The epistemic virtues of harnessing rigorous machine learning systems in ethically-sensitive domains

Thomas F Burns

Forthcoming in *Journal of Medical Ethics*

Some physicians, in their care of patients at risk of misusing opioids, use machine learning (ML)-based Prediction Drug Monitoring Programmes (PDMPs) to guide their decision-making in the prescription of opioids. This can cause a conflict: a PDMP score can indicate a patient is at a high risk of opioid abuse while a patient expressly reports oppositely. The prescriber is then left to balance the credibility and trust of the patient with the PDMP score.

Pozzi [1] argues that a prescriber who downgrades the credibility of a patient's testimony based on a low PDMP score is epistemically and morally unjustified and contributes to a form of testimonial injustice. This results in patients being silenced, excluded from decision-making processes, and subjected to structural injustices. Additionally, the use of ML systems in medical practices raises concerns about perpetuating existing inequalities, overestimating their capabilities, and displacing human authority. However, almost the very same critiques apply to human-based systems. Formalization, ML systems included, should instead be viewed positively [2], and precisely as a powerful means to begin eroding these and other problems in ethically-sensitive domains. In this case, the epistemic virtues of formalization include promoting transparency, consistency, and replicability in decision-making. Rigorous ML systems can also help ensure that models abide by express standards, constraints, constraints which are well-defined, and which are open to scrutiny and improvements both within and outside the specific domain of application. This is therefore an opportunity for less injustice, not more.

As detailed in the analysis of Fricker [3], and as Pozzi [1] well-describes in the medical domain, there are unfortunately many situations in which a person's testimony is attributed less credibility by others on the unjustified bases of, for example, gender or racial identity. These instances of testimonial injustice can occur entirely absent of ML systems. Pozzi's [1] concern is that ML systems 'can create further imbalances' [1] (p. 2). However, I view ML systems as an opportunity for correction, not amplification, of these imbalances. This may require changes in the way particular ML systems are designed or used in these domains, and changes in the way medical professionals consider and weigh the information such systems provide.

ML systems are often described as 'black box' systems, where it is impossible to know how the system's outputs relate to its inputs. While this is an open area of research [4], many advances have and continue to be made in ML interpretability and explainability, including in the medical domain [5]. This type of formalization can help to mitigate the 'black box' critique by requiring that models be built using well-defined and transparent methods, with clear inputs, outputs, and decision rules. This can help to build trust in the broader system within which these decisions are made, as stakeholders can more explicitly understand how assessments are made and thereby better assess the validity of those decisions.

Indeed, herein holds one of the key advantages of formalization vis-à-vis interpretable and explainable ML systems: by pursuing their further development (in a rigorous way), we can promote systematic transparency in decision-making processes. Instead of relying solely on the ability of an individual prescriber, we can transparently formalize salient variables – not to supplant the prescriber or the patient in their knowledge and expertise, but to supplement and debias [2].

Said differently: Can the average, individual physician outperform such a ML system in terms of transparency, consistency, and replicability? Can we identify and interrogate a physician's biases precisely at any day or time, now or in the past? If the ML system is sufficiently interpretable and explainable, then it offers an opportunity for us to formalize the expertise and critical evaluations of collectives of professionals and the people they serve.

For such ML systems to be successful in doing so, however, they need to be *rigorously* developed and applied. I consider a *rigorous ML system* to be one which, functionally, can be interrogated as accurately and fully (or better) as a reasonable person would expect to interrogate a person or group of persons performing the same task. By this definition, current ML-based PDMPs may not be rigorous. However, it is conceivable to me that they could be made rigorous. Upon doing so, we will not only gain the benefit of transparency, we also gain the benefits of consistency and replicability.

ML systems are commonly trained using large datasets, which often contain more information than any human could hope to remember or recall. By having precise access to such a large volume of data, rigorous ML systems can be trained to remain strictly consistent or neutral to chosen variables in a way humans can never practically achieve. This can therefore help to reduce errors and biases in decision-making, and improve the reliability of the broader system.

Replicability can also improve. While some ML models are trained on proprietary or sensitive data, which can make it difficult for other researchers or practitioners to reproduce or interrogate the results or build on the work, rigorous ML systems do not suffer this problem. Instead, by requiring rigor, we can facilitate the sharing and replication of models, and enable researchers and practitioners to build on each other's work to share best practices.

There are many potential vices of formalization and dangers in the use of ML systems in medicine and other ethically-sensitive domains. However, there are also virtues and opportunities which are critical for building trust in medical and healthcare systems, reducing errors and biases, and enabling the sharing and replication of successful models.

Rigorous ML systems should be seen as a necessary but not sufficient condition for ensuring ethical and responsible use of ML systems in medicine. A comprehensive approach that considers the socio-technical aspects, power relations, legal considerations, and real-world consequences of these systems will still be needed. This approach should involve interdisciplinary collaboration among technologists, ethicists, physicians, other experts, and stakeholders (including patients) to evaluate and refine the developed models and systems, ensuring that they are ethically robust and contextually appropriate.

References

1. Pozzi G. Testimonial injustice in medical machine learning, Journal of Medical Ethics, 2023.

2. Dutilh Novaes C. “The debiasing effect of formalization” (pp. 221–248) in Formal Languages in Logic, 2012.
3. Fricker M. Epistemic injustice: power and the ethics of knowing. Oxford University Press, 2007.
4. Allo P. The Interpretability Problem and the Critique of Abstraction(s), Journal of Cross-Disciplinary Research in Computational Law, 2022.
5. Petch J., Di S., Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology, Canadian Journal of Cardiology, 2022.