UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Why we do need Explainable AI for Healthcare

Cina, G.; Röber, T.; Goedhart, R.; Birbil, I.

DOI
10.48550/arXiv.2206.15363

Publication date
2022

Document Version
Submitted manuscript

License
CC BY

Link to publication

Citation for published version (APA):
Cina, G., Röber, T., Goedhart, R., & Birbil, I. (2022). *Why we do need Explainable AI for Healthcare*. (v1 ed.) ArXiv. https://doi.org/10.48550/arXiv.2206.15363

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:16 Sept 2026

arXiv:2206.15363v1 [cs.HC] 30 Jun 2022

Why we do need Explainable AI for Healthcare

Giovanni Cinà, PhD[1, 2], Tabea Röber, MSc[3], , Rob Goedhart, PhD[3], and Ilker Birbil, PhD[3]

[1]Institute of Logic, Language, and Computation, University of Amsterdam, NL
[2]Pacmed, Amsterdam, NL
[3]Business Analytics, Amsterdam Business School, University of Amsterdam, Amsterdam, NL

July 1, 2022

Abstract

The recent spike in certified Artificial Intelligence (AI) tools for healthcare has renewed the debate around adoption of this technology. One thread of such debate concerns Explainable AI and its promise to render AI devices more transparent and trustworthy. A few voices active in the medical AI space have expressed concerns on the reliability of Explainable AI techniques, questioning their use and inclusion in guidelines and standards. Revisiting such criticisms, this article offers a balanced and comprehensive perspective on the utility of Explainable AI, focusing on the specificity of clinical applications of AI and placing them in the context of healthcare interventions. Against its detractors and despite valid concerns, we argue that the Explainable AI research program is still central to human-machine interaction and ultimately our main tool against loss of control, a danger that cannot be prevented by rigorous clinical validation alone.

1 Introduction

Along with the blooming of Artificial Intelligence (AI) and the accompanying increase in model complexity, there has been a surge of interest in explainable AI (XAI henceforth), namely AI that allows humans to understand its inner workings.[1,2,3,4,5] This interest is particularly keen in safety-critical domains such as healthcare, where it is perceived that XAI can engender trust, help monitoring bias, and facilitate AI development.[1,6] XAI has already shown to improve clinicians' ability to diagnose and assess prognoses of diseases as well as assist with planning and resource allocation. For example, Letham and colleagues developed a stroke prediction model that matches the performance of the most accurate machine learning algorithms, while remaining as interpretable as conventional scoring methods used in clinical practice.[7]

Despite the enthusiasm, and a growing community of researchers devoting energy to XAI, there is currently no consensus on the reliability of XAI techniques and several researchers have cast serious doubts on whether XAI solutions should be incorporated into guidelines and standards, or even deployed at all.[8,9]

Undoubtedly, there is an inherent tension between the desire for machines performing better than humans, and the desire for machines providing human-understandable explanations. Together with their super-human capacities – such as the ability to juggle dozens or hundreds of factors – the sub-symbolic character of statistical learning techniques contributes to rendering machine learning models opaque to humans.

There is a wide variety of techniques for XAI, and many categorizations have been proposed in the literature. Techniques can roughly be grouped into local vs. global, and model-specific vs. model-agnostic approaches. Local methods aim to explain model outputs for individual samples, while global methods focus on making models more explainable at an aggregate level. Model-specific methods are tailored to explain a specific type of model, while model-agnostic methods can be applied to a range of different models. Many of the well-known techniques yield post-hoc explanations, meaning that they generate explanations for already trained, so-called 'black-box', models. Alternatively, there exist approaches that are inherently explainable, also known as white-box models, such as decision trees and linear regression models. A detailed taxonomy is beyond the scope of this paper; for an extensive overview we refer the reader to existing reviews.[10,11,12]

What is important to stress is that, in order to overcome the opaqueness of black-box models, XAI techniques often bring in additional layers of complexity, such as:

- explanations themselves can be inaccurate or deceiving,[13,9]
- since an objective ground truth is lacking, it is hard to evaluate XAI techniques,[14,15]
- trade-offs exist between explainability and other desiderata such as high performance.[16,17]

What role is left then for XAI in the medical realm, where adoption and trust of new technologies are notoriously complex? Given the wealth and heterogeneity of XAI techniques, as well as the doubts and limitations mentioned in the literature, the answer to this question is not straightforward.

In this Viewpoint we collect, review and appraise the criticisms that are leveled against XAI. We conclude that, despite several critical points of attention, there is no decisive argument to abandon the research for better XAI techniques, especially in the medical field, where a higher level of oversight is preferable.

2 Criticisms to Explainable AI

We begin by recollecting the main doubts raised about the usefulness and efficacy of XAI, in no particular order; each point is then expanded and discussed. The criticisms, along with the main counterpoints discussed below, are visually represented in Figure 1.

1. XAI techniques may be incorrect or unreliable.
2. Human users may fall prey to biases when interacting with XAI tools.
3. There are no objective measures of explainability.
4. XAI may work on an aggregate level, but not on an individual level.
5. Evidence-based assessment is the proper way to evaluate medical (AI) interventions, not XAI.
6. In the medical domain, we cannot trade performance for explainability.
7. XAI is not suitable for all users.

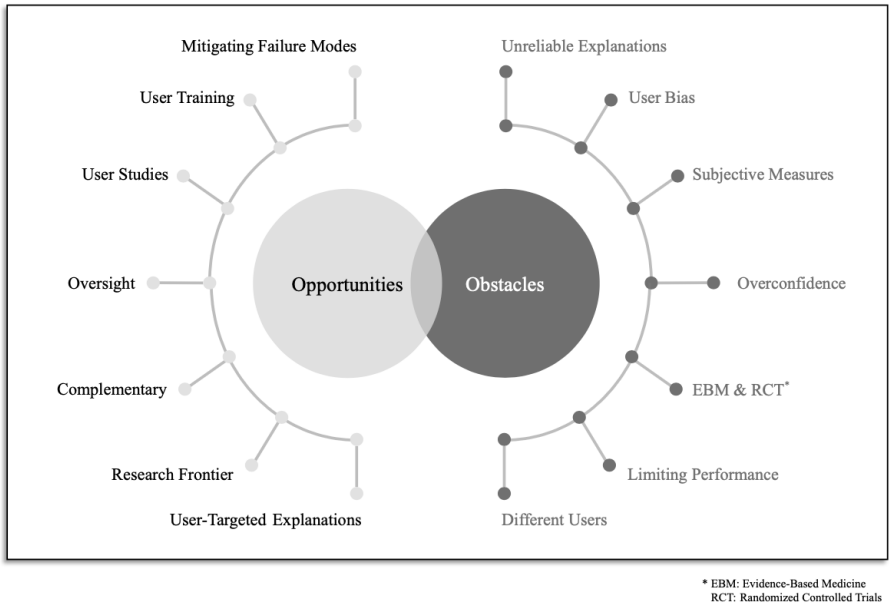

Figure 1: A visual summary of the criticisms and the related opportunities.

XAI techniques may be incorrect or unreliable.

Criticism. XAI techniques are sometimes faulty: they can be fooled with adversarial attacks or with noise.[18,19] Moreover, for the XAI techniques that approximate an original black-box model, there can be an additional discrepancy between the explanation and the real model behavior, resulting in unreliable explanations.[13]

Discussion. While these points are true, the mitigating factor here is that every technology has failure modes. As long as we have awareness and know how to counter undesired effects, such

techniques can still be useful. Furthermore, some such negative results can also be a pointer on how to improve XAI techniques; *e.g.* adversarial attacks on SHAP (one of the techniques offering individual-level explanations) suggest that out-of-distribution samples should not be employed in approximate SHAP calculations.

Human users may fall prey to biases when interacting with XAI tools.

Criticism. Users of XAI techniques struggle to understand what is a good explanation and often fall prey to biases — e.g. preferring explanations that confirm their own beliefs, the so-called confirmation bias — or misuse XAI tools.[20,21]

Discussion. Such remarks are on point and should be addressed when designing interfaces of AI tools and training programs, but they do not constitute a sufficient reason to give up on XAI. It is very well documented that many researchers misunderstand and misuse p-values[22], but this is not a good argument to abandon significance testing. A similar point could be made about conventional medical procedures, where different kinds of bias have been documented.[23] The best solution for such problems is to have the users properly trained; notably, when it comes to medical devices the training of users is quite well regulated; *e.g.* in Europe via the Medical Device Regulation and the ISO standards.[24]

There are no objective measures of explainability.

Criticism. There is no obvious ground truth for explainability, and therefore there is no straightforward way to assess how good an XAI technique really is.[14,15]

Discussion. This surely is a barrier, and more work is required to establish how the reliability of XAI tools can be tested; what is a good explanation, and what is useful to the user/developer/regulator are questions that have an intrinsic psychological nature and would benefit from more principled and widespread experimentation. Several of these concerns have been raised recently,[25] and substantial work has already been done in investigating what constitutes an appropriate and useful explanation.[18,26,27,28,29] These efforts include eliciting users' preferences with user studies, and deriving insights on the notion of explainability from the perspective of social sciences and related fields. Observing this need for an interdisciplinary structure, many scientists have recently participated in projects, workshops, discussion groups that involve colleagues working in diverse areas.

XAI may work on an aggregate level, but not on an individual level.

Criticism. Ghassemi and colleagues distinguish XAI techniques used in the specific (individual level) case and in the general case, finding that they can be of use in the latter but not in the former.[9] The reason is that users too often cannot decide if the explanation is sensible for an individual sample and have to base their judgment on an unwarranted intuition.

Discussion. While this risk is concrete, it is important to realize that this criticism mostly applies to certain kinds of data. In images, single input features (i.e. pixels) have no intrinsic meaning; only certain patterns of feature activation have meaning. When a certain area of an image is highlighted we simply do not know if what we recognize (*e.g.* the shape of a kidney) is the same as what the AI recognizes, because we have no access to the high-level representations of the machine. Hence, we cannot be sure whether the machine is recognizing the kidney as the salient part of the image at hand, or if there is some other spurious reason. There are however data types, like Electronic Health Records, where single input features have a very clear semantics. If, for instance, the patient's values of BMI and glucose are singled out as salient by an explainability technique, this may be enough for a clinical user to consider whether the AI's output is sensible.

Thus in some applications, having XAI developed on an individual level in a way that allows users to 'check' the reason(s) behind the prediction, can be a way to find and escalate possible problems of the AI at an aggregate level. Especially when the data and the AI environment are of high complexity, we need oversight mechanisms to continuously monitor and detect unforeseen failure modes, and individual users can play a role in such monitoring processes. Furthermore, the explanations of AI predictions will typically be one of the pieces of information that clinicians take into account. In current practice, for example, physicians obtain results of different tests, which at times may even be potentially conflicting, *e.g.* radiopaque lesion resembling infectious mass in a CT scan but lack of elevated infection markers. Such different factors are weighed in, each with its degree of uncertainty, to arrive at a decision. In this workflow, the input(s) of XAI techniques can be contrasted with other sources of information and be instrumental in i) achieving a deeper insight into the clinical case at hand and ii) decide whether to (not) trust the AI.

Evidence-based assessment is the proper way to evaluate medical interventions, not XAI.

Criticism. According to this criticism, randomized controlled trials (RCTs) and evidence-based assessment are better ways to evaluate medical interventions than XAI. The primary goal is to have interventions that work, even if we do not understand why. Black-box solutions are used in medicine already, so we do not need XAI as long as we have established that a tool or intervention is effective. The desire for explainability is secondary.[8]

Discussion. It is indisputable that XAI should not be conceived as a replacement of evidence-based instruments to validate model effectiveness. However, the criticism raises a false dichotomy: evidence-based evaluation tools and XAI can coexist, and we do not need to abandon one in favor of the other. We do not need to choose between a thorough evaluation and an explainable model: thorough evaluation should be a default (and it often is, given current regulations), and the need for explainability can be discussed separately.

Furthermore, framing the discourse about explainability as a binary classification between primary and secondary requirements is needlessly reductionist. Surely there can be trade-offs with accuracy that we must be cognizant of (more on this later), but the benefit of explainability may pay off in other areas, *e.g.* the discovery of biased outcomes for certain subgroups, or the disentangling of causal effects from spurious ones. As put forward by various governing bodies, explainability is

one aspect within a set of requirements that are partially interdependent and should be balanced against each other. The pursuance of explainability will have different priority depending on the application. An AI tool to speed up reconstruction of images in the back-end might not need explainability, while a decision support tool to predict risk of complications might need it if the risks are to be discussed and assessed with a patient. A blanket statement declaring explainability 'secondary' does not help in determining which case is which.

A final point should be made about validation of AI tools and evidence-based medicine (EBM). EBM can rely on a stable set of protocols that are established and widely used for the certification of new interventions. It should be noted however that, despite its indisputable importance, EBM is not a laser-precise discipline: e.g. there is a vast body of literature highlighting the fact that clinical trials are often not reflecting clinical practice, or have non-representative populations, or are incomparable with each other, or become quickly outdated.[30] Furthermore, as argued by several researchers,[31,32,30,33] RCTs do not always control for the biases they are intended to control, they do not produce reliably generalizable knowledge, or they can create unnecessary constraints on clinical testing. In addition, obtaining statistically significant evidence is often extremely time-consuming and expensive. Finally, the success rate of clinical trials for medications is famously low.[34] Against this backdrop, we should be very careful to promote the initiation of many studies or trials testing the effectiveness of tools whose inner workings are entirely unknown, i.e. black boxes. Explainability is a precious tool not just for the end user, but also for the developer: the vast majority of models are discarded before reaching the test phase because they are found to be clinically implausible thanks to explainability techniques. In this sense, while we do not dispute the high position of RCTs in the hierarchy of evidence,[35] explainability is a complementary tool that can even help expedite or improve the effectiveness of RCTs and EBM as a whole.

In the medical domain, we cannot trade performance for explainability.

Criticism. The trade-off between explainability and performance may induce people to use simpler and less accurate models, rather than more effective ones. This is a problem, as we may be exchanging a further improvement in medical outcomes with a debatable increase in transparency. There are several medical interventions (*e.g.* some drugs) for which a mechanistic explanation is not known, but they are nevertheless proven to be effective and widely used. If we are accepting of black-box interventions in healthcare, we can allow effective AI models to be opaque as well.[8]

Discussion. It is indeed the case that some explainability techniques (or simpler, more interpretable models) can result in lower performance. Whether this is desirable is a decision that should be taken case-by-case. Moreover, the analogy with medications, and the fact that we are 'fine' with them being black-boxes, seems misguided. There are already different tools in medical practice for which different levels of understanding are required. For example, as opposed to drugs, radiography machines can only be operated by specialists. Hence for complex tools we do require users to have some knowledge of the inner workings of the device. The reason underlying this choice is that there are failure modes we want to be able to control. Moreover, the fact that now we do not require explanations for some interventions (as long as they are proven to work) does not mean that this is the desirable way forward. On the contrary, in some

cases this could be construed as a problem. Arguably, we should keep pushing research to understand why certain treatments are effective. Along this line of reasoning, if we can build a tool with an intrinsic ability to explain itself – at a bearable cost in terms of trade-offs – we should definitely pursue it.

Finally, a 'simple' algorithm is not necessarily less accurate. When coupled with feature engineering or feature reduction, even those 'simple' algorithms may achieve good performances. This criticism thus overlooks an important line of work on inherently interpretable methods that are aiming at matching the performances of more complex models.[13,7] Figure 2 showcases a simple model which is easy to understand and achieves high accuracy.

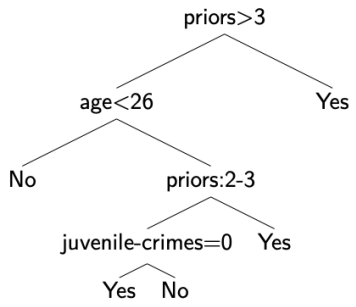

Figure 2: Example of an accurate white-box algorithm for the COMPAS recidivism problem (Figure 4 in[36]).

XAI is not suitable for all users.

Criticism. XAI techniques can be used to support the work of different professionals interacting with an AI tool. Beside regular users, other roles include the developers, the medical experts, the regulators, and so on, each with different requirements and needs for explainability. Arguably, XAI techniques could be useful for trained professionals such as developers or auditors, but could be dangerous when shown to untrained users.

Discussion. This point has some overlap with points 2 and 4, but it is nonetheless important to stress that XAI techniques will almost never offer a one-size-fits-all solution. Different kinds of interaction with the machine will be required depending on the purpose of the participating human; it is highly likely that different techniques will answer different needs.[37] Current research extends to interactive machine learning to provide user-targeted explanations.[38] As to the potential misuse or misunderstanding of XAI by simple users, the solution lies in a combination of robust methodology, proper training, and expectation management.

3 Should we downsize our efforts in XAI for healthcare?

We have reviewed some of the core criticisms offered in the literature against the widespread use of XAI techniques in medical contexts. Some of the qualms, as well as some of the counterpoints, are in fact not limited to healthcare applications and can be recast at a more general level. We should however be mindful of how the debate is influenced by the higher stakes present in safety-critical AI applications in the health realm. It is worth remarking that demoting the usefulness of explainable medical AI means stifling the efforts to get human oversight over a new technology that is still in its infancy but can be deployed at scale very rapidly, potentially affecting a large swath of the population. XAI is not a silver bullet that will solve all of AI's problems, and in particular it is not a substitute for rigorous evaluation of model performance. Nonetheless, XAI tools can be of value in several circumstances. Human oversight over machines is still an important tenet, and the introduction of more black-boxes is not going to stem the loss of human control in a quickly digitizing world. It has been argued that augmentation of human capacities is preferable to automation,[39] since the latter leads to deskilling as well as detachment and unemployment spikes. Augmentation presupposes the possibility of a human-machine synergy; in other words, a way for humans to relate and engage with AI.

Different public bodies, including the European Union, are discussing regulations for AI products. While the field of XAI may not yet be at a level of maturity where XAI techniques are embedded into guidelines in the form of hard requirements, this does not entail that XAI techniques should be sidelined or disregarded. Said techniques can be employed to exercise an additional layer of control over AI products; as with all tools, practitioners should be aware of the potential failure modes. As for guidelines, it could be meaningful to allow the use of XAI techniques – without enforcing a specific choice – and require an argument of why a certain technique was chosen.

We noted in the introduction the inherent tension in the request for explainability: while we want machines to perform at super-human level, we also want them to be understandable by humans. This is however just another case of competing objectives, not a good reason to toss one of the objectives out the window (for another example of competing objectives: saving healthcare costs and keeping people alive). The challenge is to find a trade-off that is satisfactory, a challenge that is extremely case-dependent. The question should be: while explainability may not be a requirement for healthcare interventions tout court, is AI a technology on which we think "extra" explainability is needed? This technology has peculiar aspects that set it apart, including the potential to be prejudiced against subgroups based on gender, ethnic background, and socioeconomic status,[40] the fact that it may lose reliability when the input data changes,[41] or that it may rely on spurious correlations found in the data.[42] For these reasons, coupled with the potential of rapid and impactful adoption, we argue that a higher level of oversight is preferable compared to regular interventions.

4 Conclusion

Despite its infancy, the research on XAI is growing very rapidly. Thanks to this growth, new techniques are tested extensively by many researchers in the field, and the duration to become

a time-tested approach is much shorter than it has been in the past. Given the arguments we discussed, and the aforementioned possibility to rapidly hone useful techniques, we conclude that we should step up our efforts in XAI research for healthcare, as this research program is promising and addresses a core issue. Ultimately, we still want clinicians to be the mediators that can bridge the coldness of algorithms and scientific protocols with the human experience of the patient. If clinical users of medical AI cannot interact with machines, if their role becomes that of mere executors of interventions that are proven to be effective, the sensitive task of providing care could well become de-humanizing.

References

[1] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv*, 2017.

[2] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable AI: a review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2020.

[3] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: an overview of interpretability of machine learning," *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, vol. 00, pp. 80–89, 2018.

[4] O. Biran and C. Cotton, "Explanation and justification in machine learning: a survey," *IJCAI-17 Workshop On Explainable AI (XAI)*, vol. 8, no. 1, pp. 8—13, 2017.

[5] D. Doran, S. Schulz, and T. R. Besold, "What does explainable AI really mean? A new conceptualization of perspectives," *arXiv*, 2017.

[6] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[7] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[8] L. G. McCoy, C. T. Brenna, S. Chen, K. Vold, and S. Das, "Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based," *Journal of Clinical Epidemiology*, vol. 142, pp. 252–257, 2021.

[9] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, 2021.

[10] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: a survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019.

[11] C. Molnar, *Interpretable machine learning: a guide for making black box models explainable*. Leanpub, second ed., 2022.

[12] G. Ras, N. Xie, M. V. Gerven, and D. Doran, "Explainable deep learning: a field guide for the uninitiated," *Journal of Artificial Intelligence Research*, vol. 73, pp. 329–397, 2022.

[13] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[14] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: a survey on methods and metrics," *Electronics*, vol. 10, no. 5, p. 593, 2021.

[15] F. Yang, M. Du, and X. Hu, "Evaluating explanation without ground truth in interpretable machine learning," *arXiv*, 2019.

[16] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.

[17] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Communications of the ACM*, vol. 62, no. 6, pp. 70–79, 2019.

[18] D. Slack, S. A. Friedler, C. Scheidegger, and C. D. Roy, "Assessing the local interpretability of machine learning models," *arXiv*, 2019.

[19] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," in *Advances in Neural Information Processing Systems*, vol. 32, Curran Associates, Inc., 2019.

[20] Kaur, Nori, Harmanpreet and, Jenkins, Harsha and, Caruana, Samuel and, Wallach, Rich and, and Vaughan, Hanna and Wortman, "Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning," *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2020.

[21] A. Rosenfeld, "Better metrics for evaluating explainable artificial intelligence," *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*, no. 299, pp. 45—50, 2021.

[22] R. L. Wasserstein and N. A. Lazar, "The ASA's statement on p-values: context, process, and purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016.

[23] R. C. Schwartz and D. M. Blankenship, "Racial disparities in psychotic disorder diagnosis: a review of empirical literature," *World Journal of Psychiatry*, vol. 4, no. 4, pp. 133–40, 2014.

[24] T. E. U. M. D. Regulation, "Regulation (EU) 2017/745 (EU MDR)," 2022.

[25] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: fundamental principles and 10 grand challenges," *Statistic Surveys*, vol. 16, pp. 1–85, 2022.

[26] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen, and B. Baesens, "An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models," *Decision Support Systems*, vol. 51, no. 1, pp. 141–154, 2011.

[27] T. Miller, "Explanation in artificial intelligence: insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2018.

[28] A. Jacovi, J. Bastings, S. Gehrmann, Y. Goldberg, and K. Filippova, "Diagnosing AI explanation methods with folk concepts of behavior," *arXiv*, 2022.

[29] K. Krawiec, M. Virgolin, A. D. Lorenzo, F. Randone, E. Medvet, and M. Wahde, "Model learning with personalized interpretability estimation (ML-PIE)," *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1355–1364, 2021.

[30] M. Solomon, "Just a paradigm: evidence-based medicine in epistemological context," *European Journal for Philosophy of Science*, vol. 1, no. 3, p. 451, 2011.

[31] N. Cartwright, "Presidential address: will this policy work for you? Predicting effectiveness better: how philosophy helps," *Philosophy of Science*, vol. 79, no. 5, pp. 973–989, 2012.

[32] T. Greenhalgh, J. Howick, N. Maskrey, and E. B. M. R. Group, "Evidence based medicine: a movement in crisis?," *BMJ : British Medical Journal*, vol. 348, no. jun13 4, p. g3725, 2014.

[33] J. Worrall, "Why there's no cause to randomize," *The British Journal for the Philosophy of Science*, vol. 58, no. 3, pp. 451–488, 2007.

[34] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics (Oxford, England)*, vol. 20, no. 2, pp. 273–286, 2019.

[35] B. Djulbegovic and G. H. Guyatt, "Progress in evidence-based medicine: a quarter century on," *The Lancet*, vol. 390, no. 10092, pp. 415–423, 2017.

[36] X. Hu, C. Rudin, and M. Seltzer, "Optimal sparse decision trees," *arXiv*, 2019.

[37] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[38] K. Sokol and P. Flach, "One explanation does not fit all," *KI - Künstliche Intelligenz*, vol. 34, no. 2, pp. 235–250, 2020.

[39] D. Acemoglu and P. Restrepo, "The economics of artificial intelligence: an agenda," in *Artificial Intelligence, Automation, and Work* (A. Agrawal, J. Gans, and A. Goldfarb, eds.), p. 197 – 236, University of Chicago Press, 2017.

[40] C. Fitzgerald and S. Hurst, "Implicit bias in healthcare professionals: a systematic review," *BMC Medical Ethics*, vol. 18, no. 1, p. 19, 2017.

[41] D. Ulmer, L. Meijerink, and G. Cinà, "Trust issues: uncertainty estimation does not enable reliable OOD detection on medical tabular data," in *Proceedings of the Machine Learning for Health NeurIPS Workshop*, vol. 136 of *Proceedings of Machine Learning Research*, pp. 341—354, PMLR, 2020.

[42] C. S. Calude and G. Longo, "The deluge of spurious correlations in big data," *Foundations of Science*, vol. 22, no. 3, pp. 595–612, 2017.