We might be afraid of black-box algorithms

Carissa Véliz [1], Carina Prunkl,[1] Milo Phillips-Brown,[1,2] Theodore M Lechterman [1]

INTRODUCTION

Fears of black-box algorithms are multiplying. They are said to prevent accountability,[1] to make it harder to detect bias[2] and so on. Some fears concern the epistemology of black-box algorithms in medicine and the ethical implications of that epistemology. In 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI,'[3] Juan Durán and Jongsma seek to allay such fears. While we find some of their arguments compelling, we still see reasons for fear.

The gap between epistemic and normative justification

Duránand Jongsma's main claim is that black-box algorithms can confer epistemic justification.They helpfully note the scope of this claim's implications: its truth would not alone give the ethical stamp of approval to decisions that are informed by black-box algorithms. For example, a clinician may be epistemically justified—on the basis of a black-box algorithm's diagnosis—in believing that a patient has cancer without thereby being normatively justified in (say) administering chemotherapy. The epistemic justification can inform the decision, but further, moral considerations are needed to normatively justify it. This is a general point about the different natures of epistemic and normative justification.

Yet, especially in the medical context, the distinction between epistemic and normative considerations are not always clear-cut. For example, an algorithm that looks like it plays a purely epistemic role—such as one that diagnoses a given condition—may implicitly encode normative considerations about what counts as a 'healthy' or 'normal' body.[4] Another example might be an algorithm that recommends treatments, where considerations about an algorithm's reliability bleed into the normative judgments about how to treat patients.

Understanding Durán and Jongsma on epistemic trustworthiness

Some opponents of black-box algorithms, Durán and Jongsma write, assume that transparency is necessary for 'epistemic trustworthiness'. If true, black-box algorithms would indeed be worrisome. (Imagine a doctor prescribing a medication because an algorithm she is not justified in believing diagnosed the patient with a certain condition.)

As noted, Durán and Jongsma reject this assumption, maintaining that 'computationally reliability' is sufficient for epistemic trustworthiness: '[C]omputational reliabilism, which does not require transparency and supports the reliability of algorithms, justifies the belief that results of medical AI are to be trusted' (p. 1).

While we find this thought intriguing, we would like to know more about how Durán and Jongsma think about how transparency and epistemic trustworthiness relate to one another. In a different part of the article, Durán *et al* seem to imply that transparency *is* necessary for epistemic trustworthiness: 'transparency by itself is necessary, although not sufficient for entrenching the reliability of black-box algorithms and the overall trustworthiness of their results' (p. 9).

Trust and automation complacency

Even if black-box algorithms can be epistemically trust*worthy*, they may nonetheless not engender the appropriate levels of trust in people. Trust is primarily a psychological state, and it is well documented that human beings are susceptible to being *too trusting* of algorithms. *Automation complacency* is the tendency to overestimate the reliability of decision support systems. In the medical context, automation complacency could prove particularly dangerous. While Durán *et al* give a conceptual argument that black-box algorithms are epistemically trustworthy, there is an open empirical question of whether black-box algorithms are more or less likely to engender automation complacency.

Transparency would allow practitioners to retrace algorithmic recommendations and achieve some familiarity with the algorithm's workings. We wonder whether transparency in turn could provide at least some protection against complacency, since a clinician can be held more accountable for her decisions.

When black-box algorithms hide normative assumptions

As pointed out above, epistemic and normative considerations often blur into each other in the medical setting. This means, among other things, that black-box algorithms can interfere with normative deliberation. Suppose we have a clinician and patient who are confronted with a black-box algorithm that recommends a certain course of chemotherapy. How do they proceed? Because the algorithm is not transparent, patient and clinician are unable to interrogate the basis for the recommendation; they have limited grounds to determine whether, for instance, the decision reflects particular evaluative judgments about pain tolerance, longevity, bodily integrity, resuscitation preferences and so on. In such a case, we believe, the non-transparency of the algorithm may undermine governing principles of medical ethics, including patient autonomy and informed consent.

Computational reliability over time

Let us grant Durán *et al*'s key claim that computationally reliable black-box medical algorithms are epistemically trustworthy. It must be noted that an algorithm that is computationally reliable in one setting at one time might not be everywhere and forever computationally reliable.

Machine learning algorithms typically work with historical data that might become obsolete. Imagine that a black-box algorithm is computationally reliable, at some time, in diagnosing the severity of a patient's cancer. At this time, some patients live within a certain zip code where there is a toxic factory. Whenever there is doubt in the result of, say, an MRI of a patient who lives in this zip code, the algorithm rightly diagnoses a patient with a serious form of cancer. If, 5 years later, the factory is shut down and the increased risk of severe cancer disappears, the algorithm may then be *over*diagnosing the seriousness of cancer in patients from that zip code. When patients have unnecessary surgery (eg, the removal of a tumour that would have never become a health threat),

[1]Institute for Ethics in AI, University of Oxford, Oxford, UK
[2]Jain Family Institute, New York City, USA

Correspondence to Dr Carissa Véliz, Hertford College, University of Oxford, Oxford OX1 3BW, UK; carissa.veliz@philosophy.ox.ac.uk

it is hard to identify the overtreatment. Only with randomised controlled trials[5]—and not with the markers of computational reliability that Durán *et al* argue for—can we realise that an algorithm is leading us to overtreatment.

CONCLUSION

While black-box algorithms are new, ill-understood processes have long been adopted in medicine; for example, we have used aspirin for many more years than we have understood it. More work is needed to reflect on the differences and similarities between black-box algorithms and traditional medical treatments whose workings are opaque to us. For starters, the mechanism of aspirin is constant over time, but many black-box algorithms change as they get new information. Furthermore, how aspirin works is a natural fact; how algorithms work depends on us. Better understanding the ethical implications of those differences will give us a window into just how afraid we should be of black-box algorithms in medicine.

Twitter Carissa Véliz @carissaveliz

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient consent for publication Not required.

Provenance and peer review Commissioned; internally peer reviewed.

CV, CP, MP-B and TML contributed equally.

To cite Véliz C, Prunkl C, Phillips-Brown M, *et al*. *J Med Ethics* 2021;**47**:339–340.

Received 31 March 2021
Accepted 31 March 2021

► http://dx.doi.org/10.1136/medethics-2020-106820

J Med Ethics 2021;**47**:339–340.
doi:10.1136/medethics-2021-107462

ORCID iDs
Carissa Véliz http://orcid.org/0000-0002-3189-3994
Theodore M Lechterman http://orcid.org/0000-0001-5085-2891

REFERENCES

1 O'Neil C. *Weapons of math destruction*. New York: Crown Books, 2015.
2 Challen R, Denny J, Pitt M, *et al*. Artificial intelligence, bias and clinical safety. *BMJ Qual Saf* 2019;28(3):231–7.
3 Durán JM, Jongsma KR. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical AI. *J Med Ethics* 2021;45:329–35.
4 Butler J. Judith Butler with Sunaura Taylor: Interdependence. In: Taylor A, ed. *The examined life: excursions with contemporary Thinkers*. New York: The New Press, 2009: 185–214.
5 Véliz C. Privacy and digital ethics after the pandemic. *Nat Electron* 2021;4(1):10–11.