On the risk of confusing interpretability with explicability

Christian Herzog¹

Received: 13 July 2021 / Accepted: 24 November 2021 / Published online: 9 December 2021
© The Author(s) 2021

Abstract

This Comment explores the implications of a lack of tools that facilitate an explicable utilization of epistemologically richer, but also more involved white-box approaches in AI. In contrast, advances in explainable artificial intelligence for black-box approaches have led to the availability of semi-standardized and attractive toolchains that offer a seemingly competitive edge over inherently interpretable white-box models in terms of intelligibility towards users. Consequently, there is a need for research on efficient tools for rendering interpretable white-box approaches in AI explicable to facilitate responsible use.

Keywords Explainability · Artificial intelligence · Interpretability · Explicability · Ethics

1 Introduction

Calls for rendering artificial intelligence (AI) explainable have surfaced in many proposals for AI ethics guidelines, oftentimes even elevating ‘explicability’ to an ethical principle [10]. While the status as an ethical principle may be debatable [28], explicability is usually considered paramount for responsible AI. According to Floridi et al. [8], *explicability* is richer notion that *explainability*, as it is combining demands for intelligibility and accountability as a way to enabling both those working with and those affected by AI systems to understand and challenge ways of interaction and outcomes [24]. In other terms, demands for explainability—referring to means for providing some form of explanations to users—are usually accompanied by demands for transparency—referring to the communication of (ideally) all aspects necessary to scrutinize or challenge an AI during its development, implementation, deployment and use [7, 54]. In essence, this Comment proposes to take explicability as the combination of both intelligibility and accountability seriously also when basing AI on inherently interpretable models.

In most ethics guidelines, definitions of AI are usually broadly construed, such that they also encompass white-box model approaches, such as expert or fuzzy systems, or

probabilistic (graphical) models, to name a few. Depending on their complexity and from an applied perspective, these models may be just as opaque as black-box model approaches [5]. Accordingly, calls for making explanations of AI systems humanly understandable have recently increased. From proposals to consider the nature of human-to-human explanations for effective and communicative human–machine interfaces [20, 21] to adopting a human-centric perspective on explanations, in general [6], researchers have addressed a move beyond explanations merely as a tool for debugging and gaining new scientific insights. Hence, attempts at defining ‘interpretability’, e.g., as “the degree to which a human can understand the cause of a decision” [20], are helpful, but the goal of this Comment is to rather emphasize explicability by asking ‘which human?’, ‘under which circumstances?’ and ‘for which purposes?’

2 Explicability as an ethical principle

Explicability has first been advocated as an ethical principle for AI by Floridi et al. [8], where it was essentially added to the canon of biomedical ethical principles developed by Beauchamp and Childress [4]. The principle has since been taken up by the final version of the European Commission’s Independent High-Level Expert Group on Artificial Intelligence’s Ethics Guidelines on Trustworthy AI [12]. In Floridi et al. [8] the principle of explicability is introduced as enabling the other principles of ‘beneficence’, ‘non-maleficence’, ‘autonomy’ and ‘justice’, rather than being

✉ Christian Herzog
christian.herzog@uni-luebeck.de

¹ Ethical Innovation Hub, University of Lübeck, Ratzeburger Allee 160, 23562 Lübeck, Germany

a primary principle. The authors attempt to synthesize the different but related notions of ‘transparency’, ‘understandability’, ‘interpretability’ and ‘accountability’ by explaining the principle of ‘explicability’ in both epistemological and ethical terms.

Epistemologically, Floridi et al. intend ‘explicability’ to encompass the notion of ‘intelligibility’, meaning that for intelligible AI it is in some sense possible to give an answer to the question of “how does it work?”. However, Floridi et al. stop short of making precise demands regarding the extent of explanations of the inner workings of an AI algorithm. Demands for intelligibility may well be context-specific in the sense that, e.g., in medicine fully mechanistic explanations may not be available [17], while practitioners and patients alike might be satisfied with invoking correlative evidence as the foundation for responsible shared decision making. In addition, the question of “how does it work?” may also not relate to concrete evidence leading to specific decisions (or rather suggestions), but rather to the explanation of fundamental procedural mechanisms employed and conditions relevant to arrive at them. This may entail providing information about the data bases and its provenance as well as interpretive processes [22] and algorithmic processing involved to reach an output. Again, what level of depth is satisfactory in providing details about the conditions of algorithmic (or algorithm-supported) decision making may depend very much on the context and recipient. Explaining the essential inner workings of an AI algorithm to a physician may require entirely different expositions than explaining this to a patient, or an AI system developer. Again, Floridi et al. are stopping short of demanding that an AI algorithm may be explainable in every single detail. Instead, in stating that “we must negotiate the terms of the relationship between ourselves and this transformative technology, on grounds that are readily understandable to the proverbial person ‘on the street’” Floridi et al. [8] rather propose to demand the provision of the right level of explanation, such that any particular provider and user of—or person affected by—AI technology is able and willing to take responsibility for effects resulting from an AI system’s use, or contest the associated decisions, respectively.

In ethical terms, then, Floridi et al. intend ‘explicability’ to encompass, or rather, enable ‘accountability’. Here, I suggest taking their formulation of the underlying question on “who is responsible for the way it works?” to consider the wider socio-technical system of which the ethical and—in terms of the European Commission’s guidelines—trustworthy AI system [27] is a part. This entails that we should not only be interested in the developing party as the responsible entity, but also in the commissioning, deploying and, ultimately, the utilizing parties. This, of course, still makes accountable use reliant on some form of explanation of how the AI system itself works, but it also allows a wider

recognition of, e.g., accompanying efforts that contribute to helping less AI tech-savvy users make responsible use of AI. For instance, consider physicians who are provided with a medical decision-support tool. Such a tool may be designed as, say, an expert system. It would, thus, not constitute a black box system in the sense that the algorithm is opaque even to the developers, cf. Herzog [11]. However, it may still be designed without explanatory interfaces that facilitate a flow of information from AI system to physician and on towards the patient for enabling informed consent and, ultimately, shared decision-making. On the other hand, such a system may well be designed to allow for this kind of transparent and user/recipient-centered handling of explanations, but it is employed in and may even contribute to workflow settings that favors rationalized medical treatment over actual care. I suggest that the notion of ethical or trustworthy AI should consider the entire relevant socio-technical system, instead of focusing on the AI system as an artefact alone. In this sense, Floridi et al.’s principle of ‘explicability’ is different from demanding that AI systems be fully explainable and, hence, non-opaque, but rather requires its accountable use, where actors are both sufficiently able and comfortable to take responsibility [13].

3 Arguments against ‘explicability’ as an ethical principle

Robbins [28] argues against letting ‘explicability’ constitute an ethical principle. To be fair, he does not dismiss explicability as useless, but rather makes the points that (i) explicability should attach to a concrete action rather than the entity responsible for it, (ii) the need for explicability is context-dependent, and (iii) explicability should not restrict AI to offer decision support for applications that only incorporate reasoning whose considerations are acceptable a priori. In the latter case, Robbins argues, standard automation would be sufficient. I will briefly attempt to refute these arguments against granting ‘explicability’ the status of an ethical principle, before moving on to the core perspective of this Comment: That there is a lack of effort to achieve explicability in the sense proposed by Floridi et al., which is especially profound when considering inherently interpretable models, as they seem to be perceived to not be in need of explanatory interfaces that support context- and recipient-specific intelligibility and accountability.

As with respect to (i) and (ii) Robbins elaborates that there are certain low-risk AI decisions that do not require explanations, because it is very likely that no harm can be done. Robbins provides weather forecasts as an example. Notwithstanding that mistaken weather forecasts can lead individuals into hazardous and life-threatening situations, there is undoubtedly a steep risk-gradient between medical

decision support and weather predictions. However, just as there may be AI whose risk is negligible, there is also AI whose contribution to beneficence may be highly insignificant. Does this deprive ‘beneficence’ of the status as an ethical principle? Perhaps, but then we would not be left with much of any principles at all. Rather, I suggest that all ethical principles can be relevant to either action/decision or decision maker (or the corresponding device) to variable degrees. Consequently, Robbins’ argument that because for some decisions we may demand explanations, while for others, we may need none, is not conclusive of the view that the notion of ‘explicability’ should necessarily be attached to particular decisions. An entity capable of making decisions can be required to explain in general, even though in some instances decisions may be quite inconsequential.

With respect to (iii), it appears that it follows that Robbins’ definition of AI must necessarily be a quite narrow one, perhaps even a moving target. The line between standard automation and AI is not clear and expert systems, whose underlying rationale and considerations are clearly known from the outset, would—in Robbins’ view—not constitute AI. This certainly is not the stance adopted the EU’s novel proposal for AI regulation, which defines AI systems as systems that “can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with” [26]. Clearly, this does not preclude expert systems. Furthermore, AI systems that offer decision support with underlying considerations that are not immediately evident to either developers, users or recipients, might simply require adequate time frames allowing scrutiny of said decisions. Robbins employs the example of possibly discriminatory or otherwise unacceptable explanations for accepting or rejecting loans and argues that we could as well hard-code the non-discriminatory and fair considerations for granting a loan into a system, which would then constitute standard automation. In this particular example, I completely tend to agree, even though the EU definition would potentially still allow to denote this an AI system. It can be argued that in matters of fairness—and possibly all matters relating to societal consensus on moral values—the decision principles should be almost always set out a priori, instead of being learned by a system.

However, e.g., in other contexts such as medicine, it is conceivable that the necessary considerations for the respective decision principles can reach levels of complexity that are difficult to handle and program into a system manually. For instance, consider the prospect of automated and highly current meta-analyses [19] that can support medical decision making. It may well be that the considerations, i.e., a set of steps that would guide expert analysis and could be hard-coded, could themselves constitute useful AI system output. In turn, this might be new input to the physicians.

Given the necessary resources, these would need to scrutinize the medical considerations and learn from them. Robbins anticipates this objection, only to quickly dismiss it by pointing out that newly found considerations could simply be hard-coded again into new automation algorithms (which he does not consider AI). This is a good point but leaves the benefit of the AI system revealing new correlations or reasons untouched. In light of advances in precision and individualized medicine, such AI systems may constitute important ingredients to a continuous feedback between research and practice in the vein of a learning health care system [23, 25]. Explicability plays a key role of facilitating this way towards scientific progress and epistemic gain.

Consider Fig. 1 as a simple illustration of how a medical expert system (potentially qualifying as AI in terms of the EU draft artificial intelligence act) may benefit from a, say, sub-symbolic AI meta-analysis system that identifies updates to the set of medical considerations which constitute the medical decision tree. Even though the meta-analysis system may be fed hard-coded criteria to ensure the completeness and fairness in its assessment of the medical literature, its output cannot possibly be known a priori. However, its output—consisting in updates to the set of medical considerations relevant to the decision tree—itself may constitute easy to understand criteria. As a whole, a medical decision support system designed this way would not operate on *medical* considerations that are deemed acceptable a priori. Rather both the changing set of medical considerations and criteria for the quality of the research and evidence considered together amount to items that contribute to a more complete explanation of a medical decision. This should not

Fig. 1 Schematic of a possible AI-supported medical expert system

be taken as a proposal for a blueprint on how to construct medical AI. Rather, this example should exemplify how AI system architectures may provide ingredients for catering to a justified demand for explicability, while not excluding subsystems whose reasoning is completely known a priori and could potentially be hard-coded. In the case of medical AI, updates from meta-analyses would certainly be running on a much slower time scale than the actual decision tree-based medical decision support to allow for regulatory and factual-medical checks. In that sense, the medical decision support might be running on considerations that could potentially be hard-coded a priori just as well. However, if the same architecture would be employed in a field other than medicine, both the subsystems' time scales may be much closer, while the entire system would still produce potentially intelligible explanations. There is no reason, why in less consequential fields than medicine such as, e.g., business administration, explanations may well be contested and scrutinized a posteriori, i.e., as long as the possibility of proper redress is not endangered.

4 The risk of confusing interpretability with explicability

Above, I have defended 'explicability's' status as an ethical principle and have outlined the depth of its implications for conceiving of intelligibility and accountability as important notions for a responsible and trustworthy use of AI in practice. In summary, my view of 'explicability' is that it means both more and less than explainability understood in a sense that considers only mechanistic explanations, and hence, 'interpretable models' as good foundations for responsible use. 'Explicability' means more, because in my view 'explicability' demands explanatory interfaces tailored to the recipient and use-case, incorporating a wide array of possible forms of explanations that may not be complete, but that focus on putting the respective stakeholders in a position to take responsibility. On the other hand, 'explicability' also means less—or rather is less precise—, because the principle remains flexible enough to not strictly demand mechanistic explanations at every level of usage and, perhaps, may even allow for none if not available and required for serving the other principles of 'beneficence', 'non-maleficence' and 'autonomy'. Granted, quite often, we will want to enact regulatory requirements on AI that demands full transparency and interpretability. However, my main point is that we should not mistake calls for transparency and interpretability as sufficient for achieving explicability in the sense that guarantees actual accountable use in practical situations under harsh time constraints and potentially wide ranges of technological proficiency (i.e., a lack thereof).

Hence, I will now turn to the Comment's main conjecture: There is a lack of effort to achieve true explicability in cases where models used in AI are inherently interpretable. Meanwhile, black-box approaches have received more attention from the explainable AI community with efforts often directed at providing intelligible algorithmic outputs for end-users—albeit with varying degrees of success [1]. However, systems based on interpretable models are not necessarily intelligible to other stakeholders than the developers, hindering accountability of the system users when explanation interfaces are lacking. In turn, providing domain-specific intelligible explanations for black-box systems does not necessarily achieve accountability.

As an example for the latter, consider saliency maps, see, e.g., (Simonyan et al. [30]), as a means to generate maps representing the importance of each pixel in arriving at an image classification. This approach to make sense of black-box AI-based image classification is notorious for being unreliable [2] and, hence, may undermine accountable use despite the deceptively intuitive appearance of its outputs even if applied to highly unstructured data such as medical imaging.

On the other hand, generally interpretable models, such as simple regression approaches yielding parameterized equations, or graphical models visualizing equations or statistical independence statements do not provide added value in terms of explicability when deployed without further means to transport the in-situ reasoning behind the interpretable model.

Hence, let the notion of misconstruing interpretability as implying explicability be denoted the 'interpretability-explicability confusion'. There is ample indication that confusion of interpretability and explicability is rather widespread. Take, for instance, a very recent study on 'Explainable AI' commissioned by the German Ministry for Economic Affairs and Energy [14]. Prominently positioned within its executive summary, a process diagram is intended to provide orientation on the most common tools for explainable AI. However, according to this diagram, no means of explanation to domain experts (or beyond) are suggested if a white-box model can be used without sacrificing performance. It should be clear, however, that some white-box models are neither inherently interpretable by all domain experts, nor do outputs of AI algorithms based on them typically warrant their responsible use. Rather, there might be a need for interfaces that support domain experts in the communication of algorithm outputs towards, e.g., patients (such as in healthcare).

Granted, it may be significantly easier to derive means of explanations from interpretable models in specific cases. This might explain a certain underappreciation of the task in basic research and more general tool development. However, it appears as though there is a general tendency to equate

interpretable models with explicable models, even among highly skilled data scientists. Consider, e.g., Rudin [29], who does not distinguish between the functions of an explanation to provide either scientific insight or allow for its responsible use—in the terminology of this Comment, interpretable or explicable algorithms. Quite often, both these notions may, in fact, significantly overlap. However, they are not the same, especially when algorithmic outputs need to be made sufficiently plausible (or rather need to be augmented to reveal potential implausibility) for users to challenge them in situations where, e.g., time is of the essence. Potential application domains, where this may be so, include healthcare, autonomous driving and many others.

Admittedly, Rudin is making the compelling claim that more often than not, non-interpretable models bring about more risks than benefits. More specifically, she strives to invalidate the myth of the accuracy-interpretability trade-off. However, it can still be maintained that interpretable, yet—due to their complexity—opaque models may not receive the same attention as black-box approaches to facilitate transparency by the explainable AI research community. This is also supported by recent surveys on explainable AI that categorize published literature into *post hoc* approaches that create interpretable models, explain black-box deep learning models, or explain black-box models in a model-agnostic way [15]. Comparably much fewer approaches try to render an AI algorithm explainable *ante-hoc* [31]. While, of course, many model-agnostic approaches can also be used on complex white-box model-based AI, there seems to be no significant strand of model-specific explainable AI research targeting algorithmic explicability of interpretable models directly.

Consider, for instance, large-scale probabilistic models for statistical inference. Frameworks for constructing these are often already considered with expert interpretability in mind. Factor graphs [16], for instance, allow for the hierarchical and graphical construction of large models, mix-and-match approaches to synthesize algorithms from different parts and—given proper toolboxes—methods for debugging and analysis by zooming in on parts of the graphical representation. Such approaches are not only used for signal processing, but are applied in areas like neuroscience, attempting to uncover some basic mechanisms of active inference [9]. However transparent such an approach may be to the expert, even domain experts such as physicians, who accept their need for statistical inference results to improve healthcare, may be quickly left with decision support systems, whose outputs they cannot scrutinize and challenge without further development efforts for transparent interfaces.

Simply branding approaches based on interpretable, yet highly complex, models as ‘explainable AI’ is not supportive of responsible use of this otherwise epistemologically advantageous approach. In contrast, rather obviously opaque

algorithms, such as deep learning, may receive more attention towards their transparency to facilitate their responsible use, merely, because they are not interpretable to begin with. Perhaps the lack of epistemic certainty during the development leads to a kind of humility towards the different needs of all stakeholders, while teams working with interpretable models consider themselves certain to produce correct results.

Therefore, this Comment contains the proposition to the community to not allow the epistemological benefits of white-box models remain unharnessed simply due to a lack of consideration for making the leap from an interpretable to an explicable algorithm. The specific form that this leap should take is, obviously, contextual, i.e., much depends on the actual application under consideration.

5 Conclusion and open questions

There is ample research on making so-called black-box AI approaches explainable for AI-, domain- and non-experts. While any work on tackling algorithmic opacity is commendable, a preoccupation with augmenting black-box AI approaches for responsible use may have led to a neglect for the fact that complex white-box models are also opaque to most domain- and non-experts. Just because white-box models can be interpreted by expert developers, they do not automatically yield solutions that facilitate responsible and accountable use in practice. Notwithstanding that—or perhaps, because—often white-box models hold a clear epistemological advantage over black-box models, tools to provide explanations of their output to domain- and non-experts have not nearly received the same attention as black-box approaches. Therefore, researchers should carry over the significant impetus that the discussion about ethical and trustworthy AI has brought to the field of explainable AI to encompass opaque algorithms in general by also including interpretable, yet highly complex, models. The epistemological headway of white-box approaches should not pale due to a lack of effort for making them amenable for responsible use in practice.

A confusion between explicability and interpretability carries this risk. An important contribution to this discussion is also due to Babic et al. [3], who advocate caution towards *post hoc* methods for making AI explainable. This caution is well warranted for, as it resonates with an agenda similar to this Comment’s. However, by stressing the fact that explainability does not automatically incur accountability, Babic et al. fail to see that interpretability does not, either. This is why we need to stress a demand for explicability instead, to strike the balance between performance and intelligibility of any AI-powered device without sacrificing accountability.

Black-box AI models definitely deserve their fair share of attention. They often promise a fast development cycle from data to model and, perhaps, even straight into application. An ever-growing set of tools and tutorials facilitates this. Likewise, tools for rendering black-box approaches explainable are widely available, at many times skipping model interpretability and cutting directly to outputs intended for domain experts or end-users. There is a sense that this perceived ease of employing such a readily available toolchain makes it even more attractive to skip the trouble of more involved data analytics, acquiring model domain knowledge, devising algorithms and custom-tailoring transparency-supportive means for explainability. While there certainly is value in cutting short the development cycle for bringing effective tools to the market (or point of care) more quickly, it strikes as odd that the presumably epistemologically more rewarding way of going for interpretable white-box models instead would remain underutilized due to a lack of tools that support their ethical use by means of transparent and understandable interfaces.

The challenges may potentially not even simply consist in a lack of research results. This Comment is almost certainly not doing proper justice to some prolific research groups that combine both interpretable approaches and methods for accountable communication of algorithm outputs to the user. However, in any case, an increase in visibility and availability of both interpretable and transparent AI methods for applied research appears necessary.

In addition, the malleability of AI methods making them often directly applicable to a wide range of domains seems to call for increasingly multi-disciplinary and diverse teams. Engaging in meaningful stakeholder involvement and considering societal and ethical implications from the very beginning of a development process [18] is likely to reveal a much broader spectrum of what it takes to deliver explicable algorithms in practice. This applies to both black- as well as white-box approaches.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Alufaisan, Y., Marusich, L.R., Bakdash, J.Z., Zhou, Y., Kantarcioğlu, M.: Does explainable artificial intelligence improve human decision-making? ArXiv. (2020). <https://doi.org/10.31234/osf.io/d4r9t>
2. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., et al.: Assessing the (Un)Trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiol. Imaging* (2020). <https://doi.org/10.1101/2020.07.28.20163899>
3. Babic, B., Gerke, S., Evgeniou, T., Glenn Cohen, I.: Beware explanations from AI in health care. *Science* **373**(6552), 284–286 (2021). <https://doi.org/10.1126/science.abg1834>
4. Beauchamp, T.L., Childress, J.F.: Principles of biomedical ethics. Oxford University Press (2012)
5. Burrell, J.: How the machine ‘Thinks’: understanding opacity in machine learning algorithms. *Big Data Soc.* **3**(1), 1–12 (2016). <https://doi.org/10.2139/ssrn.2660674>
6. Confalonieri, R., Coba, L., Wagner, B., Besold, T.R.: A historical perspective of explainable artificial intelligence. *Wiley Interdiscipl. Rev. Data Min. Knowl. Discov.* **11**(1), 1–21 (2021). <https://doi.org/10.1002/widm.1391>
7. Dignum, V.: Responsible Artificial Intelligence. Edited by Barry O’Sullivan and Michael Woolridge. Springer (2019). <https://doi.org/10.1007/978-3-030-30371-6>
8. Floridi, L., Cows, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., et al.: AI4People—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach.* **28**, 689–707 (2018). <https://doi.org/10.1007/s11023-018-9482-5>
9. Friston, K.J., Parr, T., De Vries, B.: The graphical brain: belief propagation and active inference. *Netw. Neurosci.* **1**(4), 381–414 (2017). https://doi.org/10.1162/netn_a_00083
10. Hagendorff, T.: The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* (2020). <https://doi.org/10.1007/s11023-020-09517-8>
11. Herzog, C. Technological opacity of machine learning in health-care. In: 2nd Weizenbaum Conference: Challenges of Digital Inequality—Digital Education, Digital Work, Digital Life. Berlin, Germany. <https://doi.org/10.34669/wi.cp/2.7> (2019)
12. Independent High-Level Expert Group on Artificial Intelligence Set Up By the European Commission. 2019. “Ethics Guidelines for Trustworthy AI.” <https://ec.europa.eu/futurium/en/ai-alliance-consultation>. Accessed 5 July 2019
13. Kiener, M.: Can ‘Taking Responsibility’ as a normative power close AI’s responsibility gap?” In: CEPE/IACAP Joint Conference 2021: The Philosophy and Ethics of Artificial Intelligence. Hamburg, Germany (2021)
14. Kraus, T., Ganschow, L., Eisenträger, M., Wischmann, S.: Erklärbare KI: VDI/VDE Innovation + Technik GmbH. Berlin (2021)
15. Linardatos, P., Papastefanopoulos, V., Kotsiantis, S.: Explainable AI: a review of machine learning interpretability methods. *Entropy* **23**(1), 1–45 (2021). <https://doi.org/10.3390/e23010018>
16. Loeliger, H.-A.: An introduction to factor graphs. *IEEE Signal Process. Mag.* **21**(1), 28–41 (2004)
17. London, A.J.: Artificial intelligence and black-box medical decisions: accuracy versus explainability. *Hastings Cent. Rep.* **49**(1), 15–21 (2019). <https://doi.org/10.1002/hast.973>
18. McLennan, S., Fiske, A., Celi, L.A., Müller, R., Harder, J., Ritt, K., Haddadin, S., Buyx, A.: An embedded ethics approach for AI development. *Nat. Mach. Intell.* **2**(9), 488–490 (2020). <https://doi.org/10.1038/s42256-020-0214-1>
19. Michelson, M., Chow, T., Martin, N.A., Ross, M., Ying, A.T.Q., Minton, S.: Artificial intelligence for rapid meta-analysis: case

- study on ocular toxicity of hydroxychloroquine. *J. Med. Internet Res.* **22**(8), e20007 (2020). <https://doi.org/10.2196/20007>
20. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* **267**, 1–38 (2019). <https://doi.org/10.1016/j.artint.2018.07.007>
  21. Mittelstadt, B.D., Floridi, L.: The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* (2015). <https://doi.org/10.1007/s11948-015-9652-2>
  22. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in AI. In: *FAT* 2019—Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 279–88. Association for Computing Machinery, Inc. <https://doi.org/10.1145/3287560.3287574> (2019)
  23. Naylor, C.D.: On the prospects for a (Deep) learning health care system. *JAMA* **320**(11), 1099–1100 (2018). <https://doi.org/10.1001/jama.2018.11103>
  24. OECD: Recommendation of the Council on Artificial Intelligence. OECD (2019)
  25. Olsen, L.A., Aisner, D., McGinnis, M.J.: The learning healthcare system. In: *IOM Roundtable on Evidence-Based Medicine—Workshop Summary*. Washington, D.C.: National Academies Press. <https://doi.org/10.17226/11903> (2007)
  26. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts. 2021. COM/2021/206 Final.
  27. Rieder, G., Simon, J., Wong, P.-H.: Mapping the stony road toward trustworthy AI: expectations, problems, conundrums. *SSRN Electron. J.* (2020). <https://doi.org/10.2139/ssrn.3717451>
  28. Robbins, S.: A misdirected principle with a catch: explicability for AI. *Mind. Mach.* **29**(4), 495–514 (2019). <https://doi.org/10.1007/s11023-019-09509-3>
  29. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**(5), 206–215 (2019). <https://doi.org/10.1038/s42256-019-0048-x>
  30. Simonyan, K., Vedaldi, A., Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv arXiv:1312.6034 [Cs]* (2014)
  31. Vilone, G., Luca, L.: Explainable artificial intelligence: a systematic review. *arXiv:2006.00093* (2020)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.