POLICY FORUM

TECHNOLOGY AND REGULATION

Beware explanations from AI in health care

The benefits of explainable artificial intelligence are not what they appear

By **Boris Babic[1,2,3], Sara Gerke[4,5], Theodoros Evgeniou[6], I. Glenn Cohen[5,7]**

Artificial intelligence and machine learning (AI/ML) algorithms are increasingly developed in health care for diagnosis and treatment of a variety of medical conditions (*1*). However, despite the technical prowess of such systems, their adoption has been challenging, and whether and how much they will actually improve health care remains to be seen. A central reason for this is that the effectiveness of AI/ML-based medical devices depends largely on the behavioral characteristics of its users, who, for example, are often vulnerable to well-documented biases or algorithmic aversion (*2*). Many stakeholders increasingly identify the so-called black-box nature of predictive algorithms as the core source of users' skepticism, lack of trust, and slow uptake (*3*, *4*). As a result, lawmakers have been moving in the direction of requiring the availability of explanations for black-box algorithmic decisions (*5*). Indeed, a near-consensus is emerging in favor of explainable AI/ML among academics, governments, and civil society groups. Many are drawn to this approach to harness the accuracy benefits of noninterpretable AI/ML such as deep learning or neural nets while also supporting transparency, trust, and adoption. We argue that this consensus, at least as applied to health care, both overstates the benefits and undercounts the drawbacks of requiring black-box algorithms to be explainable.

EXPLAINABLE VERSUS INTERPRETABLE

It is important to first distinguish explainable from interpretable AI/ML. These are two very different types of algorithms with different ways of dealing with the problem of opacity—that AI predictions generated from a black box undermine trust, accountability, and uptake of AI.

A typical AI/ML task requires constructing an algorithm that can take a vector of inputs (for example, pixel values of a medical image) and generate an output pertaining to, say, disease occurrence (for example, cancer diagnosis). The algorithm is trained on past data with known labels, which means that the parameters of a mathematical function that relate the inputs to the output are estimated from that data. When we refer to an algorithm as a "black box," we mean that the estimated function relating inputs to outputs is not understandable at an ordinary human level (owing to, for example, the function relying on a large number of parameters, complex combinations of parameters, or nonlinear transformations of parameters).

Interpretable AI/ML (which is not the subject of our main criticism) does roughly the following: Instead of using a black-box function, it uses a transparent ("white-box") function that is in an easy-to-digest form, for example, a linear model whose parameters correspond to additive weights relating the input features and the output or a classification tree that creates an intuitive rule-based map of the decision space. Such algorithms have been described as intelligible (*6*) and decomposable (*7*). The interpretable algorithm may not be immediately understandable by everyone (even a regression requires a bit of background on linear relationships, for example, and can be misconstrued). However, the main selling point of interpretable AI/ML algorithms is that they are open, transparent, and capable of being understood with reasonable effort. Accordingly, some scholars argue that, under many conditions, only interpretable algorithms should be used, especially when they are used by governments for distributing burdens and benefits (*8*). However, requiring interpretability would create an important change to ML as it is being done today—essentially that we forgo deep learning altogether and whatever benefits it may entail.

Explainable AI/ML is very different, even though both approaches are often grouped together. Explainable AI/ML, as the term is typically used, does roughly the following: Given a black-box model that is used to make predictions or diagnoses, a second explanatory algorithm finds an interpretable function that closely approximates the outputs of the black box. This second algorithm is trained by fitting the predictions of the black box and not the original data, and it is typically used to develop the post hoc explanations for the black-box outputs and not to make actual predictions because it is typically not as accurate as the black box. The explanation might, for instance, be given in terms of which attributes of the input data in the black-box algorithm matter most to a specific prediction, or it may offer an easy-to-understand linear model that gives similar outputs as the black-box algorithm for the same given inputs (*4*). Other models, such as so-called counterfactual explanations or heatmaps, are also possible (*9*, *10*). In other words, explainable AI/ML ordinarily finds a white box that partially mimics the behavior of the black box, which is then used as an explanation of the black-box predictions.

Three points are important to note: First, the opaque function of the black box remains the basis for the AI/ML decisions, because it is typically the most accurate one. Second, the white box approximation to the black box cannot be perfect, because if it were, there would be no difference between the two. It is also not focusing on accuracy but on fitting the black box, often only locally. Finally, the explanations provided are post hoc. This is unlike interpretable AI/ML, where the explanation is given using the exact same function that is responsible for generating the output and is known and fixed ex ante for all inputs.

A substantial proportion of AI/ML-based medical devices that have so far been cleared or approved by the US Food and Drug Administration (FDA) use noninterpretable black-box models, such as deep learning (*1*). This may be because black-box models are deemed to perform better in many health care applications, which are often of massively high dimensionality, such as image recognition or genetic prediction. Whatever the reason, to require an explanation of black-box AI/ML systems in health care at present entails using post hoc explainable AI/ML models, and this is what we caution against here.

[1]Department of Philosophy, The University of Toronto, Toronto, ON, Canada. [2]Department of Statistical Sciences, The University of Toronto, Toronto, ON, Canada. [3]INSEAD, Singapore. [4]Penn State Dickinson Law, Carlisle, PA, USA. [5]The Petrie-Flom Center for Health Law Policy, Biotechnology, and Bioethics at Harvard Law School, The Project on Precision Medicine, Artificial Intelligence, and the Law (PMAIL), Cambridge, MA, USA. [6]INSEAD, Fontainebleau, France. [7]Harvard Law School, Cambridge, MA, USA. Email: igcohen@law.harvard.edu.

LIMITS OF EXPLAINABILITY

Explainable algorithms have been a relatively recent area of research, and much of the focus of tech companies and researchers has been on the development of the algorithms themselves—the engineering—and not on the human factors affecting the final outcomes. The prevailing argument for explainable AI/ML is that it facilitates user understanding, builds trust, and supports accountability (*3*, *4*). Unfortunately, current explainable AI/ML algorithms are unlikely to achieve these goals—at least in health care—for several reasons.

Ersatz understanding

Explainable AI/ML (unlike interpretable AI/ML) offers post hoc algorithmically generated rationales of black-box predictions, which are not necessarily the actual reasons behind those predictions or related causally to them. Accordingly, the apparent advantage of explainability is a "fool's gold" because post hoc rationalizations of a black box are unlikely to contribute to our understanding of its inner workings. Instead, we are likely left with the false impression that we understand it better. We call the understanding that comes from post hoc rationalizations "ersatz understanding." And unlike interpretable AI/ML where one can confirm the quality of explanations of the AI/ML outcomes ex ante, there is no such guarantee for explainable AI/ML. It is not possible to ensure ex ante that for any given input the explanations generated by explainable AI/ML algorithms will be understandable by the user of the associated output. By not providing understanding in the sense of opening up the black box, or revealing its inner workings, this approach does not guarantee to improve trust and allay any underlying moral, ethical, or legal concerns.

There are some circumstances where the problem of ersatz understanding may not be an issue. For example, researchers may find it helpful to generate testable hypotheses through many different approximations to a black-box algorithm to advance research or improve an AI/ML system. But this is a very different situation from regulators requiring AI/ML-based medical devices to be explainable as a precondition of their marketing authorization.

ILLUSTRATION: BRUNO MANYOKU

Lack of robustness

For an explainable algorithm to be trusted, it needs to exhibit some robustness. By this, we mean that the explainability algorithm should ordinarily generate similar explanations for similar inputs. However, for a very small change in input (for example, in a few pixels of an image), an approximating explainable AI/ML algorithm might produce very different and possibly competing explanations, with such differences not being necessarily justifiable or understood even by experts. A doctor using such an AI/ML-based medical device would naturally question that algorithm.

Tenuous connection to accountability

It is often argued that explainable AI/ML supports algorithmic accountability. If the system makes a mistake, the thought goes, it will be easier to retrace our steps and delineate what led to the mistake and who is responsible. Although this is generally true of interpretable AI/ML systems, which are transparent by design, it is not true of explainable AI/ML systems because the explanations are post hoc rationales, which only imperfectly approximate the actual function that drove the decision. In this sense, explainable AI/ML systems can serve to obfuscate our investigation into a mistake rather than help us to understand its source. The relationship between explainability and accountability is further attenuated by the fact that modern AI/ML systems rely on multiple components, each of which may be a black box in and of itself, thereby requiring a fact finder or investigator to identify, and then combine, a sequence of partial post hoc explanations. Thus, linking explainability to accountability may prove to be a red herring.

THE COSTS OF EXPLAINABILITY

Explainable AI/ML systems not only are unlikely to produce the benefits usually touted of them but also come with additional costs (as compared with interpretable systems or with using black-box models alone without attempting to rationalize their outputs).

Misleading in the hands of imperfect users

Even when explanations seem credible, or nearly so, when combined with prior beliefs of imperfectly rational users, they may still drive the users further away from a real understanding of the model. For example, the average user is vulnerable to narrative fallacies, where users combine and reframe explanations in misleading ways. The long history of medical reversals—the discovery that a medical practice did not work all along, either failing to achieve its intended goal or carrying harms that outweighed the benefits—provides examples of the risks of narrative fallacy in health care. Relatedly, explanations in the form of deceptively simple post hoc rationales can engender a false sense of (over)confidence. This can be further exacerbated through users' inability to reason with probabilistic predictions, which AI/ML systems often provide (*11*), or the users' undue deference to automated processes (*2*). All of this is made more challenging because explanations have multiple audiences, and it would be difficult to generate explanations that are helpful for all of them.

Underperforming in at least some tasks

If regulators decide that the only algorithms that can be marketed are those whose predictions can be explained with reasonable fidelity, they thereby limit the system's de-

velopers to a certain subset of AI/ML algorithms. For example, highly nonlinear models that are harder to approximate in a sufficiently large region of the data space may thus be prohibited under such a regime. This will be fine in cases where complex models—like deep learning or ensemble methods—do not particularly outperform their simpler counterparts (characterized by fairly structured data and meaningful features, such as predictions based on relatively few patient medical records) (*8*). But in others, especially in cases with massively high dimensionality—such as image recognition or genetic sequence analysis—limiting oneself to algorithms that can be explained sufficiently well may unduly limit model complexity and undermine accuracy.

BEYOND EXPLAINABILITY

If explainability should not be a strict requirement for AI/ML in health care, what then? Regulators like the FDA should focus on those aspects of the AI/ML system that directly bear on its safety and effectiveness—in particular, how does it perform in the hands of its intended users? To accomplish this, regulators should place more emphasis on well-designed clinical trials, at least for some higher-risk devices, and less on whether the AI/ML system can be explained (*12*). So far, most AI/ML-based medical devices have been cleared by the FDA through the 510(k) pathway, requiring only that substantial equivalence to a legally marketed (predicate) device be demonstrated, without usually requiring any clinical trials (*13*).

Another approach is to provide individuals added flexibility when they interact with a model—for example, by allowing them to request AI/ML outputs for variations of inputs or with additional data. This encourages buy-in from the users and reinforces the model's robustness, which we think is more intimately tied to building trust. This is a different approach to providing insight into a model's inner workings. Such interactive processes are not new in health care, and their design may depend on the specific application. One example of such a process is the use of computer decision aids for shared decision-making for antenatal counseling at the limits of gestational viability. A neonatologist and the prospective parents might use the decision aid together in such a way to show how various uncertainties will affect the "risk:benefit ratios of resuscitating an infant at the limits of viability" (*14*). This reflects a phenomenon for which there is growing evidence—that allowing individuals to interact with an algorithm reduces "algorithmic aversion" and makes them more willing to accept the algorithm's predictions (*2*).

From health care to other settings

Our argument is targeted particularly to the case of health care. This is partly because health care applications tend to rely on massively high-dimensional predictive algorithms where loss of accuracy is particularly likely if one insists on the ability of good black-box approximations with simple enough explanations, and expertise levels vary. Moreover, the costs of misclassifications and potential harm to patients are relatively higher in health care compared with many other sectors. Finally, health care traditionally has multiple ways of demonstrating the reliability of a product or process, even in the absence of explanations. This is true of many FDA-approved drugs. We might think of medical AI/ML as more like a credence good, where the epistemic warrant for its use is trust in someone else rather than an understanding of how it works. For example, many physicians may be quite ignorant of the underlying clinical trial design or results that led the FDA to believe that a certain prescription drug was safe and effective, but their knowledge that it has been FDA-approved and that other experts further scrutinize it and use it supplies the necessary epistemic warrant for trusting the drug. But insofar as other domains share some of these features, our argument may apply more broadly and hold some lessons for regulators outside health care as well.

When interpretable AI/ML is necessary

Health care is a vast domain. Many AI/ML predictions are made to support diagnosis or treatment. For example, Biofourmis's RhythmAnalytics is a deep neural network architecture trained on electrocardiograms to predict more than 15 types of cardiac arrhythmias (*15*). In cases like this, accuracy matters a lot, and understanding is less important when a black box achieves higher accuracy than a white box. Other medical applications, however, are different. For example, imagine an AI/ML system that uses predictions about the extent of a patient's kidney damage to determine who will be eligible for a limited number of dialysis machines. In cases like this, when there are overarching concerns of justice— that is, concerns about how we should fairly allocate resources—ex ante transparency about how the decisions are made can be particularly important or required by regulators. In such cases, the best standard would be to simply use interpretable AI/ML from the outset, with clear predetermined procedures and reasons for how decisions are taken. In such contexts, even if interpretable AI/ML is less accurate, we may prefer to trade off some accuracy, the price we pay for procedural fairness.

CONCLUSION

We argue that the current enthusiasm for explainability in health care is likely overstated: Its benefits are not what they appear, and its drawbacks are worth highlighting. For health AI/ML-based medical devices at least, it may be preferable not to treat explainability as a hard and fast requirement but to focus on their safety and effectiveness. Health care professionals should be wary of explanations that are provided to them for black-box AI/ML models. Health care professionals should strive to better understand AI/ML systems to the extent possible and educate themselves about how AI/ML is transforming the health care landscape, but requiring explainable AI/ML seldom contributes to that end. ■

REFERENCES AND NOTES

1. S. Benjamens, P. Dhunnoo, B. Meskó, *NPJ Digit. Med.* **3**, 118 (2020).
2. B. J. Dietvorst, J. P. Simmons, C. Massey, *Manage. Sci.* **64**, 1155 (2018).
3. A. F. Markus, J. A. Kors, P. R. Rijnbeek, *J. Biomed. Inform.* **113**, 103655 (2021).
4. M. T. Ribeiro, S. Singh, C. Guestrin, in *KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), pp. 1135–1144.
5. S. Gerke, T. Minssen, I. G. Cohen, in *Artificial Intelligence in Healthcare*, A. Bohr, K. Memarzadeh, Eds. (Elsevier, 2020), pp. 295–336.
6. Y. Lou, R. Caruana, J. Gehrke, in *KDD '12: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2012), pp. 150–158.
7. Z. C. Lipton, *ACM Queue* **16**, 1 (2018).
8. C. Rudin, *Nat. Mach. Intell.* **1**, 206 (2019).
9. D. Martens, F. Provost, *Manage. Inf. Syst. Q.* **38**, 73 (2014).
10. S. Wachter, B. Mittelstadt, C. Russell, *Harv. J. Law Technol.* **31**, 841 (2018).
11. R. M. Hamm, S. L. Smith, *J. Fam. Pract.* **47**, 44 (1998).
12. S. Gerke, B. Babic, T. Evgeniou, I. G. Cohen, *NPJ Digit. Med.* **3**, 53 (2020).
13. U. J. Muehlematter, P. Daniore, K. N. Vokinger, *Lancet Digit. Health* **3**, e195 (2021).
14. U. Guillen, H. Kirpalani, *Semin. Fetal Neonatal Med.* **23**, 25 (2018).
15. Biofourmis, RhythmAnalytics (2020); www.biofourmis.com/solutions/.

ACKNOWLEDGMENTS

We thank S. Wachter for feedback on an earlier version of this manuscript. All authors contributed equally to the analysis and drafting of the paper. **Funding:** S.G. and I.G.C. were supported by a grant from the Collaborative Research Program for Biomedical Innovation Law, a scientifically independent collaborative research program supported by a Novo Nordisk Foundation grant (NNF17SA0027784). I.G.C. was also supported by Diagnosing in the Home: The Ethical, Legal, and Regulatory Challenges and Opportunities of Digital Home Health, a grant from the Gordon and Betty Moore Foundation (grant agreement number 9974). **Competing interests:** S.G. is a member of the Advisory Group–Academic of the American Board of Artificial Intelligence in Medicine. I.G.C. serves as a bioethics consultant for Otsuka on their Abilify MyCite product. I.G.C. is a member of the Illumina ethics advisory board. I.G.C. serves as an ethics consultant for Dawnlight. The authors declare no other competing interests.

10.1126/science.abg1834