Review

Explainable Artificial Intelligence (XAI) in Healthcare

Tim Hulsen *

Posted Date: 7 March 2023

doi: 10.20944/preprints202303.0116.v1

Keywords: XAI; AI; artificial intelligence; explainable; explainability; machine learning; deep learning; data science; big data; healthcare; medicine

Review

Explainable Artificial Intelligence (XAI) in Healthcare

Tim Hulsen

Department of Hospital Services & Informatics, Philips Research, Eindhoven, the Netherlands; tim.hulsen@philips.com

Abstract: Artificial Intelligence (AI) describes computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and language translation. Examples of AI techniques are machine learning, neural networks and deep learning. AI can be applied in many different areas, such as econometrics, biometry, e-commerce and the automotive industry. In recent years, AI has found its way into healthcare as well, helping doctors to make better decisions ('clinical decision support'), localizing tumors in magnetic resonance images, reading and analyzing reports written by radiologists and pathologists, and much more. However, AI has one big risk: it can be perceived as a 'black box', limiting trust in its reliability, which is a very big issue in an area in which a decision can mean life or death. As a result, the term Explainable Artificial Intelligence (XAI) has been gaining momentum. XAI tries to ensure that AI algorithms (and the resulting decisions) can be understood by humans. In this narrative review, we will have a look at the current status of XAI in healthcare, describe several issues around XAI, and discuss whether it can really help healthcare to advance, for example by increasing understanding and trust. Finally, alternatives to increase trust in AI are discussed, as well as future research possibilities in the area of XAI.

Keywords: XAI; AI; artificial intelligence; explainable; explainability; machine learning; deep learning; data science; big data; healthcare; medicine

1. Introduction

Artificial Intelligence (AI) is 'the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages' [1]. Examples of AI techniques are machine learning (ML), neural networks (NN) and deep learning (DL). AI can be applied in many different areas, such as econometrics (stock market predictions), biometry (facial recognition), e-commerce (recommendation systems) and the automotive industry (self-driving cars). In recent years, AI has found its way into the domain of biomedicine [2] and healthcare [3] as well. It is used to help researchers analyze big data to enable precision medicine [4] and to help clinicians to improve patient outcomes [5]. AI algorithms can help doctors to make better decisions ('clinical decision support', CDS), localize tumors in magnetic resonance (MR) images, read and analyze reports written by radiologists and pathologists, and much more. In the near future, generative AI and natural language processing (NLP) technology, such as Chat Generative Pre-trained Transformer (ChatGPT), could also help to create human-readable reports [6].

However, there are a number of barriers to the effective use of AI in healthcare. The first one is 'small' data, resulting in bias [7]. When studies are carried out on a patient cohort with limited diversity in race, ethnicity, gender, age, etc., results from these studies might be difficult to be applied to patients with different characteristics. An obvious solution for this bias is to create datasets using larger, more diverse patient cohorts, and to keep bias in mind when designing experiments. A second barrier exists in privacy and security issues. Strict regulations (such as the European GDPR, the American HIPAA and the Chinese PIPL) exist, limiting the use of personal data, and imposing large fines on leakage of such data. These issues can be solved in different ways; for example, by using federated or distributed learning. In this way, the algorithm travels to the data, and sends results back to a central repository. The data do not need to be transferred to another party, avoiding privacy

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 March 2023 doi:10.20944/preprints202303.0116.v1

and security issues as much as possible [8]. Another solution is the use of synthetic data: artificial data which might either be generated from scratch or based on real data, usually generated using AI algorithms such as Generative Adversarial Networks (GANs) [9]. A third barrier is the limited trust that clinicians and patients might have in AI algorithms. They can be perceived as a 'black box': something goes in, and something comes out, with no understanding of what happens inside. This distrust in AI algorithms, their accuracy and reliability, is a very big issue in an area in which a decision could mean life or death of the patient. As a result of this distrust, the term Explainable Artificial Intelligence (XAI) [10] has been gaining momentum as a possible solution. XAI tries to make sure that algorithms (and the resulting decisions) can be understood by humans.

XAI is being mentioned more and more in scientific publications, as can be seen in Figure 1. Its first mention in a PubMed title, abstract or keywords was in 2018, in a paper about machine learning in neuroscience [11]. Since then, it has been mentioned a total of 511 times, of which more than half (310) in papers from 2022 or the first months of 2023. The results for the Embase database show a similar trend. This shows the growing importance of XAI in (bio)medicine and healthcare. In this narrative review, we will have a look at several issues around XAI, and whether it can really help healthcare to advance, for example by increasing understanding and trust. Furthermore, this review will discuss alternatives to increase trust in AI as well as future research possibilities in the area of XAI.

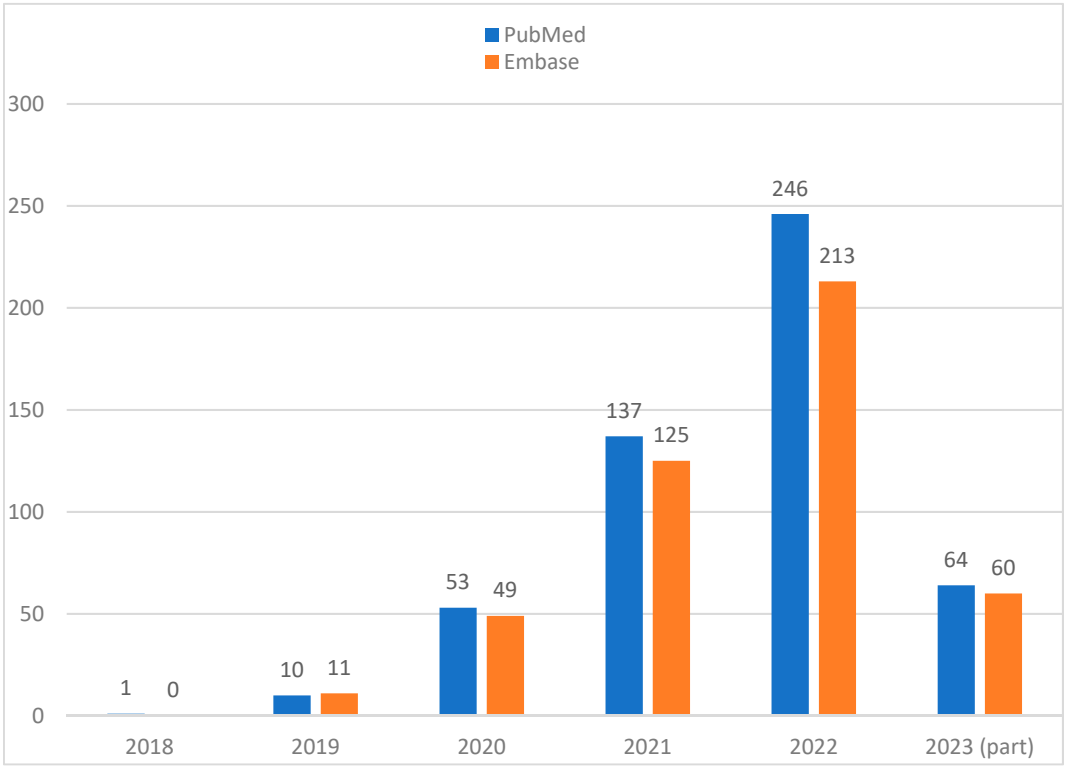

Figure 1. Number of publications containing the term "explainable artificial intelligence" in the titles, abstracts and keywords of the PubMed and Embase databases per year. Queries performed on 2023-03-06.

2. From 'black box' to '(translucent) glass box'

With explainable AI, we try to get from a 'black box' to a transparent 'glass box' [12] (sometimes also referred to as a 'white box' [13]). In a glass box model (such as a decision tree or linear regression model), all parameters are known, and we know exactly how the model comes to its conclusion, giving full transparency. In the ideal situation, the model is fully transparent, but in many situations (e.g., deep learning models) the model might be explainable only to a certain degree, which could be described as a 'translucent glass box' with an opacity level somewhere between 0% and 100%. A low

 doi:10.20944/preprints202303.0116.v1

opacity of the translucent glass box (or high transparency of the model) can lead to a better understanding of the model, which in turn could increase trust. This trust can exist on two levels: trust in the model versus trust in the prediction [14]. Data scientists are usually mostly interested in the model itself, whereas users (in healthcare: often clinicians, but sometimes patients) are mostly interested in the predictions based on that model. Therefore, trust for data scientists generally means trust in the model itself, while trust for clinicians and patients means trust in its predictions.

3. Legal and regulatory compliance

Another advantage of XAI is that it can help organizations comply with laws and regulations that require transparency and explainability in AI systems. Within the General Data Protection Regulation (GDPR) of the European Union, transparency is a fundamental principle for data processing [15]. In practice, the complexity of AI algorithms makes it difficult to adhere fully to this principle. Felzmann et al. [16] proposes that transparency as required by the GDPR in itself may be insufficient to achieve the positive goals associated with transparency, such as an increase in trust. Instead, they propose to understand transparency relationally, where information provision is conceptualized as communication between technology providers and users, and where assessments of trustworthiness based on contextual factors mediate the value of transparency communications. The EU is currently working on the Artificial Intelligence Act [17] which makes a distinction between non-high-risk and high-risk AI systems. On non-high-risk systems only limited transparency obligations are imposed, while for high-risk systems many restrictions are imposed on quality, documentation, traceability, transparency, human oversight, accuracy and robustness. Bell et al. [18] states that transparency is left to the technologists to achieve and propose a stakeholder-first approach that assists technologists in designing transparent, regulatory compliant systems, which is a useful initiative. Besides GDPR, there are other privacy laws for which XAI might be an interesting development. In the USA there is the Health Insurance Portability and Accountability Act (HIPAA) privacy rule [19], which is related to the Openness and Transparency Principle in the Privacy and Security Framework. This Openness and Transparency Principle stresses that it is "important for people to understand what individually identifiable health information exists about them, how that information is collected, used, and disclosed, and how reasonable choices can be exercised with respect to that information" [20]. The transparency of the usage of health information might point to a need for explainability of algorithms here. In China, article 7 of the Personal Information Protective Law (PIPL) prescribes that "the principles of openness and transparency shall be observed in the handling of personal information, disclosing the rules for handling personal information and clearly indicating the purpose, method, and scope of handling" [21], which also points to a need for transparency in data handling and AI algorithms.

4. Explainability: transparent or post-hoc

Arrieta et al. [22] classified studies on XAI into two approaches: some works focus on creating transparent models, while most works wrap black-box models with a layer of explainability, the so-called post-hoc models. The transparent models are based on linear or logistic regression, decision trees, k-nearest neighbours, rule-based learning, general additive models and Bayesian models. These models are considered to be transparent because they are understandable by itself. The post-hoc models need to explained by resorting to diverse means to enhance their interpretability, such as text explanations, visual explanations, local explanations, explanations by example, explanations by simplification and feature relevance explanations techniques. Phillips et al. [23] defines four principles for explainable AI systems: 1) explanation: explainable AI systems deliver accompanying evidence or reasons for outcomes and processes; 2) meaningful: provide explanations that are understandable to individual users; explanation accuracy: provide explanations that correctly reflect the system's process for generating the output; and 4) knowledge limits: a system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output. Vale et al. [24] states that machine learning post-hoc explanation methods cannot guarantee the insights they

 doi:10.20944/preprints202303.0116.v1

generate, which means that they cannot be relied upon as the only mechanism to guarantee fairness of model outcomes in high-stakes decision-making, such as in healthcare.

5. Privacy and security: a mixed bag

On the one hand, XAI can help to improve the safety and security of AI systems by making it easier to detect and prevent errors and malicious behavior [25]. On the other hand, XAI can also raise privacy and security concerns, as providing explanations for AI decisions may reveal sensitive information or show how to manipulate the system, for example by reverse engineering [26]. A fully transparent model can make a hacker feel like a kid in a candy store. Therefore, it is important to carefully consider the privacy and security implications of XAI and to take appropriate risk mitigation measures; certainly in healthcare where the protection of sensitive personal data is an important issue. Combining explainability of algorithms with privacy-preserving methods such as federated learning [27] might help. Saifullah et al. [28] argue that XAI and privacy-preserving machine learning (PPML) are both crucial research fields, but no attention has yet been paid to their interaction. They investigated the impact of private learning techniques on generated explanations for deep learning-based models and conclude that federated learning should be considered before differential privacy. If an application requires both privacy and explainability, they recommend differential private federated learning [29] as well as perturbation-based XAI methods [30]. Some research on security in combination with XAI has been done as well. Viganò and Magazzeni [31] propose the term `Explainable Security` (XSec) as an extension of XAI to the security domain. According to the authors, XSec has unique and complex characteristics: it involves several different stakeholders and is multi-faceted by nature. Kuppa and Le-Khac [32] designed a novel black box attack for analyzing the security properties (consistency, correctness and confidence) of gradient based XAI methods, which could help in designing secure and robust XAI methods.

6. Collaboration between humans and AI

It is important for clinicians (but also patients, researchers, etc.) to realize that humans can and should not be replaced by an AI algorithm [33]. An AI algorithm could outscore humans in specific tasks, but humans (at this moment in time) still have an added value with their domain expertise, broad experience and creative thinking skills. It might be the case that when the accuracy of an AI algorithm on a specific task is compared to the accuracy of the clinician, the AI gets better results. However, the AI model should not be compared to the human alone, but to the combination of AI model and human, because in the clinical practice they will almost always work together. In most cases, the combination (also known as 'AI-assisted decision making') will get the best results [34]. The combination of an AI model with the human expertise also makes the decision more explainable: the clinician can combine the explainable AI with his/her own domain knowledge. Figure 2 shows what qualities a human and an AI model can offer in clinical decision making, with the combination offering the best results.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 March 2023 doi:10.20944/preprints202303.0116.v1

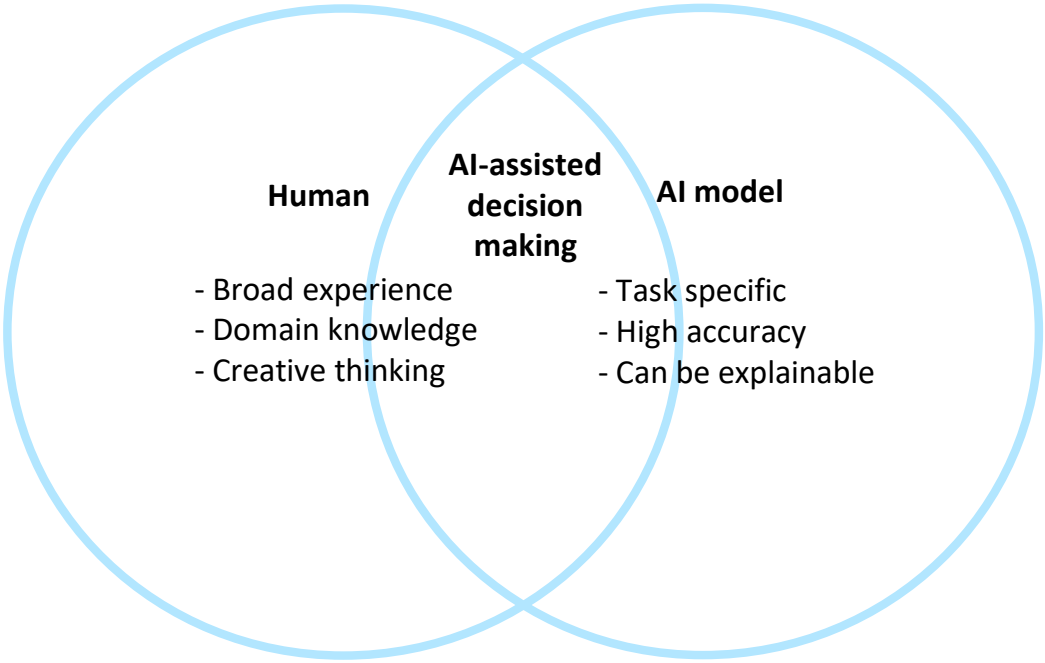

Figure 2. The combination of human and AI model can create powerful AI-assisted decision making.

7. Do explanations always raise trust?

The goal of explainability to end users of AI systems is ultimately to increase trust in the system. However, even with a good understanding of an AI system, end users may not necessarily trust the system. Druce at al. [35] show that a statistically significant increase in user trust and acceptance of an AI system can be reached by using a three-fold explanation: a graphical depiction of the systems generalization and performance in the current game state, how well the agent would play in semantically similar environments, and a narrative explanation of what the graphical information implies. Le Merrer and Trédan [36] argue that explainability might be promising in a local context, but that it cannot simply be transposed to a remote context, where a model trained by a service provider is only accessible to a user through a network and its application programming interface (API). They show that providing explanations cannot prevent a remote service from lying about the true reasons leading to its decisions, undermining the very concept of remote explainability in general. Within healthcare, trust is a fundamental issue because important decisions might be taken based on the output of the AI algorithm. Therefore, it would be good to take all necessary actions described here to increase trust in AI algorithms in healthcare.

8. Scientific Explainable Artificial Intelligence (sXAI)

Durán (2021) [37] differentiates scientific XAI (sXAI) from other forms of XAI. He states that the current approach for XAI is a bottom-up model: it consists of structuring all forms of XAI attending to the current technology and available computational methodologies, which could lead to confounding classifications (or 'how-explanations') with explanations. Instead, he proposes a bona-fide scientific explanation in medical AI. This explanation addresses three core components: 1) the structure of sXAI; 2) the role of human agents and non-epistemic beliefs in sXAI; and 3) how human agents can meaningfully assess the merits of an explanation. This author finally proposes a shift from standard XAI to sXAI, accompanied by substantial changes in the way explanation in medical AI is constructed and interpreted. Cabitza et al. [38] discusses this approach and concludes that current XAI methods fail to be bona-fide explanations: as a consequence, their framework cannot be applied to current XAI work. For sXAI to work, it needs to be integrated into future medical AI algorithms in a top-down manner.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 March 2023 doi:10.20944/preprints202303.0116.v1

9. 'Glass box' vs. 'crystal ball': balance between explainability and accuracy/performance

In some cases, the need for explainability can come at the cost of reduced performance of the model. For example, in order to make a model fully explainable (a 'glass box'), it might need to be simplified a bit. A very accurate prediction model (a 'crystal ball') might lose part of its accuracy because of this simplification. Or it needs to introduce some extra, more simple steps to make it more transparent, causing a reduction in performance. Linear models and rule-based models are very transparent, but usually have a lower performance than deep learning algorithms (Figure 5 of [39]). Therefore, in a real-world situation it might not be possible to achieve full explainability because accuracy and performance are usually considered more important. A balance needs to be maintained between the two, as is shown in Figure 3. In healthcare, this balance might shift more to the 'crystal ball' as accuracy might be considered more important than transparency and explainability. Van der Veer et al. [40] concluded that citizens may indeed value explainability of AI systems in healthcare less than in non-healthcare domains and less than often assumed by professionals, especially when weighed against system accuracy, and that citizens should therefore be actively consulted when developing policy on AI explainability.

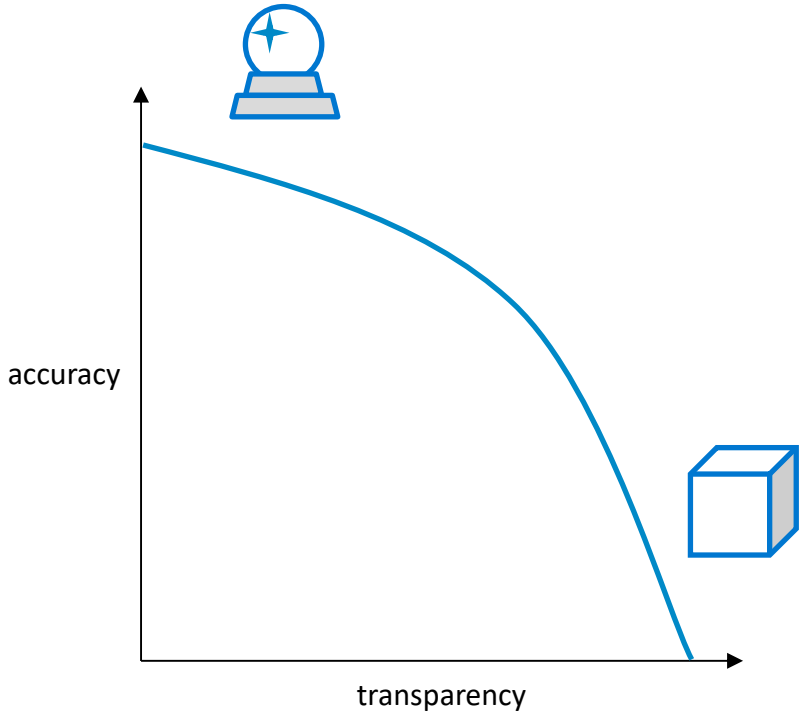

Figure 3. An increasing transparency of a (prediction) model might cause a decrease in accuracy, going from a 'crystal ball' to a 'glass box', and vice versa.

10. How to measure explainability?

Accuracy and performance can be measured easily by metrics such as specificity, selectivity and area under the Receiver Operating Characteristic (ROC) curve (AUC). Explainability is much more difficult to be measured because the quality of an explanation is somewhat subjective. Multiple researchers have tried to come up with an assessment of explainability. Sokol & Flach [41], for example, have created "explainability fact sheets" to assess explainable approaches along five dimensions: functional, operational, usability, safety and validation. This is quite an extensive approach. Most researchers measure explainability simply by evaluating how well an explanation is understood by the end user. Lipton [42] identifies three measures: 1) simulatability: can the user recreate or repeat (simulate) the computational process based on provided explanations of a system; 2) decomposability: can the user comprehend individual parts (and their functionality) of a predictive model; 3) algorithmic transparency: can the user fully understand the predictive algorithm? Hoffman

 doi:10.20944/preprints202303.0116.v1

et al. [43] uses "mental models", representations or expressions of how a person understands some sort of event, process, or system [44], as a user's understanding of the AI system. This mental model can be evaluated on criteria such as correctness, comprehensiveness, coherence and usefulness. Fauvel at al. [45] present a framework that assesses and benchmarks machine learning methods on both performance and explainability. For measuring the explainability, they look at model comprehensibility, explanation granularity, information type, faithfulness and user category. For model comprehensibility, only two categories are defined: "black-box" and "white-box" models, suggesting that these components could be further elaborated in future work. For granularity of the explanation, they use three categories: "global", "local" and "global & local" explainability. They propose a generic assessment of the information type in three categories from the least to the most informative: 1) importance: the explanations reveal the relative importance of each dataset variable on predictions; 2) patterns: the explanations provide the small conjunctions of symbols with a predefined semantic (patterns) associated with the predictions; 3) causal: the most informative category corresponds to explanations under the form of causal rules. The faithfulness of the explanation shows if the user can trust the explanation, with the two categories "imperfect" and "perfect". Finally, the user category shows the target user at which the explanation is aimed: "machine learning expert", "domain expert" and "broad audience". This user category is important because it defines the level of background knowledge they have. As suggested by the authors, all these metrics and categories can be defined in more detail in future XAI research.

11. Increasing complexity in the future

The first neural networks (using a single layer) were relatively easy to understand. With the advent of deep learning (using multiple layers) and new types of algorithms such as Deep Belief Networks (DBNs) [46] and Generative Adversarial Networks (GANs) [47], made possible by the increasing computer power, artificial intelligence algorithms are gaining complexity. In the future, with Moore's law continuing to proceed, this trend will likely continue. With algorithms getting more complex, it might also be more difficult to make them explainable. Ongoing research in the field of XAI might make it possible that new techniques will be developed that make it easier to explain and understand and complex AI models. For example, Explainability-by-Design [48] takes proactive measures to include explanation capability in the design of decision-making systems, so that no post-hoc explanations are needed. However, there is also the possibility that the complexity of AI models will overtake our ability to understand and explain them. Sarkar [49] even talks about an 'explainability crisis', which will be defined by the point at which our desire for explanations of machine intelligence will eclipse our ability to obtain them, and uses the 'five stages of grief' (denial, anger, bargaining, depression and acceptance) to describe the several phases of this crisis. The author's conclusion is that XAI is probably in a race against model complexity, but also that this may not such a big issue as it seems, as there are several ways to either reduce complexity or improve explanations. Ultimately, all will depend on the trajectory of AI development and the progress made in the field of XAI.

12. Discussion

Privacy laws such as GDPR, HIPAA and PIPL all include clauses that state that the handling of healthcare data should be transparent, which means that AI algorithms that work with these data should be transparent and explainable as well. However, making AI explainable is a difficult task, and it will be even more difficult when the complexity of AI algorithms will continue to increase. This increasing complexity might make it almost impossible for end users in healthcare (clinicians as well as patients) to understand and trust the algorithms. Therefore, perhaps we should not aim to explain AI to the end users, but to the researchers and developers deploying them, as they are most interested in the model itself. End users just want to be sure that the predictions done by the algorithm are accurate, which can be proven by showing them correct predictions from the past. Another important issue is the balance between explainability and accuracy or performance. Especially in healthcare, accuracy (and to a lesser extent performance) is crucial as it could be a matter of life and

death. Therefore, explainability might be considered of less importance in healthcare, compared to accuracy. If an algorithm's accuracy is lowered because of post-hoc explanations, it would be good to consider other methods to increase trust. For example, trust in algorithms could also be raised by ensuring robustness and by encouraging fairness [50]. Robustness of an algorithm in healthcare can be proven by presenting good results based on long-term use in different patient populations. Fairness of an AI algorithm is concurrent with bias minimization. A bias could be introduced by having a training dataset with low diversity, or by subjective responses of clinicians to a questionnaire. These biases should be identified and addressed during the validation and verification of the algorithm. Finally, algorithms (scripts, but also underlying data) should be made available for re-use when possible [51], so that results can be reproduced, increasing trust in the algorithm. Another solution to the explainability-accuracy trade-off might lie in the adoption of sXAI, in which explainability is integrated in a top-down manner into future medical AI algorithms, and Explainability-by-Design, which includes explanation capability in the design of decision-making systems. sXAI and Explainability-by-Design could be combined with ongoing research in privacy and security in AI (such as XSec) to create future-proof explainable artificial intelligence for healthcare.

Supplementary Materials: None.

Author Contributions: All work was carried out by T.H.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: None.

Conflicts of Interest: Dr. Tim Hulsen is an employee of Philips Research.

References

1. Joiner, I.A. Chapter 1 - Artificial Intelligence: AI is Nearby. In *Emerging Library Technologies*, Joiner, I.A., Ed.; Chandos Publishing: 2018; pp. 1-22.
2. Hulsen, T. Literature analysis of artificial intelligence in biomedicine. *Annals of translational medicine* **2022**, *10*, 1284, doi:10.21037/atm-2022-50.
3. Yu, K.-H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nature biomedical engineering* **2018**, *2*, 719-731.
4. Hulsen, T.; Jamuar, S.S.; Moody, A.; Karnes, J.H.; Orsolya, V.; Hedensted, S.; Spreafico, R.; Hafler, D.A.; McKinney, E. From Big Data to Precision Medicine. *Frontiers in Medicine* **2019**, doi:10.3389/fmed.2019.00034.
5. Hulsen, T.; Friedecký, D.; Renz, H.; Melis, E.; Vermeersch, P.; Fernandez-Calle, P. From big data to better patient outcomes. *Clinical Chemistry and Laboratory Medicine (CCLM)* **2022**, doi:doi:10.1515/cclm-2022-1096.
6. Biswas, S. ChatGPT and the Future of Medical Writing. *Radiology 0*, 223312, doi:10.1148/radiol.223312.
7. Celi, L.A.; Cellini, J.; Charpignon, M.-L.; Dee, E.C.; Dernoncourt, F.; Eber, R.; Mitchell, W.G.; Moukheiber, L.; Schirmer, J.; Situ, J. Sources of bias in artificial intelligence that perpetuate healthcare disparities—A global review. *PLOS Digital Health* **2022**, *1*, e0000022.
8. Hulsen, T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. *Int J Environ Res Public Health* **2020**, *17*, doi:10.3390/ijerph17093046.
9. Vega-Márquez, B.; Rubio-Escudero, C.; Riquelme, J.C.; Nepomuceno-Chamorro, I. Creation of Synthetic Data with Conditional Generative Adversarial Networks. Cham, 2020; pp. 231-240.
10. Gunning, D.; Stefik, M.; Choi, J.; Miller, T.; Stumpf, S.; Yang, G.Z. XAI-Explainable artificial intelligence. *Sci Robot* **2019**, *4*, doi:10.1126/scirobotics.aay7120.
11. Vu, M.T.; Adalı, T.; Ba, D.; Buzsáki, G.; Carlson, D.; Heller, K.; Liston, C.; Rudin, C.; Sohal, V.S.; Widge, A.S.; et al. A Shared Vision for Machine Learning in Neuroscience. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **2018**, *38*, 1601-1607, doi:10.1523/jneurosci.0508-17.2018.
12. Rai, A. Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science* **2020**, *48*, 137-141, doi:10.1007/s11747-019-00710-5.

13. Loyola-Gonzalez, O. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access* **2019**, *7*, 154096-154113.
14. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why Should I Trust You? : Explaining the Predictions of Any Classifier. *Kdd '16* **2016**, 1135–1144, doi:10.1145/2939672.2939778.
15. Consulting, I. Recital 58 - The Principle of Transparency. Available online: https://gdpr-info.eu/recitals/no-58/ (accessed on
16. Felzmann, H.; Villaronga, E.F.; Lutz, C.; Tamò-Larrieux, A. Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society* **2019**, *6*, 2053951719860542, doi:10.1177/2053951719860542.
17. Commission, E. Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (accessed on
18. Bell, A.; Nov, O.; Stoyanovich, J. Think About the Stakeholders First! Towards an Algorithmic Transparency Playbook for Regulatory Compliance. *arXiv preprint arXiv:2207.01482* **2022**.
19. Office for Civil Rights, H. Standards for privacy of individually identifiable health information. Final rule. *Federal register* **2002**, *67*, 53181-53273.
20. Services, U.S.D.o.H.H. The HIPAA Privacy Rule and Electronic Health Information Exchange in a Networked Environment - Openness and Transparency. Available online: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/special/healthit/opennesstransparency.pdf (accessed on
21. Creemers, R.; Webster, G. Translation: Personal Information Protection Law of the People's Republic of China – Effective Nov. 1, 2021. Available online: https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/ (accessed on
22. Arrieta, A.B.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; García, S.; Gil-López, S.; Molina, D.; Benjamins, R. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* **2020**, *58*, 82-115.
23. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Broniatowski, D.A.; Przybocki, M.A. Four principles of explainable artificial intelligence. *Gaithersburg, Maryland* **2020**, 18.
24. Vale, D.; El-Sharif, A.; Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: risks and limitations in non-discrimination law. *AI and Ethics* **2022**, *2*, 815-826, doi:10.1007/s43681-022-00142-y.
25. Charmet, F.; Tanuwidjaja, H.C.; Ayoubi, S.; Gimenez, P.-F.; Han, Y.; Jmila, H.; Blanc, G.; Takahashi, T.; Zhang, Z. Explainable artificial intelligence for cybersecurity: a literature survey. *Annals of Telecommunications* **2022**, *77*, 789-812, doi:10.1007/s12243-022-00926-7.
26. Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M.K.; Ristenpart, T. Stealing Machine Learning Models via Prediction APIs. In Proceedings of the USENIX security symposium, 2016; pp. 601-618.
27. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence* **2020**, *2*, 305-311, doi:10.1038/s42256-020-0186-1.
28. Saifullah, S.; Mercier, D.; Lucieri, A.; Dengel, A.; Ahmed, S. Privacy Meets Explainability: A Comprehensive Impact Benchmark. *arXiv preprint arXiv:2211.04110* **2022**.
29. Geyer, R.C.; Klein, T.; Nabi, M. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557* **2017**.
30. Ivanovs, M.; Kadikis, R.; Ozols, K. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters* **2021**, *150*, 228-234.
31. Viganò, L.; Magazzeni, D. Explainable Security. *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* **2018**, 293-300.
32. Kuppa, A.; Le-Khac, N.A. Black Box Attacks on Explainable Artificial Intelligence(XAI) methods in Cyber Security. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), 19-24 July 2020, 2020; pp. 1-8.
33. Bhattacharya, S.; Pradhan, K.B.; Bashar, M.A.; Tripathi, S.; Semwal, J.; Marzo, R.R.; Bhattacharya, S.; Singh, A. Artificial intelligence enabled healthcare: A hype, hope or harm. *Journal of family medicine and primary care* **2019**, *8*, 3461-3464, doi:10.4103/jfmpc.jfmpc_155_19.
34. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. *Fat* '20* **2020**, 295–305, doi:10.1145/3351095.3372852.
35. Druce, J.; Harradon, M.; Tittle, J. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. *arXiv preprint arXiv:2106.03775* **2021**.
36. Le Merrer, E.; Trédan, G. Remote explainability faces the bouncer problem. *Nature Machine Intelligence* **2020**, *2*, 529-539, doi:10.1038/s42256-020-0216-z.
37. Durán, J.M. Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artificial Intelligence* **2021**, *297*, 103498, doi:https://doi.org/10.1016/j.artint.2021.103498.

38. Cabitza, F.; Campagner, A.; Malgieri, G.; Natali, C.; Schneeberger, D.; Stoeger, K.; Holzinger, A. Quod erat demonstrandum? - Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications* **2023**, *213*, 118888, doi:https://doi.org/10.1016/j.eswa.2022.118888.
39. Guang, Y.; Qinghao, Y.; Jun, X. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Information Fusion* **2022**, *77*, 29-52, doi:https://doi.org/10.1016/j.inffus.2021.07.016.
40. van der Veer, S.N.; Riste, L.; Cheraghi-Sohi, S.; Phipps, D.L.; Tully, M.P.; Bozentko, K.; Atwood, S.; Hubbard, A.; Wiper, C.; Oswald, M.; et al. Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association* **2021**, *28*, 2128-2138, doi:10.1093/jamia/ocab127.
41. Sokol, K.; Flach, P. Explainability fact sheets: a framework for systematic assessment of explainable approaches. In Proceedings of the Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020; pp. 56-67.
42. Lipton, Z.C. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* **2018**, *16*, 31–57, doi:10.1145/3236386.3241340.
43. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* **2018**.
44. Klein, G.; Hoffman, R.R. Macrocognition, mental models, and cognitive task analysis methodology. *Naturalistic decision making and macrocognition* **2008**, 57-80.
45. Fauvel, K.; Masson, V.; Fromont, E. A performance-explainability framework to benchmark machine learning methods: application to multivariate time series classifiers. *arXiv preprint arXiv:2005.14501* **2020**.
46. Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y. An Empirical Evaluation of Deep Architectures on Problems with Many Factors of Variation. *Icml '07* **2007**, 473–480, doi:10.1145/1273496.1273556.
47. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Communications of the ACM* **2020**, *63*, 139-144.
48. Huynh, T.D.; Tsakalakis, N.; Helal, A.; Stalla-Bourdillon, S.; Moreau, L. Explainability-by-Design: A Methodology to Support Explanations in Decision-Making Systems. *arXiv preprint arXiv:2206.06251* **2022**.
49. Sarkar, A. Is explainable AI a race against model complexity? *arXiv preprint arXiv:2205.10119* **2022**.
50. Asan, O.; Bayrak, A.E.; Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J Med Internet Res* **2020**, *22*, e15154, doi:10.2196/15154.
51. Hulsen, T. The ten commandments of translational research informatics. *Data Science* **2019**, *2*, 341-352, doi:10.3233/DS-190020.