your preference for content with feature X (Lakkakula et al. 2010; Mrkva and Van Boven 2020; Pennycook, Cannon, and Rand 2018; Pliner 1982; Rapp and Salovich 2018; Ulusoy et al. 2021; Fazio et al. 2015; Zebrowitz, White, and Wieneke 2008). Hence, you will (at least on the margins) come to like content with feature X, per the model's predictions. Hence, content-recommendation models can (and do!) shape peoples' preferences. And they do so in a way that improves their own predictions: they are *notorious* performative predictors.[19] The present point is that PPPs are, in effect, a kind of content-recommendation engine: they recommend *medical content*.[20]

Is it ever morally permissible for the predictions of a PPP to (by *whatever* causal route) shape patient preferences simply in order to improve the accuracy of those very predictions? No. This is for the same reasons

[17]For a recent philosophical discussion, see Parmer (2023). The debate over the ethics of nudging is ongoing. For the classic source on 'nudges' see Thaler and Sunstein (2008).
[18]For technical discussion of the broad phenomenon, see (Krueger, Maharaj, and Leike 2020; C. Evans and Kasirzadeh 2022; Farquhar, Carey, and Everitt 2022; Everitt et al. 2021).
[19]See the discussion in (Perdomo et al. 2020).
[20]Thanks to an anonymous referee for encouraging clarity on this point.

as before. The improved accuracy of the predictions doesn't in the requisite way trade-off against the values at stake in shaping patients' preferences.

This doesn't mean that, as a moral matter, PPPs must never actually performatively predict in a way that turns out to shape patients' preferences (and so perhaps improve accuracy). That would be too high a bar, as the content recommendation example, together with the research on the mere exposure effect, illustrates. Simply being told that you prefer something is itself somewhat likely to make you prefer it. And if PPPs are to be *used* (rather than stuck in a drawer) then their predictions will presumably have *some* causal impact, that causal impact might involve shaping patients' preferences, and it might thereby improve the accuracy of the PPP.[21]

Equally: a clinician might *actually* shape a patient's preferences and they might do so in a way that improves a predictive model's accuracy. What's morally impermissible is shaping a patient's preferences *as a means to improving the accuracy of a predictive model*. To avoid this moral hazard in the case of a clinician, we simply detach any incentives a clinician might have for improving the accuracy of a model from the incentives they have to shape patients' preferences. In effect, we disallow or disincentivize dialogues like the one above. There are many obvious ways to do this.

How do we avoid this moral hazard in the case of a model trained using deep learning? That's a very good question. It is trivial to describe the property we want a suitable model to have. We want it to be such that it is manifestly, provably not incentivized to make performative predictions that shape patient preferences. But it turns out to be extraordinarily difficult to assure ourselves that any given model in fact has this property. There are, for example, no extant proposals in the machine learning literature for identifying the conditions under which a model has this property.[22] This is a Bad Thing. If we lack a meaningful way to assure ourselves that a particular PPP is not incentivized to shape patients'

[21]The only research that I'm aware of that approaches the question of performative prediction in the context of medical AI is a review article (Chen et al. 2021); there, the authors simply note the possibility of distributional shift (aka performative prediction).

[22]This is also the conclusion of other AI safety researchers. Compare (Hendrycks et al. 2022; C. Evans and Kasirzadeh 2022; Ashton and Franklin 2022). This is not to say that there are no proposals about how to ensure that models have *other* interesting properties related to performative prediction, e.g., can achieve various strategic equilibria; for relevant discussion see (Mendler-Dünner et al. 2020; Brown, Hod, and Kalemaj 2022; Miller, Perdomo, and Zrnic 2021). Very recent work aims to identify and penalize induced preference shifts in recommender systems (e.g. Carroll et al. 2022); that work is clearly relevant to the present problem, though it doesn't yet represent a solution.

preferences, then it seems clear that experimental, which is to say patient-involving, research on PPPs will (and should) be blocked by institutional review boards, which correctly take a dim view of this kind of moral risk. It goes without saying that PPPs should not be deployed in a clinical setting.

4 Conclusion & discussion

Let me step back from the particulars for a moment; below, I'll return to them. Readers familiar with the broader literature on AI safety will not be surprised by anything they've read. There's a concatenation of long-standing, well-known, very hard problems in AI safety that all have something like the general form of the problem I've here identified in a particular case: either we can't actually affect a model's incentives, or we can't interrogate them, or having done either of those we can't *assure* ourselves of the precise content of those incentives, etc. Despite important differences in theorizing about and technical approaches to solving these and related problems, they're often lumped under one name: the 'Alignment Problem'.[23] The alignment problem is big, fuzzy, and poorly understood. So, one way to respond to what I've said so far is to point out that it's simply an instance of a well-known (if not well-understood) problem and, moreover, *give us a minute, we're working on it*.

That reaction, I think, is a mistake given the present context. The alignment problem may be big and fuzzy, but the problem identified here is relatively small and precise.[24] ML models are being used in medicine right now, today. PPPs are being proposed not just as an interesting idea, but as a thing that should begin to be put into practice (Ferrario, Gloeckler, and Biller-Andorno 2023; Wendler 2021). We should not wait for a solution to the broadest possible description of the broadest possible AI safety problem (e.g. the Alignment Problem) to clearly articulate the moral hazards involved in particular proposed uses of the technology. I am not sure I know what progress on the Alignment Problem looks like, or even what way it is best to talk about many of the questions that researchers seem to care about in this broad area. But things are much simpler in this case. A proposal is being seriously floated in the scientific and philosophical literature to deploy technology that will, by the

[23]This is part of why no one agrees on a definition of *the* Alignment Problem.
[24]Thanks to an anonymous referee for this way of putting the contrast between the alignment problem and the problem I identify in the paper.

technical experts' own admission, not be disincentivized from bringing about what is an uncontroversially serious moral harm.[25]

Disclosure statement

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

ORCID

Nathaniel Sharadin http://orcid.org/0000-0002-4610-664X

References

Antill, Gregory. 2019. "Evidence and Self-Fulfilling Belief." *American Philosophical Quarterly* 56 (4): 319–330. https://doi.org/10.2307/48563046.

Ashton, Hal, and Matija Franklin. 2022. "The Problem of Behaviour and Preference Manipulation in AI Systems." In *CEUR Workshop Proceedings*. Vol. 3087. CEUR Workshop Proceedings.

Biller-Andorno, Nikola, and Armin Biller. 2019. "Algorithm-Aided Prediction of Patient Preferences — An Ethics Sneak Peek." *New England Journal of Medicine* 381 (15): 1480–1485. https://doi.org/10.1056/NEJMms1904869.

Brock, Dan W. 2014. "Reflections on the Patient Preference Predictor Proposal." *Journal of Medicine and Philosophy* 39 (2): 153–160.

Brown, Gavin, Shlomi Hod, and Iden Kalemaj. 2022. "Performative Prediction in a Stateful World." In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, 6045–6061. PMLR. https://proceedings.mlr.press/v151/brown22a.html.

Buchanan, Allen, and Dan W. Brock. 2019. "Deciding for Others." In *Death, Dying and the Ending of Life, Volumes I and II*, edited by Margaret Battin and Leslie Francis, 205–282.

Carroll, Micah D., Anca Dragan, Stuart Russell, and Dylan Hadfield-Menell. 2022. "Estimating and Penalizing Induced Preference Shifts in Recommender Systems." In *Proceedings of the 39th International Conference on Machine Learning*, 2686–2708. PMLR. https://proceedings.mlr.press/v162/carroll22a.html.

Chen, Irene Y., Shalmali Joshi, Marzyeh Ghassemi, and Rajesh Ranganath. 2021. "Probabilistic Machine Learning for Healthcare." *Annual Reviews of Biomedical Data Science* 4: 393–415. https://doi.org/10.48550/arXiv.2009.11087.

Ditto, P. H., and C. J. Clark. 2014. "Predicting End-of-Life Treatment Preferences: Perils and Practicalities." *Journal of Medicine and Philosophy* 39 (2): 196–204. https://doi.org/10.1093/jmp/jhu007.

[25]Thanks to Simon Goldstein, Dan Hendrycks, Jacqueline Harding, Cameron Kirk-Giannini, David Krueger, Nick Laskowski, Robert Long, Elliot Thornley, and members of the Cottage Group for helpful discussions about these and related issues.

Dresser, R. 2014. "Law, Ethics, and the Patient Preference Predictor." *Journal of Medicine and Philosophy* 39 (2): 178–186. https://doi.org/10.1093/jmp/jhu004.

Emanuel, Linda L., Michael J. Barry, John D. Stoeckle, Lucy M. Ettelson, and Ezekiel J. Emanuel. 1991. "Advance Directives for Medical Care — A Case for Greater Use." *New England Journal of Medicine* 324 (13): 889–895. https://doi.org/10.1056/NEJM199103283241305.

Evans, Charles, and Atoosa Kasirzadeh. 2022. "User Tampering in Reinforcement Learning Recommender Systems." arXiv. https://doi.org/10.48550/arXiv.2109.04083.

Evans, Owain, Andreas Stuhlmüller, Chris Cundy, Ryan Carey, Thomas McGrath, and Andrew Schreiber. 2018. "Predicting Human Deliberative Judgments with Machine Learning".

Everitt, Tom, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. "Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective." *Synthese* 198 (S27): 6435–6467. https://doi.org/10.1007/s11229-021-03141-4.

Faden, Ruth R., and Tom L. Beauchamp. 1986. *A History and Theory of Informed Consent*. Oxford: Oxford University Press.

Farquhar, Sebastian, Ryan Carey, and Tom Everitt. 2022. "*Path-Specific Objectives for Safer Agent Incentives*." arXiv. https://doi.org/10.48550/arXiv.2204.10018.

Fazio, Lisa K., Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. 2015. "Knowledge Does not Protect Against Illusory Truth." *Journal of Experimental Psychology: General* 144 (5): 993–1002. https://doi.org/10.1037/xge0000098.

Ferrario, Andrea, Sophie Gloeckler, and Nikola Biller-Andorno. 2023. "Ethics of the Algorithmic Prediction of Goal of Care Preferences: From Theory to Practice." *Journal of Medical Ethics* 49 (3): 165–174. https://doi.org/10.1136/jme-2022-108371.

Franklin, Matija, Hal Ashton, Rebecca Gorman, and Stuart Armstrong. 2022. "Recognising the Importance of Preference Change: A Call for a Coordinated Multidisciplinary Research Effort in the Age of AI." *The AAAI-22 Workshop on AI for Behavior Change*, 1–7. https://doi.org/10.48550/arXiv.2203.10525.

Gneiting, Tilmann, and Adrian E Raftery. 2007. "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association* 102 (477): 359–378. https://doi.org/10.1198/016214506000001437.

Good, Owen S. 2021. "GameStop's Stock Market Explosion, Explained." *Polygon (blog)*. January 27, 2021. https://www.polygon.com/2021/1/27/22252600/gamestop-stock-gme-why-whats-happening-explain.

Hendrycks, Dan, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2022. "Unsolved Problems in ML Safety." arXiv. https://doi.org/10.48550/arXiv.2109.13916.

Jardas, E. J., David Wasserman, and David Wendler. 2022. "Autonomy-Based Criticisms of the Patient Preference Predictor." *Journal of Medical Ethics* 48 (5): 304–410.

Jezewski, Mary Ann, Mary Ann Meeker, Loralee Sessanna, and Deborah S. Finnell. 2007. "The Effectiveness of Interventions to Increase Advance Directive Completion Rates." *Journal of Aging and Health* 19 (3): 519–536. https://doi.org/10.1177/0898264307300198.

John, S. 2014. "Patient Preference Predictors, Apt Categorization, and Respect for Autonomy." *Journal of Medicine and Philosophy* 39 (2): 169–177. https://doi.org/10.1093/jmp/jhu008.

Kim, S. Y. H. 2014. "Improving Medical Decisions for Incapacitated Persons: Does Focusing on "Accurate Predictions" Lead to an Inaccurate Picture?" *Journal of Medicine and Philosophy* 39 (2): 187–195. https://doi.org/10.1093/jmp/jhu010.

Krueger, David, Tegan Maharaj, and Jan Leike. 2020. "*Hidden Incentives for Auto-Induced Distributional Shift*." arXiv. https://doi.org/10.48550/arXiv.2009.09153.

Lakkakula, Anantha, James Geaghan, Michael Zanovec, Sarah Pierce, and Georgianna Tuuri. 2010. "Repeated Taste Exposure Increases Liking for Vegetables by low-Income Elementary School Children." *Appetite* 55 (2): 226–231. https://doi.org/10.1016/j.appet.2010.06.003.

Li, Meng, and Gretchen B. Chapman. 2020. "Medical Decision Making." In *The Wiley Encyclopedia of Health Psychology*, edited by Lee Cohen, 347–353. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119057840.ch84.

Mainz, Jakob Thrane. 2022. "The Patient Preference Predictor and the Objection from Higher-Order Preferences." *Journal of Medical Ethics* 49 (3): 221–222. https://doi.org/10.1136/jme-2022-108427.

Manson, Neil C., and O. O'Neill. 2007. *Rethinking Informed Consent in Bioethics*. Cambridge: Cambridge University Press. https://eprints.lancs.ac.uk/id/eprint/4042/.

Mendler-Dünner, Celestine, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. 2020. "Stochastic Optimization for Performative Prediction." In *Advances in Neural Information Processing Systems, 33*, 4929–4939. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/33e75ff09dd601bbe69f351039152189-Abstract.html.

Miller, John P., Juan C. Perdomo, and Tijana Zrnic. 2021. "Outside the Echo Chamber: Optimizing the Performative Risk." In *Proceedings of the 38th International Conference on Machine Learning*, 7710–7720. PMLR. https://proceedings.mlr.press/v139/miller21a.html.

Mrkva, Kellen, and Leaf Van Boven. 2020. "Salience Theory of Mere Exposure: Relative Exposure Increases Liking, Extremity, and Emotional Intensity." *Journal of Personality and Social Psychology* 118 (6): 1118–1145. https://doi.org/10.1037/pspa0000184.

Parmer, W. Jared. 2023. "Nudges, Nudging, and Self-Guidance Under the Influence." *Ergo* 9 (44): 1199–1232.

Pennycook, Gordon, Tyrone D. Cannon, and David G. Rand. 2018. "Prior Exposure Increases Perceived Accuracy of Fake News." *Journal of Experimental Psychology: General* 147 (12): 1865–1880. https://doi.org/10.1037/xge0000465.

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.

Pliner, Patricia. 1982. "The Effects of Mere Exposure on Liking for Edible Substances." *Appetite* 3 (3): 283–290. https://doi.org/10.1016/S0195-6663(82)80026-3.

Rapp, David N., and Nikita A. Salovich. 2018. "Can't We Just Disregard Fake News? The Consequences of Exposure to Inaccurate Information." *Policy Insights from the*

Behavioral and Brain Sciences 5 (2): 232–239. https://doi.org/10.1177/2372732218785193.

Rid, Annette, and David Wendler. 2014a. "Treatment Decision Making for Incapacitated Patients: Is Development and Use of a Patient Preference Predictor Feasible?" *Journal of Medicine and Philosophy* 39 (2): 130–152.

Rid, Annette, and David Wendler. 2014b. "Use of a Patient Preference Predictor to Help Make Medical Decisions for Incapacitated Patients." *Journal of Medicine and Philosophy* 39 (2): 104–129.

Salmond, Susan W., and Estrella David. 2005. "Attitudes Toward Advance Directives and Advance Directive Completion Rates." *Orthopaedic Nursing* 24 (2): 117.

Shalowitz, David I., Elizabeth Garrett-Mayer, and David Wendler. 2006. "The Accuracy of Surrogate Decision Makers: A Systematic Review." *Archives of Internal Medicine* 166 (5): 493–497.

Sharadin, Nathaniel Paul. 2018. "Patient Preference Predictors and the Problem of Naked Statistical Evidence." *Journal of Medical Ethics* 44 (12): 857–862. https://doi.org/10.1136/medethics-2017-104509.

Sharadin, Nathaniel. 2019. "Should Aggregate Patient Preference Data Be Used to Make Decisions on Behalf of Unrepresented Patients?" *AMA Journal of Ethics* 21 (7): 566–574.

Silva, Paul Jr. forthcoming. "Self-Fulfilling Beliefs: A Defense." *Australasian Journal of Philosophy*.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth and Happiness*. United Kingdom: Penguin Books.

Tretter, Max, and David Samhammer. 2023. "For the Sake of Multifacetedness. Why Artificial Intelligence Patient Preference Prediction Systems Shouldn't be for Next of kin." *Journal of Medical Ethics* 49 (3): 175–176. https://doi.org/10.1136/jme-2022-108775.

Ulusoy, Ezgi, Dustin Carnahan, Daniel E Bergan, Rachel C Barry, Siyuan Ma, Suhwoo Ahn, and Johnny McGraw. 2021. "Flooding the Zone: How Exposure to Implausible Statements Shapes Subsequent Belief Judgments." *International Journal of Public Opinion Research* 33 (4): 856–872. https://doi.org/10.1093/ijpor/edab022.

Wendler, David. 2021. "A Call for a Patient Preference Predictor." *Critical Care Medicine* 49 (6): 877–880.

Wendler, David, Bob Wesley, Mark Pavlick, and Annette Rid. 2016. "A New Method for Making Treatment Decisions for Incapacitated Patients: What Do Patients Think About the Use of a Patient Preference Predictor?" *Journal of Medical Ethics* 42 (4): 235–241.

Zebrowitz, Leslie A., Benjamin White, and Kristin Wieneke. 2008. "Mere Exposure and Racial Prejudice: Exposure to Other-Race Faces Increases Liking for Strangers of That Race." *Social Cognition* 26 (3): 259–275. https://doi.org/10.1521/soco.2008.26.3.259.