AI

MDPI

Review

Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare

Tim Hulsen

Department of Hospital Services & Informatics, Philips Research, 5656 AE 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 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 some central concepts in XAI, describe several challenges around XAI in healthcare, 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

check for updates

Citation: Hulsen, T. Explainable Artificial Intelligence (XAI): Concepts and Challenges in Healthcare. *AI* **2023**, *4*, 652–666. https://doi.org/10.3390/ai4030034

Academic Editor: Kenji Suzuki

Received: 31 May 2023
Revised: 11 July 2023
Accepted: 9 August 2023
Published: 10 August 2023

Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

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 to 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 some 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., the 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 for the 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 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 the 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 488 times, of which more than 63% (311) in papers from 2022 or from the first months of 2023. The results for the Embase database show a similar trend. A full list of the publications can be found in Supplementary Tables S1 (PubMed) and S2 (Embase). This trend shows the growing importance of XAI in (bio)medicine and healthcare. Taking this growth into consideration, the number of manuscripts that discuss the concepts and challenges of XAI in the context of healthcare remains small. In this narrative review, we will have a look at several concepts around XAI and what their importance might be for the implementation and acceptance of AI in healthcare. This review will also provide some future directions. It will not attempt to give a full overview of the current literature on this topic or explain in detail which methods exist to explain AI algorithms, as several excellent reviews on this topic already exist [12–15]. First, we will go through some central concepts of XAI. We will explain the terminologies "black box" and "glass box". Then, we will look at two approaches to explainability, transparency, and post-hoc explanations, followed by a discussion on the collaboration between humans (e.g., clinicians) and AI. The subsequent two sections introduce scientific XAI and discuss the explanation methods of granular computing and fuzzy modeling. Second, we will discuss some challenges of XAI in healthcare. The first section is about legal and regulatory compliance, which is of particular importance in healthcare, dealing with sensitive personal data. The next sections discuss the effects of XAI on privacy and security and the question of whether the explanations always raise trust. Another section discusses the balance between explainability and accuracy/performance, followed by an overview of methods to measure explainability and a contemplation on the future increasing complexity of AI algorithms. The penultimate section shows some examples of XAI applied in a healthcare setting. Finally, the discussion puts everything in a broader context and mentions some future research possibilities of XAI in healthcare.

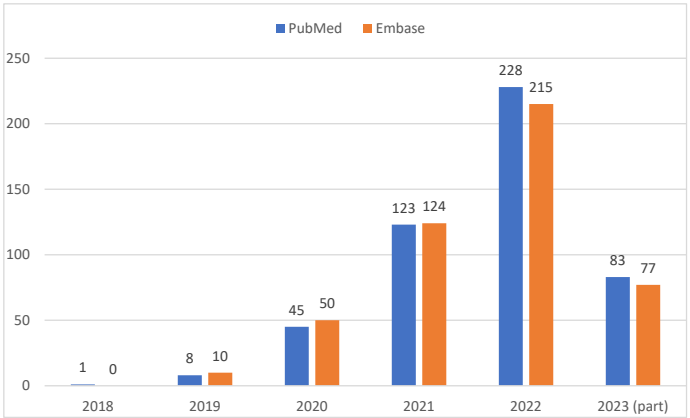

Figure 1. Number of publications containing the term "explainable artificial intelligence" in the titles, abstracts, and keywords of the PubMed and Embase databases per year. Queries performed on 26 March 2023.

2. Central Concepts of XAI

2.1. From "Black Box" to "(Translucent) Glass Box"

With explainable AI, we try to progress from a "black box" to a transparent "glass box" [16] (sometimes also referred to as a "white box" [17]). 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 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, as explained by Ribeiro et al. [18]. In healthcare, there are many different stakeholders who have different explanation needs [19]. For example, data scientists are usually mostly interested in the model itself, whereas users (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. The "trusting a prediction" problem can be solved by providing explanations for individual predictions, whereas the "trusting the model" problem can be solved by selecting multiple such predictions (and explanations) [18]. Future research could determine in which context either of these two approaches should be applied.

2.2. Explainability: Transparent or Post-Hoc

Arrieta et al. [20] 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 (Figure 2). The transparent models are based on linear or logistic regression, decision trees, k-nearest neighbors, rule-based learning, general additive models, and Bayesian models. These models are considered to be transparent because they are understandable by themselves. The post-hoc models (such as neural networks, random forest, and deep learning) need to be 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. [21] define 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; (3) 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. [22] state that machine learning post-hoc explanation methods cannot guarantee the insights they generate, which means that they cannot be relied upon as the only mechanism to guarantee the fairness of model outcomes in high-stake decision-making, such as in healthcare.

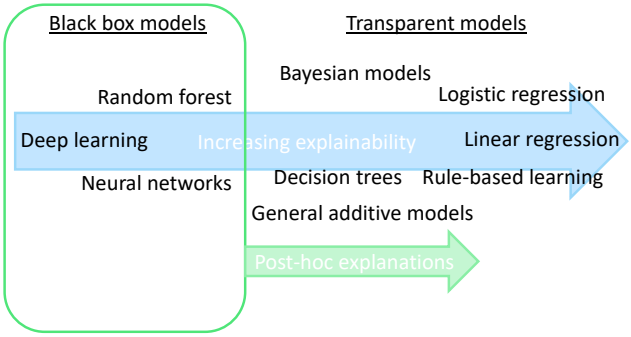

Figure 2. Black box models (needing post-hoc explanations) vs. inherently transparent models.

2.3. 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 [23]. An AI algorithm could outscore humans in specific tasks, but humans (at this moment in time) still have 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 the AI model and a human because, in clinical practice, they will almost always work together. In most cases, the combination (also known as "AI-assisted decision making") will obtain the best results [24]. The combination of an AI model with human expertise also makes the decision more explainable: the clinician can combine the explainable AI with his/her own domain knowledge. In CDS, explainability allows developers to identify shortcomings in a system and allows clinicians to be confident in the decisions they make with the support of AI. [25]. Amann et al. state that if we would move in the opposite direction toward opaque algorithms in CDSS, this may inadvertently lead to patients being passive spectators in the medical decision-making process [26]. Figure 3 shows what qualities a human and an AI model can offer in clinical decision-making, with the combination offering the best results. In the future, there might be a shift to the right side of the figure, but the specific qualities of humans will likely ensure that combined decision-making will still be the best option for years to come.

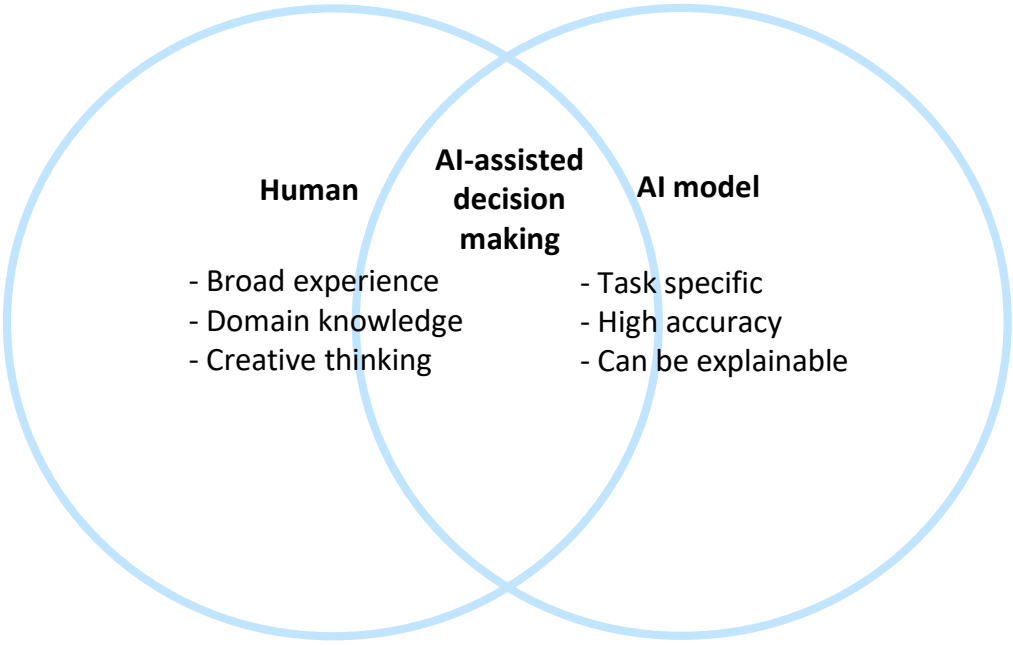

Figure 3. The combination of human and AI models can create powerful AI-assisted decision-making.

2.4. Scientific Explainable Artificial Intelligence (sXAI)

Durán (2021) [27] 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, consisting of the "explanans" (the unit that carries out an explanation), the "explanandum" (the unit that will be explained), and the "explanatory relation" (the objective relation of dependency that links the explanans and the explanandum); (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 concludes by proposing a shift from standard XAI to sXAI, together with substantial

changes in the way medical XAI is constructed and interpreted. Cabitza et al. [28] discuss this approach and conclude that existing XAI methods fail to be bona fide explanations, which is why 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. This means that algorithms should not be explained by simply describing "how" a decision has been reached, but we should also look at what other scientific disciplines, such as philosophy of science, epistemology, and cognitive science, can add to the discussion [27]. For each medical AI algorithm, the explanans, explanandum, and explanatory relation should be defined.

2.5. Explanation Methods: Granular Computing (GrC) and Fuzzy Modeling (FM)

Many methods exist to explain AI algorithms, as described in detail by Holzinger et al. [29]. There is one technique that is particularly useful in XAI because it is motivated by the need to approach AI through human-centric information processing [30], Granular Computing (GrC), which was introduced by Zadeh in 1979 [31]. GrC is an "emerging paradigm in computing and applied mathematics to process data and information, where the data or information are divided into so-called information granules that come about through the process of granulation" [32]. GrC can help make models more interpretable and explainable by bridging the gap between abstract concepts and concrete data through these granules. Another useful technique related to GrC is Fuzzy Modeling (FM), a methodology oriented toward the design of explanatory and predictive models. FM is a technique through which a linguistic description can be transformed into an algorithm whose result is an action [33]. Fuzzy modeling can help explain the reasoning behind the output of an AI system by representing the decision-making process in a way that is more intuitive and interpretable. Although FM was originally conceived to provide easily understandable models to users, this property cannot be taken for granted, but it requires careful design choices [34]. Much research in this area is still ongoing. Zhang et al. [35] discuss the multi-granularity three-way decisions paradigm [36] and how this acts as a part of granular computing models, playing a significant role in explainable decision-making. Zhang et al. [37] adopt a GrC framework named "multigranulation probabilistic models" to enrich semantic interpretations for GrC-based multi-attribute group decision-making (MAGDM) approaches.

In healthcare, GrC could, for example, help break down a CDS algorithm into smaller components, such as the symptoms, patient history, test results, and treatment options. This can help the clinician understand how the algorithm arrived at its diagnosis and determine if it is reliable and accurate. FM could, for example, be used in a CDS system to represent the uncertainty and imprecision in the input data, such as patient symptoms, and the decision-making process, such as the rules that are used to arrive at a diagnosis. This can help to provide a more transparent and understandable explanation of how the algorithm arrived at its output. Recent examples of the application of GrC and FM in healthcare are in the disease areas of Parkinson's disease [38], COVID-19 [39], and Alzheimer's disease [40].

3. Challenges of XAI in Healthcare

3.1. 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 [41]. However, transparency is difficult to adhere to because of the complexity of AI. Felzmann et al. [42] propose that transparency, as required by the GDPR in itself, may be insufficient to achieve an increase in trust or any other positive goal associated with transparency. Instead, they recommend a relational understanding of transparency, in which the provision of information is viewed as a sort of interaction between users and technology providers, and the value of transparency messages is mediated by trustworthiness assessments based on the context. Schneeberger et al. [43] discussed

the European framework regulating medical AI based on White Paper on AI from 2020 by the European Commission [44] and concluded that this framework, by endorsing a human-centric approach, will fundamentally influence how medical AI and AI, in general, will be used in Europe in the future. The EU is currently working on the Artificial Intelligence Act [45], which will make 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. [46] state 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 [47], 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" [48]. The transparency of the usage of health information might point to a need for explainability of algorithms. 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" [49], which also points to a need for transparency in data handling and AI algorithms. Since new, more AI-specific privacy laws are being introduced around the world, regulatory compliance with AI algorithms is gaining relevance and will be an important area for research in the future.

3.2. 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 [50]. 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 [51]. A fully transparent model can make a hacker feel as if they have endless possibilities. 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 the explainability of algorithms with privacy-preserving methods such as federated learning [52] might help. Saifullah et al. [53] 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 concluded that federated learning should be considered before differential privacy. If an application requires both privacy and explainability, they recommend differential private federated learning [54] as well as perturbation-based XAI methods [55]. The importance of privacy in relation to medical XAI is shown in Figure 4 of Albahri et al. [56], with keywords such as "ethics", "privacy", "security", and "trust" being the most often-occurring keywords in papers around XAI in healthcare. Some research on security in combination with XAI has been carried out as well. Viganò and Magazzeni [57] 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 [58] 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. Kiener [59] looked specifically at security in healthcare and identified three

types of security risks related to AI: cyber-attacks; systematic bias; and mismatches, all of which can have serious consequences for medical systems. Explainability can be part of the solution for all of these risks. The author specifically mentions input attacks as a type of cyber-attack that is of high risk to AI systems. Input attacks manipulate the input data (e.g., make some small changes to an MR image) so that the AI algorithm will deliver an incorrect result [60]. In an explainable model, the clinician can look at the reasoning behind the incorrect result and possibly, detect the manipulation. Systematic bias can be brought to light as well by explaining the workings of the AI algorithm. For example, it can become clearly visible that an algorithm was only trained on data from people from one ethnic background. Mismatches can occur when the AI algorithm recommends courses of action that do not match the background situation of the individual patient. The algorithm can mistake correlation for causation and suggest, for example, an incorrect treatment. In a black-box AI, such a mismatch might be undetectable, but in a transparent, explainable AI, it might be much easier to detect or at least indicate the risk of such a mismatch.

3.3. Do Explanations Always Raise Trust?

The goal of explainability to end users of AI models is ultimately to increase trust in the model. However, even with a good understanding of an AI model, end users may not necessarily trust the model. Druce et al. [61] show that a statistically significant increase in user trust and acceptance of an AI model can be reached by using a three-fold explanation: (1) a graphical depiction of the model's generalization and performance in the current game state; (2) how well the agent would play in semantically similar environments; and (3) a narrative explanation of what the graphical information implies. Le Merrer and Trédan [62] argue that explainability might be promising in a local context but that it cannot simply be transposed to a different (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 (similar to what humans could do), 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. Mistrust might result in humans discarding accurate predictions, while overtrust could lead to over-reliance on possibly inaccurate predictions. Therefore, it would be good to take all necessary actions described here to reach the correct level of trust in AI algorithms in healthcare. One of the key actions here is to create open and honest education to end users on the strengths and weaknesses of AI algorithms. For example, people should be trained to understand the difference between local context and remote context.

3.4. "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 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 lower performance than deep learning algorithms (Figure 5 [63]). Therefore, in a real-world situation, it might not be possible to achieve full explainability because accuracy and performance are usually considered to be more important. A balance needs to be maintained between the two, as shown in Figure 4. 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. [64] concluded that citizens might indeed value the explainability of AI systems in healthcare less than in non-healthcare domains, especially when weighed against system accuracy. When developing policy on the explainability of (medical) AI, citizens should be actively consulted, as they might have a different opinion than assumed by healthcare professionals. This trade-off between accuracy and transparency could be

different for each context, however, depending on the implications of a wrong decision based on the AI algorithm. Future research could be carried out on the context-specific need for explainability.

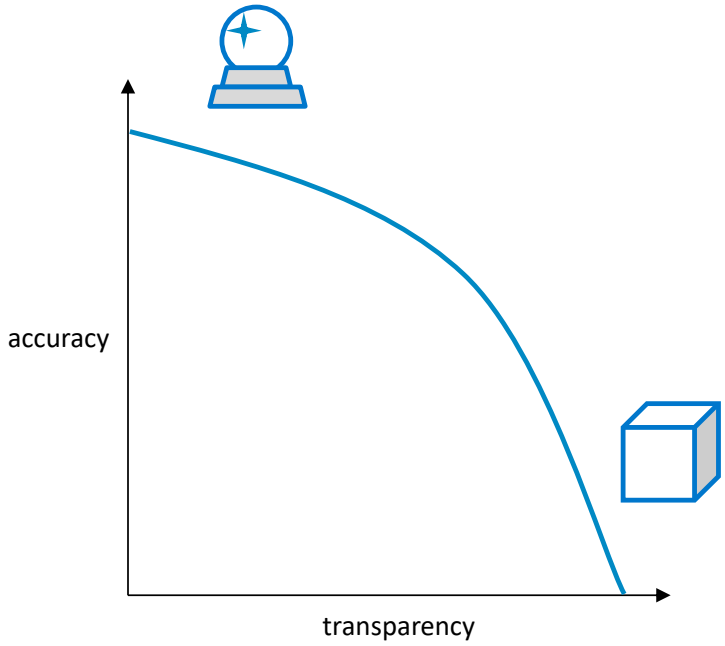

Figure 4. Increasing transparency of a (prediction) model might cause a decrease in accuracy, going from a "crystal ball" to a "glass box" and vice versa.

3.5. 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. Table 1 shows an overview of the most widely used explainability metrics from the recent literature. The four publications that introduced these metrics all look at explainability from a different angle. Sokol and Flach [65], 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 [66] 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 et al. [67] use "mental models", representations or expressions of how a person understands some sort of event, process, or system [68], 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 et al. [69] present a framework that assesses and benchmarks machine learning methods on both performance and explainability. Performance is measured compared to the state-of-the-art, best, similar, or below. For measuring 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 the granularity of the explanation, they use three categories: "global"; "local"; and "global and 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.

Table 1. Methods for assessing explainability.

Manuscript	Measures
Sokol and Flach (2020) [65]	- Functional - Operational - Usability - Safety - Validation
Lipton (2018) [66]	- Simulatability - Decomposability - Algorithmic transparency
Hoffman et al. (2018) [67]	- Correctness - Comprehensiveness - Coherence - Usefulness
Fauvel et al. (2020) [69]	- Performance: o Best o Similar o Below - Explainability: o Model comprehensibility: ▪ Black box models ▪ White box models o Explanation granularity: ▪ Global ▪ Local ▪ Global and local o Information type: ▪ Importance ▪ Patterns ▪ Causal o Faithfulness: ▪ Imperfect ▪ Perfect o User category: ▪ Machine learning expert ▪ Domain expert ▪ Broad audience

3.6. 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) [70] and Generative Adversarial Networks (GANs) [71], made possible by the increasing computer power, artificial intelligence algorithms are gaining complexity. In the future, this trend will likely continue, with Moore's law still continuing to proceed. With algorithms becoming 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 complex AI models. For example, Explainability-by-Design [72] 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 [73] 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 be such a big issue as it seems, as there are several ways to either improve explanations or reduce AI complexity. Ultimately, it all will depend on the trajectory of AI development and the progress made in the field of XAI.

4. Application Examples

XAI has been applied to healthcare in medicine in a number of ways already. AI has been very successful in improving medical image analysis, and recently, researchers have also been trying to combine this success (through high accuracy) with an increased explainability and interpretability of the models created. Van der Velden et al. [74] identified over 200 papers using XAI in deep learning-based medical image analysis and concluded that most papers in this area used a visual explanation (mostly through saliency maps [75]) as opposed to textual explanations and example-based explanations. These saliency maps highlight the most important features which can distinguish between diseased and non-diseased tissue [76]. Manresa-Yee et al. [77] describe explanation interfaces that are being used in healthcare, mostly by clinicians. They identified three main application areas for these interfaces: prediction tasks; diagnosis tasks; and automated tasks. One example of a clinician-facing explanation interface is the dashboard presented by Khodabandehloo et al. [78], which uses data from sensorized smart homes to detect a decline in the cognitive functions of the elderly in order to promptly alert practitioners.

Joyce et al. [79] studied the use of XAI in psychiatry and mental health, where the need for explainability and understandability is higher than in other areas because of the probabilistic relationships between the data describing the syndromes, outcomes, disorders, and signs/symptoms. They introduced the TIFU (Transparency and Interpretability For Understandability) framework, which focuses on how a model can be made understandable (to a user) as a function of transparency and interpretability. They conclude that the main applications of XAI in mental health are prediction and discovery, that XAI in mental health requires understandability because clinical applications are high-stakes, and that AI tools should assist clinicians and not introduce further complexity.

5. Discussion

Current privacy laws such as GDPR, HIPAA, and PIPL 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. Future privacy laws will likely be even more strict on AI explainability. However, making AI explainable is a difficult task, and it will be even more difficult when the complexity of AI algorithms continues 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 mostly interested in the model itself. End users, especially patients, mostly want to be sure that the predictions made 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 [80]. Robustness of an algorithm in healthcare can be proven by presenting good results based on long-term use in different patient populations. When a model is robust, its explanation will not change much when minor changes are made to the model [81]. The 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. XAI can help find these biases as well as mitigate them [82]. These biases can be addressed during the validation and verification of the algorithm. Finally, algorithms (scripts, but also underlying data) should be made available for reuse when possible [83] so that the results can be reproduced, increasing trust in the algorithm. GrC and FM can help increase trust as well by making models more interpretable and explainable. Another solution to the explainability–accuracy trade-off might lie in the adoption of sXAI, in which explainability is integrated into a top–down manner into future medical AI algorithms, and Explainability-by-Design, which includes explanation capability in the design of decision-making systems. GrC, FM, 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. In any case, explainability should be considered as important as other metrics, such as accuracy and robustness, as they all raise trust in AI. Future endeavors to make AI explainable should be personalized, as different end users need different levels of explanations. The explanations should be communicated to the end user in an understandable manner, for example, through an easy-to-use user interface. Explainability should also not compromise the privacy rights of the patients [84]. For XAI in healthcare to fully reach its potential, it should be embedded in clinical workflows, and explainability should be included in AI development from the start instead of adding post-hoc explanations as an afterthought.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ai4030034/s1, Table S1: PubMed publications with the search term "explainable artificial intelligence"; Table S2: Embase publications with the search term "explainable artificial intelligence".

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.

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

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
CDS	Clinical Decision Support
ChatGPT	Chat Generative Pre-trained Transformer

DBN	Deep Belief Network
DL	Deep Learning
FM	Fuzzy Modeling
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
GrC	Granular Computing
HIPAA	Health Insurance Portability and Accountability Act
MAGDM	Multi-Attribute Group Decision Making
ML	Machine Learning
MR	Magnetic Resonance
NLP	Natural Language Processing
NN	Neural Networks
PIPL	Personal Information Protective Law
PPML	Privacy-Preserving Machine Learning
ROC	Receiver Operating Characteristic
sXAI	Scientific Explainable Artificial Intelligence
XAI	Explainable Artificial Intelligence
XSec	Explainable Security

References

1. Joiner, I.A. Chapter 1—Artificial intelligence: AI is nearby. In *Emerging Library Technologies*; Joiner, I.A., Ed.; Chandos Publishing: Oxford, UK, 2018; pp. 1–22.
2. Hulsen, T. Literature analysis of artificial intelligence in biomedicine. *Ann. Transl. Med.* **2022**, *10*, 1284. [CrossRef] [PubMed]
3. Yu, K.-H.; Beam, A.L.; Kohane, I.S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2018**, *2*, 719–731. [CrossRef] [PubMed]
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. *Front. Med.* **2019**, *6*, 34. [CrossRef]
5. Hulsen, T.; Friedecký, D.; Renz, H.; Melis, E.; Vermeersch, P.; Fernandez-Calle, P. From big data to better patient outcomes. *Clin. Chem. Lab. Med. (CCLM)* **2022**, *61*, 580–586. [CrossRef] [PubMed]
6. Biswas, S. ChatGPT and the Future of Medical Writing. *Radiology* **2023**, *307*, e223312. [CrossRef] [PubMed]
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 Digit. Health* **2022**, *1*, e0000022. [CrossRef]
8. Hulsen, T. Sharing Is Caring-Data Sharing Initiatives in Healthcare. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3046. [CrossRef]
9. Vega-Márquez, B.; Rubio-Escudero, C.; Riquelme, J.C.; Nepomuceno-Chamorro, I. Creation of synthetic data with conditional generative adversarial networks. In Proceedings of the 14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019), Seville, Spain, 13–15 May 2019; Springer: Cham, Switzerlnad, 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*, eaay7120. [CrossRef]
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. *J. Neurosci.* **2018**, *38*, 1601–1607. [CrossRef]
12. Bharati, S.; Mondal, M.R.H.; Podder, P. A Review on Explainable Artificial Intelligence for Healthcare: Why, How, and When? *IEEE Trans. Artif. Intell.* **2023**. [CrossRef]
13. Sheu, R.-K.; Pardeshi, M.S. A Survey on Medical Explainable AI (XAI): Recent Progress, Explainability Approach, Human Interaction and Scoring System. *Sensors* **2022**, *22*, 8068. [CrossRef] [PubMed]
14. Tjoa, E.; Guan, C. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 4793–4813. [CrossRef] [PubMed]
15. Jung, J.; Lee, H.; Jung, H.; Kim, H. Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. *Heliyon* **2023**, *9*, e16110. [CrossRef] [PubMed]
16. Rai, A. Explainable AI: From black box to glass box. *J. Acad. Mark. Sci.* **2020**, *48*, 137–141. [CrossRef]
17. 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. [CrossRef]
18. Ribeiro, M.T.; Singh, S.; Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144. [CrossRef]
19. Gerlings, J.; Jensen, M.S.; Shollo, A. Explainable AI, but explainable to whom? An exploratory case study of xAI in healthcare. In *Handbook of Artificial Intelligence in Healthcare: Practicalities and Prospects*; Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., Jain, L.C., Eds.; Springer International Publishing: Cham, Switzeralnd, 2022; Volume 2, pp. 169–198.

20. 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. *Inf. Fusion* **2020**, *58*, 82–115. [CrossRef]
21. Phillips, P.J.; Hahn, C.A.; Fontana, P.C.; Broniatowski, D.A.; Przybocki, M.A. *Four Principles of Explainable Artificial Intelligence*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020; Volume 18.
22. Vale, D.; El-Sharif, A.; Ali, M. Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. *AI Ethics* **2022**, *2*, 815–826. [CrossRef]
23. 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. *J. Fam. Med. Prim. Care* **2019**, *8*, 3461–3464. [CrossRef]
24. Zhang, Y.; Liao, Q.V.; Bellamy, R.K.E. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 295–305. [CrossRef]
25. Antoniadi, A.M.; Du, Y.; Guendouz, Y.; Wei, L.; Mazo, C.; Becker, B.A.; Mooney, C. Current Challenges and Future Opportunities for XAI in Machine Learning-Based Clinical Decision Support Systems: A Systematic Review. *Appl. Sci.* **2021**, *11*, 5088. [CrossRef]
26. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; the Precise, Q.c. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
27. Durán, J.M. Dissecting scientific explanation in AI (sXAI): A case for medicine and healthcare. *Artif. Intell.* **2021**, *297*, 103498. [CrossRef]
28. 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 Syst. Appl.* **2023**, *213*, 118888. [CrossRef]
29. Holzinger, A.; Saranti, A.; Molnar, C.; Biecek, P.; Samek, W. Explainable AI methods—A brief overview. In Proceedings of the xxAI—Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, Vienna, Austria, 12–18 July 2020; Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 13–38.
30. Bargiela, A.; Pedrycz, W. *Human-Centric Information Processing through Granular Modelling*; Springer Science & Business Media: Dordrecht, The Netherlands, 2009; Volume 182.
31. Zadeh, L.A. Fuzzy sets and information granularity. In *Fuzzy Sets, Fuzzy Logic, and Fuzzy Systems: Selected Papers*; World Scientific: Singapore, 1979; pp. 433–448.
32. Keet, C.M. Granular computing. In *Encyclopedia of Systems Biology*; Dubitzky, W., Wolkenhauer, O., Cho, K.-H., Yokota, H., Eds.; Springer: New York, NY, USA, 2013; p. 849.
33. Novák, V.; Perfilieva, I.; Dvořák, A. What is fuzzy modeling. In *Insight into Fuzzy Modeling*; John Wiley & Sons: Hoboken, NJ, USA, 2016; pp. 3–10.
34. Mencar, C.; Alonso, J.M. Paving the way to explainable artificial intelligence with fuzzy modeling: Tutorial. In Proceedings of the Fuzzy Logic and Applications: 12th International Workshop (WILF 2018), Genoa, Italy, 6–7 September 2018; Springer International Publishing: Cham, Switzerland, 2019; pp. 215–227.
35. Zhang, C.; Li, D.; Liang, J. Multi-granularity three-way decisions with adjustable hesitant fuzzy linguistic multigranulation decision-theoretic rough sets over two universes. *Inf. Sci.* **2020**, *507*, 665–683. [CrossRef]
36. Zadeh, L.A. Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. *Fuzzy Sets Syst.* **1997**, *90*, 111–127. [CrossRef]
37. Zhang, C.; Li, D.; Liang, J.; Wang, B. MAGDM-oriented dual hesitant fuzzy multigranulation probabilistic models based on MULTIMOORA. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1219–1241. [CrossRef]
38. Zhang, C.; Ding, J.; Zhan, J.; Sangaiah, A.K.; Li, D. Fuzzy Intelligence Learning Based on Bounded Rationality in IoMT Systems: A Case Study in Parkinson's Disease. *IEEE Trans. Comput. Soc. Syst.* **2022**, *10*, 1607–1621. [CrossRef]
39. Solayman, S.; Aumi, S.A.; Mery, C.S.; Mubassir, M.; Khan, R. Automatic COVID-19 prediction using explainable machine learning techniques. *Int. J. Cogn. Comput. Eng.* **2023**, *4*, 36–46. [CrossRef]
40. Gao, S.; Lima, D. A review of the application of deep learning in the detection of Alzheimer's disease. *Int. J. Cogn. Comput. Eng.* **2022**, *3*, 1–8. [CrossRef]
41. Intersoft Consulting. Recital 58—The Principle of Transparency. Available online: https://gdpr-info.eu/recitals/no-58/ (accessed on 26 March 2023).
42. 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 Soc.* **2019**, *6*, 2053951719860542. [CrossRef]
43. Schneeberger, D.; Stöger, K.; Holzinger, A. The European legal framework for medical AI. In Proceedings of the International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Dublin, Ireland, 25–28 August 2020; Springer: Cham, Switzerland, 2020; pp. 209–226.
44. European Commission. *On Artificial Intelligence—A European Approach to Excellence and Trust*; European Commission: Brussels, Belgium, 2020.
45. European Commission. 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 26 March 2023).

46. Bell, A.; Nov, O.; Stoyanovich, J. Think about the Stakeholders First! Towards an Algorithmic Transparency Playbook for Regulatory Compliance. *arXiv* **2022**, arXiv:2207.01482. [CrossRef]
47. HHS Office for Civil Rights. Standards for privacy of individually identifiable health information—Final rule. *Fed. Regist.* **2002**, *67*, 53181–53273.
48. HHS Office for Civil Rights. 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 26 March 2023).
49. Creemers, R.; Webster, G. Translation: Personal Information Protection Law of the People's Republic of China—Effective 1 November 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 26 March 2023).
50. 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. *Ann. Telecommun.* **2022**, *77*, 789–812. [CrossRef]
51. 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, Austin, TX, USA, 10–12 August 2016; pp. 601–618.
52. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
53. Saifullah, S.; Mercier, D.; Lucieri, A.; Dengel, A.; Ahmed, S. Privacy Meets Explainability: A Comprehensive Impact Benchmark. *arXiv* **2022**, arXiv:2211.04110.
54. Geyer, R.C.; Klein, T.; Nabi, M. Differentially private federated learning: A client level perspective. *arXiv* **2017**, arXiv:1712.07557.
55. Ivanovs, M.; Kadikis, R.; Ozols, K. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognit. Lett.* **2021**, *150*, 228–234. [CrossRef]
56. Albahri, A.S.; Duhaim, A.M.; Fadhel, M.A.; Alnoor, A.; Baqer, N.S.; Alzubaidi, L.; Albahri, O.S.; Alamoodi, A.H.; Bai, J.; Salhi, A.; et al. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Inf. Fusion* **2023**, *96*, 156–191. [CrossRef]
57. Viganò, L.; Magazzeni, D. Explainable security. In Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, 7–11 September 2020; pp. 293–300.
58. 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), Glasgow, UK, 19–24 July 2020; pp. 1–8.
59. Kiener, M. Artificial intelligence in medicine and the disclosure of risks. *AI Soc.* **2021**, *36*, 705–713. [CrossRef]
60. Comiter, M. *Attacking Artificial Intelligence AI's Security Vulnerability and What Policymakers Can Do about It*; Belfer Center for Science and International Affairs: Cambridge, MA, USA, 2019.
61. Druce, J.; Harradon, M.; Tittle, J. Explainable artificial intelligence (XAI) for increasing user trust in deep reinforcement learning driven autonomous systems. *arXiv* **2021**, arXiv:2106.03775.
62. Le Merrer, E.; Trédan, G. Remote explainability faces the bouncer problem. *Nat. Mach. Intell.* **2020**, *2*, 529–539. [CrossRef]
63. 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. *Inf. Fusion* **2022**, *77*, 29–52. [CrossRef]
64. 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. *J. Am. Med. Inform. Assoc.* **2021**, *28*, 2128–2138. [CrossRef]
65. Sokol, K.; Flach, P. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, 27–30 January 2020; pp. 56–67.
66. 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. [CrossRef]
67. Hoffman, R.R.; Mueller, S.T.; Klein, G.; Litman, J. Metrics for explainable AI: Challenges and prospects. *arXiv* **2018**, arXiv:1812.04608.
68. Klein, G.; Hoffman, R.R. Macrocognition, mental models, and cognitive task analysis methodology. In *Naturalistic Decision Making and Macrocognition*; Ashgate Publishing: Farnham, UK, 2008; pp. 57–80.
69. Fauvel, K.; Masson, V.; Fromont, E. A performance-explainability framework to benchmark machine learning methods: Application to multivariate time series classifiers. *arXiv* **2020**, arXiv:2005.14501.
70. Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In Proceedings of the International Conference on Machine Learning (ICML '07), Corvallis, OR, USA, 20–24 June 2007; pp. 473–480. [CrossRef]
71. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
72. 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* **2022**, arXiv:2206.06251.
73. Sarkar, A. Is explainable AI a race against model complexity? *arXiv* **2022**, arXiv:2205.10119.
74. van der Velden, B.H.M.; Kuijf, H.J.; Gilhuijs, K.G.A.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef] [PubMed]

75. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv* **2013**, arXiv:1312.6034.
76. Chaddad, A.; Peng, J.; Xu, J.; Bouridane, A. Survey of Explainable AI Techniques in Healthcare. *Sensors* **2023**, *23*, 634. [CrossRef] [PubMed]
77. Manresa-Yee, C.; Roig-Maimó, M.F.; Ramis, S.; Mas-Sansó, R. Advances in XAI: Explanation Interfaces in Healthcare. In *Handbook of Artificial Intelligence in Healthcare: Practicalities and Prospects*; Lim, C.-P., Chen, Y.-W., Vaidya, A., Mahorkar, C., Jain, L.C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; Volume 2, pp. 357–369.
78. Khodabandehloo, E.; Riboni, D.; Alimohammadi, A. HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline. *Future Gener. Comput. Syst.* **2021**, *116*, 168–189. [CrossRef]
79. Joyce, D.W.; Kormilitzin, A.; Smith, K.A.; Cipriani, A. Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digit. Med.* **2023**, *6*, 6. [CrossRef] [PubMed]
80. Asan, O.; Bayrak, A.E.; Choudhury, A. Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians. *J. Med. Internet Res.* **2020**, *22*, e15154. [CrossRef] [PubMed]
81. Marcus, G. The next decade in AI: Four steps towards robust artificial intelligence. *arXiv* **2020**, arXiv:2002.06177.
82. Das, A.; Rad, P. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv* **2020**, arXiv:2006.11371.
83. Hulsen, T. The ten commandments of translational research informatics. *Data Sci.* **2019**, *2*, 341–352. [CrossRef]
84. Harder, F.; Bauer, M.; Park, M. Interpretable and differentially private predictions. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 4083–4090.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.