Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives

Atul Rawal [1], James McCoy [1], Danda Rawat [1], Brian Sadler [1], and Robert Amant [1]

[1]Affiliation not available

October 30, 2023

Abstract

This is a survey paper on Explainable Artificial Intelligence (XAI).

Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges and Perspectives

Atul Rawal, *Member, IEEE,* James McCoy, Danda Rawat, *Senior Member, IEEE*, Brian M. Sadler, *Life Fellow, IEEE*, and Rob St. Amant

***Abstract*—Artificial intelligence (AI) and Machine Learning (ML) have come a long way from the earlier days of conceptual theories, to being an integral part of today's technological society. Rapid growth of AI/ML and their penetration within a plethora of civilian and military applications, while successful, has also opened new challenges and obstacles. With almost no human involvement required for some of the new decision-making AI/ML systems, there is now a pressing need to gain better insights into how these decisions are made. This has given rise to a new field of AI research, Explainable AI (XAI). In this paper, we present a survey of XAI characteristics and properties. We provide an in-depth review of XAI themes, and describe the different methods for designing and developing XAI systems, both during and post model-development. We include a detailed taxonomy of XAI goals, methods, and evaluation, and sketch the major milestones in XAI research. An overview of XAI for security, and cybersecurity of XAI systems, is also provided. Open challenges are delineated, and measures for evaluating XAI system robustness are described.**

***Index Terms*—Explainable AI (XAI), Artificial Intelligence, Machine Learning, Robust AI, Explainability**

I. Introduction

Artificial intelligence (AI) and machine learning (ML) have come a long way from early conceptual theories, to being an integral part of today's technological society. Recent advances in AI and ML have resulted in the widespread application of data-driven learning systems. In many cases these advances now require almost zero human involvement/supervision, with the AI/ML systems making decisions based on the learned data. When these decision-making systems are used in ways that impact human lives, such as healthcare and military applications, there is a crucial need to understand how the AI/ML systems make decisions [1], [2].

How can we be sure that the AI/ML systems can be trusted? How can we be sure that there is no inherent bias within these systems decisions? There have been many real-world examples of AI system failures. Amazon's recruitment AI had bias against women, with preference given to male candidates, and Facebook's advertisement AI was biased against race, gender and religion [3] [4]. Within the US healthcare system, bias against people of color has been reported in many AI algorithms [5].

AI bias may be a reflection of human training or data collected by human operated systems for machine learning, but regardless of cause private companies and government agencies alike are trying to make sure that the AI/ML systems provide unbiased explainable decisions.

This has led to the creation of new policies and laws, not just in the United States, but across the world. For example, the European Union General Data Protection Regulation, which provides consumers with a "Right to Explanation" [6]. The US Algorithmic Accountability Act of 2019 dictates "assessments of high-risk systems that involve personal information or make automated decisions" [7].

Accountable AI is the solution to make sure that AI decision systems can be trusted. This has given rise to research on Explainable Artificial Intelligence (XAI). Yearly publications reflect the recent and rapid rise in XAI, Interpretable, Intelligible, and Transparent AI (Fig. 1), with XAI's emergence in 2017 along with the US DoD DARPA XAI program.

Even though there are related surveys on XAI [1], [8]–[12]., which provide great overviews of XAI, a recent and updated survey that provides a more comprehensive look at not just XAI's development, but it's goals and evaluation metrics is also needed. Further, there is a lack of studies that highlight the current state of the art, when it comes to the security of XAI systems. This survey paper aims at filling the gaps in literature by providing a comprehensive survey that looks at all

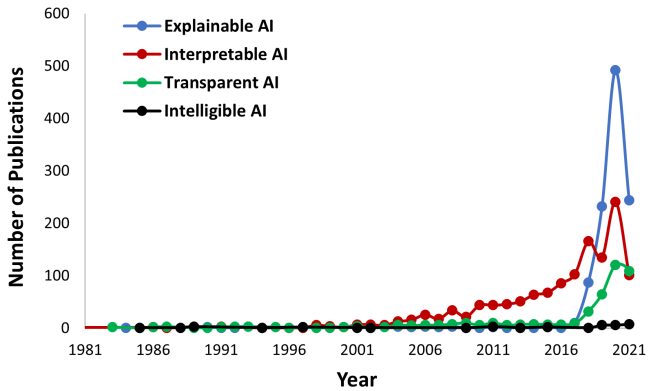

Fig. 1: Yearly publications for Explainable, Interpretable, Transparent and Intelligible AI. (Data derived from SCOPUS)

A. Rawal, J. McCoy and D. B. Rawat are with the Department of Electrical Engineering and Computer Science at Howard University, USA

B. M. Sadler and R. Amant are with the US Army Research Lab, USA

This work was supported in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory.

aspects of XAI from development to evaluation, and highlights some of the more recent breakthroughs and advances that have been made towards XAI. The main contributions of this survey include:

- We present a detailed overview of XAI by focusing on all aspects of the field from design & development to evaluation.
- We provide a comparison of the XAI development methods by characterizing them into either *transparent models* or *post-hoc models* and provide examples of current ML models that are compatible with each method.
- We summarize a comprehensive taxonomy for design/development and evaluation of XAI.
- We present the major milestones in XAI development since 1983.
- We provide an insight into the security of XAI and highlight recent advances towards secure XAI.
- We present an open discussion of challenges that still remain within the field and perspectives on recommendations for addressing them.

This paper is organized as follows. Section II presents a taxonomy and insight into the terms, design, and development methods for XAI. Section III provides a brief survey of the design and development methods for achieving explainability with AI/ML systems. Section IV describes techniques that are used for measuring the effectiveness of XAI systems, and Section V overviews XAI security. Section VI discusses open challenges and current trends in XAI research, followed by concluding remarks in Section VII.

II. OVERVIEW OF EXPLAINABLE AI (XAI)

While explanation systems have been around since the 1980's, recent years have seen a major increase in XAI research with ML/AI models. The prevalence of black-box models in most military and commercial AI/ML systems using Deep Learning (DL) and other machine learning techniques has given rise to the need for more transparent systems capable of explaining their decisions. The U.S Defense Advanced Research Projects (DARPA) Agency defined XAI as "AI systems that can explain their rationale to a human user, characterize their strengths and weakness, and convey an understanding of how they will behave in the future" [13]. Fig. 2 provides a conceptual overview of XAI. Even though there are works and studies on XAI that existed before 1983, the timeline in Fig. 4 simply presents some of the more significant milestones in XAI research since 1983.

Explainability is central to developing a trustworthy and explainable system. Proper implementation of explainability needs to ensure that the system is accountable to build public confidence in the algorithmic implementation. For a system to be accountable measures should be implemented to ensure that input biases are recognized and mitigated.

Busuioc, listed the following criteria for AI systems to be accountable (see Fig 3) [14].

- The input information and data must be free of any and all biases - As it may not be possible to rid of bias completely, AI systems must be able to recognize and mitigate biases.
- Decisions must be explainable to the end user - The absolute end user must be capable of understanding the decisions/predictions made by the AI systems.
- There must be consequences for its actions - meaningful accountability dictates the imposition of sanctions and affording redress to those negatively affected.

Although the words interpretability, understandability, comprehensibility, and explainability have been used interchangeably within the literature, we should distinguish between them in reference to XAI systems. The explanation of the system must go beyond the interpretation of the system. The system's decision-making processes and detailed steps must be comprehensible to the (possibly non-technical) end-user. The end-users must understand why and how the system came to a particular result versus alternatives; they must be able to identify incorrect results and understand why they are wrong; and they must be able to decide when to trust, and when not to trust, the system.

A. XAI Terminology

A challenge with XAI is the free and interchangeable use of terms when it comes to explanation. For example, in the literature, interpretable and explainable are often used as synonyms although generally representing two distinct concepts. Other terms such as transparency and comprehensibility have also been used as substitutes for explainability. Each of these have a pronounced definition, and therefore should not be used so freely. This section reviews the major nomenclature that has been used within the XAI research field.

Transparency is how understandable a model is without providing any insights into the actual algorithmic process of the AI system. It's the degree to which an end user can understand the function of the model, without any technical details [15], [16].

Comprehensibility is commonly associated with the complexity of the AI/ML model. It represents the algorithms' ability to portray/display its learned results in human terms [17]–[19].

Interpretability in reference to XAI is the ability of the AI/ML system to be explained in human terms [1] [9].

Explainability is a set of processes or methods that ensures that the system to capable of allowing humans to comprehend its overall decision and reasoning. Explainability can be understood as a summary of the overall working features and calculations that produce the final system output. Arrieta, et al., gave an excellent definition of explainability in terms of machine learning (ML) as, "Given a certain audience, explainability refers to the details and reasons a model gives it make its functioning clear and easy to understand" [1] [16] [20] [21].

B. XAI Goals

The explosive growth of AI enabled applications that are able to operate autonomously has increased the need to examine the effectiveness of these systems. Due to the inability of

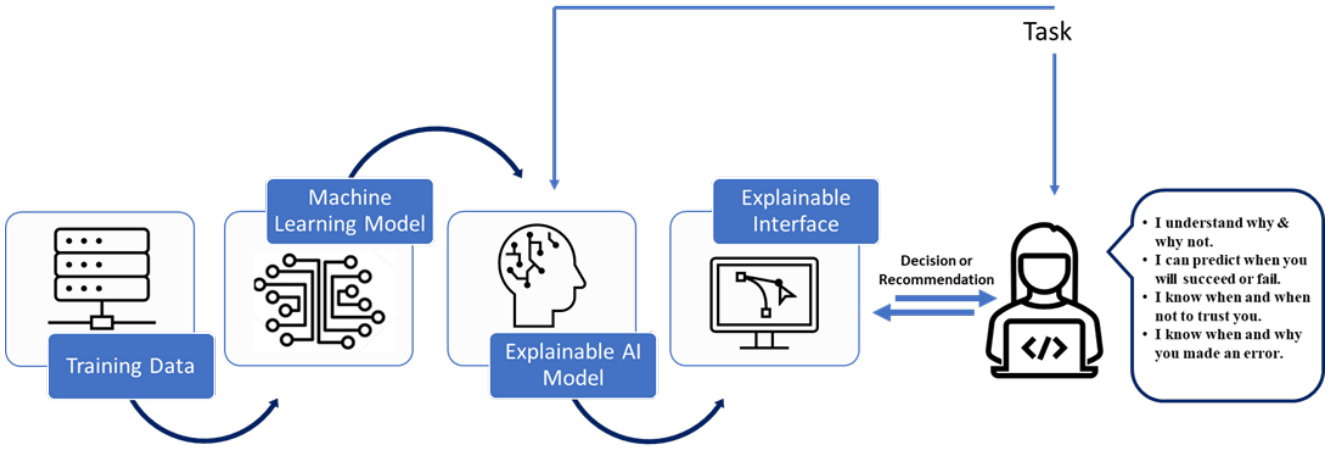

Fig. 2: Overview of the Explainable AI (XAI) concept.

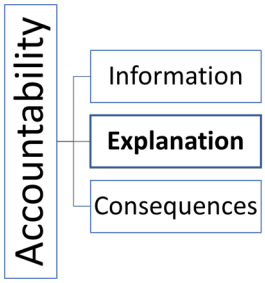

Fig. 3: Criteria for Accountable AI.

many of these autonomous or AI enabled systems to fully be expressed by humans, DARPA seeks to address these limitations through designing a suite of machine learning tools that can be implemented to make the system robust and trustworthy. These techniques aim to: [13]

1) Produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and
2) Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners."

This dictates a need for merging the understandability of the system by the user, and their trust of the system. While understandability and trust are both undoubtedly vital goals for robust XAI systems, various other goals should also be considered for the development of these systems. Goals such as transparency, fairness, bias avoidance, informativeness, causality, confidence, transferability, privacy and safety, and ease of use should also be considered essential for XAI design and development [1], [8], [15], [22]–[24].

Trust in AI/ML systems has been viewed by many as the main goal of XAI models. It is the level of confidence in the actions of an AI/ML system making decisions for specific problems. However, this should not be taken at face value as the only goal of XAI systems. While a crucial aspect of any XAI system, trust alone is not sufficient for explainability. Like the saying, "not everything that shines is gold", not every trustworthy AI/ML system is explainable [1] [8] [25]–[28].

Understandability represents the features of the AI/ML system to make an end-user understand how it works, with or without the explanation for its algorithm and decision-making processes. Its explanations should aim at improving the user experience via the understanding of the system and its decisions [1] [16] [8] [29] [10].

Fairness and bias avoidance are two very critical goals. Due to the inherent bias within today's society in a multitude of fields, XAI systems must be designed and developed without any biases to guarantee that they can be trusted to not make the mistakes of their creators (Humans). As mentioned previously, real-world examples of biased AI decision-making may have serious consequences. Some of the sources of bias within these systems arise from biased training data and feature learning. Explainable systems can provide the end-user with the choice to either trust or not trust the systems based upon improved understanding of factors that influenced the result. Explainability can thus aide in avoiding biases that cause unethical and harmful consequences [1] [6] [8] [30] [31].

Insightfulness should also be considered a crucial goal for XAI design and development. As Arrieta, et al., stated in their survey, problems being solved by AI/ML systems aren't necessarily the same as those intended by the users. Therefore, it is very important that a system's explainability helps the user to gain insight into the overall goals of the system. There is a need to extract information about the systems [1]

Causality among data is an important source of information for XAI systems. The study of causal reasoning from observed data is already a robust field of research. Several studies have presented explainable systems as an important tool for investigating and deriving causal relationships among different variables [32]–[34]. Causality in relation to explainability, gave rise to a new term *causability*. *Causability*, coined by Holzinger, et al.,is defined as "the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use" [35]. A current argument within the field of XAI states that for AI/ML systems

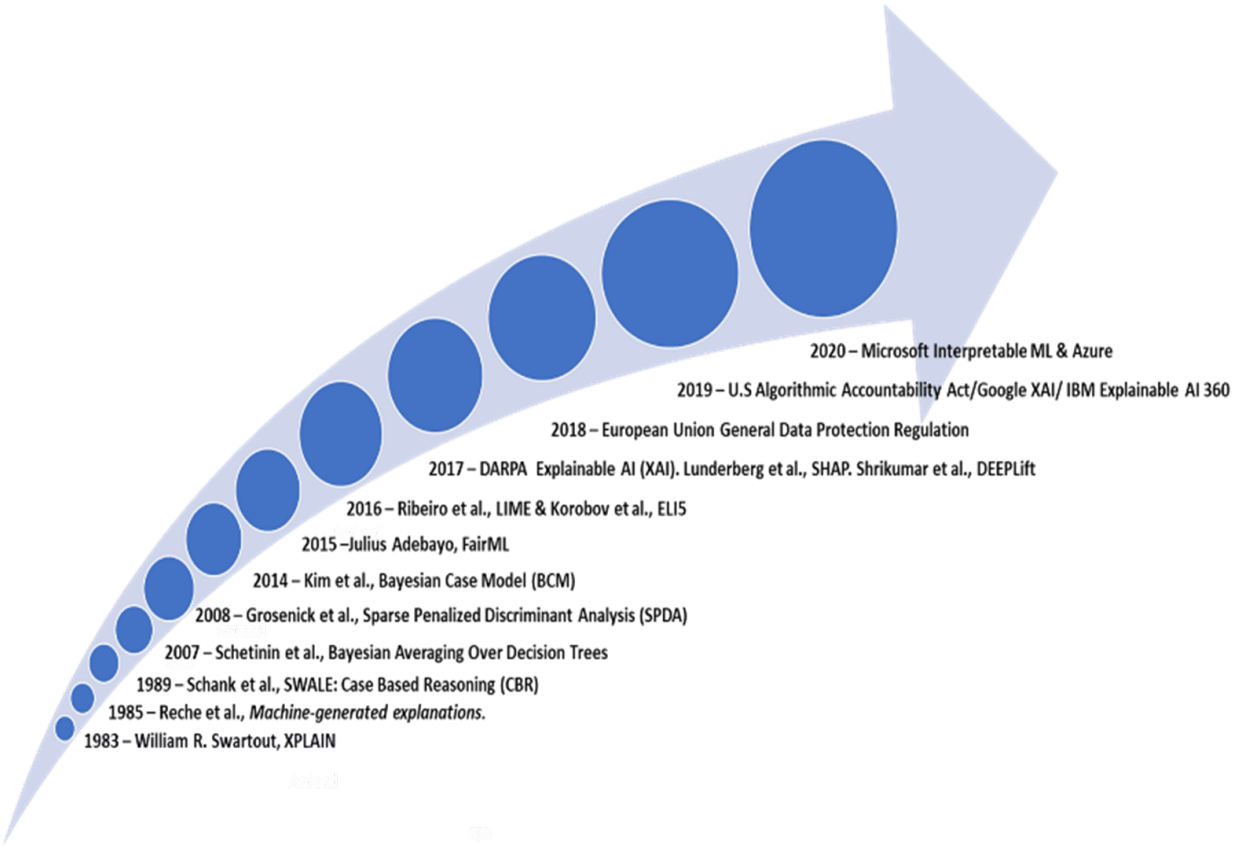

Fig. 4: Major XAI milestones since 1983.

to generate human-like explanations, they human causally understandable explanations are crucial [36]. Kilbertus, et al., emphasized the importance of causal approaches for avoiding bias and discrimination, and providing better explanations [36], [37]. Holzinger et al., introduced the System Causability Scale (SCS), a method of evaluating the quality of explanations based on causability [38].

Transferability is one of the major challenges for AI/ML. The practicality of adapting one AI/ML model for different applications is still being heavily researched for numerous models and systems. It is therefore desirable that XAI systems be able to adapt explainability to different problems/applications as well.It should however be noted that transferability does not always imply explainability, and it should not be assumed that all transferable models will be explainable [39]–[41].

Privacy protection and data protection are a major challenge in many applications. Data driven AI/ML systems must be designed with the privacy of information in mind, and XAI systems that use large datasets from the public domain must be able to protect the consumers privacy. This is a potential vulnerability if an XAI system might reveal private data in the course of explanation. To this end, XAI systems must be developed that can hide/protect sensitive data from users and developers alike [1] [8] .

Mohseni, et al., suggested characterizing the design goals of XAI systems according to the designated end users and divided them into three groups: AI/data novices, data experts, and AI experts. AI novices are defined as regular users with little or no technical knowledge of AI/ML systems. Data experts refers to researchers and scientists that make practical use of AI/ML on a daily basis for research, commercial, or military applications. AI experts are the scientists and engineers that design and develop AI/ML systems. The paper noted that while there are overlaps and similarities between the goals for the different user groups, distinctions can be made in the design methods, implementation and research objectives. (Table I) [8].

While the development of trustworthy and robust XAI models is a priority, delivery of the explanation created by the model to the end-user is also a focal goal of XAI. DARPA's XAI program put an emphasis on the explainable interface of the AI/ML systems, to not only create better explainable models, but also improve on how these explanations are relayed to the users. The integration of state-of-the-art human-computer interface (HCI) methods with the XAI principles, strategies and models, along with psychological theories of explanations is carried out to achieve more effective explanations for the user [13].

III. XAI Development and Design

XAI systems can be designed using either a transparent model approach or a post-hoc explainability approach [20]. The difference between these methods stems from systems that are inherently explainable by design, and those that need to be made explainable. Within the classification of transparent models, three specific distinctions can be made for the degree

TABLE I: XAI design goals based on user groups.

User Group	Design Goals
AI/Data Novices	Algorithmic Transparency, User and Trust Reliance, Bias Mitigation, Privacy Awareness
Data Experts	Model Visualization and Inspection, Model Tuning and Selection.
AI Experts	Model Interpretability, Model Debugging.

of transparency of the models, simulatability, decomposability, and algorithmic transparency [1]. Post-hoc explainable models can also be further characterized into text explanation, visual explanation, global explanation, local explanation, rule-based explanation, explanation by simplification, explanation by example, and feature relevance. Tables II and III provide an overview of the different methods used to achieve explainability and Fig 5. presents a taxonomy for the goals, development and evaluation of XAI. [1] [20] [8] [42].

A. Degree of Transparency

Transparent models are designed to be both interpretable and explainable. The three levels of transparency act in a "Russian doll" manner, where the highest degree encompasses all three transparency levels. Models that are simulatable maintain the highest level of transparency, followed by decomposable models and finally algorithmically transparent models.

Simulatability is the ability to be simulated by the user. Lipton, et al., defined a simutalable model as "a model where a human can take in input data together with the parameters of the model and in reasonable time step through every calculation required to produce a prediction" [15] [27].

Decomposability, also referred to as intelligibility, can be defined as the ability of the system to explain all its processes [43]. Challenges of making AI/ML systems decomposable lie in the fact that not every system can be made as such. The inherent strain in making systems decomposable is the difficulty in explaining all the parts and processes of the systems, as it requires all the input parameters and variables to be easily interpretable [1].

Algorithmic transparency, as explained by Gareth, et al., refers to the ability of the user to logically understand the AI/ML system's error surface, allowing the user to predict the system's actions in different problems or situations [40]. It is the level of a user's understanding of the AI/ML systems' operations to process the data and produce the result/decision. Algorithmic transparency is however limited to specific models, such as linear models, but is not applicable to deep learning (DL) models due the requirement of the model's comprehensibility via mathematical techniques [44] [45].

B. Post-hoc Explainability

The post-hoc explainability approach is a set of techniques that can be implemented after the system is complete to make the system more explainable. This is done using post development/design methods such as text explanation, visual explanation, and local explanation.

Text explanations encompass all explanation methods that yield symbols/texts representing model functions by mapping the algorithms' rationale to the symbols. This approach seeks to improve the overall explainability of the AI/ML system by generating text explanations of their results [31].

Visual explanations use visual representations of system behaviors to improve the systems overall explainability. This approach can be effective at explaining internal system behaviors and processes for non-technical users, and can be coupled with other procedures such as text explanations to further enhance the explanation's effectiveness.

Explanation by example provide examples of the results generated by the AI/ML systems, allowing for a better comprehension of the system. These explanations provide examples of historical situations that are similar to the current one. [42] One of the most effective types of explanations by examples, are *counterfactual* explanations. Various studies have highlighted the importance of counterfactuals as the missing link for XAI to achieve human-like intelligence and human-understandable explanations [35]. Chou, et al., defined counterfactuals as "a conditional assertion whose antecedent is false and whose consequent describes how the world would have been if the antecedent had occurred (a what-if question)." [36] They provide specific explanations to convey what features need to be changed to achieve desired prediction/decision [46], [47].Choy et al, also analyzed 18 model agnostic XAI counterfactual algorithms currently in use and classified them based on their theoretical approach as listed below: [36]

- Instance-centric algorithms
- Constraint-centric algorithms
- Genetic-centric algorithms
- Regression-centric algorithms
- Game-centric algorithms
- Probabilistic-centric algorithms
- Case-based reasoning algorithms

Rule based explanations provide "if..then.." explanations for results. Even though these methods can be used post-hoc, they can be inherently transparent for a rule-based learner. [42] [48]

Explanation by simplification methods create a new simplified version of the trained AI/ML system for explanations. This reduces complexity and can simplify the explanation as well.

Explanation by knowledge extraction Explanation via knowledge extraction is done via two common approaches; decompositional and pedagogical [49]. Decompositional approaches extract knowledge rules directly form the model's structures and weights. Whereas pedagogical approaches extract knowledge from input-output pairings [49], [50]. An excellent example of pedagogical approaches is the novel tree induction algorithm introduced by Craven, et al., TREPAN [51]. It extracts decision trees from statistical classifiers.

TABLE II: An overview of different explanation methods and machine learning models.

XAI Model	Explainability Method	Machine Learning Model
Transparent Models		Decision Trees K-Nearest Neighbors Rule-based Learners General Additive Methods Bayesian Methods
Post-Hoc Explainability	Test Explanation Visual Explanation Global & Local Explanations Explanation by Example Explanation by Simplification Feature Relevance	Neural Networks Ensemble Methods, Classifier Systems, SVM, Neural Networks Decision Trees, Rule-based learners, Neural Networks Neural Networks Rule Based Learners, Decision trees, SVM, Probabilistic methods Ensemble methods, classifiers systems, SVM, Neural Networks

TABLE III: An overview of different available post-hoc explanation methods.

Explainable Method	Explainable Model	Model Specific/Agnostic	Global/Local Explanations
Feature Relevance	Shapley Values (SHAP)	Agnostic	Local
Local Explanation	Local Interpretable Model-Agnostic Explanations (LIME)	Agnostic	Local
Global Explanation	Skater	Agnostic	Global
Visual Explanation	Individual Conditional Expectation (ICE)	Specific	Both
Text Explanation	Visual Question Answering (VQA)	Agnostic	Local

Confalonieri, et al., expanded TREPAN, with their TREPAN Reloaded algorithm which included domain knowledge to enhance the understandability of surrogate decision trees. They used ontologies that model domain knowledge to generate better explanations. [52]

Feature relevance explanations generate a relevance score of the managed variables. The approach produces a comparison of the relative scores for each variable, providing the emphasis of each of the variables on the results generated by the system.

Global explanations are model explanations that articulate the operating procedures of the entire AI/ML system. These are meant to be thorough in their explanation of the entire system model. *Local explanations*, in contrast to global explanations, provide reasoning for only a section of the AI/ML system. They explain by dissecting the solution space and providing explanations for specific input/output pairs.

Global and local explanations are a higher level concept in comparison to the aforementioned explanation methods, and are mentioned here to classify between specific explanations within the ML pipeline and the entire ML system as a whole.

Both transparent and post-hoc explainable AI systems can be achieved via numerous available ML techniques including linear regression, decision trees, support vector machines, Bayesian models, and k-nearest neighbors. Some approaches are more transparent, and linear/logistic regression, decision trees, K-nearest neighbors, rule-based learners, Bayesian models and general additive models have been used due to their various levels of transparency [1], [53]–[60].

Post-hoc explainable techniques have also been studied extensively with a plethora of ML models being used for various applications. Arrieta, et al., presented a distinction within the post-hoc models consisting of model-agnostic and model specific post-hoc explainable methods [1]. Model-agnostic methods can be used with any ML model without the challenges associated with transferability. Whereas model specific post-hoc explainable methods are designed for specific ML models, these techniques can be applied to other ML models including deep learning (DL) models.

Riberio, et al., proposed the Local Interpretable Mode-Agnostic Explanations (LIME) technique that provides interpretable and trustworthy explainability of classifier predictions. LIME uses explanation by simplification and local explanation methods to generate a local interpretable model around the prediction [27] [61].

Genetic rule-extraction (G-rex) is a method for providing explanations by simplification via rule extraction from opaque models to increase the accuracy of comprehensible representations [62] [63].

Tan, et al., presented a "distill and compare" method for explanation by simplification of black-box models. Model distillation was done by training transparent models from the original black-box model to duplicate its results [64].

Lundberg, et al., presented a unified framework for interpreting predictions, SHapley Additive exPlanations (SHAP). SHAP provides explanations via feature relevance where an importance value is assigned to each feature for specific predictions. It provides additive feature importance values for accurate and consistent explainable predictions of how much each feature was involved in the system's decision/prediction [65].

Cortez, et al., presented visual explanation techniques for black-box models by using Sensitivity Analysis (SA) based visualization. They built upon an existing SA model to propose a Global SA (GSA) that extended the method's applications to numerous visualization techniques for the assessment of input relevance [66] [67].

Hugh, et al., presented DeepSHAP for explanations of complex models. It provides layer wise propagation of SHapley values for deep learning models [11].

Vazquez, et al., developed a compact support vector machine (SVM) model called growing support vector classifier, to give explanations with high fidelity and accuracy for decisions

made by SVM systems via input space segmentation in Voronoi selections. Voronoi selections of a feature are defined as *"the set of points that are closer to that feature than to any other*." [68].

Zilke, et al., presented the explanation by simplification method for deep learning models. The Deep neural network Rule Extraction via Decision tree induction (DeepRED) algorithm to extract rules form deep neural networks by adding more decision trees and rules. [69]

Che, et al., introduced the Interpretable Mimic learning (IML) approach for deep learning. They extracted interpretable models by using gradient boosting trees with predictions as strong as the original deep learning model. Their results showed excellent performance along with explanations for clinicians [70].

Shrikumar, et al. presented the DeepLIFT (Learning Important FeaTures) method for explanation of deep neural networks. The method provides importance scores for multilayer neural networks by calculating the distinction between the each neuron's activation and its reference activation [71] .

IV. XAI Evaluation

Different measures are needed to evaluate and verify the validity and performance of explanations given by XAI systems, that may be designed with different explanation goals. To this end, DARPA's XAI program assessed XAI systems using these measures [13].

- User Satisfaction
- Mental Model
- Task Performance
- Trust Assessment

User satisfaction measured the clarity and utility of the explanation based on the views of the end-user [13].Both subjective and objective approaches have been explored to measure the usefulness, understandability/comprehensibility, and end user satisfaction. Common approaches found in the literature are user-interviews, self-reporting questioners, Likert-scale questionnaires, and expert case studies. Studies by Bunt, et al., Gedikli, et al., and Lim, et al., employed user interviews to investigate their satisfaction and the most efficient ways to provide explanations [72]–[75]. Other studies such as Coppers, et al., and Lage, et al., use a Likert-scale questionnaire to quantify the user's satisfaction [76] [77].

Mental models are derived from the philosophical, psychological, and naturalistic models of human explanatory reasoning to measure the effectiveness of an explanation, which is the user's understanding of the system and the ability to predict its decisions in different situations [13].This aids in the users' decisions to either trust or doubt the AI/ML systems decisions, based on how much they understand/comprehend the system and how it came to a specific decision. These measures focus on understanding individual decisions, the overall model, strengths and weaknesses of the system, and what/how predictions. Different approaches have been used to evaluate how effective mental models are at measuring the user-understanding of the system, prediction of the system decision/results and the failures. Lombrozo suggested the importance of the feature's explanation which impacts the categorization and is critical to the understanding of the conceptual representation [78]. Lim, et al., studied the different types of explanations expected by users in different scenarios [74]. Penney, et al. and Rader, et al., investigated the users' interpretations of the AI/ML systems and their algorithms [79] [80]. Dodge, et al., Kim, et al., Kulesza, et al, and Lakkaraju, et al., employed user interviews and questionnaires to evaluate the mental models of the explanations [81]–[84]. Model output and failure predictions were also measured to evaluate the mental model in studies by Ribeiro, et al., Nushi, et al., and Bansal, et al. [27] [85]–[87].

Task performance for the XAI system measures whether the explanation improves the user's decision making or not, and also how well the user understands the XAI systems. User task performance was the evaluation of the user's performance for the designated task supported by the system. [13] Studies from Lim, et al., Lakkaraju, et al., Kahng, et al., Groce, et al., and Kulesza, et al., investigated the performance, throughput, and the prediction accuracy of the end users. [75] [84] [88]–[90] While other studies, such as the ones from Kulesza, et al., M. Liu, et al., S. Liu, et al., Stumpf, et al., evaluated the performance of the XAI system itself by measuring the model accuracy, tuning and selection. [91]–[94] Confalonieri, et al., presented task performance evaluations based on both subjective and objective methods for their proposed Trepan Reloaded algorithm. Objective evaluations were based on syntactic complexity of a decision tree, whereas subjective evaluations were based on user performance and ratings [95]

Trust Assessment in any AI system is of the utmost importance. For XAI systems specifically, the user's trust in the system is a measure of its effectiveness. Ultimately it is the evaluation of the user's ability to know when to trust or doubt the decisions made by the XAI systems. [13] Trust in these systems has been investigated in literature in various ways, including user knowledge, confidence, competence and use over time. Studies by Nourani, et al., and Ming, et al., investigated how the system's properties such as accuracy, precision, inclusion, and level of explanation affected user's trust on the system. [96] [97] Other studies have measured the trust based on subjective and objective measures such as interviews/questionnaires (subjective) and user's understandability, compliance and their perception of systems confidence (objective). [96]–[102]

V. XAI Security

XAI and cybersecurity are closely related. On the one hand, the XAI system needs to be secure, and on the other hand XAI may aid security.However, there is relatively little work on how to make XAI systems more robust, and how to protect them from adversarial attacks [103].

As one of the most famous quotes from Marvel's cinematic universe states "With great power, comes great responsibility." Explainable AI's explanations also bring about a great deal of responsibility for AI systems to generate precise and accurate explanations. Especially in time-critical applications such as

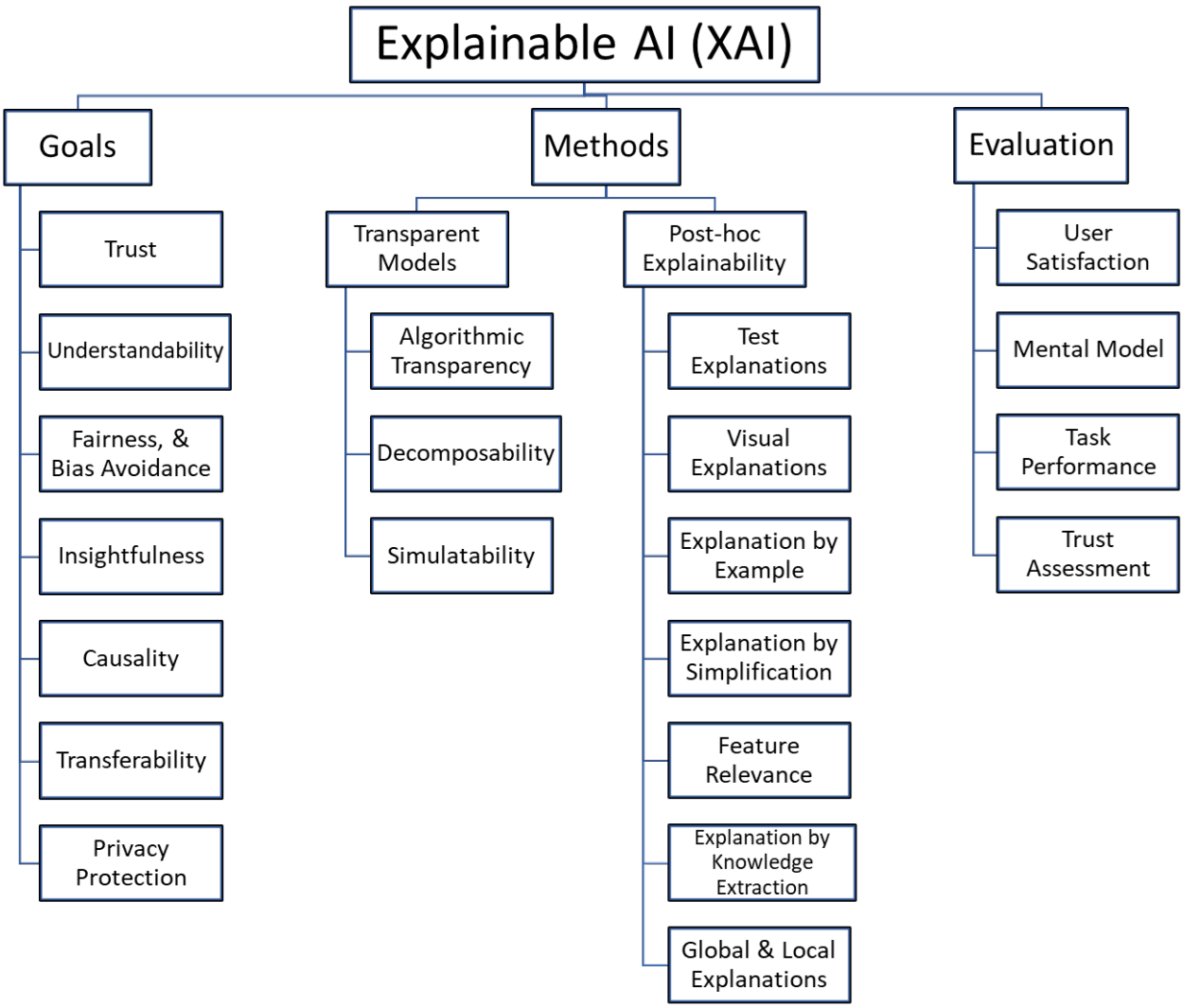

Fig. 5: Taxonomy for design and development of Explainable AI (XAI) systems.

medical or military. This is derived from the essential XAI goal of Trust, as false explanations will result in a complete loss or reduced trust in the system.

Due to the inherent white-box nature of XAI systems, whether they are transparent or post-hoc explanations, they are more susceptible adversarial attacks than black-box models. With explanations provided for not only the decisions/predictions of the AI/ML systems, but also their inner-workings, they can be easily manipulated for adversarial purposes. To this end, the security of XAI systems is of vital importance to protect them from adversarial attacks and perturbations leading to false and inaccurate explanations Therefore, it is essential to develop techniques to make them more robust and better protected against the attacks and exposure of any private/sensitive data.

The development of secure XAI systems likely requires a multi-faceted approach. Vigano, et al., introduced the concept of Explainable Security (XSec) for research on the security of XAI systems and provided a thorough review on how to secure these systems [104]. They proposed a multifaced approach for securing XAI systems using the “Six - W’s” : who, what, when, where, why and how as follows.

- Who gives and receives the explanation?
- What is explained?
- When is an explanation given?
- Where is the explanation given?
- Why is explainable security needed?
- How to explain security?

Vigano, et al., expanded on ”Six - W’s” very well, as each ”W” by itself holds major implications for the security of XAI systems. *who*, as listed above is concerned with the personnel involved with not just design and development of the XAI systems, but also the end-user, the possible adversarial attackers, the analyst for the systems and the security experts defending them against such attacks. As with any math equation, the larger the number of variables, the more complex systems become. And with almost anyone involved, becoming a vital part of the security for the XAI systems, it is a very complex topic that needs to probed further to gain valuable insights into securing XAI. For the *what*, explanations will defer in accordance to the stakeholders, aims and the level of details

needed. Several parameters will also influence the explanations such as the system model, its properties, threat model and vulnerabilities. *When* the explanation is provided will also play a vital role in XAI security. All the security aspects of the XAI system will need to be defined during all major phases of design, development, deployment, and defense. *where* the explanation is given will also impact the security of the XAI system. The explanation could be treated as its entity from the AI system and be separated and delivered. The authors believe the best-option forward, would be a "security-explanation-carrying-system", which requires a significant amount of work to secure the explanation. The *how* will depend on the XAI system itself, it will have to be explained in a method suited for the specific stakeholder. Finally, the *why* seems like an obvious question, as XAI systems will no-doubt need to be secured to protect the end-user and their privacy. [104]

Kuppa, et al., also presented a taxonomy for XAI in relation to cybersecurity. They proposed three different approaches with a) X-PLAIN – explanations of the predictions/data, b) XSP – PLAIN – explanations for security and privacy, and finally, c) XT- PLAIN – explanation for the threat models [105].

Additionally, protecting the confidentiality, integrity, and availability of XAI systems (the so-called *CIA* principles) is crucial for their practical deployment. As adversarial learning techniques grow more advanced and robust against current ML and DL techniques we must assume that attacks will be forthcoming against XAI systems. Due to their innate sensitivity, ranging from their learning datasets to the decisions/recommendations made, securing them against any and all perturbations to the data, learning models, and biases is critical for XAI [12]. Xu, et al., presented adversarial perturbations for misleading classifiers and causing variations to the network interpretability maps [106]. Ghorbani, et at., demonstrated the fragility of deep learning explanations when two identical images with minute perturbations can lead to different explanations [107]. Mittlelstadt, et al., demonstrated the vulnerabilities of the available XAI algorithms such as LIME and SHAP [108]. Kuppa, et al., presented a black-box attack on gradient-based XAI systems [105].

For a more realistic scenario where attackers don't have knowledge of the network architecture, model inputs and weights are manipulated to attack XAI. Heo, et al., demonstrated the vulnerabilities of state of the art saliency-map based systems by fooling the system with adversarial model manipulation [109]. They were able to a change the explanations given by the system without affecting its accuracy, by incorporating the explanations directly within the penalty term of the objective function. They proposed two different types of "fooling" attacks, passive and active. *Passive fooling* causes the XAI systems to generate uninformative results, whereas *active fooling* generates false explanations.

Another common attack method is to attack the input data itself to alter the explanation given by the system. Dombrowski, et al., demonstrated that adversarial manipulations of the input data can drastically change the explanation maps [110]. The authors also demonstrated methods to make the XAI systems more robust from the insights they gained by attacking. They were able to increase system resiliency to attacks by smoothing the explanation process.

As robust and trustworthy AI/ML systems require privacy and transparency as foundational pillars, the trade-off between explainability and privacy preservation is another major concern within XAI security. While the explanations help the user understand the systems decisions/predictions, privacy is of the utmost importance for protecting sensitive information. Existing studies have shown the vulnerability of transparent and explainable models to leak such sensitive data. [111]–[114] Shokri, et al., explored the privacy related risks of explainable ML models via the use of membership inference attacks. They demonstrated the significant privacy leakage from propagation-based explanations by revealing statistical information about the decision boundaries of the model. Additionally, they quantified the leakage of private information based on the model predictions and their explanations. Privacy-preserving algorithms, such as the ones by Agarwal, et al., Aggrawal et al., and Zhong, et al., for AI/ML systems will play a major role in making XAI systems more robust [115]–[117]. Harder, et al., presented simple interpretable models to approximate complex models via locally linear maps to achieve a high classification accuracy, while also preserving the privacy of the model [118]. Quantifying the trade-off between privacy and explanations will provide insightful details into how far explanations can be taken without risking the system's privacy.

Motivated by ensemble defense techniques for robust machine learning models, Rieger, et al., proposed a simple yet effective technique of combining explanation methods, AGG-Mean (Aggregated Explanations), to make the XAI system more robust adversarial manipulation [119]. Their method was effective against white-box attacks where the adversaries have the exact knowledge of the model.

While the security of these XAI systems remains a challenge for the field, the use of these systems for cybersecurity purposes also remains to be properly evaluated. Their inherent nature makes them an excellent option for securing AI/ML systems where explanations are crucial in identifying and defending against different types of attacks. If explanations are provided for adversarial attacks, they become easier to defend against.

Mahbooba, et al., demonstrated the use of XAI to reinforce an intrusion detection system (IDS) via decision trees [120]. Using simple if…then decision tree rules with logical conditions, the authors were able to distinguish between normal network traffic and malicious traffic. The rules, which are explainable aide the security personnel to take the proper course of action against incoming adversarial attacks.

Another example of XAI systems for cybersecurity is presented by Islam, et al., [121]. The authors proposed a domain knowledge aided XAI system for better explainability for an IDS. The infusion of CIA principles in the XAI-based black box model provided better explainability and generalizability. This was shown effective in detecting adversarial attacks, even unknown attacks. A major advantage of this work was the finding that it can accommodate big data.

Rao, et al., presented a novel new approach for protecting systems against the alarm flooding problem. By using explana-

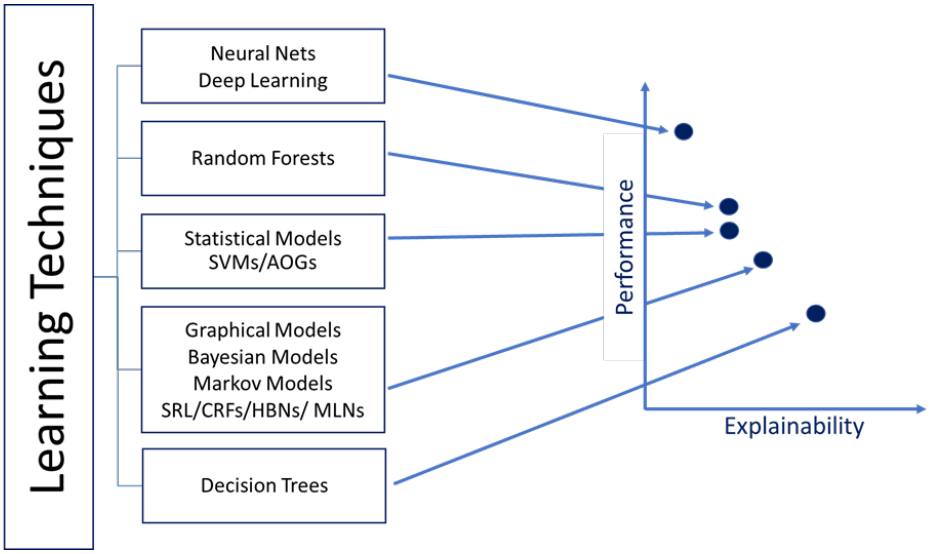

Fig. 6: A comparison of different machine learning techniques' explainability and performance (as presented by DARPA).

tions for anomalies, they applied a zero-shot method for detecting alarm labels generated by security information and event management(SIEM) and intrusion detection systems(IDS) to match them to specific adversarial attacks on the systems. XAI is used to characterize the incoming attacks into specific categories based on the attack's feature [122].

Mane, et al., presented a deep neural network model combined with XAI for intrusion detection. XAI algorithms, SHAP, LIME, Contrastive Explanations Method (CEM), ProtoDash and Boolean Decision Rules via Column Generation (BRCG) are used to generate explanations on which features influences the predictions of the IDS system for an impending attack [123]. Marino, et al., demonstrated an adversarial XAI approach for misclassifications made by IDS. Minimum perturbations to correct misclassified samples into accurate classifications are made to generate explanations for the misclassifications of the samples in the first place [124]

XAI systems have also been used to carryout various types of cyberattacks. Kuppa, et al., presented four different explanation based black-box attacks to compromise the CIA principles of the classifiers. They presented privacy attacks with Explanation-based model extraction, and Explanation-based membership inference attacks. Evasion attacks were performed via Explanation-based poisoning attacks and Explanation based adversarial sample generation attacks. Evasion attacks were demonstrated on commercial anti-virus systems, while membership inference attacks were used to extract user passwords. They also provided possible defenses against XAI-based attacks such as adversarial training, input/network randomization [125]. Garcia, et al., also demonstrated the use of XAI for adversarial attacks against host fingerprinting and biometric authentication systems. XAI was used to extract decision boundaries from an oracle, and determine the most relevant features within the model. This was done without the need or any prior information about the potential victims. [126] These types of studies have done an excellent job highlighting the risks that XAI pose to both the users and attackers.

VI. Open Challenges & Perspectives

Even though remarkable strides have been made in both AI/ML systems and XAI itself, numerous challenges still remain. These include transferability of the post-hoc explainability methods, the lack of universally adopted definition, standards, and measures for the explainability of AI/ML systems, the balance between explainability and performance, and the challenges of making deep learning models explainable.

Explainability vs Performance - The trade-off for the balance of explainability and performance is also a major issue. As deep learning models become more and more complex and successful at solving learning problems, their inherent "non-transparency" presents a major challenge in making them explainable for XAI purposes. As stated by Rudin, higher complexity does not inherently mean higher accuracy, and this has been very true for such DL models. [127] As shown in Fig 6, ML models with higher performance for prediction accuracy have the lower explainability performance. Thus, more research needs to focus on improving the performance and higher accuracy of these systems. There must be an optimal balance for which both the systems performance and explainability are accepted.

Lack of a universal standard - One of the major challenges within the field of XAI is terminology and ambiguity of definitions.As shown in the earlier sections, numerous terms are used when trying to articulate explainability to an AI/ML system. Furthermore, terms like interpretability, understandability and comprehensibility have been used as synonyms and only in the past few years have the terms taken on distinct meanings.However, a lack of a standard unified definition for the theory of explainability is noted.A unifying framework will provide common ground for researchers to contribute towards the properly defined needs and challenges of the field. Also metrics, other than simple interviews and questionnaires, are needed for measuring and evaluating the effectiveness of XAI. A study by Hoffman, et al., presented one of the only evaluation metrics for measuring the explanations of AI/ML

systems [128]. To this end, survey papers such as the ones by Arrieta, et al., Mohseni, et al., and the one presented in this paper will aid the overall development of XAI as an emerging new field [1], [8].

Fairness of AI - Another major concern for XAI coincides with one of the vital reasons/goals for the creations of such explainable systems: fairness and bias detection. As the fields of accountable AI and XAI were born out of a need for fair and unbiased decision making that affects human lives, getting rid of such biases remains a challenge within this young field. Benjamins, et al., noted that the discipline of fairness in AI inherently includes bias detection. Proposals for datasets with private and sensitive data may disproportionately affect underrepresented groups [129]. These datasets, when used for training black-box models such as DL systems, can result in biased decisions, which can cause discriminatory, unethical and unfair issues [130].

In addition to datasets, other sources of biases can include limited features, disparities in sample sizes and proxy features [131]. Different techniques have been proposed to mitigate biases within XAI systems. Kamiran, et al., proposed a pre-processing technique for the learning dataset by reweighing it, to eliminate discrimination [132]. Zemel, et al., presented a technique to achieve fairness via optimizing the representation of the data that presents the best encoding while also obscuring some parts to protect the membership information [133]. Other approaches included techniques such as adversarial de-biasing during data processing, equalized odds for post-processing, and bias detection techniques [131] [134]–[136].

Transferability - of post-hoc explanation methods is also a vital challenge. Transferability of post-hoc explainability techniques remains one of the most challenging issues. Most post-hoc techniques are designed to explain specific AI/ML models/systems. While some techniques successfully explain certain models, they may be deemed difficult and perform poorly when explaining other models. These post-hoc techniques are typically very much intertwined with the particular ML model and network architecture. There is a need for more generalized methods and AI/ML designs that are inherently explainable with different post-hoc methods. Deep Learning methods for example are very difficult to explain due to their black-box nature.

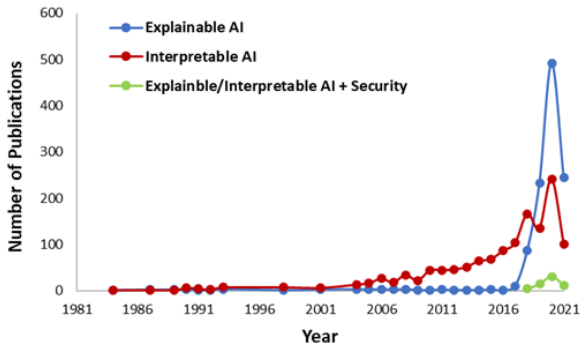

Fig. 7: A comparison of studies for XAI security against studies for XAI. (Data derived from SCOPUS)

XAI Security - Finally, as explained in the previous section, the security of XAI also remains a major challenge. Due to the infancy of the field, major work is being done in improving the explainability to bring it up to par with performance of the model. While this is a crucial step forward for the development and practical deployment of XAI systems, their security cannot be ignored. As shown in Figure 7, the amount of research being done for the security of XAI systems is very limited. The number of publications for XAI compared to XAI and Security is almost negligible. Therefore, if these systems are to be used for both civilian and military purposes, they must be made robust and resilient against adversarial attacks. The field of adversarial machine learning (AML) grows and progresses to include efficient attacks against most AI/ML models.The goal of robust AI/ML systems is coupled with the goal of making them explainable. Using the explanations from the system to detect and defend them against different adversarial attacks may play a crucial role in overall performance and successful application.

Semantics - In addition to the previously mentioned concepts, semantics also plays an integral role in XAI. Confalonieri, et al. emphasized explanations that can support common-sense reasoning when based upon ontologies, conceptual networks or knowledge graphs. They also stated the importance of these semantic methods for the development of AI/ML systems capable of providing stakeholder specific explanations [49].
Neural-symbolic learning and reasoning, in regard to semantics, also will play a major role within XAI. It is an interdisciplinary fusion of different (research subjects/topics) for generation of better explanations. Garcez, et al., stated, “neural-symbolic reasoning seeks to integrate principles from neural networks learning and logical reasoning.” [137] They state the goal of neural-symbolic reasoning is to “integrate robust connectionist learning and sound symbolic reasoning.” For neural-networks, neural-symbolic computation can provide dynamic alternatives for knowledge representation, learning and reasoning. Garcez, et al., presented the effectiveness of neural-symbolic computing by highlighting its characteristic as the ”integration of neural learning with symbolic knowledge representation and reasoning allowing for the construction of explainable AI systems.” [138] Borges, et al., presented a novel neural-computation model for neural networks that is capable of learning and representing temporal knowledge. The model extracts temporal knowledge from trained networks via effective representation, adoption of the temporal models and learning from examples [139]. de Penning, et al., introduced a novel model for online learning and reasoning in complex training environments, capable fo learning new hypotheses from observed data and making recommendations based on them via the combination of neural learning and symbolic representation [140]

VII. Summary

XAI will play an important role in the development and application of AI/ML systems.In this paper we presented a taxonomy and literature survey of Explainable AI (XAI). We

defined terms associated with the field and laid out goals and methods for the design and development of trustworthy XAI systems, including robustness and security against adversarial attack. A variety of challenges were also described.

ACKNOWLEDGMENTS

This work was supported in part by the DoD Center of Excellence in AI and Machine Learning (CoE-AIML) at Howard University under Contract Number W911NF-20-2-0277 with the U.S. Army Research Laboratory.

REFERENCES

[1] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[2] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[3] J. Dastin, J. Weber, and M. Dickerson, "Amazon scraps secret ai recruiting tool that showed bias against women," *Reuters*, 2018.

[4] K. Hao, "Facebook's ad-serving algorithm discriminates by gender and race," *MIT Technology Review*, 2019. [Online]. Available: https://www.technologyreview.com/2019/04/05/1175/facebook-algorithm-discriminates-ai-bias/

[5] H. Ledford, "Millions of black people affected by racial bias in health-care algorithms," *Nature*, vol. 574, no. 7780, pp. 608–609, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/31664201

[6] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, p. 3152676, 2017.

[7] M. MacCarthy, "An examination of the algorithmic accountability act of 2019," *SSRN Electronic Journal*, 2019.

[8] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *arXiv preprint arXiv:1811.11839*, 2018.

[9] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[10] A. Weller, "Transparency: motivations and challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 23–40.

[11] H. Chen, S. Lundberg, and S.-I. Lee, "Explaining models by propagating shapley values of local components," in *Explainable AI in Healthcare and Medicine*. Springer, 2021, pp. 261–270.

[12] F. Hussain, R. Hussain, and E. Hossain, "Explainable artificial intelligence (xai): An engineering perspective," *arXiv preprint arXiv:2101.03613*, 2021.

[13] D. Gunning and D. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[14] M. Busuioc, "Accountable artificial intelligence: Holding algorithms to account," *Public Administration Review*, 2020.

[15] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[16] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, 2018.

[17] M. Gleicher, "A framework for considering comprehensibility in modeling," *Big data*, vol. 4, no. 2, pp. 75–88, 2016.

[18] A. A. Freitas, "Comprehensible classification models: a position paper," *ACM SIGKDD Explorations Newsletter*, vol. 15, no. 1, pp. 1–10, 2014.

[19] A. Fernandez, F. Herrera, O. Cordon, M. J. del Jesus, and F. Marcelloni, "Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to?" *IEEE Computational Intelligence Magazine*, vol. 14, no. 1, pp. 69–81, 2019.

[20] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[21] L. Edwards and M. Veale, "Slave to the algorithm? why a 'right to an explanation' is probably not the remedy you are looking for," *Duke law and technology review*, vol. 16, pp. 18–84, 2017.

[22] A. Vellido, J. D. Martín-Guerrero, and P. J. Lisboa, "Making machine learning models interpretable." in *ESANN*, vol. 12. Citeseer, 2012, pp. 163–172.

[23] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[24] D. Doran, S. Schulz, and T. R. Besold, "What does explainable ai really mean? a new conceptualization of perspectives," *arXiv preprint arXiv:1710.00794*, 2017.

[25] A. Chander, R. Srinivasan, S. Chelian, J. Wang, and K. Uchino, "Working with beliefs: Ai transparency in the enterprise," in *IUI Workshops*, 2018.

[26] J. Haspiel, N. Du, J. Meyerson, L. P. Robert Jr, D. Tilbury, X. J. Yang, and A. K. Pradhan, "Explanations and expectations: Trust building in automated vehicles," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 2018, pp. 119–120.

[27] M. T. Ribeiro, S. Singh, and C. Guestrin, "" 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*, 2016, pp. 1135–1144.

[28] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[29] B. Lim, "Improving understanding, trust, and control with intelligibility in context-aware applications," *Human-Computer Interaction*, 2011.

[30] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, vol. 5, no. 2, pp. 153–163, 2017.

[31] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Díaz-Rodríguez, "Towards explainable neural-symbolic visual reasoning," *arXiv preprint arXiv:1909.09065*, 2019.

[32] H.-X. Wang, L. Fratiglioni, G. B. Frisoni, M. Viitanen, and B. Winblad, "Smoking and the occurence of alzheimer's disease: Cross-sectional and longitudinal data in a population-based study," *American journal of epidemiology*, vol. 149, no. 7, pp. 640–644, 1999.

[33] J. Pearl, "Structural and probabilistic causality," *Psychology of learning and motivation*, vol. 34, pp. 393–435, 1996.

[34] P. Rani, C. Liu, N. Sarkar, and E. Vanman, "An empirical study of machine learning techniques for affect recognition in human–robot interaction," *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, 2006.

[35] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 9, no. 4, p. e1312, 2019.

[36] Y.-L. Chou, C. Moreira, P. Bruza, C. Ouyang, and J. Jorge, "Counterfactuals and causability in explainable artificial intelligence: Theory, algorithms, and applications," *arXiv preprint arXiv:2103.04244*, 2021.

[37] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *arXiv preprint arXiv:1706.02744*, 2017.

[38] A. Holzinger, A. Carrington, and H. Müller, "Measuring the quality of explanations: the system causability scale (scs)," *KI-Künstliche Intelligenz*, pp. 1–6, 2020.

[39] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[40] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[41] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[42] J. van der Waa, E. Nieuwburg, A. Cremers, and M. Neerincx, "Evaluating xai: A comparison of rule-based and example-based explanations," *Artificial Intelligence*, vol. 291, p. 103404, 2021.

[43] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2012, pp. 150–158.

[44] A. Datta, S. Sen, and Y. Zick, "Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 598–617.

[45] K. Kawaguchi, "Deep learning without poor local minima," *arXiv preprint arXiv:1605.07110*, 2016.

[46] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 344–350.

[47] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harv. JL & Tech.*, vol. 31, p. 841, 2017.

[48] B. M. Keneni, D. Kaur, A. Al Bataineh, V. K. Devabhaktuni, A. Y. Javaid, J. D. Zaientz, and R. P. Marinier, "Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles," *IEEE Access*, vol. 7, pp. 17 001–17 016, 2019.

[49] R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A historical perspective of explainable artificial intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 1, p. e1391, 2021.

[50] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[51] M. Craven and J. Shavlik, "Extracting tree-structured representations of trained networks," *Advances in neural information processing systems*, vol. 8, pp. 24–30, 1995.

[52] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. d. P. Martín, "Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks," *arXiv preprint arXiv:1906.08362*, 2019.

[53] U. Hoffrage and G. Gigerenzer, "Using natural frequencies to improve diagnostic inferences," *Academic medicine*, vol. 73, no. 5, pp. 538–540, 1998.

[54] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *The journal of educational research*, vol. 96, no. 1, pp. 3–14, 2002.

[55] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[56] S. Jiang, G. Pang, M. Wu, and L. Kuang, "An improved k-nearest-neighbor algorithm for text categorization," *Expert Systems with Applications*, vol. 39, no. 1, pp. 1503–1509, 2012.

[57] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "An knn model-based approach and its application in text categorization," in *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2004, pp. 559–570.

[58] U. Johansson, R. König, and L. Niklasson, "The truth is in there-rule extraction from opaque models using genetic programming." in *FLAIRS Conference*. Miami Beach, FL, 2004, pp. 658–663.

[59] H. Nunez, C. Angulo, and A. Catala, "Rule-based learning systems for support vector machines," *Neural Processing Letters*, vol. 24, no. 1, pp. 1–18, 2006.

[60] B. Kim, C. Rudin, and J. A. Shah, "The bayesian case model: A generative approach for case-based reasoning and prototype classification," in *Advances in neural information processing systems*, 2014, pp. 1952–1960.

[61] M. T. Ribeiro, S. Singh, and C. Guestrin, "Nothing else matters: model-agnostic explanations by identifying prediction invariance," *arXiv preprint arXiv:1611.05817*, 2016.

[62] R. Konig, U. Johansson, and L. Niklasson, "G-rex: A versatile framework for evolutionary data mining," in *2008 IEEE International Conference on Data Mining Workshops*. IEEE, 2008, pp. 971–974.

[63] U. Johansson, L. Niklasson, and R. König, "Accuracy vs. comprehensibility in data mining models," in *Proceedings of the seventh international conference on information fusion*, vol. 1. Citeseer, 2004, pp. 295–300.

[64] S. Tan, R. Caruana, G. Hooker, and Y. Lou, "Distill-and-compare: Auditing black-box models using transparent model distillation," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 303–310.

[65] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *arXiv preprint arXiv:1705.07874*, 2017.

[66] P. Cortez and M. J. Embrechts, "Opening black box data mining models using sensitivity analysis," in *2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*. IEEE, 2011, pp. 341–348.

[67] ——, "Using sensitivity analysis and visualization techniques to open black box data mining models," *Information Sciences*, vol. 225, pp. 1–17, 2013.

[68] A. Navia-Vázquez and E. Parrado-Hernández, "Support vector machine interpretation," *Neurocomputing*, vol. 69, no. 13-15, pp. 1754–1759, 2006.

[69] J. R. Zilke, E. L. Mencía, and F. Janssen, "Deepred–rule extraction from deep neural networks," in *International Conference on Discovery Science*. Springer, 2016, pp. 457–473.

[70] Z. Che, S. Purushotham, R. Khemani, and Y. Liu, "Interpretable deep models for icu outcome prediction," in *AMIA annual symposium proceedings*, vol. 2016. American Medical Informatics Association, 2016, p. 371.

[71] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[72] F. Gedikli, D. Jannach, and M. Ge, "How should i explain? a comparison of different explanation types for recommender systems," *International Journal of Human-Computer Studies*, vol. 72, no. 4, pp. 367–382, 2014.

[73] A. Bunt, M. Lount, and C. Lauzon, "Are explanations always important? a study of deployed, low-cost intelligent interactive systems," in *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, ser. IUI '12. New York, NY, USA: Association for Computing Machinery, 2012, p. 169–178.

[74] B. Y. Lim and A. K. Dey, "Assessing demand for intelligibility in context-aware applications," in *Proceedings of the 11th international conference on Ubiquitous computing*, 2009, pp. 195–204.

[75] B. Y. Lim, A. K. Dey, and D. Avrahami, "Why and why not explanations improve the intelligibility of context-aware intelligent systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, pp. 2119–2128.

[76] S. Coppers, J. Van den Bergh, K. Luyten, K. Coninx, I. Van der Lek-Ciudin, T. Vanallemeersch, and V. Vandeghinste, "Intellingo: An intelligible translation environment," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[77] I. Lage, E. Chen, J. He, M. Narayanan, B. Kim, S. J. Gershman, and F. Doshi-Velez, "Human evaluation of models built for interpretability," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, 2019, pp. 59–67.

[78] T. Lombrozo, "Explanation and categorization: How "why?" informs "what?"," *Cognition*, vol. 110, no. 2, pp. 248–253, 2009.

[79] S. Penney, J. Dodge, C. Hilderbrand, A. Anderson, L. Simpson, and M. Burnett, "Toward foraging for understanding of starcraft agents: An empirical study," in *23rd International Conference on Intelligent User Interfaces*, 2018, pp. 225–237.

[80] E. Rader and R. Gray, "Understanding user beliefs about algorithmic curation in the facebook news feed," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 173–182.

[81] J. Dodge, S. Penney, A. Anderson, and M. M. Burnett, "What should be in an xai explanation? what ift reveals." in *IUI Workshops*, 2018.

[82] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[83] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W.-K. Wong, "Too much, too little, or just right? ways explanations impact end users' mental models," in *2013 IEEE Symposium on visual languages and human centric computing*. IEEE, 2013, pp. 3–10.

[84] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675–1684.

[85] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[86] B. Nushi, E. Kamar, and E. Horvitz, "Towards accountable ai: Hybrid human-machine analyses for characterizing system failure," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 6, no. 1, 2018.

[87] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-ai team performance," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, 2019, pp. 2–11.

[88] M. Kahng, P. Y. Andrews, A. Kalro, and D. H. Chau, "Activis: Visual exploration of industry-scale deep neural network models," *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 88–97, 2017.

[89] A. Groce, T. Kulesza, C. Zhang, S. Shamasunder, M. Burnett, W.-K. Wong, S. Stumpf, S. Das, A. Shinsel, F. Bice *et al.*, "You are the only possible oracle: Effective test selection for end users of interactive machine learning systems," *IEEE Transactions on Software Engineering*, vol. 40, no. 3, pp. 307–323, 2013.

[90] T. Kulesza, S. Stumpf, M. Burnett, W.-K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel, and K. McIntosh, "Explanatory debugging: Supporting end-user debugging of machine-learned programs," in *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 2010, pp. 41–48.

[91] T. Kulesza, M. Burnett, W.-K. Wong, and S. Stumpf, "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th international conference on intelligent user interfaces*, 2015, pp. 126–137.

[92] S. Liu, X. Wang, J. Chen, J. Zhu, and B. Guo, "Topicpanorama: A full picture of relevant topics," in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2014, pp. 183–192.

[93] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu, "Towards better analysis of deep convolutional neural networks," *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 91–100, 2016.

[94] S. Stumpf, V. Rajaram, L. Li, W.-K. Wong, M. Burnett, T. Dietterich, E. Sullivan, and J. Herlocker, "Interacting meaningfully with machine learning systems: Three experiments," *International journal of human-computer studies*, vol. 67, no. 8, pp. 639–662, 2009.

[95] R. Confalonieri, T. Weyde, T. R. Besold, and F. M. del Prado Martín, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, p. 103471, 2021.

[96] M. Yin, J. Wortman Vaughan, and H. Wallach, "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–12.

[97] M. Nourani, S. Kabir, S. Mohseni, and E. D. Ragan, "The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 7, no. 1, 2019, pp. 97–105.

[98] P. Pu and L. Chen, "Trust building with explanation interfaces," in *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 93–100.

[99] F. Nothdurft, F. Richter, and W. Minker, "Probabilistic human-computer trust handling," in *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, 2014, pp. 51–59.

[100] M. Eiband, D. Buschek, A. Kremer, and H. Hussmann, "The impact of placebic explanations on trust in intelligent systems," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–6.

[101] S. Berkovsky, R. Taib, and D. Conway, "How to recommend? user trust factors in movie recommender systems," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017, pp. 287–300.

[102] A. Bussone, S. Stumpf, and D. O'Sullivan, "The role of explanations on trust and reliance in clinical decision support systems," in *2015 international conference on healthcare informatics*. IEEE, 2015, pp. 160–169.

[103] G. Fidel, R. Bitton, and A. Shabtai, "When explainability meets adversarial learning: Detecting adversarial examples using shap signatures," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[104] L. Viganò and D. Magazzeni, "Explainable security," in *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2020, pp. 293–300.

[105] A. Kuppa and N.-A. Le-Khac, "Black box attacks on explainable artificial intelligence (xai) methods in cyber security," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[106] K. Xu, S. Liu, P. Zhao, P.-Y. Chen, H. Zhang, Q. Fan, D. Erdogmus, Y. Wang, and X. Lin, "Structured adversarial attack: Towards general implementation and better interpretability," *arXiv preprint arXiv:1808.01664*, 2018.

[107] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[108] B. Mittelstadt, C. Russell, and S. Wachter, "Explaining explanations in ai," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 279–288.

[109] J. Heo, S. Joo, and T. Moon, "Fooling neural network interpretations via adversarial model manipulation," *arXiv preprint arXiv:1902.02041*, 2019.

[110] A.-K. Dombrowski, M. Alber, C. J. Anders, M. Ackermann, K.-R. Müller, and P. Kessel, "Explanations can be manipulated and geometry is to blame," *arXiv preprint arXiv:1906.07983*, 2019.

[111] R. Shokri, M. Strobel, and Y. Zick, "On the privacy risks of model explanations," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 2021, pp. 231–241.

[112] J. Domingo-Ferrer, C. Pérez-Solà, and A. Blanco-Justicia, "Collaborative explanation of deep models with limited interaction for trade secret and privacy preservation," in *Companion Proceedings of The 2019 World Wide Web Conference*, 2019, pp. 501–507.

[113] T. Budig, S. Herrmann, and A. Dietz, "Trade-offs between privacy-preserving and explainable machine learning in healthcare."

[114] H. Chang and R. Shokri, "On the privacy risks of algorithmic fairness," *arXiv preprint arXiv:2011.03731*, 2020.

[115] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2001, pp. 247–255.

[116] C. C. Aggarwal and S. Y. Philip, "A general survey of privacy-preserving data mining models and algorithms," in *Privacy-preserving data mining*. Springer, 2008, pp. 11–52.

[117] S. Zhong, "Privacy-preserving algorithms for distributed mining of frequent itemsets," *Information Sciences*, vol. 177, no. 2, pp. 490–503, 2007.

[118] F. Harder, M. Bauer, and M. Park, "Interpretable and differentially private predictions," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 4083–4090.

[119] L. Rieger and L. K. Hansen, "A simple defense against adversarial attacks on heatmap explanations," *arXiv preprint arXiv:2007.06381*, 2020.

[120] B. Mahbooba, M. Timilsina, R. Sahal, and M. Serrano, "Explainable artificial intelligence (xai) to enhance trust management in intrusion detection systems using decision tree model," *Complexity*, vol. 2021, 2021.

[121] S. R. Islam, W. Eberle, S. K. Ghafoor, A. Siraj, and M. Rogers, "Domain knowledge aided explainable artificial intelligence for intrusion detection and response," *arXiv preprint arXiv:1911.09853*, 2019.

[122] D. Rao and S. Mane, "Zero-shot learning approach to adaptive cybersecurity using explainable ai," *arXiv preprint arXiv:2106.14647*, 2021.

[123] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable ai framework," *arXiv preprint arXiv:2103.07110*, 2021.

[124] D. L. Marino, C. S. Wickramasinghe, and M. Manic, "An adversarial approach for explainable ai in intrusion detection systems," in *IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society*. IEEE, 2018, pp. 3237–3243.

[125] A. Kuppa and N.-A. Le-Khac, "Adversarial xai methods in cybersecurity," *IEEE Transactions on Information Forensics and Security*, 2021.

[126] W. Garcia, J. I. Choi, S. K. Adari, S. Jha, and K. R. Butler, "Explainable black-box attacks against model-based authentication," *arXiv preprint arXiv:1810.00024*, 2018.

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

[128] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[129] R. Benjamins, A. Barbado, and D. Sierra, "Responsible ai by design in practice," *arXiv preprint arXiv:1909.12838*, 2019.

[130] B. d'Alessandro, C. O'Neil, and T. LaGatta, "Conscientious classification: A data scientist's guide to discrimination-aware classification," *Big data*, vol. 5, no. 2, pp. 120–134, 2017.

[131] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *arXiv preprint arXiv:1610.02413*, 2016.

[132] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, 2012.

[133] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International conference on machine learning*. PMLR, 2013, pp. 325–333.

[134] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 335–340.

[135] Y. Ahn and Y.-R. Lin, "Fairsight: Visual analytics for fairness in decision making," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1086–1095, 2019.

[136] E. Soares and P. Angelov, "Fair-by-design explainable models for prediction of recidivism," *arXiv preprint arXiv:1910.02043*, 2019.

[137] A. d. Garcez, T. R. Besold, L. De Raedt, P. Földiak, P. Hitzler, T. Icard, K.-U. Kühnberger, L. C. Lamb, R. Miikkulainen, and D. L. Silver, "Neural-symbolic learning and reasoning: contributions and challenges," in *2015 AAAI Spring Symposium Series*, 2015.

[138] A. d. Garcez, M. Gori, L. C. Lamb, L. Serafini, M. Spranger, and S. N. Tran, "Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning," *arXiv preprint arXiv:1905.06088*, 2019.

[139] R. V. Borges, A. d. Garcez, and L. C. Lamb, "Learning and representing temporal knowledge in recurrent networks," *IEEE Transactions on Neural Networks*, vol. 22, no. 12, pp. 2409–2421, 2011.

[140] H. L. H. de Penning, A. S. d. Garcez, L. C. Lamb, and J.-J. C. Meyer, "A neural-symbolic cognitive agent for online learning and reasoning," in *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.