Design publicity of black box algorithms: a support to the epistemic and ethical justifications of medical AI systems

Andrea Ferrario

ABSTRACT

In their article 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI', Durán and Jongsma discuss the epistemic and ethical challenges raised by black box algorithms in medical practice. The opacity of black box algorithms is an obstacle to the trustworthiness of their outcomes. Moreover, the use of opaque algorithms is not normatively justified in medical practice. The authors introduce a formalism, called computational reliabilism, which allows generating justified beliefs on the algorithm reliability and trustworthy outcomes of artificial intelligence (AI) systems by means of epistemic warrants, called reliability indicators. However, they remark the need for reliability indicators specific to black box algorithms and that justified knowledge is not sufficient to justify normatively the actions of the physicians using medical AI systems. Therefore, Durán and Jongsma advocate for a more transparent design and implementation of black box algorithms, providing a series of recommendations to mitigate the epistemic and ethical challenges behind their use in medical practice. In this response, I argue that a peculiar form of black box algorithm transparency, called design publicity, may efficiently implement these recommendations. Design publicity encodes epistemic, that is, reliability indicators, and ethical recommendations for black box algorithms by means of four subtypes of transparency. These target the values and goals, their translation into design requirements, the performance and consistency of the algorithm altogether. I discuss design publicity applying it to a use case focused on the automated classification of skin lesions from medical images.

In the paper 'Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI',[1] Durán and Jongsma stress out the necessity of solid epistemology of algorithms as a basis for the investigation of the ethical challenges arising from their use.[1] If it is not possible to entrench reliable knowledge from medical AI, physicians follow the AI predictions based on ill-posed reasons.[1] Moreover, they claim that transparency[i] is unfit at coping with the epistemic *desiderata* of those using medical AI (i.e. generating reliable outputs) and that epistemic opacity of the black box algorithms makes 'it impossible to ground the reliability of the algorithm and, consequently, on whether researchers, physicians and patients can trust the results of such systems'.[1] The lack of 'epistemic warrants' due to opacity of algorithms is an obstacle to their trustworthiness and the ethically justified use of medical AI systems. In fact, without epistemic warrants physicians are not justified to include inputs from opaque algorithms in their decision-making. Therefore, the introduction of methods aiming at justifying beliefs on the algorithms is sought after to generate trustworthy results. This affects the trust in the medical AI, as well.[ii]

Durán and Jongsma claim that computational reliabilism allows discussing 'the epistemic conditions for the reliability of black box algorithms, and the trustworthiness of results in medical AI'.[1] The authors argue that computational reliabilism offers reasons to justify the beliefs in the results generated by the AI system. As an outcome, this allows fostering trust in the medical AI by the justified trustworthiness of the algorithm outcomes. However, computational reliability is originally defined in the context of computer simulations, and not of black box algorithms.[2] Therefore, as noted by Durán and Jongsma, reliability indicators specific to black box algorithms have still to be introduced to efficiently tackle the epistemic challenges arising from their use in medical AI systems.

Moreover, Durán and Jongsma note that having conditions that justify epistemic beliefs on black box algorithms is a necessary but not sufficient condition for the moral justification of physicians' actions.[1] In fact, they argue that algorithms should contribute to decision making by providing inputs, only.[1] As physicians face the process of interpreting the algorithmic outputs for their decision making, this may lead to different outcomes and moral consequences.[1] Therefore, it is important to assess the alignment (or the lack of) between the values built in the algorithm and those expressed by the patient, for all possible decision-making scenarios.[1]

The authors suggest that the provision of information on the way in which algorithms are designed, implemented and maintained may support the applicability of black box algorithm in medical practice, given the aforementioned epistemic and ethical challenges.[1] They advocate for a close collaboration between physicians and informatics experts. Moreover, they argue that the algorithmic recommendations should be presented with different interpretations, highlighting decision-making options that reflect different medical scenarios.[1] These recommendations should be discussed with the patients, to identify which option would align the best with their values.

I argue that it is possible to combine the aforementioned epistemic and ethical *desiderata* of black box algorithms in an account of transparency, called 'design publicity'.[3] As opposed to algorithmic transparency or post hoc interpretability,[iii] design publicity supports both (1) the introduction of reliability indicators specific to black box algorithms and (2) the disclosure of the values behind the design of the algorithm and their implementation. It allows fostering collaboration between engineers and physicians along the end-to-end development of the medical AI. In fact, design publicity provides justifications of the (1) epistemic knowledge and (2) use of the algorithm by means of

Management, Technology and Economics, ETH Zurich, Zürich, Switzerland

Correspondence to Dr Andrea Ferrario, Management Technology and Economics, ETH Zürich, Zürich, Switzerland; aferrario@ethz.ch

[i] The term 'transparency', i.e. 'algorithmic procedures that make the inner workings of a black box algorithm interpretable to humans'[1] is used in the sense originally introduced by Lipton[7]. On the other hand, "post-hoc interpretability" refers to the provision of explanations of the algorithm outcomes.[3 7]

[ii] I note that trust and trustworthiness are distinct concepts. The former refers to a process leading to the decision to rely on a person's (or a machine's) action to achieve a predetermined goal, while the latter is a property of whom (or what) is potentially trusted. I also note that there is no consensus on whether it is possible to define the concept of trust in medical AI and discuss its trustworthiness (as opposed to sheer reliance and reliability). I refer to Hatherley and Ferrario et al.'s contributions for a recent debate on this point[8 9] and Jacovi *et al*[10] for a discussion on trust and trustworthiness from the human-machine interaction perspective.

[iii] See foot note one for a definition of transparency and post hoc interpretability of black box algorithms.

four components: value, translational, performance and consistency transparency.

Value transparency aims at disclosing the 'the standards, norms, and goal that were implemented in the system'.[3] In particular, it allows discussing these normative standards with those eventually affected by its outcomes, that is, patients. Therefore, value transparency supports the choice of normative standards and values as a result of discussions involving different stakeholders, as opposed to the implementation of values relevant to machine learning engineers, only.[3] It offers the possibility to discuss the ethical trade-offs and value scenarios stemming from shared decision making with medical AI systems. Translation transparency refers to the disclosure of the criteria behind the implementation of a given algorithmic goal in machine code.[3] This is relevant for the case of AI systems as a given goal and its accompanying values may be implemented in different ways through value-sensitive design procedures.[4] Performance transparency allows discussing the limits of the algorithm performance, the error types and their epistemic implications.[3] In particular, it discloses the conditions granting the epistemic authority[iv] of the algorithm, contributing to the evaluation of its epistemic reliability.[5] Finally, consistency transparency tackles an often neglected challenge of deployed algorithms: their dependence on time and algorithm retraining.[3] It discloses the necessity to ensure consistency in the provision of algorithm outcomes, by explaining the effects of retraining on the outcomes and the physicians' decision making.

As an example, let us consider Loi *et al*'s design publicity account for the Esteva *el al*'s skin cancer deep learning classifier.[6] This medical AI system uses a deep learning algorithm and expert knowledge (a taxonomy of skin diseases) to classify skin lesions in medical images across 757 disease classes.[6] In this example, value transparency explains the goal, standard and norms of the skin cancer classifier. The system is conceived to provide a reliable classification of a skin lesion from a medical image, returning a confidence score and a label as input for dermatologists to determine the best treatment. The system has to be reliable to avoid unnecessary treatments and false negatives. More importantly, it is conceived to provide ample segments of the population with a self-diagnosis tool in their mobile devices, fostering patients autonomy beyond the clinic.[6] Translation transparency explains how the aforementioned goal and values are implemented by providing a description of how data are curated, an overview of all possible lesions in training data, and their variability (eg, luminosity conditions and the inclusion of different ethnic groups). Moreover, it supports the explanation of how the reliability goal is achieved, by means of reliability indicators (eg, by fostering transparency and post hoc interpretability). It describes how the expert knowledge is encoded within a skin lesion taxonomy. Performance transparency supports the justification of the use of in vitro assessments of the performance of the classifier and the resulting epistemic authority of the medical AI, by disclosing the details and limitations of the performance assessment setting, including the choice of the binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses, and malignant melanomas versus benign nevi.[6] Finally, consistency transparency advocates for explanations of the dependence on time of the algorithm outcomes and its impact on patients' treatments. These explanations may provide patients with 'histories' of outcomes for their skin lesions, using old versions of the algorithm and of the taxonomy. An analysis of the 'time robustness' of the classification of a lesion may act as a mean to increase trustworthiness in a given outcome and trust in the AI system.[v]

As a result, I argue that design transparency may support the applicability of black box algorithms in medical practice by the combination of epistemic and ethical justifications, without necessarily searching for alternatives, such as the use of interpretable models (only),[7] or black box algorithms without warrants.[1]

[iv] Bjerring and Busch discuss the challenges behind the epistemic authority of medical AI systems.[5] Here, I note that the authority is often stemming from in-vitro assessments of the performance of the medical AI, where the AI and human experts are given the same type of information in a controlled environment.[6] I argue that discussions supporting performance transparency may contribute to the introduction of more realistic settings to test the epistemic authority of medical AI systems.

[v] For a patient with a skin lesion it would be probably of no interest to know that the algorithm that automatically classifies the lesion from its medical images has been designed through transfer-learning from a Google proprietary algorithm trained on millions of images from 1000 object categories.[6] On the other hand, knowing that the algorithm combines high quality medical imaging information with expert-knowledge to classify skin lesions is of relevance, as it provides the (correct) picture of a composite system that leverages the knowledge of experts in dermatology and oncology, on top of state-of-the-art algorithms for image classification.

Acknowledgements I thank Michele Loi and Eleonora Viganò for their valuable comments on the first draft of the manuscript.

Contributors AF is the sole author of this work.

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 Not commissioned; internally peer reviewed.

To cite Ferrario A. *J Med Ethics* Epub ahead of print: [please include Day Month Year]. doi:10.1136/medethics-2021-107482

Received 7 April 2021
Accepted 20 April 2021

Linked

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

J Med Ethics 2021;**0**:1–3.
doi:10.1136/medethics-2021-107482

REFERENCES

1 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. doi:10.1136/medethics-2020-106820. [Epub ahead of print: 18 Mar 2021].
2 Durán JM, Formanek N. Grounds for trust: essential Epistemic opacity and computational Reliabilism. *Minds Mach* 2018;28(4):645–66.
3 Loi M, Ferrario A, Viganò E. Transparency as design publicity: explaining and justifying inscrutable algorithms. *Ethics Inf Technol*2020:1–11.
4 van de Poel I, Poel vande I. Embedding values in artificial intelligence (AI) systems. *Minds Mach* 2020;30(3):385–409.
5 Bjerring JC, Busch J. Artificial intelligence and patient-centered decision-making. *Philos Technol*;141(1).
6 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542(7639):115–8.
7 Lipton ZC. The Mythos of model interpretability. *Queue* 2018;16(3):31–57.
8 Hatherley JJ. Limits of trust in medical AI. *J Med Ethics* 2020;46(7):478–81.
9 Ferrario A, Loi M, Viganò E. Trust does not need to be human: it is possible to trust medical AI. *J Med Ethics*;33(3):medethics-2020-106922.

10 Jacovi A, Marasović A, Miller T. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In: *Proceedings of the 2021 ACM conference on Fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery. In Press, 2021: 624–35.