Ethics and Information Technology (2025) 27:5
https://doi.org/10.1007/s10676-024-09800-7

ORIGINAL PAPER

Nullius in Explanans: an ethical risk assessment for explainable AI

Luca Nannini[1,2] · Diletta Huyskes[3] · Enrico Panai[4,5] · Giada Pistilli[6,7] · Alessio Tartaro[8]

Abstract

Explanations are conceived to ensure the trustworthiness of AI systems. Yet, relying solemnly on algorithmic solutions, as provided by explainable artificial intelligence (XAI), might fall short to account for sociotechnical risks jeopardizing their factuality and informativeness. To mitigate these risks, we delve into the complex landscape of ethical risks surrounding XAI systems and their generated explanations. By employing a literature review combined with rigorous thematic analysis, we uncover a diverse array of technical risks tied to the robustness, fairness, and evaluation of XAI systems. Furthermore, we address a broader range of contextual risks jeopardizing their security, accountability, reception alongside other cognitive, social, and ethical concerns of explanations. We advance a multi-layered risk assessment framework, where each layer advances strategies for practical intervention, management, and documentation of XAI systems within organizations. Recognizing the theoretical nature of the framework advanced, we discuss it in a conceptual case study. For the XAI community, our multifaceted investigation represents a path to practically address XAI risks while enriching our understanding of the ethical ramifications of incorporating XAI in decision-making processes.

Keywords Explainable AI (XAI) · AI governance · Ethics assessment · Risk management · Adversarial perturbation · Robustness · Epistemology

✉ Luca Nannini
l.nannini@usc.es

Diletta Huyskes
diletta.huyskes@unimi.it

Enrico Panai
enricopanai@gmail.com

Giada Pistilli
giada@huggingface.co

Alessio Tartaro
a.tartaro@phd.uniss.it

1 Minsait by Indra Sistemas, Madrid, Spain

2 CiTIUS - Centro Singular de Investigación en Tecnoloxías Intelixentes, Universidade de Santiago de Compostela, Santiago de Compostela, Spain

3 University of Milan, Milan, Italy

4 Università Cattolica del Sacro Cuore (UCSC), Milan, Italy

5 EMlyon Business School Paris, Paris, France

6 Sorbonne Université, Paris, France

7 Hugging Face, New York, USA

8 Department of Humanities and Social Sciences, University of Sassari, Sassari, Italy

Introduction

Explainable Artificial Intelligence (XAI) has emerged as a relevant area of research within the broader field of AI, as it seeks to provide human-understandable explanations for the decisions, recommendations, and predictions made by AI systems (Gunning & Aha, 2019). While the use of XAI has the potential to enhance transparency and accountability in AI-driven decision-making processes, it also raises new ethical concerns and challenges. XAI methods are generally developed to bring greater clarity to AI systems: yet such tools are evaluated primarily through quantitative measures, often without sufficient involvement from all stakeholders affected by these explanations (Kaur et al., 2020; Schemmer et al., 2022) or unclear benefits for their usefulness (Bertrand et al., 2022; Chen et al., 2023; Schemmer et al., 2022; Vasconcelos et al., 2023). Explanations bring risks that, if not properly addressed, may undermine the intended benefits of XAI and negatively impact the individuals and communities affected by AI decisions (Bertrand et al., 2022; de Bruijn et al., 2022; Janssen et al., 2022; Liao & Varshney, 2021). Indeed, if the explanations produced are not adequately vetted and validated by affected users (Langer

Published online: 14 December 2024

et al., 2021), they may be of limited informativeness, if not entirely useless or even harmful (Liao & Varshney, 2021; Robbins, 2019). In this perspective, the Royal Society's motto "*Nullius in Verba*," which translates to "*take nobody's word for it*," emphasizes the importance of verifying claims through evidence and rigorous analysis rather than relying solely on authority or assertions (McKie, 1960; The Royal Society, 1662). In the context of XAI, we propose a slight adaptation of this motto: "*Nullius in Explanans*," or "*take nobody's explanation for it.*" This rephrasing highlights the need for a comprehensive and systematic approach to assessing and mitigating the risks associated with explanations generated by AI systems. Rather than simply accepting the explanations at face value, it is then crucial to critically examine their validity, robustness, and potential vulnerabilities. Indeed, despite the growing attention to XAI risks there remains a lack of comprehensive frameworks for assessing and mitigating the diverse array of technical and sociotechnical risks associated with XAI systems.

This paper aims to address this gap by proposing a novel multi-layered risk assessment framework. We combine a literature review with thematic analysis, capturing a broad spectrum of risks and their underlying relationships. Our primary contribution lies in developing a taxonomy that classifies identified risks into two main categories: *technical* risks, related to data and architecture of XAI systems, and *contextual* risks, related to reception and deployment of explanations. From these categories, we advance a novel risk assessment framework for their identification and mitigation. To clarify, such assessment shall not be intended as a mechanism for demonstrating the "trustworthiness" of an XAI system. Instead, it constitutes a tool for critical reflection to facilitate introspection and inquiry regarding their design rationale and objectives. This paper is intended for a broad audience, including XAI practitioners, researchers, policymakers, and individuals interested in the ethical implications of AI and XAI systems. While some technical aspects of XAI methods are discussed, we aim to present the risks and the risk assessment framework in a manner accessible to readers with varying levels of technical expertise.

We begin in section "Background" by discussing relevant work that detailed desiderata and risks of explanations alongside ethical risk assessments. After, we will expose our method to retrieve and elaborate relevant research in section "Method", presenting in the following section "Categorization of risks in XAI systems" the taxonomy of technical risks in XAI (section "Technical risks") and sociotechnical ones (section "Contextual risks"). Building on this taxonomy, section "A risk assessment framework for XAI systems" introduces our multi-layered XAI risk assessment framework, that we illustrate in application through a theoretical case study in section "Use case example". Finally, section "Conclusion" concludes with a discussion of research limitations and future directions.

Background

In the realm of XAI, risks are predominantly treated as *ends*, signifying domain-specific objectives that explanations can address. When viewed as *mediums* associated with the structure of explanations, they are mostly related to the degree of fidelity concerning AI systems. Systematic reviews on XAI typically explore strategies and metrics for appraising explanations, encompassing both quantitative and qualitative evaluation methodologies, including human-centered evaluation approaches (Adadi & Berrada, 2018; Guidotti et al., 2019; Stepin et al., 2021).

A number of studies have advanced qualitative evaluation criteria, focusing on surveying acceptance and understandability of explanations by end users (Langer et al., 2021; Löfström et al., 2022; Mohseni et al., 2021). Despite the burgeoning interest in qualitative XAI evaluation criteria, there remains a dearth of contributions investigating the empirical usability of explanations (Kaur et al., 2020; Schemmer et al., 2022). The desirable cognitive properties inform these contributions of a "good explanation," taking into account human–computer interaction perspectives and concepts from social science and psychology (Lipton, 2018; Miller, 2019; Miller et al., 2017).

Trade-offs in XAI approaches To begin, the selection of XAI approaches encounters inherent technical challenges, notably when dealing with complex, high-dimensional data. For instance, *Surrogate Models* and *Rule Extraction*, while fostering model interpretability, run the risk of oversimplifying intricate models, thereby potentially compromising the accuracy of their representation (Andrews et al., 1995; Craven & Shavlik, 1995; Freitas, 2013; Mohseni et al., 2018). Further, several XAI methods, including *Partial Dependence Plot* (PDP), *Individual Conditional Expectations Plot* (ICE), and *Global Variable Importance* (GVI) measures[1], often grapple with the delicate issue of feature interactions and correlations (Fisher et al., 2019; Friedman, 2001; Goldstein et al., 2015). These dependencies can not only

[1] PDP is graphical visualization that shows the marginal effect of a feature on the predicted outcome of a machine learning model, while accounting for the average effect of all other features (Friedman, 2001); ICE is similar to PDP but shows the dependence of the predicted outcome on a feature for each instance separately, allowing for the identification of heterogeneous relationships (Goldstein et al., 2015); GVI quantifies the overall importance of each feature in a model's predictions, typically by calculating the increase in the model's prediction error after permuting the values of the feature (Fisher et al., 2019).

result in misleading representations but also limit the scope of the insights provided, affecting their utility, particularly in high-stakes contexts. Even approaches like Accumulated *Local Effects Plots* (ALE) and *Counterfactual Explanations*, designed to mitigate some of these issues by offering localised insights or presenting alternative scenarios respectively, encounter their own challenges. ALE plots might struggle with visualising feature interactions (Sorokina et al., 2008), whereas generating meaningful counterfactuals tend to be instance-based and might not provide an overarching understanding of the model (Stepin et al., 2021; Wachter et al., 2017). These challenges underscore the importance of an informed and judicious choice of XAI methods, contingent on the requirements of users and specific contexts.

Designing contextual explanations The imperative to comprehend explanations within the ecosystem where XAI solutions are developed has been underscored, particularly with regard to their epistemological value (Robbins, 2019). This pertains to the usability of explanations for a diverse array of end users (Schemmer et al., 2022), rather than solely their developers (Kaur et al., 2020). In response to this demand, a nascent subcurrent has emerged, concentrating on providing tangible approaches to tailor explanations for multiple users, aspiring to enhance their effectiveness by proffering design and evaluation guidelines (Mohseni et al., 2021). This includes deliberating on the type of explanations (Cabitza et al., 2023) or the sociocultural context of interaction among recipients (Dazeley et al., 2021). Other framework contributions, such as the survey from Löfström and Hammar, delineated subjective criteria of qualitative evaluation, advancing a model of explanation quality aspects (Löfström et al., 2022). Moreover, scholars such as Rudin advocated for inherently interpretable AI system designs when high stakes envelop their decisions (Rudin, 2019). In this vein, explainability desiderata shall inform and anticipate the design of XAI solutions, critically inquiring over the need for explanations concerning stakes and context of deployment of AI systems.

Proactive approaches and ethical risk assessments Despite the ongoing discourse surrounding the implementation of explanations in AI systems, alternative validation instruments for AI system explanations, such as impact assessment or risk management procedures, may offer valuable yet unexplored benefits (Floridi, 2018; Moss et al., 2021). Some XAI scholars persist in referencing the "right to explanation" in the EU GDPR (European Commission, 2016) to justify the benevolence of their research studies. Yet, due to the casuistry and debate over the enactment of such as a *right*, rather than benevolence, their statements potentially indicate limited policy knowledge over requirements for establishing a legal baseline to implement XAI services (Nannini, 2024). This concern might be further exacerbated by the heterogeneous policy landscape and the challenges policymakers confront in harmonizing regulations and guidelines with XAI research (Hacker & Passoth, 2022; Nannini et al., 2023). Given the potentially loose legislative baseline and the profusion of disparate "best practices" for ideal explanation properties, a proactive approach concentrating on quantifying the risks of explanations may be desirable to address policy and operationalization requirements of explanations. Recent work in AI governance and risk management, particularly *Ethical Risk Assessments* (ERA), can be instrumental in structuring the development of useful explanations (Hasan, et al., 2022; Mökander & Floridi, 2022; Moss et al., 2021; Selbst, 2021; Tartaro et al., 2024). ERAs provide valuable insights into both theoretical governance and its effectiveness within practical case studies. These assessments are not independent, but they constitute valuable internal evaluations that focus on the potential negative impacts on stakeholders' rights and interests while also considering positive benefits. ERAs involve two main stages: identification of potential harms and their prioritization. Such assessments transcend legal compliance and serve as the primary mechanism for analyzing social impacts and anticipating future audit or assurance requirements in the evolving regulatory landscape (Hasan, et al., 2022).

Related work & current gap To the best of our knowledge, no research has yet embarked on taking such a proactive and structured approach toward XAI risk assessment. The only framework for systematically assessing explainable approaches is advanced by Sokol and Flach (2020). The proposed taxonomy facilitates the systematic comparison of explainability approaches and offers insights into their capabilities and discrepancies between their theoretical qualities and implementation properties. The work of de Bruijn et al. (2022) provide a comprehensive list of objections to XAI, including the difficulty of explaining AI to the public, the non-neutrality of explanations, the dynamic nature of algorithms, and other issues. Alongside pitfalls, they propose corresponding strategies to mitigate these risks at the governance level, emphasizing the importance of managing and addressing these concerns proactively. The recent survey by Baniecki and Biecek (2024) provides a comprehensive overview of adversarial attacks and defense mechanisms in XAI. While their work shares some commonalities with ours in addressing the security and trustworthiness of XAI systems, our research takes a broader perspective considering also contextual risks. To sum, the current research benefits from these works, yet stresses a perspective on XAI grounded in risk assessment, not just relying on XAI model selection or unstructured recommendations. By adopting this proactive approach to explanations design, we aim to anticipate not just the technical limitations of XAI, but also the risks stemming from sociotechnical considerations.

Method

We first performed a research literature retrieval grounded on concerns and vulnerabilities of XAI, from where we identified key technical risks. This preliminary analysis constituted the bedrock from which we departed our thematic analysis. As a second step, our search strategy through citation chaining and snowballing incorporated diverse disciplinary perspectives, including computer science, cognitive science, psychology, law, ethics, sociology, and others, ensuring a comprehensive view of the contextual risks associated with explanations in AI. This approach was inspired by social sciences studies informing the field of XAI (Lipton, 2017, 2018; Miller, 2019; Wilkenfeld & Lombrozo, 2015). This allowed us to garner a deeper understanding of how explanations function in non-AI contexts, enriching our understanding of potential risks when these concepts are transposed into the XAI domain.

Research retrieval & filtering We began targeting various the Scopus academic database and then expanding to other peer-reviewed sources such as ACM Digital Library and IEEE Explore. For search strings, keywords or concepts such as *explainable*, *XAI*, *interpretable ML* were incorporated with terms as *vulnerabilities*, *adversarial attacks*, *robustness*, *data poisoning* and others. Terms were chosen based on our prior knowledge of common challenges and threats faced by AI systems in general and XAI systems in particular. The departing Scopus queries were:

1. Query (1) targeted technical risks related to the robustness of XAI methods, including their vulnerability to adversarial attacks, model manipulation, and input perturbations.[2]
2. Query (2) focused on fairness risks in XAI, covering topics such as algorithmic bias, discrimination, disparate impact, and various fairness metrics and constraints.[3]
3. Query (3) addressed privacy and security risks associated with XAI, including information leakage, model inversion attacks, membership inference attacks, model extraction, and risks to intellectual property.[4]

To ensure a comprehensive search, we also included synonyms and related terms for each keyword. For example, when searching for *adversarial attacks*, we also used terms like *adversarial examples*, *adversarial perturbations*, and *adversarial manipulations*. This approach helped capture a wider range of relevant literature that may use slightly different terminology to describe similar concepts.

Selection criteria & analysis To ensure the relevance and quality of the articles included in our analysis, we included papers: (I°.) Published in a peer-reviewed journal, conference proceedings, or book chapters; (II°.) Focused on explainable AI from a perspective informed by risk assessment, associated vulnerabilities, or AI ethics frameworks; (III°.) Presented a theoretical or empirical analysis of risks related to XAI explanations, system architectures, or data; (IV°.) Written in English.

In addition to the structured search of XAI-specific literature, from our paper pool we expanded to similar works through citation chaining and snowballing incorporated diverse disciplinary perspectives, including computer science, cognitive science, psychology, law, ethics, sociology, and others. We deliberately included papers from non-XAI/AI contexts, particularly from the period before the establishment of the XAI program by DARPA in 2016 (Gunning & Aha, 2019). This decision was motivated by the recognition that the study of explanations has a long and rich history in fields such as psychology, cognitive science, philosophy, and human–computer interaction—e.g., (Clark & Brennan, 1991; Harman, 1965; Hempel & Oppenheim, 1948; Keil et al., 2000; Lombrozo, 2012; Salmon, 1984, 1989; Trout, 2002; Wilson & Keil, 1998). By drawing from this diverse body of knowledge, we aimed to gain a more comprehensive understanding of the potential risks and challenges associated with explanations in human communication, and to identify foundational concepts and theories that have shaped the current understanding of explainability in AI (Confalonieri et al., 2021).

Data extraction and analysis In analyzing this collection of papers, we adopted an iterative and reflexive process. We derived key themes directly from the literature and honed through continuous comparison with our expanding

[2] (1) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“robustness” OR “adversarial attacks” OR “adversarial examples” OR “adversarial perturbations” OR “model manipulation” OR “saliency maps” OR “counterfactual explanations” OR “concept activation vectors” OR “input perturbations”) AND (“risks” OR “vulnerabilities” OR “challenges” OR “issues”))

[3] (2) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“fairness” OR “bias” OR “discrimination” OR “disparate impact” OR “demographic parity” OR “equal opportunity” OR “algorithmic fairness” OR “fairness metrics” OR “fairness constraints” OR “fairness-aware learning”) AND (“risks” OR “vulnerabilities” OR “challenges”))

[4] (3) TITLE-ABS-KEY((“explainable AI” OR “XAI” OR “interpretable machine learning”) AND (“privacy” OR “security” OR “information leakage” OR “model inversion” OR “membership inference” OR “model extraction” OR “gradient leakage” OR “intellectual property” OR “trade secrets” OR “privacy-preserving” OR “secure multiparty computation”) AND (“risks” OR “vulnerabilities” OR “challenges”))”

dataset.[5] In particular, the thematic analysis was conducted in six phases following the guidelines proposed by Braun and Clarke (2006):

1. *Familiarization with the data* The researchers read the selected papers to gain an understanding of the content.
2. *Generating initial codes* Each researcher independently coded a subset of the papers, identifying initial themes and patterns related to XAI risks.
3. *Searching for themes* Through an iterative process of discussion and refinement, the researchers developed a preliminary set of themes and subthemes that captured the key risks associated with XAI systems.
4. *Reviewing themes* The researchers independently reviewed the preliminary themes and subthemes, checking their coherence and consistency against the coded data and the original papers. The researchers then met to discuss their findings and refine the themes and subthemes accordingly.
5. *Defining and naming themes* The researchers collaboratively defined and named the final set of themes and subthemes, ensuring that each theme captured a distinct and meaningful aspect of XAI risks.

We clarify that this partitioning into themes and subthemes is inherently interpretive and adaptive. We acknowledge that due to the complexity of the field and the variable lexicon used across the literature, certain papers may resonate with multiple subthemes or themes.

Categorization of risks in XAI systems

We developed a taxonomy categorizing the identified risks into two primary domains: *Technical Risks* (section "Technical risks"), related to the data and models of XAI systems, and *Contextual Risks* (section "Contextual risks"), associated with the informativeness and reception of XAI explanations. The interested reader can visualize the taxonomy in Table 2 in Appendix A. Risks reported are to be considered as not mutually excluding.[6]

Technical risks

In this subsection, we examine risks through a holistic lens rather than the more traditional approach of examining individual targets such as input data or the model itself. Our approach is centered on a comprehensive understanding of risks related to properties of the XAI models, such as model selection trade-offs, robustness against adversarial or unintentional perturbations, technical fairness, and privacy risks, as well as design evaluation.

Robustness risks

The trustworthiness of an explanation, and thus the overall XAI system, depends on its robustness to various types of uncertainties and perturbations. *Robustness Risks* relate to the stability and reliability of explanations in the presence of uncertainties, perturbations, or adversarial attacks. Robustness risks arise when explanations are sensitive to small changes in the input data, model parameters, or explanation methods, leading to inconsistent or misleading interpretations. Two primary dimensions of robustness risks in XAI can be identified as *adversarial attacks* and *discrepancies*.

Adversarial attacks are deliberate attempts to manipulate or mislead an XAI system (Carlini & Wagner, 2017b; Dombrowski et al., 2019; Goodfellow et al., 2015; Szegedy et al., 2014; Zhang et al., 2020). They can be targeted toward model explanations or the model's predictions themselves. These types of attacks are designed to be subtle, often involving minor, carefully crafted changes to the input data or the model parameters that lead to significant alterations in the output or explanations (Dombrowski et al., 2019; Zhang et al., 2020). Such attacks can greatly undermine the credibility and utility of an XAI system. Adversaries can manipulate input samples at will, and they might even have details about the model's parameters and architecture at their disposal (Biggio & Roli, 2018; Carlini & Wagner, 2017a; Ilyas et al., 2018; Madry et al., 2018; Papernot et al., 2017; Shafahi et al., 2019; Tramèr et al., 2020; Zhang et al., 2019).

[5] The thematic analysis was conducted by a team of three researchers with diverse expertise in XAI, AI ethics, and qualitative research methods. This interdisciplinary team composition ensured a comprehensive and rigorous analysis of the data. Researcher 1 (R1) has a background in computer science and XAI, with extensive experience in developing and evaluating XAI methods. Researcher 2 (R2) specializes in AI ethics and has published on the social and ethical implications of AI systems. Researcher 3 (R3) is an expert in qualitative research methods.
In the initial code generation phase, approximately 30% of the papers were analyzed by all three researchers independently. After defining the initial themes, the remaining papers were divided among the researchers for separate analysis. Regular meetings were held to discuss and refine the themes based on new insights. This iterative process continued until all papers were analyzed and thematic saturation was reached.

[6] We decided to arbitrarily adopt a categorization that reflects both the themes of literature retrieval and filtering exposed before, as well as citation chaining. We consider thus some of these risks mutual e.g., adversarial attacks can be used to manipulate the input data of the underlying AI system, which in turn can affect the fairness of the explanations generated by the XAI system; biased sociotechnical explanations (e.g., essentialism) might be used to justify unfair data distributions; technical privacy risks easily overlap with gaming opportunities, etc.

Explanation discrepancies occur when different explanation methods provide conflicting interpretations for the same model prediction or input. This lack of consistency includes variations in the underlying model, differences in the explanation algorithms, or noise in the data. Model manipulations, which could influence a large group of inputs at once, have been used for adversarial purposes (Dimanov et al., 2020; Heo et al., 2019). Manipulations require an adversary to be able to influence the training process/data or even control the model. This is enabled by poisoning attacks or constituted with query-based access only (Dong et al., 2021; Gu et al., 2019; Jagielski et al., 2018; Liu et al., 2018; Severi et al., 2021; Shafahi et al., 2018). These manipulations can either preserve the original model's functionality or focus on maintaining high accuracy, potentially improving the overall performance. The manipulated model might provide nearly the same predictions, but sensitive target features receive low relevance scores in the explanations. So-called backdooring attacks or Trojan attacks can evoke a target label when the input carries a certain trigger pattern (Gao et al., 2019; Gu et al., 2019; Jia et al., 2022; Liu et al., 2018; Severi et al., 2021). Among others, Robustness risks comprise:

(**T-RR-1**) *Attacks on saliency-based explanation methods* Saliency-based methods such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg & Lee, 2017) are vulnerable to adversarial attacks that aim to manipulate or obscure the true feature importance. Slack et al. (2020) demonstrate that these methods can be fooled by crafting adversarial classifiers that hide discriminatory behavior while appearing innocuous to LIME and SHAP. Similarly, Zhang et al. (2018) show that saliency maps can be perturbed in detectable ways by adversarial examples, and propose a detection technique based on training a classifier with both original data and saliency maps. Potential solutions to mitigate these risks include robust saliency estimation techniques (Adebayo et al., 2018), self-explaining neural networks (Alvarez-Melis & Jaakkola, 2018) that directly incorporate explanations into their architecture, adversarial training to improve model stability (Tang et al., 2022; Zhang et al., 2020), and leveraging adversarial explanations to gain a deeper understanding of the model's behavior (Woods et al., 2019).

(**T-RR-2**) *Manipulation of counterfactual explanations* Counterfactual explanations (Stepin et al., 2021; Wachter et al., 2017), which provide minimal changes to obtain a different outcome, are also susceptible to adversarial manipulation. Slack et al. (2021a) demonstrate that counterfactuals are sensitive to small input perturbations and introduce a technique to train seemingly fair adversarial models that provide low-cost recourse under perturbations, effectively deceiving users or obscuring biases. Virgolin and Fracaros (2023) propose robustness definitions for sparse counterfactuals and show that accounting for robustness helps reduce the cost of recourse under adverse perturbations. Further research focuses on detecting and mitigating manipulation effects, such as improving counterfactual plausibility (Keane & Smyth, 2020; Kenny & Keane, 2021), incorporating additional constraints (Keane et al., 2021; Kuhl et al., 2022) to ensure realistic counterfactuals, and evaluating robustness in specific application domains (Mishra et al., 2021).

(**T-RR-3**) *Attacks on concept-based explanation methods* Concept-based explanation methods, like TCAV[7] Kim et al. (2018), are vulnerable to adversarial attacks that can corrupt or misrepresent concepts. Ghorbani et al. (2019) demonstrate that interpretations of neural networks are fragile and can be altered by small, carefully crafted perturbations to the input data. They show this fragility applies to several widely-used feature importance interpretation methods. Brown and Kvinge (2023) further highlight the vulnerability of concept-based methods, specifically TCAV. They introduce "token pushing" attacks which manipulate the concept examples to induce misinterpretations, such as making irrelevant concepts appear important or hiding the importance of truly relevant concepts. Sinha et al. (2022) conduct a systematic study on the security vulnerabilities of concept-based models. Potential defenses include detection methods for adversarial examples (Ghorbani et al., 2019), careful curation and expansion of concept examples to cover potential gaps (Brown & Kvinge, 2023), and adversarial training to improve robustness (Sinha et al., 2022).

(**T-RR-4**) *Adversarial data perturbations affecting explanations* Perturbations in input data, such as those affecting PDP (Baniecki et al., 2022), can significantly alter explanations, reducing their reliability. Techniques to enforce or mitigate the effects of adversarial data perturbations include data poisoning attack strategies or frameworks targeting fairness measures or decision boundaries (Mehrabi et al., 2021; Solans et al., 2020; Zhang et al., 2021). Nanda et al. (2021) examine robustness bias, and Tang et al. (2022) propose a new training scheme called Adversarial Training on EXplanations (ATEX) to improve explanation stability.

(**T-RR-5**) *Explanation-aware backdoors* Explanation-aware backdoors are a type of malicious modification to an AI system's training data or model, specifically designed to manipulate the explanations generated by the model (Noppel et al., 2023). Unlike traditional backdoors that aim to manipulate the model's predictions (e.g., Chen et al. 2017, Gu et al. 2019, Veldanda et al. 2021), explanation-aware backdoors target the model's explanations directly. These backdoors can be used to conceal or obfuscate the true behavior of the model. For instance, an adversary could

[7] TCAV, or *Testing with Concept Activation Vectors* is a technique for interpreting the internal representation of a neural network by quantifying the degree to which a user-defined concept is important to a classification result (Kim et al., 2018).

craft a backdoor that makes a model's explanations highlight innocuous features when a trigger is present, while the actual prediction is based on sensitive or discriminatory features. Noppel et al. (2023) demonstrate several variants of explanation-aware backdoors.

(**T-RR-6**) *Debugging challenges* The effectiveness of post-hoc model explanations for diagnosing model errors has been challenged (Adebayo et al., 2020, 2022). There are indications that many explanation methods are ineffective in identifying various models, data, and test-time contamination bugs. Further, Dai et al. (2022) emphasized that disparities in explanation quality may arise in complex and non-linear models, suggesting an unexplored risk of unfairness in real-world decision-making introduced by post-hoc explanation methods.

(**T-RR-7**) *Transferability of adversarial attacks* Adversarial attacks targeting one explanation method may also affect other methods, potentially compromising the overall robustness of XAI systems. This transferability risk has been highlighted by several studies, including Lakkaraju et al. (2020), who demonstrated that adversarial examples can transfer across different explanation methods, and Sinha et al. (2021), who showed that adversarial attacks can be transferable across different natural language processing models and explanation techniques.

Fairness risks

Fairness risks concern the potential for explanations to reflect, introduce, or amplify biases and discrimination against certain individuals or groups based on sensitive attributes such as race, gender, or age. These risks can perpetuate or amplify existing societal biases and lead to unjust treatment of disadvantaged populations. Different typologies of "*fairness attacks*" in XAI systems are outlined:

(**T-FR-1**) *Fairwashing* Fairwashing involves the manipulation of explanations to present an unfair ML model not as such (Aïvodji et al., 2019, 2021). This deceptive practice distorts fairness metrics, creating a misleading impression of fairness. Fairwashing attacks can be particularly challenging to detect, as they often involve subtle changes to the explanations that are difficult to distinguish from legitimate ones. Aïvodji et al. (2019) demonstrated that fairwashing attacks can be effective in fooling both human users and automated fairness auditing tools, highlighting the need for more robust fairness evaluation methods in XAI systems.

(**T-FR-2**) *Biased sampling* Biased sampling deceives fairness auditing tools by producing datasets that portray an unfair model as unbiased (Fukuchi et al., 2020; Laberge et al., 2022). This strategy helps to mask the unfairness of a model. By carefully selecting a subset of the data that appears to be fair, biased sampling attacks can manipulate the explanations generated by XAI methods, making it difficult to identify the underlying biases in the model. Fukuchi et al. (2020) introduced a stealthily biased sampling procedure that can effectively fool fairness auditing tools, emphasizing the importance of developing more robust sampling techniques and fairness evaluation metrics.

(**T-FR-3**) *Adversarial poisoning* Adversarial poisoning corrupts training data to induce unfair classification disparities, particularly regarding sensitive attributes (Mehrabi et al., 2021; Solans et al., 2020). This deception results in skewed accuracy metrics. By carefully crafting adversarial examples and injecting them into the training data, adversarial poisoning attacks can manipulate the learned decision boundaries and explanations, leading to unfair outcomes. Mehrabi et al. (2021) and Solans et al. (2020) demonstrated the effectiveness of adversarial poisoning attacks in inducing unfairness in machine learning models, highlighting the need for more robust training procedures and fairness-aware data preprocessing techniques.

(**T-FR-4**) *Manipulation of post-hoc explanations* The manipulation of post-hoc explanations, as revealed in studies by Dimanov et al. (2020), Laberge et al. (2022), and Merrer and Trédan (2020), involves masking the role of sensitive features and undermining the reliability of remote explainability, thus affecting race, gender,or other sensitive attributes. By carefully perturbing the input data or the model parameters, an attacker can manipulate the post-hoc explanations generated by XAI methods, hiding the true importance of sensitive features and making the model appear fairer than it actually is.

(**T-FR-5**) *Explanation disparity risks* Other studies highlight the potential for explanation methods to introduce or echo unfairness during model evaluation. Dai et al. (2022) stress the importance of high-quality explanations, pointing out increased disparities with more complex models. Balagopalan et al. (2022) discovered significant differences in explanation model fidelity across protected subgroups during a quality audit. They underscore the importance of user awareness regarding fidelity gaps and draw attention to biased explanation models as an uncharted challenge. These findings suggest that explanation methods themselves can introduce or perpetuate unfairness, even when the underlying model is fair.

Evaluation risks

Evaluation risks regards the challenges and limitations in assessing, validating, and interpreting the quality, reliability, and effectiveness of explanations. Evaluation risks arise when the metrics, methods, or assumptions used to evaluate explanations are flawed, incomplete, or susceptible to manipulation, leading to incorrect conclusions or decisions based on the explanations. Examples of evaluation risks include:

(**T-ER-1**) *Dependence on model assumptions* The validity and effectiveness of explanations and robustness measures are profoundly impacted by the assumptions made during the modeling process (Noack et al., 2021). These assumptions may include, but are not limited to, linearity, feature independence, or the absence of interactions among variables. When these assumptions are violated, the explanations generated by the XAI system may be misleading or fail to capture the true underlying relationships in the data. If the underlying model assumptions are incorrect or overly simplified, the explanations or robustness measures derived from the model could be misleading or incorrect. Arora et al. (2022) highlighted how the limitations of specific explanation techniques could result in a failure to improve understanding or manipulation of complex models, such as BERT-based classifiers[8]

(**T-ER-2**) *Evaluation manipulation and deception* There exists a risk of malicious actors manipulating the evaluation of explanations to deceive users or system administrators (Warnecke et al., 2020). This risk could lead to incorrect decision-making or potential system vulnerabilities, particularly in high-stakes applications such as cybersecurity or healthcare. Further complicating this issue, Adebayo et al. (2022) showed that post-hoc explanation methods might not be effective in detecting a model's reliance on spurious signals in the training data, particularly when the spurious signal to be detected is unknown at test-time.

(**T-ER-3**) *Robustness-explainability trade-off* Even if contested (Rudin, 2019), a potential trade-off might arise between accuracy and interpretability in AI models (Noack et al., 2021). This complexity suggests that the relationship between robustness and explainability is not entirely understood. As an example, in the context of Graph Neural Networks (GNNs), Agarwal et al. (2022) pointed out the violation of several desirable properties, such as faithfulness, stability, and fairness preservation, indicating that not all explanation methods may be reliable.

(**T-ER-4**) *Reliability and consistency of interpretation methods* The effectiveness of various interpretation methods has been questioned (Hooker et al., 2019; Tomsett et al., 2020). These studies found inconsistencies in the reliability of saliency metrics and interpretability methods, raising concerns about their validity and usage. In a similar vein, the work of Huber et al. (2022) and Kim et al. (2022) both indicated a need for computational evaluation and comparison of different perturbation-based saliency map approaches.

[8] Machine learning models that use the BERT (Bidirectional Encoder Representations from Transformers) architecture, which is designed to pre-train deep bidirectional representations from unlabeled text, for various natural language processing tasks such as text classification (Devlin et al., 2019).

Contextual risks

Contextual risks in XAI systems encompass a broad range of potential issues that can arise when these systems are deployed in real-world contexts. These risks go beyond the technical aspects of the systems themselves and include security vulnerabilities, accountability challenges, cognitive biases and heuristics, argumentative and logical fallacies, epistemological issues of underdetermination and overdetermination, problematic conceptualizations such as reification and essentialism, and ethical concerns. While these risks are diverse in nature, they share some common characteristics. They all have the potential to undermine the effectiveness, trustworthiness, and fairness of XAI systems, and they can lead to unintended consequences or harms for individuals and society. These risks often involve complex interactions between the technical, psychological, social, and ethical dimensions of XAI systems, requiring an interdisciplinary approach to understanding and mitigating them.

Security risks

Security risks in XAI systems encompass vulnerabilities that can be exploited by malicious actors to compromise the integrity, confidentiality, or availability of the system and its explanations. These risks can have severe consequences, such as privacy breaches, intellectual property theft, or system manipulation.

(**CT-SR-1**) *Privacy Vulnerabilities* still on a technical level, Quan et al. (2022) highlight the risks associated with post-hoc explanations, revealing that they amplify the vulnerabilities of ML models to various attacks. These explanation methods can act as information-rich side-channels, enabling adversaries to conduct evasion, membership inference, and model extraction attacks. These insights emphasize the complexity of the privacy-explainability trade-off. Shokri et al. (2021) analyze feature-based model explanations to show how they might inadvertently leak sensitive information about a model's training set through membership inference attacks. This leakage indicates the existence of individual data in a model's training set, underscoring a challenging trade-off between data privacy and explanation quality. Echoing these findings, Duddu and Boutet (2022) alert to attribute inference attacks. In their study, sensitive attributes such as race or sex can be inferred from model explanations, reinforcing the understanding of model explanations as a potent attack surface and a threat to data privacy. Similarly, Liu et al. (2022) propose an approach based on Rényi differential privacy (RDP), ensuring robust interpretation through top-k robustness and offering a balance between robustness and computational efficiency.

(**CT-SR-2**) *Instrumentalization* Value theory, which considers transparency as an extrinsic value, suggests that transparency has utility only when it serves as a means to fulfill an intrinsic value. In some scenarios, transparency may be inconsistent when juxtaposed with intrinsic values such as the protection of privacy over personal information (Ronnow-Rasmussen, 2015). Despite being often viewed as a desirable outcome of explainability for its potential to enhance understanding and trust in the system, transparency carries its risks. One such risk is the potential for instrumentalization, where explanations allows the gaming intentions of recipients. Disclosing detailed information can enable individuals or organizations to exploit loopholes or vulnerabilities for personal gain (Agre, 2014). Explanations can inadvertently provide insight into sensitive intellectual property or trade secrets, allowing competitors or malicious actors to gain an advantage. As extensively detailed within technical risks, other concerns include the potential for adversarial attacks and reverse engineering of models upon disclosing explanations (Kuppa & Le-Khac, 2020; Oh et al., 2019), as well as the possibility of jeopardizing the security of individuals or organizations through the disclosure of sensitive information (Weitzner et al., 2008).

Accountability risks

Accountability is a crucial aspect of explanations, referring to the responsibility and justification that explainers have for their claims and actions. Ensuring accountability in XAI systems, however, can be particularly challenging due to several factors (de Bruijn et al., 2022).

(**CT-ACCR-1**) *Traceability of explanation design* The inherent complexity of AI systems as well as the supply chain related to data lineage and deployment can obscure the agent making assumptions underlying an explanation, making it difficult to trace the reasoning or actions derived from their outputs (Cobbe et al., 2023). This obscurity can be exacerbated when AI systems are deployed maliciously or manipulated to deceive, for example, by using them outside of controlled contexts to attack or pollute the informational sphere (Weidinger et al., 2022).

(**CT-ACCR-2**) *Appraising explainers* Epistemic authority, or the perceived expertise and credibility of an explainer, may project a false sense of certainty or completeness over explanations, fostering unwarranted trust in the explainer's authority and judgments. This phenomenon can lead to deference to authority, where recipients accept explanations without critical evaluation or consideration of alternative perspectives (Kruglanski et al., 2005; Zagzebski, 2012).

(**CT-ACCR-3**) *Explainer's overconfidence* Epistemic arrogance, where explainers overestimate their knowledge or abilities, can lead to overconfidence or dismissal of alternative perspectives or evidence (Kruglanski, 1989). Judgmental overconfidence concerning explanatory understandings engenders inflated self-assessments among both explainers and recipients (Kruger & Dunning, 2000; Yates et al., 1997). This cognitive bias can stifle open-mindedness and critical thinking necessary for effective explanations, potentially leading to misguided or harmful decisions.

Heuristics & reception risks

Heuristics and reception risks in XAI systems arise when explanations are influenced by cognitive biases or heuristics, or misinterpreted by recipients (Horton & Keysar, 1996). These risks can lead to the oversimplification or misrepresentation of complex issues, the reinforcement of existing biases, or the misinterpretation of the explanations' implications. Explanations carry the risk of being perceived as a panacea or placebo, leading to a false sense of understanding. People experience cognitive satisfaction when they feel they understand something, often called a "visceral rush of understanding" (Gopnik, 1998). This can lead to an overestimation of one's own understanding, a bias known as the "illusion of explanatory depth" (Rozenblit & Keil, 2002). Furthermore, explanations that are framed in a certain way, such as by invoking neuroscience or other technical jargon, can be particularly seductive, even if the information is irrelevant or misleading (Weisberg et al., 2008). Such risks can distort comprehension of the subject matter, predominantly due to:

(**CT-HRR-1**) *Cognitive heuristics* Heuristics are cognitive shortcuts that might lead to biased or incomplete reasoning. Two main heuristics potentially distort explanations. The *availability heuristic*, according to Tversky and Kahneman (1973), might result in misjudged likelihoods or importance due to reliance on easily retrievable information. On the other hand, the *representativeness heuristic* could contribute to stereotyping or discrimination by judging events' likelihood based on their fit into specific categories or stereotypes (Kahneman & Tversky, 1972).

(**CT-HRR-2**) *Implications of language and semantic framing* The choice of language and framing can unintentionally oversimplify or misrepresent explanations. Ambiguous language might cause misunderstandings or misinterpretations (Levinson, 2000), while information framing could shape perceptions and understanding, potentially leading to diverse conclusions or attitudes (Kahneman & Tversky, 1984).

(**CT-HRR-3**) *Cognitive biases* Prior beliefs and biases can influence how information is interpreted and presented, leading to oversimplification or misrepresentation. Confirmation bias-the tendency to seek and interpret information that validates existing beliefs-might result in a narrow understanding of the subject (Nickerson, 1998). Simultaneously, the illusion of explanatory depth, which is the overestimation of one's understanding of a topic, could lead to overconfidence in the provided explanations despite possible knowledge gaps or inaccuracies (Rozenblit & Keil, 2002). Lastly, the recency effect considers how the most recent explanations are given more weight than older ones, even when the older ones may be more accurate or relevant. This bias can be counterbalanced by consistently emphasizing the most relevant or accurate explanations, irrespective of their recency (Tubbs et al., 1990; Tversky & Kahneman, 1973).

Argumentative & logical risks

Heuristics and reception risks are primarily concerned with how cognitive biases, heuristics, and the recipient's interpretation can influence the understanding and impact of explanations. These risks arise from the interaction between the explanations and the human recipients, and they are largely shaped by the recipients' cognitive processes, prior knowledge, and contextual factors. On the other hand, argumentative and logical risks focus on the internal structure, reasoning, and argumentation of the explanations themselves. These risks stem from flaws in explanations' logical construction, like fallacies, circularity, or weak inferences. While these risks can also impact the recipients' understanding and acceptance of the explanations, they are primarily rooted in the explanations' inherent logical and argumentative qualities.

An example is brought by *aporia*, an argumentative fallacy where the recipient is confronted with a situation or explanation that contains an insoluble internal contradiction or paradox, resulting in confusion or bewilderment (Latour, 1988). Another is *non-sequitur*, where the explanation fails to logically follow the premises or provide a reasonable conclusion (Walton, 2010). In some cases, explanations may even induce a situation of *Obscurum per obscurius, ignotum per ignotius* (Translatable as "The obscure through the more obscure, the unknown through the more unknown"), an attempt to explain something by using concepts or terms that are even more obscure or unfamiliar to the recipient (Galilei, 1953; Wikipedia, 2023).

Circularity and tautology, as fallacies, hinder the transmission of new information and obstruct a deeper comprehension of the subject matter. They are primarily self-referential, offering no informative value.

(**CT-ALR-1**) *Circular reasoning* A form of fallacy, circularity or "begging the question", arises when the conclusion of an argument is repackaged as one of its premises. This fallacy creates a loop of self-justifying statements that lack external validation and meaningful depth (Hahn, 2011; Walton, 1994). In the context of AI explanations, circularity may manifest as an overreliance on the model's internal logic or mechanisms, devoid of external corroborative evidence or a broader understanding of the problem context. Mitigating circular reasoning in explanations requires grounding assertions in data, external findings, and the broader context of the problem addressed.

(**CT-ALR-2**) *Tautology* Tautology is another form of fallacy that surfaces as redundant repetition in logic or language, where a statement is framed as inherently true without conveying additional insight (Meibauer, 2008). Tautologies present as excessive use of jargon or technical terms that obscure the true mechanism or contribute to the illusion of explanatory depth. Strategies to avoid tautology involve the use of precise and accessible language, avoidance of redundancies, and inclusion of explicit detail to highlight unique concepts or processes.

To counter these argumentation risks, explainers shall strive to design explanations that are clear, logical, and based on familiar concepts and argumentation style (Keil, 2006; Keysar & Bly, 1995; Walton, 2008). Avoiding circularity and tautology extends beyond mere linguistic precision and logical structure, encompassing a critical assessment of assumptions and beliefs underpinning explanations. Thus in scientific disciplines, including AI, explanations should be empirically grounded, testable, and open to revision based on new evidence (Popper, 2014; Stanford, 2006).

Underdetermination & overdetermination

On an epistemological level, the phenomena of underdetermination and overdetermination can pose multifaceted challenges in the domain of explanatory practice, giving rise to potential pitfalls in developing and presenting explanations.

(**CT-DETR-1**) *Underdetermination* Philosophical discourse in the field of science extensively addresses underdetermination, particularly in the context of theory selection (Kuhn, 1981; Stanford, 2006). The dilemma arises when there exist several theories with comparable plausibility, all capable of explaining the same observed phenomena but with no decisive criteria available for preferring one over the others. This inherent ambiguity often ignites controversy among scientists and may culminate in an impasse or lack of consensus in the scientific community. The so-called *Rashomon effect* is illustrative of underdetermination, as it underscores the possible multiplicity and subjectivity in the interpretation of the same event (Derrida, 2016; Leventi-Peetz & Weber, 2022).

(**CT-DETR-2**) *Overdetermination* Conversely, overdetermination becomes pertinent in disciplines such as psychology and cognitive science. It is observed when numerous causes or factors are invoked to explain a single phenomenon, even when they may not all be necessary or directly pertinent. Consequently, an explanation becomes mired in excessive complexity, obscuring rather than illuminating the understanding of the phenomenon in question (Waldmann, 2000). An essential strategy for mitigating underdetermination and overdetermination involves careful scrutiny and evaluation of the evidence at hand, along with a pursuit of coherence and parsimony in the explanatory model (Lombrozo, 2011).

Reification & essentialism

Reification and essentialism are closely related risks that arise when explanations oversimplify or misrepresent complex social constructs or reinforce stereotypical assumptions about individuals or groups. Reification and essentialism have been studied in various fields, including social psychology, cognitive psychology, and philosophy.

(**CT-RER-1**) *Reification* It can be intended a social psychology risk, associated with explanations occurring when abstract concepts or constructs are treated as if they are concrete entities with fixed identities and values. This oversimplification or misrepresentation of a phenomenon can hinder further inquiry and understanding (Schank, 2004). For example, the reification of mental disorders as discrete entities with clear boundaries can obscure the complexity and variability of mental health experiences, which may lead to misdiagnosis or inappropriate treatment (Hyman, 2010). In philosophy, it has been used to describe how abstract concepts, such as justice or freedom, can be treated as if they are concrete entities with a clear definition and identity (Vandenberghe, 2015). In psychology, reification is linked to overgeneralizing from limited observations and relying on stereotypes and heuristics rather than critical thinking and empirical evidence (Heft, 2003).

(**CT-RER-2**) *Essentialism* On the other hand, it occurs when an explanation attributes inherent or immutable characteristics to a particular entity or group, based on preconceived notions or assumptions. This can lead to stereotyping or discrimination, and may be used to justify harmful or unjust practices or policies. Essentialism has been studied extensively in social psychology and has been shown to contribute to intergroup conflicts and inequalities (Devine, 1989; McGarty et al., 2002; Rhodes & Moty, 2020). Moreover, the use of essentialist language in scientific explanations can have negative consequences for marginalized groups, reinforcing biases and perpetuating stereotypes (Inbar & Lammers, 2012). For instance, essentialist explanations of mental health conditions that attribute certain traits or behaviors to particular genders or ethnic groups can perpetuate harmful stereotypes and contribute to disparities in access to care and treatment (Halpern, 2000; Rossnan, 2006).

Both reification and essentialism can pose significant risks to the quality and effectiveness of explanations. From a social psychology perspective, deployers of XAI critically evaluate the language and concepts they use to avoid the superimposition of distorted frames over complex phenomena (Keil, 2006). Similarly, concepts and constructs shall be recognized in their complexity and potential for variation across contexts and individuals (Gopnik et al., 2001), avoid making unwarranted assumptions about the inherent characteristics of individuals or groups (Medin & Ortony, 1989). Some approaches to counter the risks of reification and essentialism include using probabilistic or fuzzy concepts that acknowledge the variability and complexity of phenomena and recognizing the role of social and cultural factors in shaping experiences and identities (Haslam et al., 2000; Medin, 1989).

Ethical concerns

To conclude, we stress how explanations carry ethical implications, especially when they involve decisions impacting individuals or groups. In legal or medical contexts, for instance, explanations can significantly affect people's lives and well-being, contributing to systemic biases and injustices that might stem from biased data, flawed algorithms, or misinterpretations by human decision-makers (Angwin et al., 2016; de Bruijn et al., 2022; Shokri et al., 2021). Not only related to essentialism, explanations can perpetuate harmful or discriminatory narratives with the presumption of algorithmic accuracy, reinforcing views of certain sub-populations and exacerbating the marginalization and oppression of already disadvantaged groups (Eubanks, 2018; Harding, 1991; Rahman, 2020).

To recognize such ethical concerns necessitates diverse perspectives and voices in discussions around explainability and its ethical implications, including public engagement and participatory design (Cheng et al., 2019; Ehsan et al., 2022; Langer et al., 2021). In terms of public or business deliberation, it is important to acknowledge the potential limitations and trade-offs associated with integrating ethical considerations into XAI systems. As an example, certain explanations might be geared to justify not just opposite ethical instances, but rather highlight the pros and cons of each.

A risk assessment framework for XAI systems

Building upon the comprehensive categorization of technicals and contextual risks in XAI systems, we propose a multi-layered risk assessment framework designed to guide the identification, prioritization, and mitigation of these risks in practice. The proposed multi-layered risk assessment framework for XAI systems draws inspiration from ERA methodology, which has gained traction in the AI governance and risk management domain (Hasan, et al., 2022; Mökander & Floridi, 2022; Moss et al., 2021; Selbst, 2021). The framework consists of three key layers: the *Intervention Layer* (section "Intervention layer: *risk prioritization & mitigation*"), which focuses on risk prioritization and the implementation of targeted mitigation strategies; the *Management Layer* (section "Management layer: *iterative risk assessment process*"), which emphasizes continuous monitoring, adaptive risk reassessment, and feedback-driven improvement; and the *Information Layer* (section "Information layer: *documentation & communication*"), which ensures transparency through comprehensive documentation and communication. The framework—visually summarized in Table 3 of the Appendix A—provide a structured approach for proactively managing XAI risks and fostering responsible development and deployment of XAI systems.

Intervention layer: *risk prioritization & mitigation*

We depart with a tiered intervention mechanism, facilitating the effective allocation of resources in response to perceived risks, with primary emphasis on those holding the highest likelihood and potential impact. We envision this risk prioritization as an adaptable process, shifting focus according to emerging challenges within the context of XAI system deployment and development. Our risk mitigation strategies are bespoke in nature, tailored specifically to the context, needs, and identified risks within the XAI system under consideration. The Intervention Layer aligns with the risk prioritization stage of ERAs, where identified risks are assessed based on their likelihood and potential impact (Selbst, 2021).

Development of a risk matrix

The creation of a risk matrix provides a visual representation of risks based on their likelihood and impact. This enables effective prioritization of mitigation efforts. The risk matrix should be updated dynamically as new risks are identified or the XAI system evolves. Risk identification comprises the following components:

- *Categories* Risks should be segmented into meaningful categories. The categorization of risks proposed in section "Categorization of risks in XAI systems" and visually represented in A.1 can serve as a touchstone that users of the framework can employ. Risks could be categorized first as technical or contextual, and then further specified into more detailed categories, such as robustness risks (**T-RR**), fairness risks (**T-FR**), evaluation risks (**T-ER**), security risks (**CT-SR**), accountability risks (**CT-ACCR**), and so on.
- *Ownership* When possible, clearly defined responsibility for each risk should be allocated to individuals or teams, taking into account the concept of distributed morality for accountability (Floridi, 2013, 2016a).
- *Scores* A standardized scoring system should be used to assess the likelihood and impact of each risk. The scoring system should be based on a combination of quantitative and qualitative factors, considering the potential consequences of each risk on the XAI system's performance, fairness, security, and overall trustworthiness.

Implementation of mitigation actions

For each identified risk, specific mitigation actions are devised to reduce the probability or severity of the risk. These mitigation actions can be broadly categorized into: *Technical* mitigation actions involving the implementation of strategies to enhance robustness, fairness, and privacy; *organizational* actions such as forming a governance committee; *procedural* actions like scheduling regular internal assessments or external audits.

Technical mitigation actions

XAI systems face various risks, including robustness, fairness, security, privacy, and evaluation challenges. To address these risks, a range of technical mitigation actions can be employed at different stages of the XAI pipeline:

Data preprocessing Data preprocessing techniques, such as re-sampling or re-weighting (Ghalebikesabi et al., 2021; Vreš & Robnik-Šikonja, 2022), can help mitigate data biases and enhance model fairness (**T-RR-(1-5)**, **T-FR-2**). However, it is essential to be aware of potential data poisoning attacks that can manipulate the training data to influence model behavior and explanations (Baniecki & Biecek, 2022; Baniecki et al., 2022). To mitigate these risks, practitioners can employ data sanitization techniques to identify and remove poisoned data points, and use robust aggregation methods for global explanations (Liu et al., 2022; Rieger & Hansen, 2020). As a side consideration, stealthily biased sampling (Fukuchi et al., 2020) can be used to manipulate fairness metrics and conceal biases. To counter this, statistical tests can be used to detect significant differences between the original and sampled data distributions, and multiple fairness metrics should be compared across different subgroups to identify hidden biases (Fukuchi et al., 2020).

Model training and explanation generation

- Adversarial training (Lakkaraju et al., 2020; Madry et al., 2018), minimax optimization (Lakkaraju et al., 2020), and certifiably robust explanations (Cohen et al., 2019; Liu et al., 2022; Virgolin & Fracaros, 2023; Wicker et al.,

2022) can improve the model's resilience against adversarial attacks and backdoors (**T-RR-(1-5)**, **T-ER-3**).

- Fairness-aware explanation methods, such as those considering sensitive attributes and incorporating fairness constraints (Carmichael & Scheirer, 2023; Ferry et al., 2022; Weerts et al., 2023), can help mitigate biases in explanations (**T-FR-(1-5)**). However, achieving perfect fairness may not always be possible and may involve trade-offs with other desirable properties of explanations (Dai et al., 2022; Mehrabi et al., 2022).
- Focused sampling and on-manifold explainability techniques (Ghalebikesabi et al., 2021; Vreš & Robnik-Šikonja, 2022) can improve the robustness of LIME and SHAP explanations (**T-RR-(1-5)**, **T-FR-2**), but their effectiveness may depend on the quality of the sampling process and the characteristics of the data and model.

Explanation validation and evaluation

- Explanation validation methods, such as those proposed by Adebayo et al. (2018) and Zhang et al. (2018), Dai et al. (2022), can assess the fidelity, coherence, and stability of explanations (**T-RR-(1-5)**, **T-FR-1**, **T-FR-4**, **T-ER-(1-3)**). Nevertheless these methods can be computationally expensive and may not guarantee the absence of all biases or vulnerabilities.
- Model and data debugging techniques (Adebayo et al., 2020, 2022; Baniecki et al., 2022) can help diagnose errors and enhance robustness (**T-RR-(1-5)**), but their effectiveness may depend on the availability of appropriate tools and expertise.
- Uncertainty quantification frameworks, like MeTFA (Gan et al., 2022), can provide a measure of explanation uncertainty and increase stability in adversarial scenarios (**T-RR-(1-5)**). This still considering that quantifying uncertainty may not always be straightforward and may depend on the quality of the hypothesis tests and assumptions.

Security and privacy

- Data reconstruction attacks can exploit explanations to retrieve sensitive information about the training data (Ferry et al., 2022). Defenses against such attacks include limiting the granularity of explanations and applying differential privacy techniques (Dwork, 2006; Liu et al., 2022; Patel et al., 2022).
- Explanations can be used to perform membership inference attacks, breaching the privacy of individuals whose data was used to train the model (Shokri et al., 2021). Regularization techniques (Chen et al., 2019; Dombrowski et al., 2022) and knowledge distillation (Papernot et al., 2016) can help mitigate these risks, but may impact explanation quality.

Emerging techniques

- Concept-based explanations, such as TCAV (Kim et al., 2018), can provide human-understandable explanations but may face challenges in terms of robustness and generalizability (Brown & Kvinge, 2023).
- Counterfactual explanations (Stepin et al., 2021) can offer actionable insights but may be sensitive to adversarial perturbations (Keane & Smyth, 2020; Keane et al., 2021; Kuhl et al., 2022; Slack et al., 2021a). Techniques such as robust optimization (Cohen et al., 2019; Lakkaraju et al., 2020; Virgolin & Fracaros, 2023) and recourse invalidation rate minimization (Pawelczyk et al., 2023) can help improve their robustness.

While this list of mitigation actions covers a wide range of strategies, it is not exhaustive, and future research should aim to expand upon this framework as the field of XAI evolves. Practitioners should carefully consider the trade-offs and limitations associated with each technique and select the most appropriate strategies for their specific use case (Baniecki & Biecek, 2024; de Bruijn et al., 2022).

Organizational mitigation actions

- *Establishing a governance committee* Forming a committee comprising experts from different domains can improve risk management. This committee oversees the risk assessment process and ensures adherence to regulatory and ethical standards. This committee could, for instance, ensure that technical risks are mitigated effectively, while, for contextual risks, oversee the disclosure of information to prevent instrumentalization (**CT-SR-2**) or deploy measures such as obfuscation, abstraction, and pseudonymization to protect sensitive information.
- *Defining accountability* Explicit roles and responsibilities in managing risk, such as in **CT-ACCR-1** explanation design traceability, can enhance accountability and promote coordinated action (Floridi, 2016a). To address accountability risks, explainers should be mindful of their own epistemic limitations and to recognize the value of diverse perspectives and knowledge. Yet, even when systems are complex and assigning responsibility individually is not feasible, it is important to devise a method to assign it collectively using a distributed morality (Floridi, 2013, 2016a): within this lens, a consequence can be seen as a product of a series of interconnected actions produced by a network of agents. Our first step should be to recognize these nodes of "distributed moral actions". Leveraging the idea of "faultless accountability" or "strict liability", full moral responsibility is

bestowed on all agents within the relevant causal network: essentially, we consider all nodes as "responsible by default". Subsequently, an "overridability clause" may be employed to reassign responsibility in varying degrees, or even remove it completely, if an agent can prove they had no participation in the interactions. Lastly, it should be implemented a recurring adjustment mechanism until reaching a level that is axiologically satisfactory.

- *Promoting a risk-aware culture* Fostering a culture that is conscious of and proactive towards risk management can help to address the **CT-DETR-1** underdetermination and **CT-DETR-2** overdetermination phenomena. Regular training sessions can emphasize the importance of pursuing coherence and parsimony in explanatory models while mitigating risks associated with uninformative, misleading, or discriminating explanations (**CT-RER-1**, **CT-RER-2**, **CT-HRR-1**, **CT-HRR-2**, **CT-HRR-3**).

Procedural mitigation actions

- *Dynamic risk assessment* A continuously updated risk assessment is crucial in managing the dynamic and complex nature of XAI systems. Having an iterative process that can trace explanation design and appraise explainers can help to prevent risks like overconfidence and epistemic arrogance (**CT-ACCR-2**, **CT-ACCR-3**). Moreover, a recurring adjustment mechanism, such as an "overridability clause" in assigning responsibility, could be an important part of this assessment process (Floridi, 2013).
- *Ethical considerations* XAI systems have the potential to significantly impact individuals and society, making it crucial to integrate ethical considerations into their design and deployment. To address these concerns, it is recommendable for XAI designers to be aware of potential ethical implications over explanations' impact and strive to integrate ethical considerations into the design and deployment of explainable systems (Floridi, 2016b; Robbins, 2019). Practical guidelines, like ethical impact assessments, ethics committees, or Value Sensitive Design (VSD) principles, can provide actionable guidance for developers and policymakers to operationalize ethical considerations in XAI design (Friedman & Kahn, 2002; Hagendorff, 2019; Morley et al., 2023). During deployment, subjecting these systems to ongoing evaluation and scrutiny is crucial to ensure that ethical considerations are effectively integrated and maintained (Löfström et al., 2022; Sokol & Flach, 2020).

Management layer: *iterative risk assessment process*

Building upon the risk mitigation strategies established in the Intervention Layer, the Management Layer emphasizes continuous monitoring, adaptive risk reassessment, and feedback-driven improvement. This layer aligns with the iterative nature of ERAs, which require ongoing monitoring and updating of risk assessments as the AI system evolves and new risks emerge (Mökander & Floridi, 2022; Morley et al., 2023; Tartaro et al., 2024).

Continuous monitoring and adaptive risk reassessment

Rigorous, systematic auditing and monitoring practices are established, alongside a flexible approach to risk reassessment that adjusts in response to system evolution or environmental changes. Automated risk assessment tools that adapt to changes in the system or its operating environment are employed, using dynamic risk assessment methods (Raji et al., 2020; Raveendran et al., 2022; Tartaro et al., 2024).

Feedback-driven improvement

Mechanisms to gather and integrate feedback from various stakeholders are established, refining the system and its processes in a user-centric manner.

- *Feedback collection* User surveys, stakeholder meetings, and open forums are conducted to collect feedback on the system's operation, explanation generation, and potential areas of concern, following user-centered design principles (Cabitza et al., 2023; Ehsan et al., 2022; Langer et al., 2021; Liao & Varshney, 2021).
- *System refinement* The collected feedback is used to refine the explanation generation process, enhance system security, and address other areas of concern. For example, if users find the explanations too technical, adjustments are made to simplify the language used or provide additional contextual information. This can help tackle the **CT-DETR-2** overdetermination risk by focusing improvements on actual user needs and concerns.

Information layer: *documentation & communication*

The final layer of the risk assessment framework, the Information Layer, ensures transparency through comprehensive documentation and communication. This layer aligns with the importance of transparency and stakeholder engagement in ERAs (Moss et al., 2021).

Integration of intrinsic values

Transparency is integrated with other core values, such as accessibility and reproducibility. Relevant information is made readily available and comprehensible to a diverse array of stakeholders. Risk assessment findings are presented in a format that is easily digestible and understandable, regardless of the stakeholder's technical expertise, helping to bridge the gap between experts and non-experts, fostering informed decision-making, and promoting stakeholder engagement.

Documentation and reporting

Develop comprehensive documentation on the XAI system, including its architecture, data sources, algorithms, and explanation techniques, making it accessible to authorized stakeholders.

- *Comprehensive documentation* The pivotal function of documentation extends beyond record-keeping to delineating the intended and unintended uses of a particular AI system. Throughout the development and deployment AI pipeline, the concept of model cards is introduced (Mitchell et al., 2018). These comprehensive documents, widely employed today by developers, researchers, and industries, detail the technical specifications of a specific AI model, employing language that is as accessible as possible to a diverse array of stakeholders, ranging from policymakers to individuals with more technical backgrounds. Concurrently, considerable effort is devoted to documenting the dataset upon which a given AI model has been trained. The research conducted by Gebru et al. (2021) highlights the advantages not only for the technical and social appraisal of certain datasets but also for understanding their societal implications. For instance, the potential under-representation or over-representation of specific populations or languages within a dataset can have significant technical and social consequences.
- *Performance reports* Reports on the system's performance, identified risks, and mitigation measures are regularly published, ensuring that authorized stakeholders are informed of the system's ongoing development and impact. These reports can be dual in nature: internal reports serve as follow-ups on issues specific to the team, while external reports seek to inform a particular stakeholder group or a broader group. In either case, a timeline must be set and met, and most importantly, these reports should be informed by the requirements set by the documentation of the specific artifact.
- *System limitations and assumptions* Information on the XAI system's limitations and assumptions is shared, enabling stakeholders to understand and account for potential uncertainties in the explanations, maintaining transparency and verifiability (Gan et al., 2022; Papernot et al., 2016; Slack et al., 2021b).

Use case example

To illustrate the practical implications of our risk assessment framework, we present an hypothetical use case involving the application of an XAI system for fraud detection in benefit applications. In recent years, several countries have automated welfare distribution and fraud detection processes by employing risk scoring algorithms, such as Denmark (Jørgensen, 2023), the United States (Eubanks, 2018), and even World Bank programs (Human Rights Watch, 2023). In these scenarios, especially because of its public relevance, agencies and governments are increasingly being asked to provide explanations with respect to automated decisions and their impacts on people. This is particularly true in the Netherlands, where in the wake of several scandals related to the use of algorithms to detect fraud against the state in applying for benefits the country is now increasing transparency measures and process monitoring (Bekker, 2020; Hadwick & Lan, 2021; Wieringa, 2023).

Recently, an investigation revealed that the city of Rotterdam had been using risk scoring techniques to determine the risk of fraud in benefit applicants (Nast, 2023). The model employed indicators such as gender, age, and knowledge of Dutch language, effectively penalizing and flagging women, younger individuals, and people with migratory backgrounds as high-risk. Despite not having explicit XAI systems in place, this case exemplifies the potential ethical issues that could arise if explanations were provided without proper risk assessment and mitigation measures.

The supervised machine learning system used by Rotterdam from 2017 to 2021, a Gradient Boosting Machine, relied on 315 variables, including mental health history, personal relationships, and languages spoken, to assess the risk of fraud. Experts described this approach as amplifying historical discrimination, creating a dehumanizing environment for beneficiaries that extended beyond biases in the training data, permeating the choice of variables, model design, and policy process (Nast, 2023).

In the context of Rotterdam's risk scoring system, potential risks included (Table 1):

CT-SR-1 (Privacy Vulnerabilities) and **CT-SR-2 (Instrumentalization)** were lower likelihood risks, but privacy vulnerabilities could still have a medium impact, necessitating robust mitigation measures by AI Engineers.

T-FR-2 (Biased Sampling) and **T-FR-1 (Fairwashing)** were crucial fairness risks. Biased sampling, a high likelihood risk, could have a medium impact on model fairness, while fairwashing, a medium likelihood risk, could potentially mislead users about the model's fairness, having a high impact. These risks would fall under the responsibility of the AI Ethics Committee.

CT-RER-2 (Essentialism), **CT-ALR-1 (Circular Reasoning)**, and **CT-DETR-2 (Overdetermination)** were high likelihood risks associated with explanation quality, with varied impacts. Essentialism and overdetermination could significantly mislead interpretation due to biased fairness measures, having a high impact. Circular reasoning, although likely, generally posed a low impact. The AI Governance Board would be responsible for mitigating these risks, ensuring high-quality and comprehensible explanations.

The application of our risk assessment framework in this case would have prioritized these risks based on their likelihood and impact, allocating resources to address the most significant ones first. Each layer of the framework plays a crucial role in mitigating different aspects of the identified risks:

- The *intervention layer* would have focused on immediate risk mitigation strategies. For instance, to address **T-FR-2 (Biased Sampling)**, it would have implemented data preprocessing techniques such as re-sampling or re-weighting. To mitigate evaluation risks, it would have established robust explanation validation procedures. The layer would also have set up a governance committee to oversee the ethical deployment of the system, directly addressing accountability risks **CT-ACCR-(1-3)**.
- The *management layer* would have ensured the long-term effectiveness of these interventions through regular monitoring, adaptive risk reassessment, and feedback-driven improvements. This ongoing process would have been crucial in identifying and addressing emerging risks or changes in the operational environment (Raji et al., 2020; Tartaro et al., 2024). For example, it could have detected shifts in the prevalence of **CT-DETR-2 (Overdetermination)** or **CT-DETR-1 (Underdetermination)** risks over time, allowing for timely adjustments to the explanation generation process.
- The *information layer* would have complemented these efforts by focusing on comprehensive documentation, transparent communication with stakeholders, and inclusive stakeholder engagement. This layer would have been particularly effective in mitigating **CT-RER-2 (Essentialism)** risks by ensuring that the system's limitations and potential biases were clearly communicated. It would also have facilitated early detection of flaws and promoted trust in the system (Langer et al., 2021), further reinforcing the accountability measures of the Intervention Layer.

This case demonstrates how our framework's layers work together to address specific risks. The Intervention Layer would directly tackle biased sampling through data preprocessing. The Management Layer would continually monitor for emerging risks like overdetermination in explanations. The Information Layer would ensure transparent communication about the system's limitations, mitigating essentialism risks. Rotterdam's experience highlights the critical need for comprehensive risk assessment in public sector XAI. Technical explanations alone are insufficient; they must be coupled with social context. An XAI system reiterating biased criteria could perpetuate discrimination, potentially discouraging the use of automated risk assessment altogether. This underscores the importance of our multi-layered approach in developing trustworthy, fair XAI systems for public use.

Conclusion

This study introduces a novel risk assessment framework for XAI systems, offering a multi-layered approach to identify, prioritize, and mitigate technical and contextual risks. The framework enables tailored mitigation strategies, continuous monitoring, feedback-driven improvement, and transparent documentation. By proactively managing risks through a holistic, iterative process, the framework promotes the ethical, accountable, and trustworthy deployment of XAI systems. We conclude by discussing our current research limitations and directions.

Table 1 Updated risk matrix with main risks highlighted

Likelihood \ Impact	Low impact	Medium impact	High impact	Risk owner
Low likelihood	**CT-SR-2**: Instrumentalization	**CT-SR-1**: Privacy Vulnerabilities		AI Engineers
Medium likelihood			**T-FR-1**: Fairwashing	AI Engineers
High likelihood	**CT-ALR-1**: Circular Reasoning	**T-FR-2**: Biased Sampling	**CT-RER-2**: Essentialism **CT-DETR-2**: Overdetermination	AI Ethics Committee Board

Limitations

Despite the comprehensive nature of our risk assessment framework, we acknowledge several limitations:

- The rapidly evolving landscape of XAI makes it challenging to provide an exhaustive catalog of all potential risks. While we do not provide an exhaustive risk list for XAI, our study's goal is rather to foster an ongoing dialogue on the identification, understanding, and mitigation of these risks across diverse contexts. We encourage other researchers to adapt our methodology and risk categorization to their unique circumstances and refine them as required.
- Our methodology, while structured, is more qualitative compared to systematic literature reviews. This approach was necessary to capture the inherent complexity of sociotechnical risks associated with XAI explanations, which may not be easily reduced to a set of predefined keywords or a narrower focus on technical issues that might likely arise from contextual risks.
- The dynamic nature of the XAI field suggests that multiple XAI applications may interact in unforeseen ways, giving rise to new risks that resist fixed categorization. Examining risks from multiple perspectives is crucial, as they often exist in a complex web of interconnections where the implications of one issue can cascade into another (Cobbe et al., 2023; Floridi, 2016a; Sambasivan et al., 2021).

Research directions

Building upon the limitations identified, we outline several research directions to further develop and validate our XAI risk assessment framework. Firstly, we plan to transition our framework from a theoretical model to an empirically validated tool, as reflected in the growing attention towards operationalizing AI ethics impact assessments (Brown et al., 2021; Hasan, et al., 2022; Mökander & Floridi, 2022; Moss et al., 2021). This will involve applying the framework to real-world XAI systems and assessing its effectiveness in identifying and mitigating risks, also regularly updating the framework. To comprehensively address the multifaceted complexities of XAI risks, we will actively seek to incorporate diverse perspectives from a range of stakeholders, including developers, end-users, policymakers, and domain experts (Ehsan et al., 2022; Langer et al., 2021). Finally, to support the practical application of our framework, we aim to develop standardized tools and metrics for XAI risk assessment, similar to the efforts made by (Arnold et al., 2019; Gebru et al., 2021; Mitchell et al., 2018; Sokol & Flach, 2020). This will include creating risk assessment templates, checklists, and guidelines that can be easily adapted to different XAI use cases and domains, enhancing the framework's robustness, applicability, and practical impact. As a final consideration, our research revealed a scarcity of structured attempts to proactively address both technical and sociotechnical risks in XAI. This observation aligns with the current state of AI ethics research which—as denoted by (Hickok, 2021)—is transitioning from principle affirmation to operationalization (Hagendorff, 2019; Morley et al., 2023). As initiated by Kaur et al. (2020) and Schemmer et al. (2022), we urge the XAI community to focus on developing and evaluating solutions that align with stakeholders' needs, industry requirements, and regulatory norms (Ehsan et al., 2022; Nannini et al., 2023), rather than solely advancing technical constructs.

Appendix A: tables

A.1: categories of risk

See Table 2.

Table 2 Categorization of risks

Category	Subcategory	References
Technical		
Robustness	**T-RR-1**: Attacks on saliency-based explanation methods	Adebayo et al. (2018), Alvarez-Melis and Jaakkola (2018), Lundberg and Lee (2017), Ribeiro et al. (2016), Slack et al. (2020), Tang et al. (2022), Woods et al. (2019), Zhang et al. (2018, 2020)
	T-RR-2: Manipulation of counterfactual explanations	Keane and Smyth (2020), Keane et al. (2021), Kenny and Keane (2021), Kuhl et al. (2022), Mishra et al. (2021), Slack et al. (2021a), Stepin et al. (2021), Virgolin and Fracaros (2023), Wachter et al. (2017)
	T-RR-3: Attacks on concept-based explanation methods	Brown and Kvinge (2023), Ghorbani et al. (2019), Kim et al. (2018), Sinha et al. (2022)
	T-RR-4: Adversarial perturbations affecting explanations	Baniecki et al. (2022), Mehrabi et al. (2021), Nanda et al. (2021), Solans et al. (2020), Tang et al. (2022), Zhang et al. (2021)
	T-RR-5: Explanation-aware backdoors	Noppel et al. (2023)
	T-RR-6: Debugging challenges	Adebayo et al. (2020, 2022), Dai et al. (2022)
	T-RR-7: Transferability of adversarial attacks	Lakkaraju et al. (2020), Sinha et al. (2021)
Fairness	**T-FR-1**: Fairwashing	Aïvodji et al. (2019, 2021)
	T-FR-2: Biased sampling	Fukuchi et al. (2020), Laberge et al. (2022)
	T-FR-3: Adversarial poisoning	Mehrabi et al. (2021), Solans et al. (2020)
	T-FR-4: Manipulation of post-hoc explanations	Dimanov et al. (2020), Laberge et al. (2022), Merrer and Trédan (2020)
	T-FR-5: Explanation disparity risks	Balagopalan et al. (2022), Dai et al. (2022)
Evaluation	**T-ER-1**: Dependence on model assumptions	Arora et al. (2022), Noack et al. (2021)
	T-ER-2: Evaluation manipulation and deception	Adebayo et al. (2022), Warnecke et al. (2020)
	T-ER-3: Robustness-explainability trade-off	Agarwal et al. (2022), Noack et al. (2021), Rudin (2019)
	T-ER-4: Reliability of interpretation methods	Hooker et al. (2019), Huber et al. (2022), Kim et al. (2022), Tomsett et al. (2020)
Contextual		
Security	**CT-SR-1**: Privacy vulnerabilities	Duddu and Boutet (2022), Liu et al. (2022), Quan et al. (2022), Shokri et al. (2021)
	CT-SR-2: Instrumentalization	Agre (2014), Dwork (2006), Kuppa and Le-Khac (2020), Metcalf and Crawford (2016), Oh et al. (2019), Patel et al. (2022), Ronnow-Rasmussen (2015), Weitzner et al. (2008)
Accountability	**CT-ACCR-1**: Traceability of explanation design	Cobbe et al. (2023), Weidinger et al. (2022)
	CT-ACCR-2: Appraising explainers	Kruglanski et al. (2005), Zagzebski (2012)
	CT-ACCR-3: Explainer's overconfidence	Floridi (2013, 2016a), Kruger and Dunning (2000), Kruglanski (1989), Yates et al. (1997)
Heuristics & reception	**CT-HRR-1**: Cognitive heuristics	Kahneman and Tversky (1972), Tversky and Kahneman (1973)
	CT-HRR-2: Implications of language and semantic framing	Kahneman and Tversky (1984), Levinson (2000)
	CT-HRR-3: Cognitive biases	Nickerson (1998), Rozenblit and Keil (2002), Tubbs et al. (1990), Tversky and Kahneman (1973)
Argumentative & logical	**CT-ALR-1**: Circular reasoning	Hahn (2011), Walton (1994)
	CT-ALR-2: Tautology	Meibauer (2008), Popper (2014), Stanford (2006)
Under & over determination	**CT-DETR-1**: Underdetermination	Derrida (2016), Kuhn (1981), Leventi-Peetz and Weber (2022), Stanford (2006)
	CT-DETR-2: Overdetermination	Lombrozo (2011), Waldmann (2000)

Table 2 (continued)

Category	Subcategory	References
Reification & essentialism	**CT-RER-1**: Reification	Heft (2003), Hyman (2010), Lakoff (2008), Lakoff et al. (1999), Schank (2004), Searle (1979), Vandenberghe (2015), Watson (2019)
	CT-RER-2: Essentialism	Devine (1989), Inbar and Lammers (2012), McGarty et al. (2002), Rhodes and Moty (2020), Rossnan (2006)

A.2: framework layers

See Table 3.

Table 3 Details of the XAI Risk Management Framework

	Components	Subcomponents	Description
Intervention	Risk Matrix Development	Risk Categories	Segment into categories and subcategories following technical and contextual risks
		Risk Ownership	Define and allocate responsibilities for each risk to individuals or teams to promote accountability
		Risk Scores	Employ a standardized scoring system to assess the likelihood and impact of each risk using methods
	Implement Mitigation	Technical Actions	Implement technical solutions (e.g., data pre-processing, adversarial training)
		Organizational Actions	Form a governance committee, define clear roles and responsibilities and promoting risk communication
		Procedural Actions	Implement a dynamic risk assessment process
Management	Continuous Monitoring & Adaptation	System Audits	Regularly assess the performance, fairness, and security of the XAI system using methods for fairness auditing and robustness to adversarial perturbations
		Adaptive Risk Reassessment	Employ automated risk assessment tools that adapt to changes in the system or its operating environment using dynamic risk assessment methods
		Mitigation Strategy Adjustment	Adjust the mitigation strategies by adopting new encryption standards or incorporating additional adversarial training methods as per audit findings
	Feedback-Driven Improvement	Feedback Collection	Conduct user surveys or adopt any other strategy to collect feedback following user-centered design principles
		System Refinement	Use the collected feedback to refine the system as per usability engineering models
Information	Integration of Core Values	Accessibility	Ensure that information is readily available and comprehensible
		Reproducibility	Ensure that techniques are verifiable and can be replicated through comprehensive documentation
	Documentation & Reporting	Comprehensive Documentation	Develop comprehensive documentation on the XAI system, following software documentation best practices
		Performance Reports	Regularly report over system's performance, identified risks, and mitigation measures
		System Limitations	Share information on the system's limitations and assumptions

Acknowledgements Funding contribution from the ITN project NL4XAI Natural Language for Explainable AI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No 860621. This document reflects the views of the author(s) and does not necessarily reflect the views or policy of the European Commission. The REA cannot be held responsible for any use that may be made of the information this document contains.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of the submitted draft.

Ethical approval We have also obtained the necessary ethical consent from the responsible authorities (i.e., the University of Santiago de Compostela) where the research has been conducted.

Research involving human participants and/or animals No experiment was conducted with human participants or animals for this study.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I. J., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (pp. 9525–9536). https://proceedings.neurips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

Adebayo, J., Muelly, M., Abelson, H., & Kim, B. (2022). Post hoc explanations may be ineffective for detecting unknown spurious correlation. In *The tenth international conference on learning representations, ICLR 2022, virtual event, April 25-29, 2022*. OpenReview.net. https://openreview.net/forum?id=xNOVfCCvDpM

Adebayo, J., Muelly, M., Liccardi, I., & Kim, B. (2020). Debugging tests for model explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M.-F. Balcan, & H.-T. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/075b051ec3d22dac7b33f788da631fd4-Abstract.html

Agarwal, C., Zitnik, M., & Lakkaraju, H. (2022). Probing GNN explainers: A rigorous theoretical and empirical analysis of GNN explanation methods. In G. Camps-Valls, F. J. R. Ruiz, & I. Valera (Eds.), *International conference on artificial intelligence and statistics, AISTATS 2022, 28-30 March 2022, virtual event, proceedings of machine learning research* (Vol. 151, pp. 8969–8996). PMLR. https://proceedings.mlr.press/v151/agarwal22b.html

Agre, P. E. (2014). Toward a critical technical practice: Lessons learned in trying to reform AI. In *Social science, technical systems, and cooperative work* (pp. 131–157). Psychology Press. https://doi.org/10.4324/9781315805849

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, proceedings of machine learning research* (Vol. 97, pp. 161–170). PMLR. http://proceedings.mlr.press/v97/aivodji19a.html

Aïvodji, U., Arai, H., Gambs, S., & Hara, S. (2021). Characterizing the risk of fairwashing. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (pp. 14822-14834). https://proceedings.neurips.cc/paper/2021/hash/7caf5e22ea3eb8175ab518429c8589a4-Abstract.html

Alvarez-Melis, D., & Jaakkola, T. S. (2018). Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (pp. 7786–7795). https://proceedings.neurips.cc/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

Andrews, R., Diederich, J., & Tickle, A. B. (1995). Survey and critique of techniques for extracting rules from trained artificial neural networks. *Knowledge-Based Systems, 8*(6), 373–389. https://doi.org/10.1016/0950-7051(96)81920-4

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. In *Ethics of data and analytics* (pp. 254–264). Auerbach Publications.

Arnold, M., Bellamy, R. K. E., Hind, M., Houde, S., Mehta, S., Mojsilovic, A., Nair, R., Ramamurthy, K. N., Olteanu, A., Piorkowski, D., Reimer, D., Richards, J. T., Tsay, J., & Varshney, K. R. (2019). Factsheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development, 63*(4/5), 6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

Arora, S., Pruthi, D., Sadeh, N. M., Cohen, W. W., Lipton, Z. C., & Neubig, G. (2022). Explain, edit, and understand: Rethinking user study design for evaluating model explanations. In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22-March 1, 2022* (pp. 5277–5285). AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/20464

Balagopalan, A., Zhang, H., Hamidieh, K., Hartvigsen, T., Rudzicz, E., & Ghassemi, M. (2022). The road to explainability is paved with bias: Measuring the fairness of explanations. In *FAccT '22: 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea, June 21-24, 2022* (pp. 1194–1206). ACM. https://doi.org/10.1145/3531146.3533179

Baniecki, H., & Biecek, P. (2022). Manipulating SHAP via adversarial data perturbations (student abstract). In *Thirty-sixth AAAI conference on artificial intelligence, AAAI 2022, thirty-fourth conference on innovative applications of artificial intelligence, IAAI 2022, the twelveth symposium on educational advances in artificial intelligence, EAAI 2022 virtual event, February 22-March 1, 2022* (pp. 12907–12908). AAAI Press. https://doi.org/10.1609/AAAI.V36I11.21590.

Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion, 107*, 102303. https://doi.org/10.1016/j.inffus.2024.102303

Baniecki, H., Kretowicz, W., & Biecek, P. (2022). Fooling partial dependence via data poisoning. In M. Amini, S. Canu, A. Fischer, T. Guns, P. K. Novak, & G. Tsoumakas (Eds.), *Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2022, Grenoble, France, September 19-23, 2022, proceedings, part III, lecture notes in computer* *science* (Vol. 13715, pp. 121–136). Springer. https://doi.org/10.1007/978-3-031-26409-2_8

Bekker, S. (2020). *Fundamental rights in digital welfare states: The case of SyRI in the Netherlands, T.M.C. Netherlands Yearbook of International Law* (pp. 289–307). Asser Press.https://doi.org/10.1007/978-94-6265-403-7_24

Bertrand, A., Belloum, R., Eagan, J. R., & Maxwell, W. (2022). How cognitive biases affect XAI-assisted decision-making: A systematic review. In V. Conitzer, J. Tasioulas, M. Scheutz, R. Calo, M. Mara, & A. Zimmermann (Eds.), *AIES '22: AAAI/ACM conference on AI, ethics, and society, Oxford, United Kingdom, May 19-21, 2021* (pp. 78–91). ACM.https://doi.org/10.1145/3514094.3534164

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101. https://doi.org/10.1191/1478088706qp063oa

Brown, D., & Kvinge, H. (2023). Making corgis important for honeycomb classification: Adversarial attacks on concept-based explainability tools. In *IEEE/CVF conference on computer vision and pattern recognition, CVPR 2023—Workshops, Vancouver, BC, Canada, June 17-24, 2023* (pp. 620–627). IEEE. https://doi.org/10.1109/CVPRW59228.2023.00069,

Brown, S., Davidovic, J., & Hasan, A. (2021). The algorithm audit: Scoring the algorithms that score us. *Big Data & Society, 8*(1), 2053951720983865. https://doi.org/10.1177/2053951720983865

Cabitza, F., Campagner, A., Malgieri, G., Natali, C., Schneeberger, D., Stoeger, K., & Holzinger, A. (2023). Quod erat demonstrandum?–Towards a typology of the concept of explanation for the design of explainable AI. *Expert Systems with Applications, 213*, 118888. https://doi.org/10.1016/j.eswa.2022.118888

Carlini, N., & Wagner, D. A. (2017a). Adversarial examples are not easily detected: Bypassing ten detection methods. In B. Thuraisingham, B. Biggio, D. M. Freeman, B. Miller, & A. Sinha (Eds.), *Proceedings of the 10th ACM workshop on artificial intelligence and security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017* (pp. 3–14). ACM. https://doi.org/10.1145/3128572.3140444

Carlini, N., & Wagner, D. A. (2017b). Towards evaluating the robustness of neural networks. In *2017 IEEE symposium on security and privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017* (pp. 39–57). IEEE Computer Society. https://doi.org/10.1109/SP.2017.49

Carmichael, Z., & Scheirer, W. J. (2023). Unfooling perturbation-based post hoc explainers. In B. Williams, Y Chen, & J. Neville (Eds.), *Thirty-seventh AAAI conference on artificial intelligence, AAAI 2023, thirty-fifth conference on innovative applications of artificial intelligence, IAAI 2023, thirteenth symposium on educational advances in artificial intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023* (pp. 6925–6934). AAAI Press. https://doi.org/10.1609/AAAI.V37I6.25847

Chen, J., Wu, X., Rastogi, V., Liang, Y., & Jha, S. (2019). Robust attribution regularization. In: [214] (pp. 14300–14310). https://proceedings.neurips.cc/paper/2019/hash/172ef5a94b4dd0aa120c6878fc29f70c-Abstract.html

Chen, V., Liao, Q. V., Vaughan, J. W., & Bansal, G. (2023). Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW2), 1–32. https://doi.org/10.1145/3610219

Chen, X., Liu, C., Li, B., Lu, K., & Song, D. (2017). Targeted backdoor attacks on deep learning systems using data poisoning. CoRR abs/1712.05526. http://arxiv.org/abs/1712.05526

Cheng, H. F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., & Zhu, H. (2019). Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders. In S. A. Brewster, G. Fitzpatrick, A. L. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI conference on human factors in computing systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019* (p. 559). ACM. https://doi.org/10.1145/3290605.3300789

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, B. L, M. John, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 13–1991). American Psychological Association. https://doi.org/10.1037/10096-006

Cobbe, J., Veale, M., & Singh, J. (2023). Understanding accountability in algorithmic supply chains. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023* (pp. 1186–1197). ACM. https://doi.org/10.1145/3593013.3594073

Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, proceedings of machine learning research* (Vol. 97, pp. 1310–1320). PMLR. http://proceedings.mlr.press/v97/cohen19c.html

Confalonieri, R., Coba, L., Wagner, B., & Besold, T. R. (2021). A historical perspective of explainable artificial intelligence. *WIREs Data Mining and Knowledge Discovery*. https://doi.org/10.1002/WIDM.1391

Craven, M. W., & Shavlik, J. W. (1995). Extracting tree-structured representations of trained networks. In D. S. Touretzky, M. Mozer, & M. E. Hasselmo (Eds.), *Advances in neural information processing systems 8, NIPS, Denver, CO, USA, November 27-30, 1995* (pp. 24–30). MIT Press. http://papers.nips.cc/paper/1152-extracting-tree-structured-representations-of-trained-networks

Dai, J., Upadhyay, S., Aïvodji, U., Bach, S. H., & Lakkaraju, H. (2022). Fairness via explanation quality: Evaluating disparities in the quality of post hoc explanations. In V. Conitzer, J. Tasioulas, M. Scheutz, R. Calo, M. Mara, & A. Zimmermann (Eds.), *AIES '22: AAAI/ACM conference on AI, ethics, and society, Oxford, United Kingdom, May 19-21, 2021* (pp. 203–214). ACM. https://doi.org/10.1145/3514094.3534159

Dazeley, R., Vamplew, P., Foale, C., Young, C., Aryal, S., & Cruz, F. (2021). Levels of explainable artificial intelligence for human-aligned conversational explanations. *Artificial Intelligence, 299*, 103525. https://doi.org/10.1016/J.ARTINT.2021.103525

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly, 39*(2), 101666. https://doi.org/10.1016/J.GIQ.2021.101666

Derrida, J. (2016). *Dissemination*. Bloomsbury Publishing.

Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5. https://doi.org/10.1037/0022-3514.56.1.5

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/V1/N19-1423

Dimanov, B., Bhatt, U., Jamnik, M., & Weller, A. (2020). You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In H. Espinoza, J. Hernández-Orallo, X. C. Chen, S. S. ÓhÉigeartaigh, X. Huang, M. Castillo-Effen, R. Mallah, & J. A. McDermid (Eds.), *Proceedings of the workshop* *on artificial intelligence safety, co-located with 34th AAAI conference on artificial intelligence, SafeAI@AAAI 2020, New York City, NY, USA, February 7, 2020, CEUR workshop proceedings* (Vol. 2560, pp. 63–73). CEUR-WS.org. https://ceur-ws.org/Vol-2560/paper8.pdf

Dombrowski, A., Alber, M., Anders, C. J., Ackermann, M., Müller, K., & Kessel, P. (2019). Explanations can be manipulated and geometry is to blame. In: [214] (pp. 13567-13578). https://proceedings.neurips.cc/paper/2019/hash/bb836c01cdc9120a9c984c525e4b1a4a-Abstract.html

Dombrowski, A., Anders, C. J., Müller, K., & Kessel, P. (2022). Towards robust explanations for deep neural networks. *Pattern Recognition, 121*, 108194. https://doi.org/10.1016/J.PATCOG.2021.108194

Dong, Y., Yang, X., Deng, Z., Pang, T., Xiao, Z., Su, H., & Zhu, J. (2021). Black-box detection of backdoor attacks with limited information and data. In *2021 IEEE/CVF international conference on computer vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021* (pp. 16462–16471). IEEE. https://doi.org/10.1109/ICCV48922.2021.01617

Duddu, V., & Boutet, A. (2022). Inferring sensitive attributes from model explanations. In M. A. Hasan, & L. Xiong (Eds.), *Proceedings of the 31st ACM international conference on information & knowledge management, Atlanta, GA, USA, October 17-21, 2022* (pp. 416–425). ACM. https://doi.org/10.1145/3511808.3557362

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, I., & Wegener (Eds.), *Automata, languages and programming* (pp. 1–12). Springer Berlin Heidelberg. https://doi.org/10.1007/11787006_1

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022). Human-centered explainable AI (HCXAI): Beyond opening the black-box of AI. In *Extended abstracts of the 2022 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI EA '22*.https://doi.org/10.1145/3491101.3503727,

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St: Martin's Press.

European Commission. (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) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2016/679/oj

Ferry, J., Aïvodji, U., Gambs, S., Huguet, M., & Siala, M. (2022). Exploiting fairness to enhance sensitive attributes reconstruction. CoRR abs/2209.01215. https://doi.org/10.48550/ARXIV.2209.01215

Fisher, A., Rudin, C., & Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research, 20*, 177:1–177:81.

Floridi, L. (2013). Distributed morality in an information society. *Science and Engineering Ethics, 19*, 727–743. https://doi.org/10.1007/s11948-012-9413-4

Floridi, L. (2016a). Faultless responsibility: On the nature and allocation of moral responsibility for distributed moral actions. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374*(2083), 20160112. https://doi.org/10.1098/rsta.2016.0112

Floridi, L. (2016b). Tolerant paternalism: Pro-ethical design as a resolution of the dilemma of toleration. *Science and Engineering Ethics, 22*(6), 1669–1688.

Floridi, L. (2018). Soft ethics and the governance of the digital. *Philosophy & Technology, 31*, 1–8. https://doi.org/10.1007/s13347-018-0303-9

Freitas, A. A. (2013). Comprehensible classification models: A position paper. *SIGKDD Explorations, 15*(1), 1–10. https://doi.org/10.1145/2594473.2594475

Friedman, B., & Kahn, P. H. (2002). Human values, ethics, and design. In *The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications* (pp. 1177–1201). L. Erlbaum Associates Inc.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fukuchi, K., Hara, S., & Maehara, T. (2020). Faking fairness via stealthily biased sampling. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 412–419). AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/5377

Galilei, G. (1953). *Dialogue concerning the two chief world systems*. Ptolemaic and Copernican: University of California Press.

Gan, Y., Mao, Y., Zhang, X., Ji, S., Pu, Y., Han, M., Yin, J., & Wang, T. (2022). "is your explanation stable?": A robustness evaluation framework for feature attribution. In H. Yin, A. Stavrou, C. Cremers, & E. Shi (Eds.), *Proceedings of the 2022 ACM SIGSAC conference on computer and communications security, CCS 2022, Los Angeles, CA, USA, November 7-11, 2022* (pp. 1157–1171). ACM. https://doi.org/10.1145/3548606.3559392

Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., & Nepal, S. (2019). STRIP: A defence against Trojan attacks on deep neural networks. In D. Balenson (Eds.), *Proceedings of the 35th annual computer security applications conference, ACSAC 2019, San Juan, PR, USA, December 09-13, 2019* (pp. 113–125). ACM. https://doi.org/10.1145/3359789.3359790

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H. M., Daumé, H., III., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Ghalebikesabi, S., Ter-Minassian, L., DiazOrdaz, K., & Holmes, C.C (2021) On locality of local explanation models. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (pp. 18395-18407). https://proceedings.neurips.cc/paper/2021/hash/995665640dc319973d3173a74a03860c-Abstract.html

Ghorbani, A., Abid, A, & Zou, J. Y. (2019). Interpretation of neural networks is fragile. In *The thirty-third AAAI conference on artificial intelligence, AAAI 2019, the thirty-first innovative applications of artificial intelligence conference, IAAI 2019, the ninth AAAI symposium on educational advances in artificial intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27-February 1, 2019* (pp. 3681–3688). AAAI Press. https://doi.org/10.1609/aaai.v33i01.33013681

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics, 24*(1), 44–65. https://doi.org/10.1080/10618600.2014.907095

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Y. Bengio, & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings*. http://arxiv.org/abs/1412.6572

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines, 8*(1), 101–118. https://doi.org/10.1023/A:1008290415597

Gopnik, A., Sobel, D. M., Schulz, L. E., & Glymour, C. (2001). Causal learning mechanisms in very young children: Two-, three-, and four-year-olds infer causal relations from patterns of variation and covariation. *Developmental Psychology, 37*(5), 620.

Gu, T., Liu, K., Dolan-Gavitt, B., & Garg, S. (2019). BadNets: Evaluating backdooring attacks on deep neural networks. *IEEE Access, 7*, 47230–47244. https://doi.org/10.1109/ACCESS.2019.2909068

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 93:1–93:42. https://doi.org/10.1145/3236009

Gunning, D., & Aha, D. W. (2019). Darpa's explainable artificial intelligence (XAI) program. *AI Magazine, 40*(2), 44–58. https://doi.org/10.1609/AIMAG.V40I2.2850

Hacker, P., & Passoth, J. H. (2022). Varieties of AI explanations under the law. From the GDPR to the AIA, and beyond. In A. Holzinger, R. Goebel, R. Fong, T. Moon, K. Müller, & W. Samek (Eds.), *xxAI—Beyond explainable AI: International workshop, held in conjunction with ICML 2020, July 18, 2020, Vienna, Austria, revised and extended papers. Lecture notes in computer science* (pp. 343–373). Springer International Publishing. https://doi.org/10.1007/978-3-031-04083-2_17

Hadwick, D., & Lan, S. (2021). Lessons to be learned from the Dutch childcare allowance scandal: A comparative review of algorithmic governance by tax administrations in the Netherlands. *France and Germany. World Tax Journal-Amsterdam, 13*(4), 609–645.

Hagendorff, T. (2019). The ethics of AI ethics—An evaluation of guidelines. CoRR abs/1903.03425. http://arxiv.org/abs/1903.03425

Hahn, U. (2011). The problem of circularity in evidence, argument, and explanation. *Perspectives on Psychological Science, 6*(2), 172–182. https://doi.org/10.1177/1745691611400240

Halpern, D. F. (2000). Sex differences in cognitive abilities. *Psychology Press*. https://doi.org/10.4324/9781410605290

Harding, S. (1991). *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review, 74*(1), 88–95.

Hasan, A., Brown, S., Davidovic, J., Lange, B., & Regan, M. (2022). Algorithmic bias and risk assessments: Lessons from practice. *Digital Society, 1*(2), 14. https://doi.org/10.1007/s44206-022-00017-z

Haslam, N., Rothschild, L., & Ernst, D. (2000). Essentialist beliefs about social categories. *British Journal of social psychology, 39*(1), 113–127. https://doi.org/10.1348/014466600164363

Heft, H. (2003). Affordances, dynamic experience, and the challenge of reification. *Ecological Psychology, 15*(2), 149–180. https://doi.org/10.1207/S15326969ECO1502_4

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science, 15*(2), 135–175. https://doi.org/10.1086/286983

Heo, J., Joo, S., & Moon, T. (2019). Fooling neural network interpretations via adversarial model manipulation. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 2921–2932). https://proceedings.neurips.cc/paper/2019/hash/7fea637fd6d02b8f0adf6f7dc36aed93-Abstract.html

Hickok, M. (2021). Lessons learned from AI ethics principles for future actions. *AI Ethics, 1*(1), 41–47. https://doi.org/10.1007/s43681-020-00008-1

Hooker, S., Erhan, D., Kindermans, P., & Kim B. (2019). A benchmark for interpretability methods in deep neural networks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 9734–9745). https://proceedings.neurips.cc/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*(1), 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Huber, T., Limmer, B., & André, E. (2022). Benchmarking perturbation-based saliency maps for explaining atari agents. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2022.903875

Human Rights Watch. (2023). Automated neglect—hrw.org. Retrieved June 27, 2023, from https://www.hrw.org/report/2023/06/13/automated-neglect/how-world-banks-push-allocate-cash-assistance-using-algorithms

Hyman, S. E. (2010). The diagnosis of mental disorders: The problem of reification. *Annual Review of Clinical Psychology*, *6*(Volume 6, 2010):155–179. https://doi.org/10.1146/annurev.clinpsy.3.022806.091532

Ilyas, A., Engstrom, L., Athalye, A., & Lin, J. (2018). Black-box adversarial attacks with limited queries and information. In J. G. Dy, & A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, proceedings of machine learning research* (Vol. 80, pp. 2142-2151). PMLR. http://proceedings.mlr.press/v80/ilyas18a.html

Inbar, Y., & Lammers, J. (2012). Political diversity in social and personality psychology. *Perspectives on Psychological Science, 7*(5), 496–503. https://doi.org/10.1177/1745691612448792

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy, SP 2018, proceedings, 21-23 May 2018, San Francisco, California, USA* (pp. 19–35). IEEE Computer Society. https://doi.org/10.1109/SP.2018.00057

Janssen, M., Hartog, M., Matheus, R., Ding, A. I., & Kuk, G. (2022). Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government. *Social Science Computer Review, 40*(2), 478–493. https://doi.org/10.1177/0894439320980118

Jia, J., Liu, Y., & Gong, N. Z. (2022). Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *43rd IEEE symposium on security and privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022* (pp. 2043–2059). IEEE. https://doi.org/10.1109/SP46214.2022.9833644

Jørgensen, R. F. (2023). Data and rights in the digital welfare state: The case of Denmark. *Information, Communication & Society, 26*(1), 123–138. https://doi.org/10.1080/1369118X.2021.1934069

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*(3), 430–454. https://doi.org/10.1016/0010-0285(72)90016-3

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist, 39*(4), 341. https://doi.org/10.1037/0003-066X.39.4.341

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., & Jennifer, W. V. (2020). Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI conference on human factors in computing systems. Association for Computing Machinery,* *New York, NY, USA, CHI '20* (pp. 1–14). https://doi.org/10.1145/3313831.3376219

Keane, M. T., Kenny, E. M., Delaney, E., & Smyth, B. (2021). If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques. In Z. Zhou (Ed.), *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021* (pp. 4466–4474). ijcai.org.https://doi.org/10.24963/ijcai.2021/609

Keane, M. T., & Smyth, B. (2020). Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable AI (XAI). In I. Watson, R. O. Weber (Eds.), *Case-based reasoning research and development—28th International conference, ICCBR 2020, Salamanca, Spain, June 8-12, 2020, proceedings, lecture notes in computer science* (Vol. 12311, pp. 163–178). Springer. https://doi.org/10.1007/978-3-030-58342-2_11

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology, 57*, 227–254. https://doi.org/10.1146/annurev.psych.57.102904.190100

Keil, F. C., Wilson, R. A., & Wilson, R. A. (2000). *Explanation and cognition*. MIT Press.

Kenny, E. M., & Keane, M. T. (2021). On generating plausible counterfactual and semi-factual explanations for deep learning. In *Thirty-fifth AAAI conference on artificial intelligence, AAAI 2021, thirty-third conference on innovative applications of artificial intelligence, IAAI 2021, the eleventh symposium on educational advances in artificial intelligence, EAAI 2021, virtual event, February 2-9, 2021* (pp. 11575–11585). AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/17377

Keysar, B., & Bly, B. (1995). Intuitions of the transparency of idioms: Can one keep a secret by spilling the beans? *Journal of Memory and Language, 34*(1), 89–109. https://doi.org/10.1006/jmla.1995.1005

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., & Sayres, R. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In J. G. Dy, A. Krause (Eds.), *Proceedings of the 35th international conference on machine learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, proceedings of machine learning research* (Vol. 80, pp. 2673–2682). PMLR. http://proceedings.mlr.press/v80/kim18d.html

Kim, J. S., Plumb, G., & Talwalkar, A. (2022). Sanity simulations for saliency methods. In K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, & S. Sabato (Eds.), *International conference on machine learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA, proceedings of machine learning research* (Vol. 162, pp. 11173–11200). PMLR. https://proceedings.mlr.press/v162/kim22h.html

Kruger, J., & Dunning, D. (2000). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*, 1121–34. https://doi.org/10.1037/0022-3514.77.6.1121

Kruglanski, A. (1989). The psychology of being right: The problem of accuracy in social perception and cognition. *Psychological Bulletin, 106*, 395–409. https://doi.org/10.1037/0033-2909.106.3.395

Kruglanski, A., Raviv, A., Bar-Tal, D., Raviv, A., Sharvit, K., Ellis, S., Bar, R., Pierro, A., & Mannetti, L. (2005). Says who?: Epistemic authority effects in social judgment. *Advances in Experimental Social Psychology, 37*, 345–392. https://doi.org/10.1016/S0065-2601(05)37006-7

Kuhl, U., Artelt, A., & Hammer, B. (2022). Keep your friends close and your counterfactuals closer: Improved learning from closest rather than plausible counterfactual explanations in an abstract setting. In *FAccT '22: 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea, June 21-24, 2022* (pp. 2125–2137). ACM. https://doi.org/10.1145/3531146.3534630

Kuhn, T. S. (1981). textitObjectivity, value judgment, and theory choice (pp. 320–339). Duke University Press.

Kuppa, A., & Le-Khac, N. (2020). Black box attacks on explainable artificial intelligence (XAI) methods in cyber security. In *2020 international joint conference on neural networks, IJCNN 2020, Glasgow, United Kingdom, July 19-24, 2020* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9206780

Laberge, G., Aïvodji, U., & Hara, S. (2022). Fooling SHAP with stealthily biased sampling. CoRR abs/2205.15419. https://doi.org/10.48550/arXiv.2205.15419

Lakkaraju, H., Arsov, N., & Bastani, O. (2020). Robust and stable black box explanations. In *Proceedings of the 37th international conference on machine learning, ICML 2020, 13-18 July 2020, virtual event, proceedings of machine learning research* (Vol. 119, pp. 5628–5638). PMLR. http://proceedings.mlr.press/v119/lakkaraju20a.html

Lakoff, G. (2008). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Lakoff, G., Johnson, M., & Sowa, J. F. (1999). Review of philosophy in the flesh: The embodied mind and its challenge to western thought. *Computational Linguistics, 25*(4), 631–634.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from explainable artificial intelligence (XAI)?–A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence, 296*, 103473. https://doi.org/10.1016/J.ARTINT.2021.103473

Latour, B. (1988). The politics of explanation: An alternative. *Knowledge and Reflexivity: New Frontiers in the Sociology of Knowledge, 10*, 155–176.

Leventi-Peetz, A., & Weber, K. (2022). Rashomon effect and consistency in explainable artificial intelligence (XAI). In K. Arai (Ed.), *Proceedings of the future technologies conference, FTC 2022, virtual event, 20-21 October 2022, Volume 1, lecture notes in networks and systems* (Vol. 559, pp. 796–808). Springer. https://doi.org/10.1007/978-3-031-18461-1_52

Levinson, S. C. (2000). Presumptive meanings: The theory of generalized conversational implicature. *MIT Press*. https://doi.org/10.7551/mitpress/5526.001.0001

Liao, Q. V., & Varshney, K. R. (2021). Human-centered explainable AI (XAI): From algorithms to user experiences. CoRR abs/2110.10790. http://arxiv.org/abs/2110.10790

Lipton, P. (2017). Inference to the best explanation. In: W. H. Newton-Smith (Ed.), *A companion to the philosophy of science* (pp. 184–193). Blackwell. https://doi.org/10.1002/9781405164481.ch29

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue, 16*(3), 31–57. https://doi.org/10.1145/3236386.3241340

Liu, A., Chen, X., Liu, S., Xia, L., & Gan, C. (2022). Certifiably robust interpretation via Rényi differential privacy. *Artificial Intelligence, 313*, 103787. https://doi.org/10.1016/j.artint.2022.103787

Liu, Y., Ma, S., Aafer, Y., Lee, W., Zhai, J., Wang, W., & Zhang, X. (2018). Trojaning attack on neural networks. In *25th annual network and distributed system security symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society. http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-5_Liu_paper.pdf

Löfström, H., Hammar, K., & Johansson U. (2022). A meta survey of quality evaluation criteria in explanation methods. In: J. D. Weerdt, & A. Polyvyanyy (Eds.), *Intelligent information systems—CAiSE forum 2022, Leuven, Belgium, June 6-10, 2022,* *proceedings, lecture notes in business information processing* (Vol. 452, pp. 55–63). Springer. https://doi.org/10.1007/978-3-031-07481-3_7

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass, 6*(8), 539–551. https://doi.org/10.1111/j.1747-9991.2011.00413.x

Lombrozo, T. (2012). Explanation and abductive inference. *The Oxford Handbook of Thinking and Reasoning*. https://doi.org/10.1093/oxfordhb/9780199734689.013.0014

Lundberg, S. M., & Lee, S. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017, December 4-9, 2017, Long Beach, CA, USA* (pp. 4765–4774). https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. In *6th international conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30-May 3, 2018, conference track proceedings*. OpenReview.net. https://openreview.net/forum?id=rJzIBfZAb

McGarty, C. E., Yzerbyt, V. Y., & Spears, R. E. (2002). Stereotypes as explanations: The formation of meaningful beliefs about social groups. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511489877

McKie, D. (1960). The origins and foundation of the Royal Society of London. *Notes and Records of the Royal Society of London, 15*(1), 1–37. https://doi.org/10.1098/rsnr.1960.0001

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*(12), 1469. https://doi.org/10.1037/0003-066X.44.12.1469

Medin, D., & Ortony, A. (1989). *Comments on part I: Psychological essentialism* (pp. 179–196). Cambridge University Press. https://doi.org/10.1017/CBO9780511529863.009

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 115:1–115:35. https://doi.org/10.1145/3457607

Mehrabi, N., Naveed, M., Morstatter, F., & Galstyan, A. (2021). Exacerbating algorithmic bias through fairness attacks. *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(10), 8930–8938. https://doi.org/10.1609/aaai.v35i10.17080

Meibauer, J. (2008). Tautology as presumptive meaning. *Pragmatics & Cognition, 16*(3), 439–470.

Merrer, E. L., & Trédan, G. (2020). Remote explainability faces the bouncer problem. *Nature Machine Intelligence, 2*(9), 529–539. https://doi.org/10.1038/s42256-020-0216-z

Metcalf, J., & Crawford, K. (2016). Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society, 3*(1), 205395171665021. https://doi.org/10.1177/2053951716650211

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. CoRR abs/1712.00547. http://arxiv.org/abs/1712.00547

Mishra, S., Dutta, S., Long, J., & Magazzeni, D. (2021). A survey on the robustness of feature importance and counterfactual explanations. CoRR abs/2111.00358. http://arxiv.org/abs/2111.00358

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2018). Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*.

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A survey of evaluation methods and measures for interpretable machine learning. CoRR abs/1811.11839. http://arxiv.org/abs/1811.11839

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems*. https://doi.org/10.1145/3387166

Mökander, J., & Floridi, L. (2022). Operationalising AI governance through ethics-based auditing: An industry case study. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00191-3

Morley, J., Kinsey, L., Elhalal, A., Garcia, F., Ziosi, M., & Floridi, L. (2023). Operationalising AI ethics: Barriers, enablers and next steps. *AI & Society, 38*(1), 411–423. https://doi.org/10.1007/S00146-021-01308-8

Moss, E., Watkins, E. A., Singh, R., Elish, M. C., & Metcalf, J. (2021). Assembling accountability: Algorithmic impact assessment for the public interest. *SSRN*. https://doi.org/10.2139/ssrn.3877437

Nanda, V., Dooley, S., Singla, S., Feizi, S., & Dickerson, J. P. (2021). Fairness through robustness: Investigating robustness disparity in deep learning. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, FAccT '21* (pp. 466-477). https://doi.org/10.1145/3442188.3445910

Nannini, L. (2024). Habemus a right to an explanation: So what?—A framework on transparency-explainability functionality and tensions in the EU AI act. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (pp. 1023-1035). https://ojs.aaai.org/index.php/AIES/article/download/31700/33867/35764

Nannini, L., Balayn, A., & Smith, A. L. (2023). Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023* (pp. 1198–1212). ACM. https://doi.org/10.1145/3593013.3594074

Nast, C. (2023). Inside the suspicion machine. wired.com. Retrieved June 27, 2023, from, https://www.wired.com/story/welfare-state-algorithms/

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Noack, A., Ahern, I., Dou, D., & Li, B. (2021). An empirical study on the relation between network interpretability and adversarial robustness. *SN Computer Science, 2*(1), 32. https://doi.org/10.1007/s42979-020-00390-x

Noppel, M., Peter, L., & Wressnegger, C. (2023). Disguising attacks with explanation-aware backdoors. In *2023 2023 IEEE symposium on security and privacy (SP) (SP)* (pp. 664–681). Los Alamitos, CA, USA: IEEE Computer Society. https://doi.org/10.1109/SP46215.2023.00057

Oh, S. J., Schiele, B., & Fritz, M. (2019). Towards reverse-engineering black-box neural networks. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen & K. Müller (Eds.), *Explainable AI: Interpreting, explaining and visualizing deep learning, lecture notes in computer science* (Vol. 11700, pp. 121-144). Springer. https://doi.org/10.1007/978-3-030-28954-6_7

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., & Swami, A. (2017). Practical black-box attacks against machine learning. In R. Karri, O. Sinanoglu, A. Sadeghi, & X. Yi (Eds.), *Proceedings of the 2017 ACM on Asia conference on computer and communications security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017* (pp. 506–519). ACM. https://doi.org/10.1145/3052973.3053009

Papernot, N., McDaniel, P. D., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE symposium on security and privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016* (pp. 582–597). IEEE Computer Society.https://doi.org/10.1109/SP.2016.41

Patel, N., Shokri, R., & Zick, Y. (2022). Model explanations with differential privacy. In *FAccT '22: 2022 ACM conference on fairness, accountability, and transparency, Seoul, Republic of Korea, June 21-24, 2022* (pp. 1895–1904). ACM. https://doi.org/10.1145/3531146.3533235

Pawelczyk, M., Datta, T., van den Heuvel, J., Kasneci, G., & Lakkaraju, H. (2023). Probabilistically robust recourse: Navigating the trade-offs between costs and robustness in algorithmic recourse. In *The eleventh international conference on learning representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. https://openreview.net/pdf?id=sC-PmTsiTB

Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. Routledge.

Quan, P., Chakraborty, S., Jeyakumar, J. V., & Srivastava, M. B. (2022). On the amplification of security and privacy risks by post-hoc explanations in machine learning models. CoRR abs/2206.14004. https://doi.org/10.48550/arXiv.2206.14004

Rahman, A. (2020). Algorithms of oppression: How search engines reinforce racism. *New Media & Society*. https://doi.org/10.1177/1461444819876115

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, & G. Zanfir-Fortuna (Eds.), *FAT* '20: Conference on fairness, accountability, and transparency, Barcelona, Spain, January 27-30, 2020* (pp. 33–44). ACM. https://doi.org/10.1145/3351095.3372873

Raveendran, A., Renjith, V., & Madhu, G. (2022). A comprehensive review on dynamic risk analysis methodologies. *Journal of Loss Prevention in the Process Industries, 76*, 104734. https://doi.org/10.1016/j.jlp.2022.104734

Rhodes, M., & Moty, K. (2020). What is social essentialism and how does it develop? *Advances in Child Development and Behavior, 59*, 1–30. https://doi.org/10.1016/bs.acdb.2020.05.001

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should I trust you?": Explaining the predictions of any classifier. In B. Krishnapuram, M. Shah, A. J. Smola, C. C. Aggarwal, D. Shen, & R. Rastogi (Eds.), *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, USA, August 13-17, 2016* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778

Rieger, L., & Hansen, L. K. (2020). A simple defense against adversarial attacks on heatmap explanations. CoRR abs/2007.06381. http://arxiv.org/abs/2007.06381

Robbins, S. (2019). A misdirected principle with a catch: Explicability for AI. *Minds and Machines, 29*(4), 495–514. https://doi.org/10.1007/S11023-019-09509-3

Ronnow-Rasmussen, T. (2015). Intrinsic and extrinsic value. In *The Oxford handbook of value theory* (pp. 29–43). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199959303.013.0003

Rossnan, S. (2006). *Overcoming math anxiety. Mathitudes, 1*(1), 1–4.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*(5), 521–562. https://doi.org/10.1207/s15516709cog2605_1

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Salmon, W. C. (1989). Four decades of scientific explanation. *Minnesota Studies in the Philosophy of Science, 13*, 3–219.

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., & Aroyo, L. M. (2021). "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '21*. https://doi.org/10.1145/3411764.3445518

Schank, R. C. (2004). *Making minds less well educated than our own*. Routledge.

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-AI decision-making. In *Proceedings of the 2022 AAAI/ACM conference on AI, ethics, and society. Association for Computing Machinery, New York, NY, USA, AIES '22* (pp. 617–626). https://doi.org/10.1145/3514094.3534128

Searle, J. R. (1979). Expression and meaning: Studies in the theory of speech acts. *Cambridge University Press*. https://doi.org/10.1017/CBO9780511609213

Selbst. A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology*, *35*(1). https://ssrn.com/abstract=3867634

Severi, G., Meyer, J., Coull, S. E., & Oprea, A. (2021). Explanation-guided backdoor poisoning attacks against malware classifiers. In M. Bailey, & R. Greenstadt (Eds.), *30th USENIX security symposium, USENIX security 2021, August 11-13, 2021* (pp. 1487–1504). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/severi

Shafahi, A., Huang, W. R., Najibi, M., Suciu, O., Studer, C., Dumitras, T., & Goldstein, T. (2018) Poison frogs! Targeted clean-label poisoning attacks on neural networks. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems 31: Annual conference on neural information processing systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada* (pp. 6106–6116). https://proceedings.neurips.cc/paper/2018/hash/22722a343513ed45f14905eb07621686-Abstract.html

Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J. P., Studer, C., Davis, L. S., Taylor, G., & Goldstein, T. (2019). Adversarial training for free! In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada* (pp. 3353–3364). https://proceedings.neurips.cc/paper/2019/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html

Shokri, R., Strobel, M., & Zick, Y. (2021). On the privacy risks of model explanations. In M. Fourcade, B. Kuipers, S. Lazar, & D. K. Mulligan (Eds.), *AIES '21: AAAI/ACM conference on AI, ethics, and society, virtual event, USA, May 19-21, 2021*. ACM, pp. 231–241. https://doi.org/10.1145/3461702.3462533.

Sinha, S., Chen, H., Sekhon, A., Ji, Y., & Qi, Y. (2021). Perturbing inputs for fragile interpretations in deep natural language processing. In J. Bastings, Y. Belinkov, E. Dupoux, M. Giulianelli, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the fourth BlackboxNLP workshop on analyzing and interpreting neural networks for NLP, BlackboxNLP@EMNLP 2021, Punta Cana, Dominican Republic, November 11, 2021* (pp. 420–434). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.blackboxnlp-1.33

Sinha, S., Huai, M., Sun, J., & Zhang A. (2022). Understanding and enhancing robustness of concept-based models. CoRR abs/2211.16080. https://doi.org/10.48550/arXiv.2211.16080

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020) Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM conference on AI,* *ethics, and society. Association for Computing Machinery, New York, NY, USA, AIES '20* (pp. 180–186). https://doi.org/10.1145/3375627.3375830

Slack, D., Hilgard, A., Lakkaraju, H., & Singh S. (2021a). Counterfactual explanations can be manipulated. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (pp. 62-75). https://proceedings.neurips.cc/paper/2021/hash/009c434cab57de48a31f6b669e7ba266-Abstract.html

Slack, D., Hilgard, A., Singh, S., & Lakkaraju, H. (2021b) Reliable post hoc explanations: Modeling uncertainty in explainability. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang & J. W. Vaughan (Eds.), *Advances in neural information processing systems 34: Annual conference on neural information processing systems 2021, NeurIPS 2021, December 6-14, 2021, virtual* (pp. 9391-9404). https://proceedings.neurips.cc/paper/2021/hash/4e246a381baf2ce038b3b0f82c7d6fb4-Abstract.html

Sokol, K., & Flach, P. (2020). Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, FAT* '20* (pp. 56-67). https://doi.org/10.1145/3351095.3372870

Solans, D., Biggio, B., & Castillo, C. (2020). Poisoning attacks on algorithmic fairness. In F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Eds.), *Machine learning and knowledge discovery in databases—European conference, ECML PKDD 2020, Ghent, Belgium, September 14-18, 2020, proceedings, part I, lecture notes in computer science* (Vol. 12457, pp. 162–177). Springer. https://doi.org/10.1007/978-3-030-67658-2_10

Sorokina, D., Caruana, R., Riedewald, M., & Fink, D. (2008). Detecting statistical interactions with additive groves of trees. In W. W. Cohen, McCallum, A., & S. T. Roweis (Eds.), *Machine learning, proceedings of the twenty-fifth international conference (ICML 2008), Helsinki, Finland, June 5-9, 2008, ACM international conference proceeding series* (Vol. 307, pp. 1000-1007). ACM.https://doi.org/10.1145/1390156.1390282

Stanford, P. K. (2006). Exceeding our grasp: Science, history, and the problem of unconceived alternatives. *Oxford University Press*. https://doi.org/10.1093/0195174089.001.0001

Stepin, I., Alonso, J. M., Catalá, A., & Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access, 9*, 11974–12001. https://doi.org/10.1109/ACCESS.2021.3051315

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., & Fergus, R. (2014). Intriguing properties of neural networks. In: Y. Bengio & Y. LeCun (Eds.), *2nd international conference on learning representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, conference track proceedings*. http://arxiv.org/abs/1312.6199

Tang, R., Liu, N., Yang, F., Zou, N., & Hu, X. (2022). Defense against explanation manipulation. *Frontiers Big Data, 5*, 704203. https://doi.org/10.3389/fdata.2022.704203

Tartaro, A., Panai, E., & Cocchiaro, M. Z. (2024). Ai risk assessment using ethical dimensions. *AI and Ethics*. https://doi.org/10.1007/s43681-023-00401-6

The Royal Society. (1662). First charter. History of the Royal Society. https://royalsociety.org/about-us/who-we-are/history/

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., & Preece, A. D. (2020). Sanity checks for saliency metrics. In *The thirty-fourth AAAI conference on artificial intelligence, AAAI 2020, the thirty-second innovative applications of artificial intelligence conference, IAAI 2020, the tenth AAAI symposium on educational advances in artificial intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020* (pp. 6021–6029). AAAI Press. https://ojs.aaai.org/index.php/AAAI/article/view/6064

Tramèr, F., Carlini, N., Brendel, W., & Madry A. (2020). On adaptive attacks to adversarial example defenses. In: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan & H. Lin (Eds.), *Advances in neural information processing systems 33: Annual conference on neural information processing systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. https://proceedings.neurips.cc/paper/2020/hash/11f38f8ecd71867b42433548d1078e38-Abstract.html

Trout, J. D. (2002). Scientific explanation and the sense of understanding. *Philosophy of Science, 69*(2), 212–233. https://doi.org/10.1086/341050

Tubbs, R. M., Messier, W. F., & Knechel, W. R. (1990). Recency effects in the auditor's belief-revision process. *The Accounting Review, 65*(2), 452–460.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Vandenberghe, F. (2015). *Reification: History of the concept* (pp. 203–206). https://doi.org/10.1016/B978-0-08-097086-8.03109-3

Vasconcelos, H., Jörke, M., Grunde-McLaughlin, M., Gerstenberg, T., Bernstein, M. S., & Krishna, R. (2023). Explanations can reduce overreliance on AI systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction, 7*(CSCW1), 1–38. https://doi.org/10.1145/3579605

Veldanda, A. K., Liu, K., Tan, B., Krishnamurthy, P., Khorrami, F., Karri, R., Dolan-Gavitt, B., & Garg, S. (2021). Nnoculation: Catching badnets in the wild. In N. Carlini, A. Demontis, & Y. Chen, (Eds.), *AISec@CCS 2021: Proceedings of the 14th ACM workshop on artificial intelligence and security, virtual event, Republic of Korea, 15 November 2021* (pp. 49–60). ACM. https://doi.org/10.1145/3474369.3486874

Virgolin, M., & Fracaros, S. (2023). On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence, 316*, 103840. https://doi.org/10.1016/j.artint.2022.103840

Vreš, D., & Robnik-Šikonja, M. (2022). Preventing deception with explanation methods using focused sampling. *Data Mining and Knowledge Discovery*. https://doi.org/10.1007/s10618-022-00900-w

Wachter, S., Mittelstadt, B. D., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. CoRR abs/1711.00399. http://arxiv.org/abs/1711.00399

Waldmann, M. (2000). Competition among causes but not effects in predictive and diagnostic learning. *Journal of Experimental Psychology Learning Memory and Cognition, 26*, 53–76. https://doi.org/10.1037/0278-7393.26.1.53

Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., & Garnett, R (Eds.). (2019). *Advances in neural information processing systems 32: Annual conference on neural information processing systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. https://proceedings.neurips.cc/paper/2019

Walton, D. N. (1994). Begging the question as a pragmatic fallacy. *Synthese, 100*(1), 95–131. https://doi.org/10.1007/bf01063922

Walton, D. (2008). *Informal logic: A pragmatic approach*. Cambridge University Press.

Walton, D. (2010). *The place of emotion in argument*. Penn State Press.

Warnecke, A., Arp, D., Wressnegger, C., & Rieck, K. (2020). Evaluating explanation methods for deep learning in security. In *IEEE European symposium on security and privacy, EuroS &P 2020, Genoa, Italy, September 7-11, 2020* (pp. 158–174). IEEE. https://doi.org/10.1109/EuroSP48549.2020.00018

Watson, D. S. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines, 29*(3), 417–440. https://doi.org/10.1007/s11023-019-09506-6

Weerts, H. J. P., Dudík, M., Edgar, R., Jalali, A., Lutz, R., & Madaio, M. (2023) Fairlearn: Assessing and improving fairness of AI systems. *Journal of Machine Learning Research*, *24*, 257:1–257:8

Weidinger, L., Uesato, J, Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022) Taxonomy of risks posed by language models. In *2022 ACM conference on fairness, accountability, and transparency. Association for Computing Machinery, New York, NY, USA, FAccT '22* (pp. 214–229). https://doi.org/10.1145/3531146.3533088

Weisberg, D., Keil, F., Goodstein, J., Rawson, E., & Gray, J. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience, 20*, 470–477. https://doi.org/10.1162/jocn.2008.20040

Weitzner, D. J., Abelson, H., Berners-Lee, T., Feigenbaum, J., Hendler, J. A., & Sussman, G. J. (2008). Information accountability. *Communications of the ACM, 51*(6), 82–87. https://doi.org/10.1145/1349026.1349043

Wicker, M., Heo, J., Costabello, L., & Weller, A. (2022). Robust explanation constraints for neural networks. CoRR abs/2212.08507. https://doi.org/10.48550/arXiv.2212.08507

Wieringa, M. (2023). "hey syri, tell me about algorithmic accountability'': Lessons from a landmark case. *Data & Policy*. https://doi.org/10.1017/dap.2022.39

Wikipedia. (2023). Ignotum per ignotius. https://en.wikipedia.org/wiki/Ignotum_per_ignotius

Wilkenfeld, D., & Lombrozo, T. (2015). Inference to the best explanation (IBE) versus explaining for the best inference (EBI). *Science & Education*. https://doi.org/10.1007/s11191-015-9784-4

Wilson, R. A., & Keil, F. (1998). The shadows and shallows of explanation. *Minds and Machines, 8*(1), 137–159. https://doi.org/10.1023/A:1008259020140

Woods, W., Chen, J., & Teuscher, C. (2019). Adversarial explanations for understanding image classification decisions and improved neural network robustness. *Nature Machine Intelligence, 1*(11), 508–516. https://doi.org/10.1038/s42256-019-0104-6

Yates, J., Lee, J. W., & Bush, J. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and "reality''. *Organizational Behavior and Human Decision Processes, 70*(2), 87–94. https://doi.org/10.1006/obhd.1997.2696

Zagzebski, L. T. (2012). *Epistemic authority: A theory of trust, authority, and autonomy in belief*. Oxford University Press.

Zhang, C., Yang, Z., & Ye, Z. (2018). Detecting adversarial perturbations with saliency. CoRR abs/1803.08773. http://arxiv.org/abs/1803.08773

Zhang, H., Gao, J., & Su, L. (2021). Data poisoning attacks against outcome interpretations of predictive models. In F. Zhu, B. C. Ooi & C. Miao (Eds.), *KDD '21: The 27th ACM SIGKDD conference on knowledge discovery and data mining, virtual event, Singapore, August 14-18, 2021* (pp. 2165–2173). ACM. https://doi.org/10.1145/3447548.3467405

Zhang, H., Yu, Y., Jiao, J, Xing, E. P., El Ghaoui, L., & Jordan, M. I. (2019) Theoretically principled trade-off between robustness and accuracy. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, proceedings of machine learning research* (Vol. 97, pp. 7472–7482). PMLR. http://proceedings.mlr.press/v97/zhang19p.html

Zhang, X., Wang, N., Shen, H., Ji, S., Luo, X., & Wang, T. (2020) Interpretable deep learning under fire. In S. Capkun & F. Roesner (Eds.), *29th USENIX security symposium, USENIX security 2020, August 12-14, 2020* (pp. 1659–1676). USENIX Association. https://www.usenix.org/conference/usenixsecurity20/presentation/zhang-xinyang

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.