False Sense of Security in Explainable Artificial Intelligence (XAI)

Neo Christopher Chung[1,*], **Hongkyou Chung**[2,3], **Hearim Lee**[4], **Lennart Brocki**[1], **Hongbeom Chung**[2,5] and **George Dyer**[6,*]
[1]Institute of Informatics, University of Warsaw
[2]School of Law, Seoul National University, Seoul, Republic of Korea
[3]Shin & Kim LLC, Seoul, Republic of Korea
[4]Dongin Law Group, Seoul, Republic of Korea
[5]Wesang Co., Ltd., Seoul, Republic of Korea
[6]Informatism LLC, New Mexico, United States
[*]n.chung@uw.edu.pl, george@informatism.com

Abstract

A cautious interpretation of AI regulations and policy in the EU and the USA places explainability as a central deliverable of compliant AI systems. However, from a technical perspective, explainable AI (XAI) remains an elusive and complex target where even state-of-the-art methods often reach erroneous, misleading, and incomplete explanations. "Explainability" has multiple meanings which are often used interchangeably, and there are an even greater number of XAI methods – none of which presents a clear edge. Indeed, there are multiple failure modes for each XAI method, which require application-specific development and continuous evaluation. In this paper, we analyze legislative and policy developments in the United States and the European Union, such as the Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, the AI Act, the AI Liability Directive, and the General Data Protection Regulation (GDPR) from a right to explanation perspective. We argue that these AI regulations and current market conditions threaten effective AI governance and safety because the objective of trustworthy, accountable, and transparent AI is intrinsically linked to the questionable ability of AI operators to provide meaningful explanations. Unless governments explicitly tackle the issue of explainability through clear legislative and policy statements that take into account technical realities, AI governance risks becoming a vacuous "box-ticking" exercise where scientific standards are replaced with legalistic thresholds, providing only a false sense of security in XAI.

1 Introduction

Recent advancements in AI rely on end-to-end training of deep neural networks (DNNs), which act like black box algorithms. While the superb performance of predictive and generative models led to their wide applications in all areas of society, from social media to healthcare, developers, and users do not have a full grasp of their operating characteristics. The output (e.g., classification of an image) is obtained without understanding the AI's decision-making process. Yet, governance of AI demands revealing and reporting this mechanism as much as possible. Explanations are necessary because fundamentally, humans need to understand automated decision-making for meaningful governance. Furthermore, laws are being proposed and enacted to demand that AI companies provide such explanations.

There's a general consensus that increasing explainability will lead to transparent, reliable, and accountable AI, which contributes to greater safety. Yet, these terms – such as explainability, interpretability, and transparency – remain polysemantic and under-defined. These nuances are lost in recent legislative and policy texts leaving a large gap between technical feasibility and practical implementation on the one hand, and compliance thresholds on the other. Furthermore, these upcoming laws have prompted for-profit companies to eagerly market various XAI, governance, and risk management solutions that are half-baked. In the current landscape, we are concerned that the explosive development of XAI could paradoxically lead to a false sense of security.

There has been a rapid growth in research and development in XAI [Jacovi, 2023]. What started as a niche topic a decade ago has become prominent in all major machine learning (ML) and AI conferences, including specialized workshops on XAI (IJCAI, ICML, ECAI, and others). A dedicated conference on XAI, called the World Conference on Explainable Artificial Intelligence, started last year [Longo, 2023]. Overall, the number of papers studying and developing XAI has increased exponentially within the top 12 ML and AI conferences [Nauta *et al.*, 2023]. Including a wide range of journals, workshops, and other venues would increase this number drastically. While there are attempts to categorize and taxonomize assumptions and aims of XAI under development, often a bird eye view on different paradigms of explainability is lacking in the field of AI governance and policy.

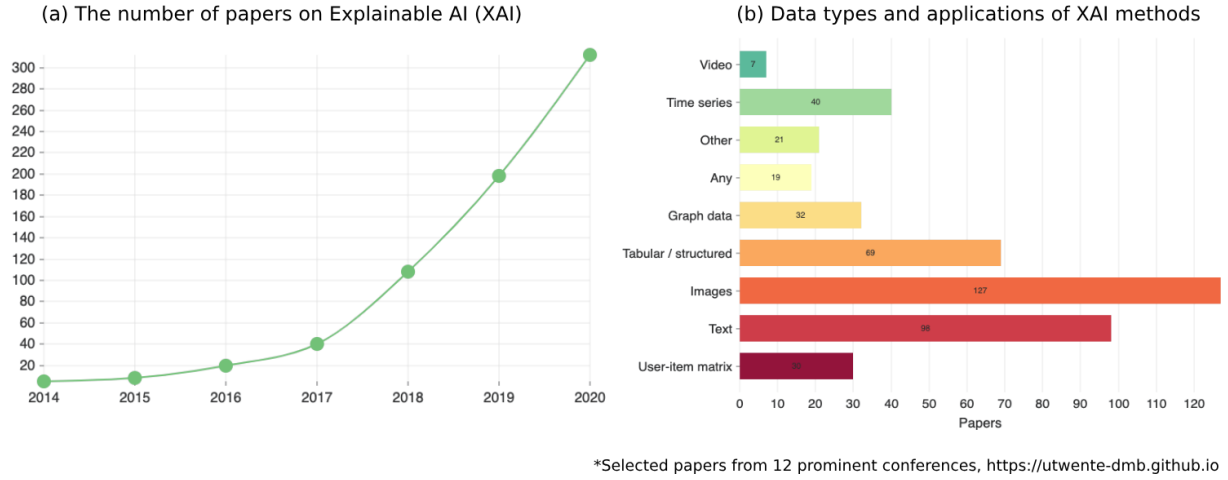

Figure 1: Explosive growth of XAI research. The number of XAI papers in 12 AI/ML conferences are counted and categorized by Nauta *et al.*, 2023.

2 Legislative and Policy Developments

Generally, AI policy statements focus on making AI safe, secure, transparent, trustworthy, and explainable. These concepts, however, receive less attention or detailed specification in the laws enacted to implement these policies. This led to ongoing debate about the extent to which a right to explanation exists. Based on relevant developments, we argue that some form of explainability will likely become a *de facto* requirement for critical AI systems deployed on the global market.

2.1 United States of America

In the United States, Joe Biden's October 2023 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (14110) calls upon the executive agencies under his authority to study and propose regulations that can effectively govern the safe and transparent use of AI in both the private and public sectors [Biden, 2023]. The National Institute of Standards and Technology (NIST), Department of Energy (DoE), National Science Foundation (NSF), and other agencies are crafting relevant resources, guidelines, and plans with industry and stakeholders.

Yet, the executive order is largely a statement of policy as it has no binding effect on deployers of AI systems in the U.S. The order mentions "safe", "secure", and their derivative words around 80 times. "Trust" and its derivatives appear a dozen times while "transparent" and associated terms are mentioned four times. Explainability is referred to only once, in the context of the U.S. government's need for "emphasizing or clarifying requirements and expectations related to the transparency of AI models and regulated entities' ability to explain their use of AI models."

Under U.S. law, there is currently only one binding directive relating to explainability: the Equal Credit Opportunity Act (ECOA), which requires creditors to notify applicants who are denied credit with specific reasons for an adverse credit decision. Specifically, §1002.9(b)(2) provides:

> The statement of reasons for adverse action required by paragraph (a)(2)(i) of this section must be specific and indicate the principal reason(s) for the adverse action. Statements that the adverse action was based on the creditor's internal standards or policies or that the applicant, joint applicant, or similar party failed to achieve a qualifying score on the creditor's credit scoring system are insufficient[U.S. Department of Justice, 1974].

Enacted in 1974, the ECOA was not written with generative AI in mind, but rather the deterministic decision-making systems in use at the time. Applied to contemporary technology, it nonetheless sets an actionable standard by ruling out the use of any black box AI systems that do not permit the isolation of the primary factors that contributed to their output. Such a clear-line rule is effective and practicable, but only because the law applies to a narrow scope of activity. It remains to be seen whether the continuation of this piecemeal legislative approach to the use of algorithmic systems is desirable, as the multiplication of possible use cases of AI will almost certainly outpace the ability of the U.S. legislature to address them. At some point, broader regulations will be needed.

2.2 European Union

Broader and more ambitious initiatives are now being promulgated within Europe: the AI Act [European Commission, 2021], the General Data Protection Regulation (GDPR) [European Parliament and Council of the European Union, 2016b], and the AI Liability Directive [European Commission, 2022]. Like the GDPR, the AI Act will be a regulation when it comes into force over the coming months. This means that they are considered as controlling law in every country within the EU. A directive, on the other hand, directs member states to adopt legislation that follows its principles, which may result in some differences among the implementing laws of each member state [Directorate-General for Communication,].

AI Act

The AI Act was endorsed by members of the European Parliament (MEP) in March 2024 [European Parliament, 2024], marking the first foray into the generalized regulation of AI. The Act and its relative merits have been discussed in detail elsewhere, and for the purposes of this paper, it is sufficient to note that it does not promulgate any substantive and precise explainability thresholds for AI developers. In fact, the AI Act itself does explicitly not reference explainability, focusing instead on the need for transparency in high-risk AI models.

Transparency is addressed in Article 13(1) [European Commission, 2021], which states that "High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately. An appropriate type and degree of transparency shall be ensured." The ethics guidelines of the European Commission (2019) clarify this requirement insofar as the decision-making process must be recorded and documented, and an "understandable explanation of the algorithmic decision-making processes" must be published [European Parliament, 2019]. According to Articles 6 and 7, high-risk systems would include those deployed within the education, health, recruitment, critical infrastructure management, law enforcement, and justice domains because such use cases could result in threats to the health, safety, or fundamental rights of individuals [European Commission, 2021].

As written, the AI Act does not provide any tangible thresholds for the transparency of high-risk models. Instead, the specific technical requirements of the models affected by Article 13 will be defined through a collaborative approach by the governing institutions created by the Act: the AI Office, the European Artificial Intelligence Board, the Advisory Forum, and the Scientific Panel of Independent Experts. This decentralized approach is commendable because it represents a tacit admission from the European legislators that the issue of AI regulation is too complex for a single body, much less a political one, to take on alone. Only through a deliberative (and almost certainly iterative) process will specific thresholds be defined for high-risk AI models.

On the other hand, the silence on the issue of explainability is deafening. After several years of study and deliberation, the inability to seriously address the concept does not bode well for future efforts, possibly camouflaging its intractable issues. More precise yet possibly misaligned guidance from the AI Act may have been more useful than no guidance at all, as stakeholders across the globe are having to guess as to how their models may or may not meet future EU regulatory standards. One particular area of uncertainty is whether an expectation of explainability will not apply to non-high-risk models in Europe. Such an assumption could be misguided, because a cautious reading of the GDPR and the AI Liability Directive shows that explainability will be – in some shape or form – a *de facto* compliance requirement for all AI models deployed in the EU.

General Data Protection Regulation (GDPR)

To ensure fair and transparent processing of personal data, GDPR Articles 13, 14, and 15 require the controller to provide the prospective or existing data subject with information relating to: "the existence of automated decision-making, including profiling, referred to in Article 22(1) and (4) and, at least in those cases, meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject" [European Parliament and Council of the European Union, 2016b]. Processing is defined in Article 4(2) as "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means."

A narrow right to explanation potentially exists within GDPR Article 22 (3) insofar as it requires the data controller to "implement suitable measures to safeguard the data subject's rights and freedoms and legitimate interests, at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision," when an individual is submitted to automated processing through contract or consent. Since Articles 13, 14, and 15 give a person a right to knowledge about the "logic involved" in the processing they are subject to, it follows that this a "legitimate interest" that data controllers should safeguard. This interpretation is favored by paragraph 71 of the GDPR's recitals, which suggests the legislative intent was to give the data subject a right "to obtain an explanation of the decision reached" [European Parliament and Council of the European Union, 2016a]. [Goodman and Flaxman, 2017] argues that GDPR does indeed establish a right to explanation.

On the other hand, it has been argued that, since recitals have no operative effect beyond aiding courts in interpreting ambiguous portions of law, the recital should be ignored because "[t]here are no ambiguities in the language that would require further interpretation with regard to the minimum requirements that must be met by data controllers" [Wachter *et al.*, 2017]. This prediction is plausible, but as the authors themselves admit, it is "only one possible future" that depends on European courts maintaining the same course of statutory interpretation during a technological and societal sea change. And even in this strict reading of the GDPR, where each article is read in isolation and without reference to recitals, the authors deduce, at minimum, a "right to explanation of the functionality of the system, a so-called 'right to be informed."'

AI Liability Directive

The final and potentially the most potent source of a right to explainability springs from the AI Liability Directive [European Commission, 2022]. In this proposed directive, the European Parliament and Council have laid out a burden-shifting scheme applying to fault-based liability claims for damages resulting from the output of AI systems. To avoid the harshest legal outcomes under this proposed burden-shifting, deployers of AI models of all risk profiles should make their systems transparent and explainable.

Generally, establishing fault-based liability requires the claimant to prove that the defendant's fault has caused their damages. However, the AI Liability Directive would create

under some conditions a presumption that a claimant's damages have been caused by the defect or fault in the AI system deployed by the defendant. This effectively shifts the burden to the defendant to prove that their system *did not* cause the claimed damages. This policy is an acknowledgment that placing the burden of proving causation on the claimant would drastically reduce their chances of recovery in almost all cases because of the inherent difficulty in "connecting the dots" between an AI system's inner workings and its outputs.

The presumption of causation always applies to claims against high-risk AI systems, except "where the defendant demonstrates that sufficient evidence and expertise is reasonably accessible for the claimant to prove the causal link" [European Commission, 2022](art 4.4). To limit the risk of liability, the deployers of high-risk models should ensure that they are not using black box models, and publish documentation and other assets that adequately explain the high-risk model's behavior. Doing so would presumably shift the burden of causation back to potential claimants.

For non-high-risk models, the presumption of causality would only apply where, according to Article 4.5, "the national court considers it excessively difficult for the claimant to prove the causal link" [European Commission, 2022]. Since there is no consensus as to what would be a scientifically or legally sufficient threshold of explaining the outputs of a generative AI system, it's unclear if or how anyone could establish with clear and convincing evidence that a fault in an AI system has proximately caused its harmful outputs. As a result, national courts in practice might interpret this section (as enacted locally) in a manner more sympathetic to claimants than intended by the EU legislature. This is a key area where again, deployers of non-high risk models can avoid potential liability by pro-actively taking steps to make their systems explainable.

Despite the lack of a clear directive requiring explainability from all categories of AI models deployed in Europe, the penumbra of "a right to explainability" can reasonably be inferred from several sections and recitals of the AI Act [European Commission, 2021], the GDPR [European Parliament and Council of the European Union, 2016b], and the proposed AI Liability Directive [European Commission, 2022]. Proposed and existing laws in the United States emphasize explainability as a core requirement for compliant algorithmic systems. For these reasons, companies seeking to limit their legal and regulatory risk in this uncertain environment will likely focus substantial resources on developing and deploying explainable AI models – even if such models do not actually improve accuracy or safety.

3 Polysemy of "Explainability"

The term "explainability" is used in a wide array of scenarios. Similarly, trustworthiness, safety, and related terms are used, often without clear guidelines. While there have been efforts to taxonomize qualities, formats, data types, and purposes of XAI, those conceptual definitions are difficult to apply and may not clearly delineate technical development in practice. Especially seeing from how legislative texts are written, "explainability" has become a catch-all phrase, which is up for broad interpretation. For comprehensive reviews and taxonomical organizations, see [Stepin *et al.*, 2021; Barredo Arrieta *et al.*, 2020; Nauta *et al.*, 2023; Saeed and Omlin, 2023]. Here we highlight five areas that any explainability regulation should address specifically:

First, there are two approaches to explaining AI; global and local. Global explanations attempt to answer how the AI model works, as an overall characterization obtained before a specific sample has been processed, *ex ante*. Furthermore, global explanations look at overall relationships between variables, which are especially critical when dealing with non-linear relationships. Local explanations, on the other hand, refer to how the AI model worked on a particular sample *post hoc*. Often, a non-linear relationship can be characterized linearly in a small window (local) of some feature values. Or, in classifying and predicting based on a specific sample image, the model may have used a small subset of features.

Under our reading, Article 13 of the GDPR [European Parliament and Council of the European Union, 2016b] would require deployers to provide global *ex ante* explanations at the moment the subject's personal data is collected for the purpose of informing them about the broad characteristics of an automated decision process. Articles 15 and 22 likely reference the need for local *post hoc* explanations upon the request of the subject. Similarly, the Equal Credit Opportunity Act, which requires that adverse automated credit decisions be accompanied by the reasons for the decision, is an example of a requirement for local, *post hoc* explainability.

Third, generally speaking, the current XAI methodology can't give comprehensive yet human-understandable explanations for black box AI. Only the full model and weights can truly describe the "complete mechanisms" of deep neural networks. In general, when explanations are complete, they could simply be used as an efficient and interpretable model with equivalent performance. In practice, post-hoc explainability methods substantially reduce the model mechanisms and characterize different aspects of the model, e.g., by estimating and visualizing the relative importance of input features (also known as saliency maps).

Fourth, the necessity and inevitability of simplification when explaining AI becomes a question about for whom explainability is intended. There would exist different levels of explanations similar to how complex phenomena may be explained to an expert, educated adult, and high schooler. The target audience and context for explainability should be considered the central element. However, that is currently not the case. When citizens have a right to explanation from high-risk AI systems, should the AI operators provide uniform and identical explanations? Furthermore, the delivery and interactivity of explanations are crucial if they are to be of practical use to the average person. Currently, operationalizing this aspect of XAI is missing.

Practically, the explainability requirements of consumer protection regulations such as the U.S. Equal Credit Opportunity Act [U.S. Department of Justice, 1974] and the GDPR [European Parliament and Council of the European Union, 2016b] imply that the provided explanation should be under-

standable by the average person. On the other hand, explanations published by corporations hoping to avoid the presumption of causality promulgated under the EU AI Liability Directive [European Commission, 2022] would need to be much more detailed. Overcoming a legally mandated presumption relating to causality cannot be achieved through pedestrian means, likely requiring the publication of data sets and methodologies that illuminate exactly how the claimant's damages were (not) caused by the AI system's alleged fault.

Fifth, being able to explain the model's decision-making process is related to the ability to control and interact with the AI model in a meaningful way [Teso and Kersting, 2019][Schmid and Finzel, 2020]. While explanations may (truthfully) characterize the model, how to change its behavior according to our desire is not always feasible. One may correct a misbehavior (e.g., misclassification) by re-labeling data, re-training the model, and even including certain pre- and post-processing steps, but this remains an unsolved challenge. We may encourage the model to work closer to how we want it to behave, but short of compiling hard-coded rules[1], there's no guarantee these guardrails will be effective.

The aforementioned five aspects of explainability are closely related. For example, local explanations may be reasonable and faithful for a given sample; however, they may not provide sufficient information to control the model. Furthermore, the simplification of AI's decision-making processes, which is necessary for most complex non-linear models (e.g., DNNs), must be informed by the target audience. Beyond these five categories, there exist other definitions of explainability. Mechanistic explanations focus on parts of the model (e.g., layers, sub-networks), whereas functional explanations are about the model's functions [Páez, 2019]. [Ginsberg, 1986; Grahne, 1998] examine counterfactual explanations which can be used to explain black box models. Overall, these different interpretations of explainability would significantly impact what must be explained, assuming the technical challenges are solved. Even with the best intentions and improved technical capabilities, misalignment between the intention of laws and the implementation of AI companies will create confusing and uninformative processes.

4 Failure Modes of XAI

Obtaining explanations from black box AI models remains highly challenging, with a number of competing methods and algorithms. There are multiple ways that explainers can give misleading information. Here, we present five critical failure modes for XAI, that are applicable to DNNs and beyond.

Robustness

First, XAI methods may not be robust to small changes in inputs [Alvarez-Melis and Jaakkola, 2018]. For example, in Figure 2, three samples are based on the same image of '3'. For this experiment, small and independent noise is added. A simple neural network with one convolutional layer is trained on the MNIST database [Deng, 2012]. The trained CNN model correctly classifies all three images with high confidence. Nonetheless, explanations by Local Interpretable Model-agnostic Explanations (LIME; on the second row) [Ribeiro *et al.*, 2016] are substantially varied. The positive and negative influence of pixels, as estimated Figure 2, would present drastically different interpretations of how the classification has been made.

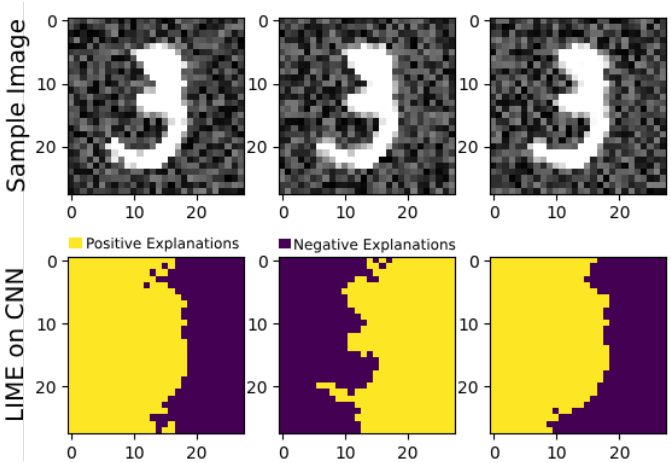

Figure 2: The first row shows the same image of 3, with different noise. CNN trained on MNIST correctly classifies these sample images as '3'. Explanations from LIME are shown in the second row.

Relatedly, [Adebayo *et al.*, 2018] investigated the robustness of saliency maps, which are designed to quantify and highlight important pixels with respect to the class label. They demonstrated that many advanced saliency maps show similar explanations even with data randomization. In the worst cases, some saliency maps, which are supposed to provide local explanations based on the model, resemble edge detection, which is by design independent of the model or the classification [Adebayo *et al.*, 2018]. Generally, these types of problems may be mitigated by optimizing for robustness [Alvarez-Melis and Jaakkola, 2018] or selecting appropriate methods [Adebayo *et al.*, 2018].

Adversarial Attacks

Second, adversarial attacks on XAI have been investigated, such that even well-performing methods give nonsensical explanations [Dombrowski *et al.*, 2019]. Instead of independent noise added in the first failure mode, the attackers can specifically design imperceptible changes to the input (e.g., the image looks identical to a human before and after modification), which completely change the explanations. This type of vulnerability mirrors well-known adversarial attacks on DNNs where carefully designed imperceptible changes to the input can cause the model to misbehave and misclassify. Mitigation strategies for adversarial attacks include explanation aggregation and model regularization [Baniecki and Biecek, 2023]. Nonetheless, attacks and defenses surrounding explanations will continue to evolve as XAI becomes more mainstream and public-facing.

[1]Rule-based algorithms exist and are useful in a wide range of areas. Those algorithms are not considered AI in this position paper.

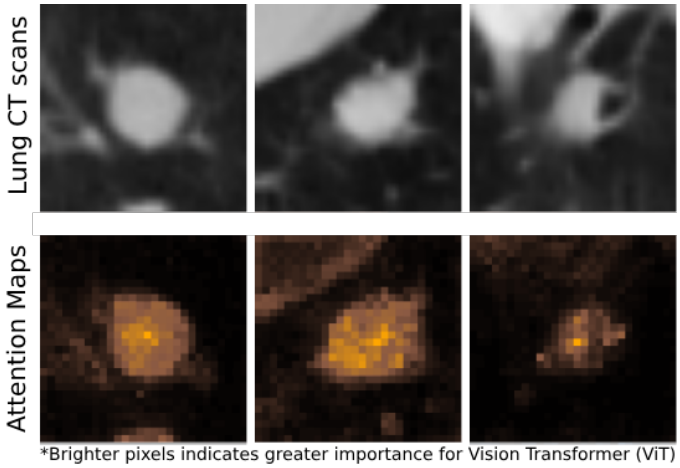

Figure 3: ViT model is fine-tuned on lung CT scans, to classify benign vs. malignant tumors. Attention maps are popular explanations for ViT.

Partial Explanations

Third, explanations can be *too incomplete* to give real insight. Explanations of complex AI models such as DNNs necessitate simplification of underlying mechanisms. While there are multiple evaluation frameworks to measure the faithfulness of XAI methods [Brocki and Chung, 2023b; Brocki and Chung, 2023a; Dawoud *et al.*, 2023], these measures are nonetheless relative without ground truth. Explaining a small fraction of complex behaviors may lead to overconfidence in both AI and XAI models. For example, we fine-tuned the ViT model from pre-trained DINO [Caron *et al.*, 2021], on lung CT scans [Armato *et al.*, 2011]. The model performs well classifying benign vs. malignant cases with an accuracy 86%. When we obtained attention maps (i.e., de facto standard explanations for ViT), we observe that they highlight lung nodules (Figure 3). At first glance, clinicians may be content with these explanations as AI has seemingly used the tumors for classification. However, without further distinction (e.g., higher resolution; signed importance), we don't actually learn which parts of a tumor are used for the model's inference. Saliency maps can tell us where the model is looking but not what the model sees [Rudin, 2019].

Data and Concept Drift

Fourth, data drift [Maimon and Rokach, 2010] and concept drift [Widmer and Kubat, 1996] refer to the change in data distribution and in applicable domains, respectively. Once trained and deployed, AI systems with explainers tend to remain fixed. Nonetheless, data and their relationship to underlying concepts fed to the systems may change over time (Figure 4). For example, an AI model trained with data from African hospitals may end up being used in Asian countries. Patients of a different ethnicity impact how AI is performing, as well as what explanations are given. XAI that worked faithfully in Africa might gives misleading explanations in Asia. On the other hand, over time, new concepts may appear, that originally did not exist in the training data (Figure 4(d)). For example, malignant tumors among Asian patients might exhibit novel characteristics in medical images, unseen among African patients. Then, XAI methods that were trained and evaluated on African populations might not work on medical images from Asian patients. To overcome these challenges, we must evaluate XAI for the application under consideration and continuously update our methods over time. Continuous data monitoring, collection, training, evaluation and deployment would involve significant costs and effort.

Anthropomorphization

Fifth, humans tend to anthropomorphize explanations by assuming a closer alignment of the model's and human's inference process than is actually the case. Starting with the birth of neural networks, an implicit or explicit connection between "artificial" and "biological" is prevalent. Mistaking that DNNs work like our brains wrongly gives an impression that they must be utilizing similar features and logic to arrive at their final decision [Inie *et al.*, 2024]. For example, when applied on natural images where objects are classified, humans may like explanations that focuses on regions of classified objects. However, it is entirely possible that the AI classifier, in fact, derives information from background pixels. Nonetheless, this type of erroneous qualitative evaluation occurs regularly, as ground truth on explanations is fundamentally lacking. Anthropomorphization is particularly problematic for XAI as the seemingly plausible explanations can easily trick humans to believe in the system.

Overall, these failure modes demonstrate the great risks that arise from blindly adapting XAI methods. At the very minimum, careful consideration and evaluation for applications are needed to reduce erroneous conclusions from explanations. There are likely other potential pitfalls, while more robust and complete XAI methods are being developed. Nonetheless, in our perspective, these failure modes will continue to persist in the current paradigm of XAI.

5 Market Conditions

In the market, both established and start-up companies are rushing to offer XAI solutions. Most of the products that we have surveyed do not interact with the underlying AI model in a meaningful way. The concept of XAI is already co-opted to provide false sense of security, where turn-key products added on a top of existing AI infrastructure may seem sufficient to end-users.

Arthur AI, Inc. claims to be "the very first company to launch ML performance and data drift monitoring for computer vision models, including in-depth explainability" [Arthur.AI,]. As one of the most well-funded start up in this space, its Series B funding was US$42M in 2022. As a specific example, for its computer vision product, it markets "Explainability and Monitoring for Your CV Models". Their documentation repeatedly states "explainability" without providing details, let alone advantages and disadvantages[2]. In the Arthur SDK (Software Development Kit) Python API (Application Programming Interface) Reference[3], it becomes clear

[2]https://docs.arthur.ai/docs Accessed April 26, 2024

[3]https://sdk.docs.arthur.ai/ Accessed April 26, 2024

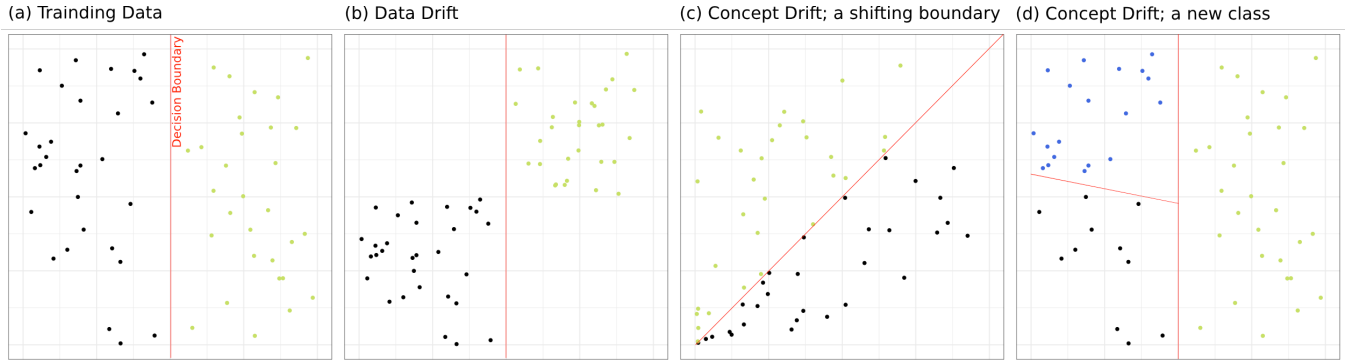

Figure 4: *(a)* The training data is initially used to produce a model, whose decision boundary classifies data in production. *(b)* Data drift occurs when the distribution of data has significantly changed. *(c)* The relationship between data and concept (i.e., class) may change over time, resulting in a concept drift. *(d)* Concept drift could arise from introduction of a new class, that was not included in the training data.

that Arthur implemented LIME (Local Interpretable Model-Agnostic Explanations) under the hood. While LIME is one of the most popular XAI methods, its limitations should be up front to ensure appropriate applications.

Google is one of the largest and the most advanced companies in AI with a market capitalization of ∼US$2 trillion. Google would advertise Vertex AI[4] with a number of explainability methods, which claim to be authoritative and to avoid the pitfalls of black box AI. Information on the limitations of their methods seems to focus on narrow aspects, instead of the fundamental intractable issue of attempting to explain a black box. While Google DeepMind publishes excellent methods and papers on XAI that dive deeper into technical details, potential clients likely do not read academic publications.

Note that over-confident marketing of XAI is not at all unique to these two companies. Explainability methods have been developed in a research environment, often within narrowly defined applications and assumptions. As those methods are packaged in a for-profit environment, technical nuances may have been glanced over.

6 Discussion and Conclusion

Considering the variety of XAI methods with a wide range of goals and operating characteristics, the use of these perfunctory explanation methods and flawed compliance tools may trigger many failure modes of XAI and provide misleading or unreliable explanations. This affects the following three groups interfacing with AI systems. Feedback loops among these groups amplify misleading impressions about XAI:

Regulators and politicians are under the impression that explainability is an issue of implementation, instead of a fundamental research problem. Law and policy documents fail to specify what are and how to obtain explanations, while simultaneously requiring certain AI operators to provide explanations. The AI Act in EU [European Commission, 2021] and the Executive Order in the USA [Biden, 2023] demand safe, trustworthy, and explainable AI. These will force tech companies to implement and outsource explainability for their AI systems, regardless of their accuracy, reliability, or utility.

AI developers and operators may view "explanations" given by existing XAI methods as golden standards. Under legislative and economic pressures, companies utilizing AI systems may adapt suboptimal XAI that are not meaningfully explaining the decision making process. Explanations that seem to match human expectations may garner approval, and companies will become falsely confident that their systems are working as intended. Without detailed regulations that define and evaluate standards for XAI, ad hoc implementations could lead a right to explanation astray.

Clients and citizens are conditioned to rely and trust the AI systems on the basis of sweeping laws and AI systems promising explanations. Whereas a black box AI raises an inherent concern about their unknowable operating characteristics, legally mandated and confidently marketed XAI, is provided with a false aura of trustworthiness. User studies show inclusion of explanations increases user trust and decreases decision-making time. But at the same time, XAI (specifically, LIME [Ribeiro *et al.*, 2016], SHAP [Lundberg and Lee, 2017], and TreeInterpreter [Saabas, 2014]) substantially decreased the decision accuracy of human analysts in a fraud detection study [Jesus *et al.*, 2021]. Regardless of the model performance, users of AI become more reliant when explanations are provided [de Brito Duarte *et al.*, 2023].

From both technical and legislative perspectives, explainability plays an important role in AI governance. While it is intuitive to present and advocate for the need to understand, interpret, and explain a model's decision-making process, the methodologies of XAI require thorough consideration of their applications, goals, and failure modes. We must dispel the false sense of security in XAI arising from over-confident products that seemingly satisfy legislative and policy texts that vaguely imply a right to explanation.

Ethical Statement

There are no ethical issues.

[4] https://cloud.google.com/vertex-ai/docs/explainable-ai/overview Accessed April 26, 2024

Acknowledgments

This work was funded by the ERA-Net CHIST-ERA grant [CHIST-ERA-19-XAI-007] long term challenges in ICT project INFORM (ID: 93603), by the National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071].

References

[Adebayo *et al.*, 2018] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*, October 2018.

[Alvarez-Melis and Jaakkola, 2018] David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*, June 2018.

[Armato *et al.*, 2011] Samuel G. Armato, Geoffrey McLennan, Luc Bidaut, and et al. The lung image database consortium (lidc) and image database resource initiative (idri): A completed reference database of lung nodules on ct scans. *Medical Physics*, 38(2):915–931, January 2011.

[Arthur.AI,] Arthur.AI. 6 challenges in operationalizing computer vision models. White Paper.

[Baniecki and Biecek, 2023] Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion, vol. 107, 102303, 2024*, 107:102303, July 2023.

[Barredo Arrieta *et al.*, 2020] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, June 2020.

[Biden, 2023] Joseph Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. The White House (United States), October 2023. Executive Order 14110.

[Brocki and Chung, 2023a] Lennart Brocki and Neo Christopher Chung. Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. *Pattern Recognition Letters*, 176:131–139, December 2023.

[Brocki and Chung, 2023b] Lennart Brocki and Neo Christopher Chung. Fidelity of Interpretability Methods and Perturbation Artifacts in Neural Networks. In *11th International Conference on Learning Representations (ICLR) Tiny Papers*. ICLR, March 2023.

[Caron *et al.*, 2021] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. April 2021.

[Dawoud *et al.*, 2023] Karam Dawoud, Wojciech Samek, Peter Eisert, Sebastian Lapuschkin, and Sebastian Bosse. Human-centered evaluation of xai methods. *ICDMW (2023) 912-921*, December 2023.

[de Brito Duarte *et al.*, 2023] Regina de Brito Duarte, Filipa Correia, Patrícia Arriaga, and Ana Paiva. Ai trust: Can explainable ai enhance warranted trust? *Human Behavior and Emerging Technologies*, 2023:1–12, October 2023.

[Deng, 2012] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[Directorate-General for Communication,] Directorate-General for Communication. Types of eu law. Technical report, European Commission. Accessed April 23, 2024.

[Dombrowski *et al.*, 2019] Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Conference on Neural Information Processing Systems*, June 2019.

[European Commission, 2021] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act). European Union Law, 2021. COM/2021/206 final.

[European Commission, 2022] European Commission. Proposal for a directive of the european parliament and of the council on adapting non-contractual civil liability rules to artificial intelligence (ai liability directive). European Union Law, 2022. COM/2022/496 final.

[European Parliament and Council of the European Union, 2016a] European Parliament and Council of the European Union. Recital 71 for article 22 - automated individual decision-making, including profiling of general data protection regulation (regulation (eu) 2016/679). Official Journal of the European Union, April 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[European Parliament and Council of the European Union, 2016b] European Parliament and Council of the European Union. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). Official Journal of the European Union, April 2016.

[European Parliament, 2019] European Parliament. Eu guidelines on ethicsin artificial intelligence:context and implementation. Briefing, 2019.

[European Parliament, 2024] European Parliament. Artificial intelligence act: Meps adopt landmark law. Press Release, March 2024.

[Ginsberg, 1986] Matthew L. Ginsberg. Counterfactuals. *Artificial Intelligence*, 30(1):35–79, October 1986.

[Goodman and Flaxman, 2017] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision making and a "right to explanation". *AI Magazine*, 38(3):50–57, September 2017.

[Grahne, 1998] G Grahne. Updates and counterfactuals. *Journal of Logic and Computation*, 8(1):87–117, February 1998.

[Inie *et al.*, 2024] Nanna Inie, Stefania, Druga, Peter Zukerman, and Emily M. Bender. From 'ai' to probabilistic automation: How does anthropomorphization oftechnical systems descriptions influence trust? *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2024)*, 2024.

[Jacovi, 2023] Alon Jacovi. Trends in explainable ai (xai) literature. January 2023.

[Jesus *et al.*, 2021] Sérgio Jesus, Catarina Belém, Vladimir Balayan, João Bento, Pedro Saleiro, Pedro Bizarro, and João Gama. How can i choose an explainer?: An application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021.

[Longo, 2023] Luca Longo, editor. *Explainable Artificial Intelligence: First World Conference, xAI 2023, Lisbon, Portugal, July 26–28, 2023, Proceedings, Part I*. Springer Nature Switzerland, 2023.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[Maimon and Rokach, 2010] Oded Maimon and Lior Rokach. Data mining and knowledge discovery handbook, 2010.

[Nauta *et al.*, 2023] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, July 2023.

[Páez, 2019] Andrés Páez. The pragmatic turn in explainable artificial intelligence (xai). *Minds and Machines*, 29(3):441–459, May 2019.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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*, pages 1135–1144, 2016.

[Rudin, 2019] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[Saabas, 2014] Ando Saabas. Interpreting random forests. https://github.com/andosa/treeinterpreter, 2014.

[Saeed and Omlin, 2023] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems*, 263:110273, March 2023.

[Schmid and Finzel, 2020] Ute Schmid and Bettina Finzel. Mutual explanations for cooperative decision making in medicine. *KI - Künstliche Intelligenz*, 34(2):227–233, January 2020.

[Stepin *et al.*, 2021] Ilia Stepin, Jose M. Alonso, Alejandro Catala, and Martin Pereira-Farina. A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001, 2021.

[Teso and Kersting, 2019] Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19. ACM, January 2019.

[U.S. Department of Justice, 1974] U.S. Department of Justice. Equal credit opportunity act (title vii of the consumer credit protection act) 15 u.s. code, chapter 41 consumer credit protection subchapter iv - equal credit opportunity§ 1691 - scope of prohibition, October 1974. https://www.law.cornell.edu/uscode/text/15/1691.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2):76–99, May 2017.

[Widmer and Kubat, 1996] Gerhard Widmer and Miroslav Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, April 1996.