Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 45–52
DOI: 10.15439/2024F0001
ISSN 2300-5963 ACSIS, Vol. 39

Attentiveness on criticisms and definition about Explainable Artificial Intelligence

Francisco Herrera
Dept. of Computer Science and Artificial Intelligence
Andalusian Institute of Data Science and Computational Intelligence (DaSCI)
University of Granada, Spain.
Email: herrera@decsai.ugr.es

***Abstract*—The emergence of deep learning at the beginning of the last decade has driven the advancement of complex models, culminating in the development of large language models and generative AI. These models represent the summit of size and complexity. Explainability should be an option that plays a key role in enabling understandable the AI-assisted decision-making and ensuring accountability. This contribution delves into the complexities of explainable artificial intelligence (XAI) through various perspectives, considering the extensive and growing body of literature. Our discussion begins by addressing the challenges posed by complex data, models, and high-risk scenarios. Given the rapid growth of the field, it is essential to tackle the criticisms and challenges that emerge as it matures, requiring thorough exploration. This contribution explores them, along with three aspects that may shed light on them. First, it is focused on the lack of definitional cohesion, examining how and why is defined XAI from the perspectives of audience and understanding. Second, it explores XAI explanations, bridging the gap between complex AI models and human understanding. Third, it is crucial to consider how to analyze and evaluate the maturity level of explainability, from a triple dimension, practicality, governance and auditability.**

***Index Terms*—eXplainable Artificial Intelligence, explanations, metrics, audience.**

I. Introduction

IN RECENT years, the rapid advancement of artificial intelligence (AI) has led to the development of increasingly complex models capable of performing tasks with remarkable accuracy. However, the opacity of these models, often referred to as "black-box AI" [1], has raised significant concerns regarding their interpretability and trustworthiness. We work with complex data, complex black box models and complex scenarios dealing with high risks problems. Explainable AI (XAI) has emerged as a critical field of research aimed at addressing these concerns by providing transparent and understandable explanations for the AI-assisted decision-making. AI-assisted decision-making refers to the process where AI systems provide recommendations or insights to help humans make decisions.

The European Union greenlit the first major AI law, AI Act[1] in december 2023, approved in march 2024, and published on july 2024. It will regulate the development, use, and application of AI. Its goal is to ensure AI systems used and developed in the EU are safe and trustworthy. "The adoption of the AI Act is a significant milestone for the European Union. This landmark law, the first of its kind in the world, addresses a global technological challenge that also creates opportunities for our societies and economies. With the AI Act, Europe emphasizes the importance of trust, transparency and accountability when dealing with new technologies while at the same time ensuring this fast-changing technology can flourish and boost European innovation." said recently on the occasion of the approval Mathieu Michel, Belgian secretary of state for digitisation, administrative simplification, privacy protection, and the building regulation[2].

Explainability is both in the fundamental principles associated with the European trustworhty AI definition[3] (respect for human autonomy, prevention of harm, equity, and explainability), and in UNESCO's ethical principle[4], number 7; Transparency and explainability. It is as well as being part of the European transparency requirement for high risk problems: "The behavior of AI systems must be able to be monitored or traced, or in other words, record all their procedures, from the data acquisition and annotation process, to each of the decisions they make. It is therefore vital that AI systems are explainable, in order to understand the decisions they make based on certain input data. It is clear that making AI processes and decisions explainable is essential." In January 2023, the National Institute of Standards and Technology (NIST) published the Artificial Intelligence Risk Management Framework (AI RMF 1.0)[5], which includes a similar list of trustworthy AI characteristics (it uses the term characteristic with a similar meaning to requirement), highlighting characteristics such as "safety and resilience", "explainability and interpretability" (separate from "transparency").

This is a general scenario under which we analyze the usability, utility and future of AI. Therefore, XAI is recognized as a crucial area with significant potential to foster trust, en-

This work was supported by the national project PID2023-150070NB-I00, Ministry of Science, Innovation and Universities

[1]AI Act, https://artificialintelligenceact.eu/the-act/

[2]https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

[3]Ethics Guidelines for Trustworthy Artificial Intelligence. HLEG-AI, 2019 https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[4]Recommendation on the Ethics of Artificial Intelligence, https://unesdoc.unesco.org/ark:/48223/pf0000381137

[5]AIRMF-NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1

sure accountability, and enable informed AI-assisted decision-making across various high-risks domains [2], [3], including healthcare, finance, and autonomous systems, public services, among others.

We can read the vast literature on XAI, from the current state of maturity [4] to its challenges [5] and highlighted criticisms [6], [7], [8]. The scientific literature is very prolific, I don't know if too much, it sheds a lot of light, many results, but also many questions. It is essential to approach XAI context with nuance and conduct in-depth analysis to ensure progress is made in the right direction. This exercise aims to briefly address various aspects and questions by analyzing authors' opinions, criticisms, established working areas, and XAI evaluation.

This contribution explores the complexities of XAI by analyzing various discussions within the literature. It begins by addressing the challenges posed by complex data, models, and high-risk scenarios. As the field matures, it is crucial to thoroughly examine the criticisms and challenges that arise. The paper then focuses on the lack of definitional cohesion, emphasizing the importance of defining XAI from the perspectives of audience and understanding, using an existing definition. We explore XAI techniques and discuss the crucial element of XAI explanations. Finally, we highlight the importance of evaluating the maturity level of explainability and how to measure it, from a triple dimension, practicality, governance and auditability.

The contribution is organized into sections based on the mentioned studies. It features concluding remarks on long way to go to enhance the usefulness of XAI, and also mentioning some topics that have been left untouched or barely explored.

II. Complex data, complex models and high risk scenarios

The emergence of deep learning at the beginning of the last decade led us to begin the advance in complex models, up to the large language models and generative AI as the summation of size and complexity. Also during these years, proposals have been made for deep structure neural networks models with different architectures that process various types of data, such as images, video, time series, text and multimodal data. Feature engineering is an essential methodology when working with tabular data and well-structured data, but it falls short when working with the complexity of the data mentioned above. It is not feasible to associate features with images, for example.

This puts us in a context of increasing complexity, which makes us understand less how AI models work, leading non-experts into the abyss of "Without understanding AI, observing the magic of AI". Therefore the desire for explainability becomes a universally accepted goal.

On the other hand, Europe has established the first law on AI, the AI Act[6] , published on 13 July 2024. AI deployment will be graded on a risk-based scale. Technologies with an unacceptable risk of causing direct harm will be banned. Where AI impacts fundamental human rights or critical systems such as essential infrastructure, public transport, healthcare or wellbeing, it will be classified as "high risk" and subject to increased levels of oversight and accountability. This regulation describes the concept of high-risk based AI systems, as those AI systems that are used in any of the following eight high-risk scenarios:

- biometric identification and categorisation of natural persons,
- management and operation of essential infrastructure,
- education and vocational training,
- employment, management of workers and access to self-employment,
- access to and enjoyment of essential public and private services and their benefits,
- matters related to law enforcement,
- management of migration, asylum and border control, or
- administration of justice and democratic processes.

In a few lines we have shown a global context that has been consolidated in recent years, and where it is necessary to advance in trustworthy AI technologies for the design of responsible AI systems [9], and XAI is a cornerstone.

III. Examining the Troubles and Criticisms in XAI

Given the maturity that XAI is gaining, reflections are raised on the path it follows and the associated problems. There are works in the literature that criticize XAI for various reasons, criticisms of its relevance in the current context.

Among them, we find specific criticisms about a concrete question as the use of certain XAI measures, for example in [10] is presented arguments demonstrating that Shapley values for explainability can produce misleading information regarding relative feature important. The authors state emphatically in [11] that: "The continued practical use of tools that approximate SHAP scores should be a reason of concern in high-risk and safety-critical domains".

On the other hand, we find deep dicussions on the XAI troubles. This contribution focuses on three articles with deep and thoughtful criticisms by their authors [6], [7], [8]. They will be developed below, to end with a brief position on these criticisms.

The paper [6], with the striking title "Dear XAI Community, We Need to Talk!", highlights and discusses eight misconceptions in XAI research. Authors argue on the lack of solid grounds due to:

- *"Proposals for new interpretation techniques that serve no clear purpose;*
- *anecdotal evidence from intuitive-looking heatmaps or "benchmarks" on seemingly relevant criteria are used as a substitute for a clear motivation;*
- *explanations are generated that mislead humans into trusting ML models without the models being trustworthy".*

[6]AI Act, https://artificialintelligenceact.eu/the-act/

The misconceptions are collected under the following titles:

1) *"Explanation Methods are Purpose-Free"*. A that explanation techniques in XAI should serve at least one practical purpose. Authors emphasize the importance of clearly demonstrating how an explanation technique fulfills its intended purpose. Without a widely accepted definition of explainability or interpretability, the purpose is the key to connecting these techniques to real-world applications. Techniques that lack practical motivation should be viewed with skepticism. If an explanation cannot be shown to help its intended audience, it is likely not useful and should be discarded.
2) *"One Explanation Technique to Rule Them All"*. Authors argue that XAI community members belief that a single best explanation technique, like SHAP, can provide perfect understanding for all purposes. Authors emphasize that the goals of explanations in XAI are diverse, such as auditing models, understanding phenomena, debugging, or enabling users to contest decisions. Different goals require different techniques and hyperparameters. They use counterfactual explanations to illustrate conflicts and trade-offs. For example, age may be excluded in counterfactuals for recourse but included for contesting decisions. But, counterfactuals may not be suitable for understanding the model as they offer limited insights.
3) *"Benchmarks do not Need a Ground-Truth"*. Authors discuss the challenges of benchmarking in XAI. While benchmarks have been successful in ML due to the presence of a ground truth, XAI lacks this central element, making objective comparisons difficult. Authors suggest two ways to progress in XAI: either abandon benchmarks and focus on qualitative evaluation or define benchmarks based on the explanation's purpose. However, some in the XAI community have taken a less rigorous approach, optimizing explanations for specific properties without clear motivation. This undermines the validity of benchmarks, turning them into promotional tools rather than objective standards.
4) *"We Should Give People Explanations They Find Intuitive"*. Authors criticize the practice of tailoring explanations in XAI to fit human intuition, which may not be faithful to the actual model. They argue that XAI should aim to make the model's mechanisms transparent rather than convincing people to trust the system. They distinguishes between explanations (actual reasons for decisions) and justifications (good reasons for decisions), noting that they often diverge in XAI. They emphasize the need for explanations that are faithful to the causal decision-making process, rather than those designed to be compelling or intuitive.
5) *"Current Deep Nets Accidentally Learn Human Concepts"*. Authors challenge the assumption that deep neural networks learn the same concepts as humans. They argue that while early layers may learn low-level concepts and later layers high-level concepts, this does not mean the model's reasoning aligns with human logic. They highlight several issues, such as the distributed representation enforced by regularization techniques and the limited impact of manipulating specific neurons. They also point out that effective communication, a key reason for shared human concepts, is not a constraint in ML training. They conclude that techniques like activation maximization may produce misleading results, and it is doubtful that humans will ever fully understand the concepts used by ML models.
6) *"Every XAI Paper Needs Human Studies"*. Authors emphasize the importance of human studies in evaluating explanations in XAI. They highlight two key questions: what conceptually counts as an explanation for a phenomenon, and which explanations are good for specific explainees. While human studies are essential for the latter, the former can be addressed through conceptual analysis and formal tools. Conceptual definitions help narrow down the vast space of possible explanations, guiding the search for good ones. Not all XAI purposes require human studies; for example, formal evaluations can be justified if human studies have already been conducted for that type of explanation.
7) *"XAI Methods can be Wrong"*. Authors discusses the limitations and challenges of saliency-based and model-agnostic explanation techniques like SHAP, LIME, and counterfactuals in XAI. They highlight that while these techniques can be manipulated to provide desired explanations, this does not necessarily mean they are wrong. Instead, they underscore the need for diverse XAI techniques, each illuminating different aspects of a model. They emphasize the importance of developing XAI techniques at various levels of abstraction to provide a comprehensive understanding of model behavior and address real-world purposes.
8) *"Extrapolating to Stay True to the Model"*. Authors discuss how most XAI techniques probe ML models, often in areas where the model has not seen any data, leading to extrapolation. Techniques like LIME, SHAP, and counterfactuals rely on probing the model, but ML models are generally poor at extrapolating to unseen instances. Explanations based on extrapolation may not be reliable. They argue that the explanations should focus on areas where the model is qualified, as probing outside the data manifold makes interpretation blurry and problematic. They emphasize the need for XAI techniques that provide insights within the data manifold for most purposes.

The paper is accompanied by four steps forward to take (section 4), sharing authors thoughts and intuitions about how they think the field should evolve to become a more substantive discipline. Their steps forward are:

- *Go from purpose to benchmark,*
- *Be clear what you need to explain and by what,*

- *Give clear instructions for how to interpret explanation Techniques, and*
- *XAI needs interdisciplinarity and expertise.*

We must consider this paper as a fairly in-depth analysis of the problems in XAI.

John Zerilli raises an interesting reflection from a philosophical prism in [7]: " *XAI has been forced to prioritise interpretability at the expense of completeness, and even realism, so that its explanations are frequently interpretable without being underpinned by more comprehensive explanations faithful to the way a network computes its predictions. While this has been taken to be a shortcoming of the field of XAI, I argue that it is broadly the right approach to the problem.". He concludes "for deeper and more comprehensive explanations of automated decisions is urgent, as in some cases it may be, we should naturally expect them, in whatever form is considered practicable by the standards of XAI. But where no such necessity arises, a satisficing explanation of an automated decision ought to suffice for assessing its credentials.*"

In [8] Authors reflect on several criticisms that need to be addressed.

- *Disagreements on the scope of XAI.* As for the causes of this disagreement, authors hypothesize that both interdisciplinarity and lack of rigor may have played a role.
- *Lack of definitional cohesion, precision, and adaption.* The title defines the criticism.
- *Misleading motivations for XAI research.* It is usually based on three statements: 1) People do not trust black box AI methods; 2) The inability to reveal their inner workings is what causes people not to trust black box AI methods; and 3) Explanations promote trust. There is insufficient evidence supporting these motivating hypotheses argue the authors.
- *Limited and inconsistent evaluations.* Although several ways to evaluating XAI methods have been proposed, no approach has been broadly adopted.

XAI as an interdisciplinary field in a mature or premature point depending on how you look at it, with a large number of publications. But, it needs to mature further in the fundamental aspects of theoretical and practical formalization. I agree XAI must evolve towards a discipline of complete utility to the important problem it addresses. It needs to explain why, what, and what, for each study or proposal.

Finally, I must highlight a progress regarding a collective discussion made by several renowned authors in the field in the following paper [5], discussing a Manifesto XAI 2.0, with the aim to define and briefly describe the open challenges in the field to face. The Manifesto is a mechanism for shaping our shared visions about research in the field of XAI, and it is the outcome of the engagement of diverse expertise and different experiences by its authors. This was summarized in nine points of interest to analyze, which raise and address as a future plan, and converges with the analysis of weaknesses and problems discussed in this section. They nine points are: *1) Creating Explanations for New Types of AI, 2) Improving (and Augmenting) Current XAI Methods, 3) Clarifying the Use of Concepts in XAI, 4) Evaluating XAI Methods and Explanations, 5) Supporting the Human-Centeredness of Explanations: To create human-understandable explanations, 6) Supporting the Multi-Dimensionality of Explainability, 7) Adjusting XAI Methods and Explanations, 8) Mitigating the Negative Impact of XAI and 9) Improving the Societal Impact of XAI.*

These nine points reflect many of the criticisms emphasized previously. This collective paper highlight explanations and also audience as important elements. But of course, they are still challenges that need to be addressed.

In the next three sections we discuss some of the mentioned critiques, from 3 prisms, definition, explanations and evaluation. Obviously, many other prisms and lenses need to be studied together with global reflections that approach the correct direction of investigation and advances in XAI.

IV. On the XAI definition

The mentioned criticism is clear. The lack of definitional cohesion in the field of explainable XAI has led to significant challenges in focusing the definition and scope of the discipline. This ambiguity limits the development of standardized methodologies and metrics, making it difficult for researchers and practitioners to evaluate and compare different XAI approaches. Consequently, the absence of a unified definition can result in fragmented efforts and hinder the progress towards achieving truly interpretable and trustworthy AI systems.

In [12] is analized XAI from the terms used along the literature: transparency, intelligibility, interpretability and explainability. Authors use the dictionary definitions to get a departure point.

- The word “transparent” refers to something that is “easily seen through, recognized, understood, detected; manifest, evident, obvious, clear” (Oxford English Dictionary,).
- An “intelligible” system should “capable of being understood; comprehensible” (Oxford English Dictionary).
- The word “interpret” definition is “to expound the meaning of (something abstruse or mysterious); to render (words, writings, an author, etc.) clear or explicit; to elucidate; to explain” (Oxford English Dictionary).
- For the word “explain”, the following definitions are used: “to provide an explanation for something to make plain or intelligible” (Oxford English Dictionary), “to make something clear or easy to understand by describing or giving information about it” (Cambridge Dictionary).

It is continuously repeated a word or idea, “understand" or “easily understood”. We already have a convergent term, but we have to ask ourselves another question. Does “understand" mean the same to the designer of the AI model, to the person who uses it, or to the person who is the recipient of its usage? Let’s think about the medical field, the designer, the programmer, the owner company, the doctor, the patient or society in general. It is certain that their vision of understanding an AI system is different.

In [13], it was placed audience (see Figure 1) as a key aspect to be considered when explaining an AI model. It was

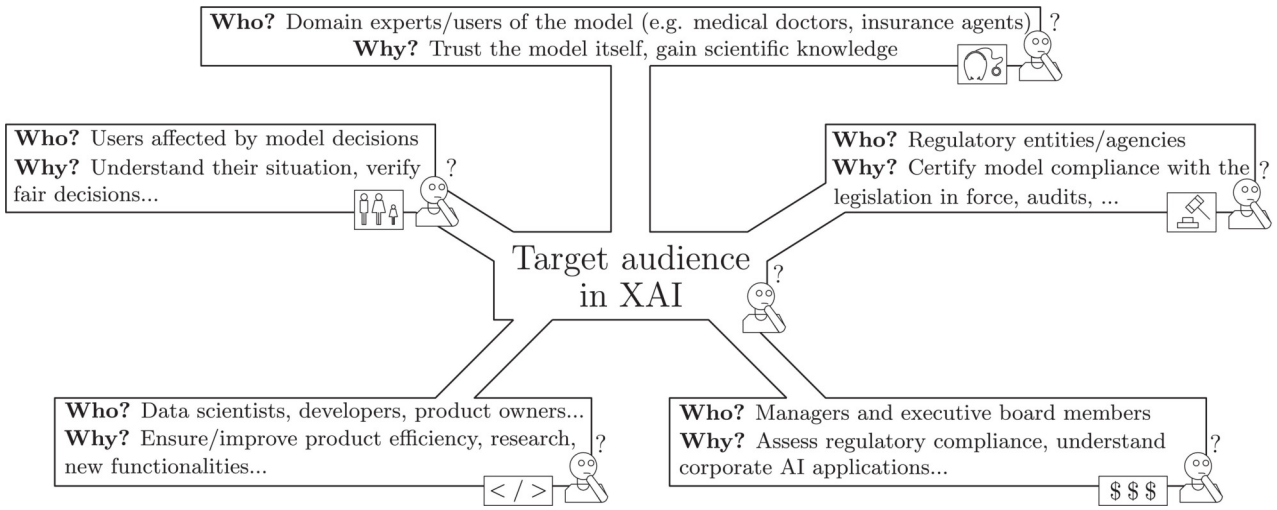

Fig. 1. Diagram showing different audience profiles (From Arrieta et al., 2220)

also elaborated on the diverse purposes sought when using XAI techniques, from trustworthiness to regulatory compliance, which round up the claimed importance of purpose and targeted audience (experts, users, developers, regulatory entities, and managers) in model explainability. In [14] was established a stakeholder interest map. It includes six levels of audience: Developer, Designer, Owner, User, Regulator, Society. This raises an interesting discussion. How would we answer the following four questions?

- Are they equal for "understanding"?
- What is their "understood" requirement?
- What is an "explanation" for them?
- What stakeholder is observing?

Under these considerations, the following definition considers both discussed elements.

Definition. [13] *Given an audience, an explainable AI is one that produces details or reasons to make its functioning clear or easy to understand.*

Regarding the lack of definitional cohesion, I believe that the definition provided in [13] serves as a solid convergence point for the topic. This definition encompasses two essential aspects, understanding and audience.

In [15], [16] have been introduced studies along which explainability approaches aim to satisfy stakeholders' desiderata and roles from stakeholders' desirements. Recent studies highlight the importance of stakeholders in different areas, and in many cases involving different stakeholders, such as, autonomous systems [16], medicine [17], [18], [19] and education [20] among others.

To finish with the definition and without going into the XAI taxonomy in depth, we must distinguish between two kind of models, interpretable models versus black box AI ones [21], [13]. Models that are interpretable per se, that introduce comprehensibility on the knowledge and the inference action, for example rule base systems or decision trees with few variables (local rule-based explainers produce logical rules which are close to human reasoning and make them suitable for non-experts). These are as opposed to black boxes, as boosting or neural networks or among others, whose difficulty of explanation increases with the neural networks number of layers. Black box AI models require post-hoc analysis. In [13], it was introduced a complete taxonomy on the post-hoc approaches with a conceptual diagram showing the different post-hoc techniques available for a machine learning (ML) model. This is an important aspect to consider in the post-hoc analysis of the black box ML models that needs connected with stakeholders' desirements and needs.

V. From data type explanations to Local linear explanations, concept-based explanations and prototype-based ones

In the context of XAI, explanations play a pivotal role in bridging the gap between complex ML models and human understanding. By providing clear and interpretable insights into how AI systems make decisions, explanations enhance transparency, build trust, and facilitate accountability. Addressing the importance of explanations, we delve into the discussed criticisms, and we must explore various techniques and methodologies that aim to make AI models more comprehensible. We move from a local linear explanations, the most popular approach, and beyond measuring features contribution, to the general idea of explanations based on the data type, the concept-based explanations, and the use of prototypes as potential element for explain decisions in complex problems/models.

We like to mention two papers. The paper [6], it includes a deep description of some XAI techniques: SHAP: SHapley Additive exPlanations, DiCE: Diverse Counterfactual Explanations, Transformers Interpret (TI) (for language models), Grad-CAM (image classification), Layer-wise Relevance (explain image), Logic Tensor (Neural-Symbolic AI), and TS4NLE

(for natural language explanations). The paper [22], it reports extensive examples of the various explanations for each data type, highlighting similarities and discrepancies of returned explanations through. It includes a website with a software repository, called *XAI Live Survey*[7], that authors mantain to keep pace with newly emergent methods.

Local linear explanations are among the most widely used methods in XAI. This approach involves approximating the behavior of a complex black-box model in the vicinity of a specific instance by using a simpler, more interpretable model, such as linear regression. Two well-known methods for generating local linear explanations are LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (SHapley Additive exPlanations). These methods help in understanding how individual features contribute to a model's prediction for a specific instance, making it easier to validate the model's decisions. LEAF framework was proposed for the evaluation and comparison of local linear explanations, with four different metrics to evaluate different desirable qualitative aspects of explanations. In [23] authors focus on the proposal of the REVEL framework (Robust Evaluation VEctorized Local-linear-explanation), whose main contribution is to offer a consistent and theoretically robust analysis of the black-box generated explanations, as well as being useful at a practical level for the evaluation of explanations. REVEL takes advantage of the existing state of the art and develops a series of theoretical improvements on the generation and evaluation methods. It redefines and proposes different quantitative measures to robustly assess different qualitative aspects of the explanations.

On the other hand, a categorization based on the data type and explanation type is fundamental to structure the area and to follow the advances. In [22], authors provide an explanation-based taxonomy with a comprehensive ontology of the explanations returned, taking into account the most popular data formats and associated approaches (see section 3, Figure 1, page 1724): tabular data (Feature Importance (FI) and Rule-Based (RB)), image (Saliency Maps (SM) and Concept Attribution (CA)), text (Sentence Highlighting (SH) and Attention Based (AB)), time series (Series Highlighting Attention Based) and graphps (Node Highlighting and Edge Highlighting). It also includes two transversal approaches that we will discuss later, prototypes (the user is provided with a series of examples that characterize a class of the black box) and counterfactuals (the user is provided with a series of examples similar to the input query but with different class prediction). This overview presents an exercise carried out to address a first study on this subject. They also report the most popular Python toolkits, AIX360 [24].

As a different explanation approach, concept-based explanations offer a compelling alternative by providing a more holistic view of the model's inner workings. Concept-based explanations better resemble the way humans reason and provide more intuitive and human-understandable insights by linking model decisions to high-level concepts. This approach helps users relate AI decisions to familiar ideas or categories, making the explanations more accessible [25]. Therefore, it have emerged as a powerful new XAI paradigm, providing model explanations in terms of human-understandable units, rather than individual features, pixels, or characters. This approach enhances the explainability by aligning explanations with concepts that are meaningful to humans. By leveraging concept-based explanations, stakeholders can gain deeper insights into the AI-assisted decision-making, making it easier to identify and address potential biases, errors, and ethical concerns. Furthermore, concept-based explanations facilitate more effective communication between AI developers and end-users, fostering trust and collaboration. As a result, concept-based explanations play a crucial role in advancing the transparency, accountability, and overall trustworthiness of AI systems in an open world. An overview on concept learning is described in [26].

The Prototype-based XAI techniques are an under utilized approaches that can provide inherently interpretable ML alternatives. Prototype selection for nearest neighbor classification has a long history in the field of ML, being highlighted as an essential tool to drive improvements in nearest neighbor techniques [27]. Prototypes must play a crucial role in the landscape of XAI, as it is discussed in [13], [28] among others, serving as a bridge between traditional explanations and concept-based explanations. Prototypes fit into this spectrum by offering concrete examples that represent typical instances of a particular class or concept. They help users understand what the model considers as a "typical" example, thereby making the model's behavior more explainable. Human reasoning is often prototype-based, using representative examples as a basis for categorization and AI-assisted decision-making. For instance, in image classification, a prototype might be a representative image that the model associates with a specific label. This can be particularly useful in identifying and understanding the characteristics that the model uses to make its decisions. By providing tangible examples, prototypes enhance the explainability of both traditional and concept-based explanations, making them a valuable tool in the XAI toolkit. We have also mentioned an approaches associated to prototype, but with with opposite use. the counterfactual explanations [29]. With a counterfactual explanation the user is provided with a series of examples similar to the input query but with different class prediction. In [13] was introduced the idea of counterfactual fairness, it tries to interpret the causes of bias using.

To finish with a challenge, XAI can be integrated as a technical objective for designers, as suggested in [30], to enhance its utility by aligning it with its intended purpose. For instance, a technical objective could be to leverage explanations to improve AI safety, as proposed in [8] with the concept of RED XAI. For example, XAI can be valuable in addressing the out-of-distribution detection problem [31], [32].

Addressing the criticisms, explanations are essential for building trust and transparency in AI systems. Concept-based

[7]https://kdd-lab.github.io/XAISurvey/

explanations, for instance, bridge the gap between complex models and human understanding by aligning model behavior with human-recognizable concepts. Prototype-based XAI techniques enhance interpretability by providing concrete instances that illustrate how the model operates. A comprehensive theory of explanations should encompass data, model, and post-hoc explainability, as discussed in [4].

While there may be insufficient evidence to universally support this claim, explanations play a crucial role in promoting trust in specific contexts. For instance, in high-risk scenarios like healthcare or finance, understanding the rationale behind AI decisions is essential for users to trust and accept those decisions. Additionally, explanations can help users learn from AI systems, thereby improving their own decision-making processes. Therefore, let's adopt the title of the paper by Hen et al. [33] as an aphorism and fundamental goal: *"Understanding the role of human intuition on reliance in human-AI decision-making with explanations"*.

VI. On the maturity level of explainability

Assessing the maturity level of explainability techniques in XAI involves evaluating various aspects to ensure they are effective, reliable, and useful. But we have to ask ourselves, from where and how?

The maturity level of explainability in AI can be assessed through several dimensions, aligned with the AI regulation debate. I propose a triple dimension, practicality, governance and auditability. In the following are shortly described the key points:

1) Practicality: Explainability in AI is becoming more practical as tools and techniques are developed to make AI systems more transparent. This includes the creation of interpretable models and the use of surrogate models to explain complex AI systems in real problems. There are a lot of practical studies, healthcare [34], finance [35], among many others applied areas, but it is necessary a methodology for a wide practical use.
2) Governance: AI governance refers to the frameworks, processes, rules, and standards that ensure AI systems are safe, ethical, and aligned with societal values. It is crucial for several reasons, ethical development, compliance and innovation, among others. From a governance perspective, frameworks need being established to ensure that AI systems are explainable. This includes guidelines for AI governance, for ethical AI and the development of metrics to assess the explainability of AI systems [36].
3) Auditability: AI auditability refers to the ability to assess and verify AI systems' algorithms, models, data, and design processes. Explainability is also crucial for the auditability of AI systems. Being able to explain AI decisions allows for better oversight and accountability, which is essential for building trust in AI technologies [37].

There is still room for improvement in development methodologies from the above-mentioned perspectives.

VII. Concluding remarks

Focused on the aforementioned criticisms and troubles, and from the perspective of a great theoretical development and not practical, I recognize that there is still a long way to go to enhance the usefulness of XAI. Many of these perspectives have been highlighted in this brief discussion. I have focused the attention on the XAI definition based on the audience, a fundamental element to advance toward useful XAI development.

I do not want to conclude without mentioning that some topics have been left unaddressed or barely explored. For instance, the impact of generative AI from the dual perspective of explainability and the use of large language models to enhance explainability. Additionally, the risks of overconfidence in explanations, which can increase decision-makers' tendency to rely on AI predictions even when the AI system is wrong, have not been addressed. Nor has there been a discussion on how XAI itself can be useful in guiding XAI-based model improvement, or the impact of XAI on various trustworthy AI requirements. Furthermore, a more in-depth examination of metrics is needed, including their pros and cons, and how they can advance practicality, governance, and auditability. Ensuring that AI systems are effective, reliable, and useful remains paramount.

References

[1] D. Castelvecchi, "Can we open the black box of AI? (News Feature)," *Nature*, vol. 538, pp. 20–23, 2016.

[2] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido *et al.*, "The role of explainable AI in the context of the AI Act," in *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*, 2023, pp. 1139–1150.

[3] L. Nannini, J. Alonso-Moral, A. Catala, M. Lama, and S. Barro, "Operationalizing Explainable AI in the EU Regulatory Ecosystem," *IEEE Intelligent Systems*, 2024.

[4] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence," *Information fusion*, vol. 99, p. 101805, 2023.

[5] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. Del Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger *et al.*, "Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions," *Information Fusion*, vol. 106, p. 102301, 2024.

[6] T. Freiesleben and G. König, "Dear XAI community, we need to talk! Fundamental misconceptions in current XAI research," in *World Conference on Explainable Artificial Intelligence*. Springer, 2023, pp. 48–65.

[7] J. Zerilli, "Explaining machine learning decisions," *Philosophy of Science*, vol. 89, no. 1, pp. 1–19, 2022.

[8] R. O. Weber, A. J. Johs, P. Goel, and J. M. Marques-Silva, "XAI is in trouble," *AI Magazine*, 2024.

[9] N. Díaz-Rodríguez, J. Del Ser, M. Coeckelbergh, M. L. de Prado, E. Herrera-Viedma, and F. Herrera, "Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation," *Information Fusion*, vol. 99, p. 101896, 2023.

[10] X. Huang and J. Marques-Silva, "On the failings of shapley values for explainability," *International Journal of Approximate Reasoning*, p. 109112, 2024.

[11] J. Marques-Silva and X. Huang, "Explainability is not a game," *Communications of the ACM*, vol. 67, no. 7, pp. 66–75, 2024.

[12] M. A. Clinciu and H. F. Hastie, "A survey of explainable AI terminology," in *1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence 2019*. Association for Computational Linguistics, 2019, pp. 8–13.

[13] 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.

[14] K. Haresamudram, S. Larsson, and F. Heintz, "Three levels of AI transparency," *Computer*, vol. 56, no. 2, pp. 93–100, 2023.

[15] M. Langer, D. Oster, T. Speith, H. Hermanns, L. Kästner, E. Schmidt, A. Sesing, and K. Baum, "What do we want from Explainable Artificial Intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," *Artificial Intelligence*, vol. 296, p. 103473, 2021.

[16] R. R. Hoffman, S. T. Mueller, G. Klein, M. Jalaeian, and C. Tate, "Explainable ai: roles and stakeholders, desirements and challenges," *Frontiers in Computer Science*, vol. 5, p. 1117848, 2023.

[17] H. V. Subramanian, C. Canfield, and D. B. Shank, "Designing explainable ai to improve human-ai team performance: a medical stakeholder-driven scoping review," *Artificial Intelligence in Medicine*, p. 102780, 2024.

[18] M. Kim, S. Kim, J. Kim, T.-J. Song, and Y. Kim, "Do stakeholder needs differ?-designing stakeholder-tailored explainable artificial intelligence (xai) interfaces," *International Journal of Human-Computer Studies*, vol. 181, p. 103160, 2024.

[19] M. Bergquist, B. Rolandsson, E. Gryska, M. Laesser, N. Hoefling, R. Heckemann, J. F. Schneiderman, and I. M. Björkman-Burtscher, "Trust and stakeholder perspectives on the implementation of ai tools in clinical radiology," *European Radiology*, vol. 34, no. 1, pp. 338–347, 2024.

[20] A. J. Karran, P. Charland, J. Martineau, A. O. de Guinea, A. Lesage, S. Senecal, and P.-M. Leger, "Multi-stakeholder perspective on responsible artificial intelligence and acceptability in education," *arXiv preprint arXiv:2402.15027*, 2024.

[21] M. Atzmueller, J. Fürnkranz, T. Kliegr, and U. Schmid, "Explainable and interpretable machine learning and data mining," *Data Mining and Knowledge Discovery*, pp. 1–25, 2024.

[22] F. Bodria, F. Giannotti, R. Guidotti, F. Naretto, D. Pedreschi, and S. Rinzivillo, "Benchmarking and survey of explanation methods for black box models," *Data Mining and Knowledge Discovery*, vol. 37, no. 5, pp. 1719–1778, 2023.

[23] I. Sevillano-García, J. Luengo, and F. Herrera, "Revel framework to measure local linear explanations for black-box models: Deep learning image classification case study," *International Journal of Intelligent Systems*, vol. 2023, no. 1, p. 8068569, 2023.

[24] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović *et al.*, "One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques," *arXiv preprint arXiv:1909.03012*, 2019.

[25] 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.

[26] E. Poeta, G. Ciravegna, E. Pastor, T. Cerquitelli, and E. Baralis, "Concept-based explainable artificial intelligence: A survey," *arXiv preprint arXiv:2312.12936*, 2023.

[27] S. Garcia, J. Derrac, J. Cano, and F. Herrera, "Prototype selection for nearest neighbor classification: Taxonomy and empirical study," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 3, pp. 417–435, 2012.

[28] A. Narayanan and K. Bergen, "Prototype-Based Methods in Explainable AI and Emerging Opportunities in the Geosciences," in *ICML 2024 AI for Science Workshop*, 2024.

[29] R. Guidotti, "Counterfactual explanations and how to find them: literature review and benchmarking," *Data Mining and Knowledge Discovery*, pp. 1–55, 2022.

[30] V. Chen, J. Li, J. S. Kim, G. Plumb, and A. Talwalkar, "Interpretable machine learning: Moving from mythos to diagnostics. queue 19, 6 (jan 2022), 28–56," 2022.

[31] H. Liu, V. Lai, and C. Tan, "Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW2, pp. 1–45, 2021.

[32] J. Choi, J. Raghuram, R. Feng, J. Chen, S. Jha, and A. Prakash, "Concept-based explanations for out-of-distribution detectors," in *International Conference on Machine Learning*. PMLR, 2023, pp. 5817–5837.

[33] V. Chen, Q. V. Liao, J. Wortman Vaughan, and G. Bansal, "Understanding the role of human intuition on reliance in human-ai decision-making with explanations," *Proceedings of the ACM on Human-computer Interaction*, vol. 7, no. CSCW2, pp. 1–32, 2023.

[34] M. Frasca, D. La Torre, G. Pravettoni, and I. Cutica, "Explainable and interpretable artificial intelligence in medicine: a systematic bibliometric review," *Discover Artificial Intelligence*, vol. 4, no. 1, p. 15, 2024.

[35] P. Weber, K. V. Carl, and O. Hinz, "Applications of explainable artificial intelligence in finance—a systematic review of finance, information systems, and computer science literature," *Management Review Quarterly*, vol. 74, no. 2, pp. 867–907, 2024.

[36] M. A. Camilleri, "Artificial intelligence governance: Ethical considerations and implications for social responsibility," *Expert systems*, vol. 41, no. 7, p. e13406, 2024.

[37] L. Waltersdorfer, F. J. Ekaputra, T. Miksa, and M. Sabou, "AuditMAI: Towards An Infrastructure for Continuous AI Auditing," *arXiv preprint arXiv:2406.14243*, 2024.