Computational Intelligence

WILEY

Computational Intelligence

SPECIAL ISSUE ARTICLE OPEN ACCESS

Survey on AI Ethics: A Socio-Technical Perspective

Dave Mbiazi[1,2] | Meghana Bhange[2,3] | Maryam Babaei[2,3] | Ivaxi Sheth[4] | Patrik Kenfack[2,3] | Samira Ebrahimi Kahou[2,5,6]

[1]Computer and Software Engineering, Polytechnique Montréal, Montreal, Quebec, Canada | [2]Mila, Montreal, Quebec, Canada | [3]Software and Information Technology Engineering, ÉTS Montréal, Montreal, Quebec, Canada | [4]CISPA-Helmholtz Center for Information Security, Saarland, Germany | [5]Electrical and Software Engineering, University of Calgary, Calgary, Alberta, Canada | [6]Canada CIFAR AI Chair, Calgary, Canada

Correspondence: Maryam Babaei (maryam.babaei.1@ens.etsmtl.ca)

Received: 31 October 2024 | **Revised:** 23 July 2025 | **Accepted:** 31 July 2025

Keywords: accountability | AI ethics | fairness | interpretability | privacy | responsibility | security | trustworthiness

ABSTRACT

The past decade has observed a significant advancement in AI, with deep learning-based models being deployed in diverse scenarios, including safety-critical applications. As these AI systems become deeply embedded in our societal infrastructure, the repercussions of their decisions and actions have significant consequences, making the ethical implications of AI deployment highly relevant and essential. The ethical concerns associated with AI are multifaceted, including challenging issues of fairness, privacy and data protection, responsibility and accountability, safety and robustness, transparency and explainability, and environmental impact. These principles together form the foundations of ethical AI considerations that concern every stakeholder in the AI system lifecycle. In light of the present ethical and future x-risk concerns, governments have shown increasing interest in establishing guidelines for the ethical deployment of AI. This work unifies the current and future ethical concerns of deploying AI into society. While we acknowledge and appreciate the technical surveys for each of the ethical principles concerned, in this paper, we aim to provide a comprehensive overview that not only addresses each principle from a technical point of view but also discusses them from a social perspective.

1 | Introduction

As AI becomes ubiquitous in our lives moving forward in this decade, focusing on the ethical implications of AI is not just essential but extremely pressing. Several ethical guidelines and principles have been released around the world by governments [1], organizations [2], and companies [3]. Among these ethical considerations, there are common principles that should be promoted in the development of AI systems [4–6]. These principles form an initial consensus of features or components that should be embedded in AI systems to make their use more socially acceptable. Figure 1 showcases the most common principles found in existing ethical guidelines [3]. These principles include *privacy and data protection*, to ensure the privacy preservation of sensitive information about individuals (Section 2); *safety and robustness*, to promote the robustness and reliability of AI systems in different real-world scenarios (Section 2); *transparency and explainability*, to uncover information about how the system works and explain the decisions made (Section 3); *fairness*, which promotes bias-free and non-discrimination in AI systems when used to make decisions in high-stakes scenarios (Section 4); *responsibility and accountability*, to promote processes and rules to enforce ethical considerations throughout the entire lifecycle of AI systems to limit unintended outcomes (Section 5); *environmental impact*, to study the impact the growing deployment and use of AI systems might have on the environment along with how these systems can be leveraged for environmental protection (Section 6). All these principles play a key role in fostering trust among all stakeholders in the AI system lifecycle. In this paper, we walk through these principles and present recent

efforts, solutions, and regulations for implementing each ethical guideline in AI systems.

Despite the vast research on AI Ethics, most existing surveys fail to provide a thorough overview of AI's negative social impacts and technical solutions to mitigate them. This work provides a sociotechnical perspective on ethical AI. We cover the social impact and value of each guideline and discuss technical contributions from the literature to address them. We extend our investigation from classic AI to newly emerged foundation models and compare how these new models inherit concerning attributes of the classic models, along with the new concerns that have emerged with them.

1.1 | Methodology

This work was originally developed for the Kaggle AI Report 2023 [7], with the primary goal of covering major works across all domains of AI ethics. We focused on identifying the major seminal works in the AI ethics subdomains and high-impact conferences and journals such as FAccT, AAAI, ICML, NeurIPS, and the IEEE Security and Privacy conference. Our selection criterion was to include foundational works that established key concepts in AI ethics subdomains, including fairness, explainability, privacy, security, AI regulation, and accountability. To do so, we searched academic databases with search terms such as, but not limited to, "AI regulation," "ML privacy," "ML security," "Privacy risks in LLMs," "AI Fairness," "AI governance," "Generative AI ownership," "Foundation models," and "AI accountability". In addition, methods such as forward and backward snowballing were employed to increase the coverage of the papers. Backward snowballing involved examining reference lists of selected papers to find earlier foundational works. Forward snowballing involved looking at citations to identify newer papers that built on the seminal work we identified. We included papers that addressed core AI ethics domains, introduced key concepts or methods, and came from recognized conferences and journals. We excluded articles that did not focus primarily on AI ethics and duplicates.

1.2 | Related Works

To position our work within the growing body of literature on AI ethics, we compare it with several recent and widely cited papers in the field. As shown in Table 1, our survey provides a broader and more technically detailed treatment of AI ethics topics. While existing works often focus on specific dimensions such as social impacts, auditability, or environmental concerns, our paper integrates these perspectives while also expanding the discussion to include underrepresented yet critical topics like existential risks, technical fairness definitions, multi-scale governance, and how these ethical challenges have evolved with the emergence of foundation models such as large language models. This comparative breadth and depth position our survey as a more comprehensive resource for all stakeholders.

TABLE 1 | Comparison with prior AI ethics survey papers.

Paper	Additional topics covered by our paper
Jiao et al. [8]	We additionally cover environmental and existential risks.
Laine et al. [9]	Their focus is limited to auditing, whereas we take a broader view.
Correa et al. [10]	Primarily focused on social dimensions; we include technical aspects as well.
Prem [11]	Emphasizes social ethics; we also examine technical definitions (e.g., fairness metrics).
Radanliev et al. [12]	We address important topics such as climate and ownership and AI existential risk.
Khan et al. [13]	Lacks in-depth coverage of technical components, which we provide.

2 | Privacy and Data Protection

Ensuring the security and privacy of machine learning models has become a crucial issue as they are widely used in various fields. To provide trustworthy AI, it is important to include safety, privacy, and security in its lifecycle. Based on some definitions, safety means reducing the probability of expected and unexpected harm [14]. According to this definition, a machine-learning model should be trained and released to be robust against different kinds of uncertainty; in other words, in case of an accident, the model should be able to continue its expected normal behavior. To prevent safety problems, first, the designer should be able to specify the correct objective function and have a method to evaluate it; second, sufficient data, time, and infrastructure should be available to train and evaluate the model [15]. Two main attributes of AI safety are safe exploration and robustness to distributional shift [15]. If one of these attributes is lacking, the model will be vulnerable to different attacks against its privacy and security.

Several studies [16–22] have indicated that machine learning models are susceptible to attacks on their privacy and security at different stages of their lifecycles. Since these models are trained on vast amounts of data, model training is sometimes outsourced, or pre-trained models are obtained from untrusted sources, which makes them vulnerable to attacks during the training phase. Additionally, machine learning models provided as a service contain valuable information about their training data and hyperparameters, making them attractive targets for attacks during the test and deployment phases [23]. Attacks on privacy commonly aim to compromise the confidentiality of various machine learning model components, while security attacks target the model's integrity and availability.

2.1 | Privacy and Security Attacks and Defenses Overview

Different attacks are possible depending on the model architecture and attackers' capabilities. In terms of security, attackers may aim to gain access to the model, steal its information, or disrupt its normal functioning. For example, an attacker may target a spam detector to make it unable to classify spam correctly by poisoning training data [24].

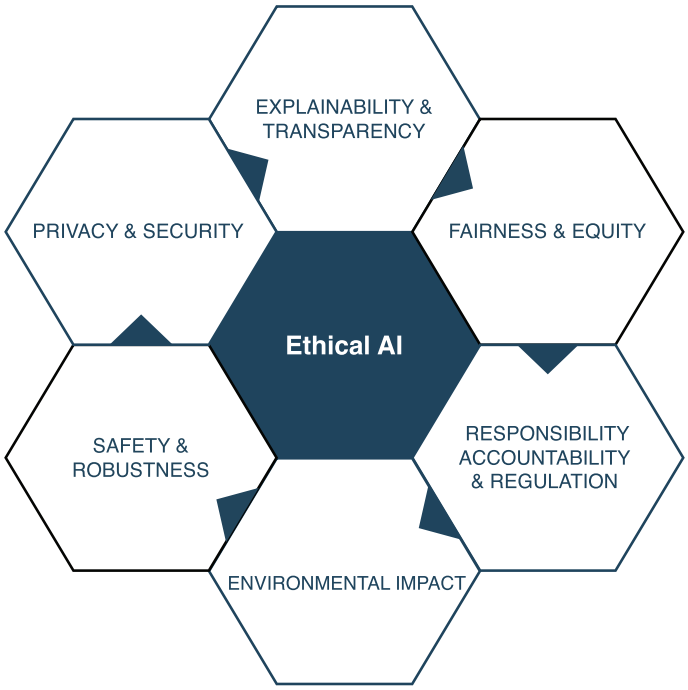

FIGURE 1 | Ethical AI principles.

The attacks can be categorized into three main categories: black-box, partial white-box, and white-box attacks, according to the attacker's knowledge of the machine learning model [25]. In black-box attacks, the attacker has no information about the training dataset or model's architecture; in white-box attacks, the attacker has full access to the model and all information about it, including the training dataset, model parameters, model architecture, prediction vectors, etc. Partial white-box attacks stand between these extremes, meaning the attacker has some information about the model's architecture and training data distribution. The most common security and privacy attacks against machine learning models are described in this section.

2.1.1 | Membership Inference

The membership inference is an attack in which an attacker attempts to determine if a particular data sample x is part of a model M's training dataset [26–28]. This attack is often carried out using black-box techniques to query the model. Different querying techniques are used to optimize the attack to gain more information about the membership of individual records in the training dataset. One of the first Membership inference attacks, implemented by Shokri et al. [29], which could achieve high accuracy in their inference, was performed on Google and Amazon's APIs that provide machine learning as a service (MLaaS). Quan et al. [30] showed that having additional knowledge about the model or training dataset distribution can improve the attack's success rate.

2.1.2 | Model Inversion

In a model inversion attack, the adversary tries to get information from the target model to reconstruct some representation of its input dataset. The first category aims to generate an actual data reconstruction [31]. In contrast, the second group of attacks tries to create class representatives or probable values of sensitive features that may not belong to the training dataset [32]. Several attacks have been implemented based on different assumptions. Some attacks assumed to have information about data and sensitive features, and some had query access to the model to get a prediction for an input X. Two main categories of this attack are performed.

2.1.3 | Property Inference

A property inference attack attempts to deduce information about the characteristics of a training dataset that are not explicitly represented in the features. Revealing such properties can result in privacy breaches since they may be considered confidential. Property inference attacks are designed to identify dataset-wide properties [33] or detect common characteristics among a subset of the training data [34]. For instance, a classification model may be trained to distinguish between criminals and non-criminals. An attacker can estimate the proportion of men and women in the dataset by conducting a property inference attack. This information was not meant to be disclosed and was learned by the model unintentionally.

2.1.4 | Model Extraction

In a model extraction attack [22], the attacker aims to develop a model replicating the target model's behavior. This is typically done when the attacker lacks information about the target model's architecture and training dataset. To achieve this, the attacker generates a training set for the attack model by sending queries to the target model and uses the predictions generated

by the target model as labels for its data points. A successful model extraction attack, as shown by Tremer et al. [22], allows the attacker to produce a model that can be used for inference or to extract information about the training dataset. However, if the attacker has some knowledge about the training set's distribution, the attack's success rate can be improved. Selecting the data to query the target model is a critical aspect of the attack methodology, which significantly impacts the accuracy and fidelity of the attack model.

2.1.5 | Poisoning

In certain situations, a machine learning model designer may not generate or thoroughly examine the data used to train the model. Generally, this may occur when data generation is outsourced to third parties, or pre-trained models are fine-tuned for a specific task. In such cases, the model becomes vulnerable to poisoning attacks [35] that target the training dataset. In a sample poisoning attack, the attacker injects malicious data into the training dataset, causing the model to learn patterns unrelated to the classification task specified for the model. These patterns can be exploited during the inference phase as a backdoor, resulting in incorrect decision-making by the model when provided with data containing these malicious patterns [36, 37].

2.1.6 | Evasion

Evasion attacks [38] occur when an adversary attempts to cause a machine learning model to misclassify a data sample during the inference phase. This attack typically occurs after the model is trained and deployed. The attacker aims to generate a data sample similar to the original ones but misclassified by maximizing a loss function based on their attack objectives. There are two main categories of evasion attacks: targeted and untargeted. While in a targeted attack, the adversary wants the manipulated sample to be classified as a specific class; in an untargeted attack, the attacker is not concerned with which class the manipulated input is classified in. The manipulated data samples created in evasion attacks are known as adversarial examples in the literature. Several techniques, such as projected gradient descent, have been developed to generate adversarial examples [38–40].

2.1.7 | Manipulation

Manipulation attacks occur when the attacker tries to explain the model's decisions based on some criteria that were not used. Here, explanations help users understand complex models' behavior (explanations are described in Section 3). However, to gain the users' trust, it is not enough to have an explanation; it should be robust, too. Manipulation attacks are the type of attacks that exploit the fragility of the explanations. It means that while the model makes decisions based on some criteria if manipulated, explanations can pretend the model is using other, more reasonable measures [41, 42]. This type of attack is performed to achieve several goals. One of its main objectives is fairwashing [43–45], which means the model designer has manipulated the explanation methods to hide the unfairness of their model's decisions (Fairness is described in detail in Section 4).

2.2 | An Overview of Defense Techniques

Researchers have proposed some defense strategies in response to security and privacy attacks. Some of the most common defense techniques are described here in summary. The k-anonymity [46] is a technique usually used to prevent privacy attacks. It guarantees that each data instance is indistinguishable from at least $k-1$ other data points. However, it may not be sufficient to prevent identification when additional information or external knowledge is available. Differential privacy is one of the most popular defenses against privacy attacks on machine learning models [47]. Differential privacy guarantees that no more privacy risks will be introduced to the data after being used to train the model, compared with when they are not used for this purpose. One of the main formalizations of differential privacy is ϵ-differential privacy, which means that if two models are training on the two datasets (D, D') that are only different in one record, for every S in the domain of the model, the probability of their distributions differs at most with the ratio of $\exp(\epsilon)$. Differential privacy can be applied in different phases of the machine learning lifecycle. It is possible to use it to make training data private [48], or during the training phase, using differentially private training techniques like PATE [49] and DPSGD[1] [50]. Furthermore, model owners can apply differential privacy techniques after training the model in the inference phase to make the model respond to the queries without privacy leakage [51].

In addition to defenses against privacy attacks, some techniques have been suggested to protect the model against security attacks. As a countermeasure to evasion attacks, adversarial training [52] has been suggested to make the model robust by exposing it to some adversarial examples in the training phase, allowing it to learn correct behavior against them.

Applying each of the above-mentioned defense techniques is a critical decision that has to be made based on the model owners, data owners, and model users' priorities. Increasing the privacy and robustness of the models comes with the cost of reducing their accuracy. Here is where the stakeholders should decide to what degree they will make their models immune against privacy and security attacks and, at the same time, how much accuracy loss they can afford to achieve this degree of immunity.

2.3 | Privacy and Security in the Emergence of Foundation Models

Now that Large Language Models (LLMs) and generative models have wide applications in all areas of science and industry and even daily use by regular users, it is important to investigate the privacy and security risks imposed by these models.

2.3.1 | LLMs for Data Security and Privacy

Researchers have shown that LLMs can be used for security objectives. For instance, ChatGPT-4.0 has been used to generate security tests for evaluating how vulnerable library dependencies impact software applications [53]. OpenAI's GPT-4 has also been effectively used for software vulnerability detection [54]. LLMs could also successfully detect vulnerabilities in specialized

domains like blockchain [55, 56] and ransomware detection [57]. Siddiq and Santos [54] introduce the SALLM framework, consisting of a new dataset specified for security and an evaluation environment. They introduce novel metrics for systematically evaluating LLMs' ability in secure coding. Researchers have also explored using LLMs to enhance privacy in some studies. For example, Vats et al. [58] utilized LLMs to deidentify textual data.

2.3.2 | Negative Impacts on Security and Privacy

In addition to their effectiveness in improving privacy and security, LLMs could be exploited for adversarial purposes. The application of LLMs for side-channel attacks has been analyzed in [59]. LLMs have also been used to analyze vulnerabilities in virtual machines and propose and automatically execute OS-level attacks against them [60].

Beckerich et al. [61] proposed using ChatGPT to distribute malicious software while avoiding detection. LLMs have also been employed to carry out network-level attacks, such as phishing [62]. Researchers have demonstrated that LLMs enhance user-level attacks, including misinformation [63], social engineering [64], and fraud [65].

2.3.3 | Vulnerabilities and Defenses in LLMs

While LLMs can be utilized to enhance privacy and security, they also pose a risk of exploitation. Adversarial users can exploit vulnerabilities in the models themselves to simulate adversarial scenarios, leading to potentially harmful activities. Research has indicated that LLMs may carry certain AI-inherited vulnerabilities. Exploiting these vulnerabilities, various attacks have been conducted against LLMs, including data poisoning [66, 67] to push the model to return malformed responses and backdoor attacks [68] employing prompt injection to manipulate the model's behavior. While these vulnerabilities have been exploited for malicious purposes, they can also be exploited to ensure copyright protection for artists and content creators. For instance, some researchers [69, 70] employed watermark embeddings to restrict diffusion models and generative adversarial networks (GANs) from exploiting copyrighted content in violation of copyright regulations. The efficacy of watermarking LLMs has been investigated in the context of text generation [71].

Similar to traditional ML models, LLMs and generative models are also susceptible to inference-time attacks in addition to vulnerabilities and attacks during the training phase. These include Attribute Inference Attacks [64], Membership Inferences [72, 73] (These attacks have also been used to find out if generative models have used copyrighted content in their training), Bias and Unfairness Exploitation [74–76], Adversarial Attacks (Instruction Tuning Attacks) [77, 78] and Prompt Injection [79, 80], Denial of Service [81, 82] and Remote Code Execution (RCE) [83].

2.3.4 | Defense Techniques

OWASP, the leading organization in software security, has recommended the OWASP Top 10 for Large Language Model Applications [84]. It is critical for LLM developers and users to thoroughly review these recommendations prior to deploying their models. Additionally, researchers have proposed various safeguards to address existing vulnerabilities, thereby mitigating the risk of malicious users executing successful attacks against LLMs. The initial step entails mitigating certain properties from training data during its generation, collection, and cleaning processes. Research has been conducted in this domain, including debiasing [85], deidentification [86], and detoxifying [87].

Within the LLM pipeline, optimization techniques can also influence the ethical alignments of LLMs. Methods such as adversarial training [88] and robust finetuning [89] can enhance the resilience of LLMs against certain adversarial attacks. Following the training of models, it is imperative to implement techniques to safeguard LLMs against adversarial attacks. These defenses encompass a range of approaches, including pre-processing techniques for analyzing prompts [90, 91], in-processing techniques for detecting malicious behaviors that could lead to responses that violate ethical regulations or disclose private information [92, 93], and post-processing techniques that analyze generated responses by LLMs before returning them to users to mitigate their potential toxicity [94, 95].

3 | Transparency and Explainability

As mentioned in Section 2, it is necessary for AI users to understand how the model works and makes its decisions. Explainability, a crucial aspect of machine learning, bridges the gap between predictive accuracy and human understanding. It has emerged as a critical aspect of machine learning models, addressing the need to understand the reasoning behind their decisions. While predictive accuracy has long been the primary focus, the growing demand for transparency in models has motivated academic research in interpreting black-box neural networks [96]. Explanations can potentially allow the deployment of machine learning models while maintaining high ethical standards [97].

In various areas of AI applications, the need to have explanations for AI model decisions has raised and played an important role in motivating researchers to focus on making complicated models explainable.

For example, in medical applications, using an explainable model for patient screening not only identifies high-risk individuals but also helps understand disease causes like cancer [98, 99]. This model could additionally provide insights into predictive factors, relationships between risk elements, and significant contributors to a diagnosis, such as biomarkers, genetic predispositions, lifestyle factors, or environmental exposures [100–102]. Interpretability in AI proves crucial in the finance sector [103, 104], aiding in understanding loan rejections, credit score calculations, and fraud detection. This helps stakeholders identify biases, errors, and discriminatory practices, ensuring transparency and accountability [105, 106]. Compliance and regulatory bodies also benefit from interpretability, as financial institutions must elucidate AI-driven decisions. Furthermore, explainability may enhance AI model accuracy and bolster customer trust [107]. Therefore, interpretability is not merely desirable but necessary in financial AI applications. Further, explainability

enhances understanding of customer behavior, enabling personalized offerings and improved customer experience [108, 109]. It fosters customer satisfaction and loyalty by tailoring products and services to meet customer needs. Transparent explanations educate customers on AI-driven decisions, alleviating concerns and fostering understanding, thereby building trust and strengthening customer relationships [110, 111]. The case of AlphaGo's [112] "move 37" exemplifies AI's potential to surpass human intuition. This neural network model, trained to play Go, made a move that initially perplexed experts due to its deviation from traditional strategies. However, as the game unfolded, it became clear that AlphaGo foresaw the potential of this unconventional move, which proved pivotal in its ultimate victory. This case underscores the need for techniques to explain AI decision-making processes.

Regulators and policymakers recognize the need for mechanisms to shed light on AI models' inner workings, enabling stakeholders to understand and assess the justifiability of AI-driven decisions. Regulatory bodies across industries emphasize transparency, fairness, and non-discrimination in AI systems. Explanations play a pivotal role in meeting these regulatory requirements, elucidating how AI models arrive at their predictions or decisions. Legal and ethical concerns surrounding AI technologies necessitate the integration of explanations into AI systems.

3.1 | Stakeholders in XAI

Most experts agree that explainability and/or interpretability are crucial for artificial intelligence (AI) and machine learning systems. However, there is not a universal understanding of what "explainable" and "interpretable" mean. As a result, analyzing the opinion of stakeholder communities surrounding explainable AI is of great importance. The majority of stakeholder-related debates merely distinguish between end users and system developers. This can be seen in [97] through the following points

- Prediction-recipients who are directly impacted by the ML-based system's predictions despite not using it themselves since the prediction is typically mediated by an expert user.
- End users who utilize the ML-based system directly, and are consequently directly impacted by it.
- Expert users who directly use the ML-based system though they are not immediately impacted by its forecasts. Since they could be held responsible (both legally and ethically) for the results of predictions put into action, they are indirectly impacted.
- Attorneys and the courts are interested in determining who is responsible for damage caused by an ML-based system.
- The Financial Services Authority, the Vehicle Certification Agency, and the Medical and Healthcare Products Regulation Agency are regulatory bodies that are not the system's immediate users nor its direct beneficiaries. They, however, safeguard the interests of prediction recipients and end users [113].

A compressed version of the latter was suggested by [114], where the authors show a "Users Chart" of stakeholder groups as developed by the Defense Advanced Research Projects Agency (DARPA). They summarized some "Sensemaking Needs" and "Explanation Requirements" for some Stakeholder Groups (such as Policy Makers and Regulators, Developers: AI Experts, test Operators, and lastly Operations: Military, Legal, and Transportation).

3.2 | Properties of Explanations

Explanations, to be able to respond to their stakeholders' needs, should involve several properties:

Clarity refers to the quality of understanding the explanation. It involves presenting explanations in a clear, concise, and interpretable manner. A clear explanation avoids unnecessary jargon or technical complexity, making it accessible to the intended stakeholders. It should be structured and organized to facilitate understanding and promote effective communication of underlying concepts or factors.

Fidelity is a crucial property of explanations, emphasizing the importance of providing correct and truthful information [115, 116]. An accurate explanation aligns with the underlying data or model, ensuring that the explanation reflects the actual reasons and factors influencing a decision. Explanations must not misrepresent or distort the information but rather provide a faithful representation of the relevant aspects of the data.

Completeness relates to the extent to which an explanation provides a comprehensive account of the relevant factors or features contributing to a decision or result. A complete explanation includes all the necessary information, leaving no significant gaps or missing components. It should cover both the primary factors and any secondary or indirect factors that may have influenced the outcome. A complete explanation helps to avoid ambiguity or misunderstanding by providing a holistic view of the situation.

Consistency emphasizes an explanation's coherence and logical consistency [117]. A consistent explanation should not contain contradictory statements or conflicting information. It should maintain a coherent narrative that aligns with the underlying data or model. Consistency ensures that the explanation is internally coherent and does not introduce confusion or ambiguity in the interpretation of the provided information.

Causality explores the causal relationships between variables or factors. A causal explanation seeks to identify and explain the cause-effect relationships that lead to a particular outcome or prediction [118]. It provides insights into the mechanisms and processes that drive the observed phenomena. Causal explanations help to uncover the underlying reasons behind a decision or result, shedding light on the factors that directly or indirectly influence the outcome.

Transparent relates to the openness and accessibility of an explanation. A transparent explanation is understandable, interpretable, and accessible to the intended audience. It avoids

unnecessary complexity or obfuscation and allows individuals to examine and verify the reasoning behind a decision or outcome. Transparency promotes trust, accountability, and scrutiny, enabling individuals to assess the reliability and fairness of the provided explanation.

Contextuality considers an explanation's relevance and contextual appropriateness [119]. A contextual explanation takes into account the specific circumstances, background knowledge, and contextual factors that may impact the interpretation and understanding of the explanation. It adapts the level of detail, language, or content to suit the specific context or audience, ensuring that the explanation is meaningful and relevant within the given context.

Granular explanations can vary in their level of granularity. They can range from high-level summaries to detailed explanations at the feature or instance level. The granularity of explanations should align with the users' needs and their level of understanding, striking a balance between simplicity and depth.

User-Centric [120] emphasizes the importance of tailoring explanations to the needs, preferences, and cognitive abilities of the intended users. A user-centric explanation is designed to effectively communicate information to the target audience, taking into account their background knowledge, expertise, and information processing capabilities. It considers the user's perspective and provides explanations adapted to their specific requirements and level of understanding. User-centric explanations enhance the usability and utility of the provided information.

3.3 | Explainability Techniques

3.3.1 | Classical Methods

A subset of algorithms that inherently produces interpretable models is a straightforward approach to achieving interpretability.

Logistic regression [121–124], an extension of linear regression, models binary outcomes based on input variables. Extensions like ridge, lasso, and elastic net regression improve performance and interpretability. Decision trees [125–127], intuitive models that split data based on input features, capture non-linear relationships. Decision rules provide transparent decision-making representations. The RuleFit algorithm [128] combines decision trees and linear regression, capturing complex interactions and incorporating interpretable linear components. However, these models' simplicity may limit capturing complex relationships and handling high-dimensional data. The choice of model depends on the problem, data characteristics, and desired interpretability level.

While these models offer interpretability, it is important to note that their simplicity and transparency come at the cost of potential limitations in capturing complex relationships and handling high-dimensional data. Additionally, the choice of model depends on the specific problem, data characteristics, and the desired level of interpretability.

3.3.2 | Post Hoc Methods

With the emergence of deep learning and the need for highly accurate large neural networks, local explanations—post hoc explanations—have emerged to be highly useful. Instead of explaining the entire model, post hoc methods explain a particular decision.

Individual Conditional Expectation [129] curves provide a fine-grained view of how changing a specific feature affects the model's prediction for an individual instance. Unlike PDP, which shows the average effect, ICE curves show the predicted outcome for each instance as the feature value varies. Feature importance-based explanations [130–133] such as Local Interpretable Model-agnostic Explanations (LIME) [134] explain individual predictions by approximating the complex model locally with a simpler, interpretable model. It generates a surrogate model that is more easily explainable and uses it to understand the reasoning behind the prediction for a specific instance. Similarly, Shapley values [133] are an attribution method that fairly allocates the prediction value to individual features. SHAP [131] is a computation method for Shapley values that combines the individual feature contributions to explain the model's predictions. It not only provides insights at the feature level but also proposes global interpretation methods by considering combinations of Shapley values across the data. Scoped rules, also known as anchors [135], are rule-based explanations that identify specific feature values that anchor or lock a prediction in place. LORE (LOcal Rule-based Explanations) [136] creates interpretable rules by using two types of perturbations, in a genetic algorithm, to find the minimal changes that would alter the prediction. Counterfactual explanations [137–139] explore what changes in the feature values would be required to achieve a desired prediction outcome. Counterfactual explanations help understand the model's decision boundaries and provide insights into how different features impact the predicted outcome by identifying the necessary modifications to the features. Concept attribution attributes the final prediction of a model to align with the high-level concept of the input [140–142].

3.3.3 | Ante-Hoc Methods

Rudin [96] highlights the main challenges of explainable models, stating the importance of learning explainable features during model training itself. Ante-hoc explainability methods involve the learning of concepts during the training phase. Early concept-based models that involved the prediction of concepts prior to the classifier were widely used in few-shot learning settings [143, 144]. Unsupervised concept learning methods [145, 146] use a concept encoder to extract the concepts and a relevance network for final predictions. Although these methods are useful when pre-defined concepts are absent, they do not enable effective interventions. Concept whitening [147] was introduced as a method to plug an intermediate layer in place of the batch normalization layer of a CNN to assist the model in concept extraction. Koh et al. [148], Espinosa Zarlenga et al. [107] extend the idea by decomposing the task into two stages: concept prediction through a neural network from inputs, and then target prediction from the concepts. Such concept-based models [149, 150]

have been further utilized to facilitate human-model interaction, a useful feature for model editing.

3.4 | Explainability in the Emergence of Foundation Models

In a rising number of different tasks, LLMs such as Gemini, Claude and GPT-4 [151–154] have shown outstanding performance. However, these models have become known as "black boxes" due to their inability to be understood clearly. Due to their opacity, they are no longer useful in high-risk fields like medicine and policymaking. Explainability is essential for LLMs since it enables users to comprehend how the model generates its predictions [155, 156].

In this line, Cífka and Liutkus [157] developed a model-agnostic explanation technique based on tracking the model's predictions as a function of the number of context tokens available for causal (autoregressive) language models. Each time a new token is introduced, there is an increment in context length, and the authors proposed a metric, *Differential Importance Score* to quantify this change. These scores appear to have the ability to find long-range dependencies (LRDs), which is particularly intriguing because they are intended to highlight information not already covered by shorter contexts, unlike attention maps, for instance.

Deep neural networks are trained to recognize very particular structural and perceptual attributes of inputs. Techniques for locating neurons that react to certain idea categories, such as textures, are readily available in computer vision. Nevertheless, the scope of these methods is constrained, labeling only a tiny portion of the neurons and behaviors in every network. To solve this, Hernandez et al. [158] proposed *MILAN* (*m*utual-*i*nformation-guided *l*inguistic *a*nnotation of *n*eurons) which generates descriptions (that capture categorical, relational, and logical structure in learned features) of neuron behavior in vision models using patch-level information about visual characteristics.

Another approach to black box language models is through text modules. Singh et al. [156] introduced the *S*ummarize *a*nd *S*core (SASC) approach that takes a text module as input and outputs a natural language explanation of the module's selectivity coupled with a reliability score. They show better interpretability for LLMs may be attained by the SASC, which can enhance automated analysis of LLM submodules such as attention heads. Bills et al. [159] offer a SASC-like technique for explaining individual neurons in an LLM by forecasting token-level neuron activations. They used an automated approach to solve the issue of scaling an interpretability approach to each neuron in an LLM which is expected to assess the trustworthiness of the models before deployment. The method clarifies how textual patterns trigger neuron activation through: explaining neuron's activation using GPT-4, simulating activations conditioned on the explanation, and scoring the explanation [160, 161].

Nevertheless, Zhao et al. [162] provide a classification of explainability methods and a systematic summary of strategies for elucidating Transformer-based language models. These techniques are categorized according to the LLM training approaches: the traditional fine-tuning method and the prompting-based method. They also delve into metrics for assessing the quality of generated explanations and explore how these explanations aid in troubleshooting to enhance model performance.

In addition to the above methods, the field of mechanistic interpretability seeks to reverse-engineer the internal computations of LLMs by identifying circuits, patterns of neuron activations, and interpretable algorithmic structures within the network [163]. Representation engineering, another active area, involves modifying or constructing internal representations, such as editing activations or directions in embedding spaces to induce or analyze specific behaviors in the model [164]. Probing techniques are also widely used, where lightweight classifiers are trained on hidden representations to determine whether specific linguistic or semantic information is encoded at various layers [165, 166].

4 | Fairness and Equity

As discussed in Sections 2 and 3, adding transparency to the AI systems enables different stakeholders to find the system's deviations from desired behavior. Fairness is one of the main requirements that many AI systems fail to meet. Fairness can be defined as the absence of any prejudice or favoritism toward an individual or a group of individuals based on their inherent or acquired characteristics, such as race, gender, religion, etc. There are numerous examples of AI applications that exhibit unwanted discriminatory behaviors. This is alarming for the need to take action to overcome the potential bias that might be embedded in AI systems. For instance, the Compas system is a software used in the US to assess the recidivism risk of defendants. Julia Angwin investigated [167] the software and showed that compared with white defendants, black defendants are predicted to be twice as likely to re-offend although they do no subsequent offenses. Another example is the AI-based hiring system [168] used by Amazon for assessing the resumes of job applicants. It was observed that the evaluations assigned to applicants' resumes exhibited gender bias. These examples of AI bias have triggered the need for developing more inclusive AI models to make their use more socially acceptable. A biased AI system does not only have a negative impact on the end-users but the organization deploying the system can suffer reputation damage, user distrust, and judicial liability. Initial important steps to mitigate these issues focused on mathematically defining and quantifying bias in AI systems.

4.1 | Definitions of Bias and Fairness

The concept of fairness has a variety of definitions depending on the domain considered [169–171]. It can be defined according to *political philosophy, areas of education, in the legal domain and according to the general public's perception* [172]. Unfairness or discrimination is divided into two categories:

- Disparate treatment: Which is defined as "intentionally treating individuals differently based on their membership in a demographic group (direct discrimination)" [173]
- Disparate impact: which is "negatively affecting members of a demographic group more than others even if by a seemingly neutral policy (indirect discrimination)" [173]

In fact, AI systems can exhibit *disparate treatment* if the model heavily relies on sensitive attributes to make predictions. However, in general, discriminatory outcomes of AI systems are not intended, but due to different sources of bias, the system will provide *disparate outcome*s over different demographic groups considered.

4.2 | Source of Bias and Fairness Notions

Unfairness in machine learning originates from three main sources of biases: biases due to the data, those from the algorithm, and user interaction. Machine learning systems and AI systems are data-driven since they rely on data to be trained, making them an integral part of the system. As a result, if the algorithm is trained on a biased dataset, then these biases are likely to be portrayed in the model's outcome.

4.2.1 | Source of Bias

There are a wide variety of sources of bias in data, and some important ones are highlighted here as reported by [172].

4.2.1.1 | Omitted-Variable Bias (OVB). OVB occurs when the dataset fails to incorporate one or more relevant variables/features. The authors in [174] showed that in a model which explains the relationship between dependent and independent variables, omitting a relevant variable leads to biased estimates. They also showed that OVB leads to statistical relationships that can be indicated as larger, smaller, or opposite to their actual value, which inflates error rates.

4.2.1.2 | Measurement Bias. Also called Reporting or Recall bias, which is a result of how important features are measured [172].

4.2.1.3 | Aggregation Bias. It occurs when it is erroneously believed that individual data points follow the trends found in aggregated data. Aggregation bias frequently happens in research because it is sometimes assumed incorrectly that patterns that exist at an aggregate level must also appear at an individual level. Sadly, as the preceding illustration showed, this is not always true. A study's results may derive incorrect conclusions due to aggregate bias, which is misleading. This kind of bias is especially damaging when it comes to the correlations between different variables.

4.2.1.4 | Representation Bias. It occurs when certain segments of the target population are underrepresented in the training data and, consequently, do not generalize well. Data representation bias may be due to biases introduced after the data was obtained, either historically, cognitively, or statistically, or it may result from how (and where) the data was initially collected [175]. Selection bias can cause representation bias, which occurs when just a small percentage of the population is sampled, the population of interest has changed, or the population of interest differs from the population used to train the model. For instance, if a poll measuring the illegal drug use of teenagers only includes high school students and leaves out homeschooled children or dropouts, it may be biased [176]. The skewness of the underlying distribution is another possible explanation for representation bias. Let's say that adults aged 18–60 years are the target demographic for a specific medical dataset. Within this community, there are minority groups; for instance, pregnant women may constitute only 5% of the target population. Because the model has fewer data points to learn from for the group of pregnant people, it is susceptible to being less robust even with perfect sampling and an identical population [176].

4.2.1.5 | Algorithmic Bias. Another common source of bias can be the algorithm itself. The machine learning model, trained on a biased dataset, can reproduce and amplify the biases in the model's output. Even if trained using an unbiased dataset, machine learning algorithms throughout their architecture have the ability to demonstrate biased behavior [172]. It arises solely as a result of the design characteristics and model architecture, such as the choice of the regularizer, and loss functions.

Having identified the origins of bias in the AI lifecycle, auditing AI systems for biases assessment requires metrics to quantify them and to evaluate the efficiency of intervention methods. In this regard, various fairness metrics (definitions) have been defined to capture different aspects of fairness.

4.2.2 | Fairness Definitions

Fairness notions can be categorized into group, individual and subgroup types [177]. Group fairness suggests that different groups are treated equally. In its widest sense, group fairness splits a population into groups defined by *protected attributes/features* (such as gender, religion, caste) and desires some statistical quantities to be equal across different groups. Mehrabi et al. [172], Weerts et al. [178] discuss a wide variety of group fairness metrics and the following are the most significant ones:

4.2.2.1 | Demographic Parity. The metric seeks to guarantee that a model's predictions are not related to one's belonging to a vulnerable group. Demographic parity refers to equal selection rates for each group in the binary classification scenario [177, 179]. Equal selection, for instance, in the context of a resume screening approach, would imply that the proportion of candidates chosen for a job interview should be the same across groups.

4.2.2.2 | Equalized Odds. This metric aims to guarantee that a machine learning model works equally effectively for all groups. It is more stringent than demographic parity because it demands that groups have the same true positive and false positive rates as well as independent predictions from the machine learning model regardless of membership in sensitive groups [179, 180]. This distinction is crucial because even if a model achieves demographic parity (i.e., its predictions are not dependent on a subject's membership in a sensitive group), it may nevertheless provide more false positive predictions for a particular group.

4.2.2.3 | Equal Opportunity. This can be understood as requiring that both protected and unprotected group members have an equal chance of being allocated to a positive result if they

belong to a positive class [172]. In other words, the equal opportunity definition states that the true positive rates for protected and unprotected groups should be equal.

4.2.2.4 | Disparate Impact (DI). This notion can also be viewed as the ratio between the two groups' rates of accurate predictions and so a high value of the ratio guarantees that the percentage of accurate predictions is consistent across groups. Nevertheless, one of the major drawbacks of *disparate impact* and *demographic parity* is that a perfectly accurate classifier may be viewed as being unfair when the proportion of real positive outcomes of the various groups is noticeably different [173].

4.2.2.5 | Individual Fairness. In contrast to group fairness, individual fairness is focused on how each individual is treated [181]. This notion requires similar individuals to receive similar outcomes from the model. Individual fairness is beneficial because it is a highly specific way of defining fairness and also because people tend to care more about individuals than large groups [182].

4.2.2.6 | Fairness Through Unawareness. This notion states that a model is fair as long as it is not trained using the sensitive attributes [172]. However, a significant weakness of this notion is its failure to consider non-sensitive features that may correlate with sensitive ones. When the model uses these non-sensitive features as proxies for the unused sensitive features, it can result in discriminatory outcomes [183].

4.2.2.7 | Counterfactual Fairness. It concerns the root causes of differences. A sensitive trait would be replaced in practice, affecting everything that occurred due to that sensitive feature down the line [184]. In the hiring scenario, one would alter a sensitive attribute, such as race, if counterfactual fairness were applied. As a result, subsequent outcomes should not be altered. Based on the counterfactual, the decision of a classifier as to whether to hire the candidate should remain the same.

4.3 | Fairness-Enhancing Methods

Fairness-enhancing methods are grouped into three main categories based on the stage of the pipeline where the fairness constraint is enforced, that is, at the data level before training the model (pre-processing techniques), during the model training (in-processing techniques), or after training the model (post-processing techniques).

4.3.1 | Pre-Processing Techniques

A model that relies on sensitive attributes (e.g., gender, race, nationality) to make predictions can lead to discrimination or unfair results. Pre-processing techniques are used to remove the influence of sensitive attributes from the data before training the model. The main advantage of these techniques is that they are model-agnostic. The transformed dataset or representation learned can then be used in downstream tasks (classification, regression, etc.) without any change to provide "fairer" outcomes. We group approaches to mitigate biases at the data level into three main categories:

- Fair representation learning: learn a fair representation of the data that obfuscates information about the sensitive attributes [185–190].
- Dataset transformation: Modify the training data by relabeling or reweighing data points [191] or apply data augmentation by interpolating samples from different group [192].
- Sampling: find a distribution close to the empirical distribution of the dataset subject to fairness constraints [191].

4.3.2 | In-Processing Techniques

These techniques are used when we have access to the model training, and it is not costly to retrain an existing model. In a nutshell, the loss function is transformed to add a loss/regularization term that penalizes the model's disparities across groups. Therefore, the model is forced to optimize for accuracy and fairness. The classification problem becomes a constrained optimization problem where the goal is to minimize the classification error (maximize the accuracy) while satisfying a given fairness constraint. However, this optimization problem is nonconvex and difficult to enforce. Therefore, existing in-processing techniques are reformulated in different ways or dual problems are solved. They can be grouped as follows:

- **Reduction approach**: The Exponentiated Gradient [193] and AdaFair [194] approaches for fairness transform any binary classification problem into a cost-sensitive classification problem, that can yield a randomized classifier having the lowest error while satisfying fairness constraints.
- **Adversarial-based approach**: Adversarial network is a method that involves two competitive neural networks, commonly used in generative models like GANs (Generative Adversarial Networks) [195]. This method is also applied to mitigate bias. A popular application is Adversarial debiasing [196]. It involves an adversary network that tries to predict the sensitive attribute while the classifier tries to defeat the adversary, thus enforcing the independence of the outcome and sensitive attribute.
- **Regularization-based approach**: Regularization is generally used in ML to prevent overfitting by penalizing the model's weights using L_1 or L_2 norms [197]. A similar technique can be employed to add regularization terms to the loss function to penalize the model for disparities over demographic groups [198–204]. For instance, Kamishima et al. [202] introduced *prejudice remover regularizer*, a regularization term for fairness that minimizes the mutual information between the model's output and the sensitive attributes.
- **Sampling-based approach**: Bias can be mitigated using oversampling [205, 206] or subsampling [207, 208]. For instance, FairBatch [208] is a batch selection process that enforces a given fairness metric by sampling mini-batches to transform the Empirical Risk Minimization (ERM) problem into a weighted ERM to incorporate fairness constraints. In a nutshell, FairBatch modifies the ratio of each demographic group in the minibatch by increasing the representation of the group of samples mostly misclassified (discriminated against) in the previous batch.

4.3.3 | Post-Processing Techniques

This group of methods treats the model as a black box and enforces fairness constraints over the model's output. Most existing post-processing methods consist of post hoc modification of the model's outputs to satisfy a given fairness metric [209–212]. In particular, Hardt et al. [211] formalized an optimization problem over the model's output to derive a classifier that satisfies the fairness constraint while minimizing the classification loss. The derived classifier depends on four parameters that measure the probability of positive outcomes given the current classifier output and the sensitive attribute. The optimization problem is thus defined as a constrained linear optimization problem. When the model output is continuous (a score function), the derived classifier is based on a threshold of each demographic group such that it maximizes the classification loss while satisfying fairness constraints, that is, equal opportunity and equalized odds. Similar methods are proposed in the literature, and they differ mainly in the way the optimization problem is defined.

4.3.4 | Pre-in-Post Process: Where Should We Enforce Fairness?

Each group of fairness-enhancing methods has its pros and cons. There are different settings where they can be applied and settings where their use is more challenging or not possible. We summarize the advantages and disadvantages of each group of methods as follows:

- Pre-processing methods can work with any type of model and machine learning tasks. As the fairness intervention is done at the data level, the downstream task can be of any type, however, it becomes difficult to control the tradeoffs, and the algorithmic bias that might arise in the downstream task is not controlled.
- In-processing methods allow control over the fairness-accuracy tradeoff that the model can achieve. Having access to the optimization problem with fairness constraints provides more flexibility in the tradeoffs; however, there is little flexibility over the type of models used, that is, the constraint optimization is model-specific.
- Similarly to pre-processing techniques, post-processing methods can be applied to any type of model (classifier), which is treated like a black box. The output of the model is modified to satisfy a given fairness metric. However, changing the model's output comes at a significant cost of accuracy. Moreover, these methods can yield unfair outcomes against certain individuals as the model output is changed to satisfy a certain fairness metric.

Overall, as shown by Friedler et al. [213], there is no consensus in the literature about which group of methods performs best. None of the methods consistently outperforms others, and their performances depend on the fairness metric and datasets.

Fairness definitions and fairness-enhancing methods presented above have been mainly applied to classical machine learning setups, where a single model is trained for a specific task. With the emergence of foundation models, new evaluation and mitigation strategies have been proposed to target the new learning paradigm.

4.4 | Fairness and Equity in the Emergence of Foundation Models

Foundation models such as GPT-3.5 [151] are pretrained on massive amounts of data without a specific task in mind, learning various complex patterns that can be adapted to a range of downstream tasks [214]. Specifically, a foundation model fine-tuned on a small, specific task often performs better than a task-specific model trained from scratch. This new paradigm is not free from bias since foundation models can capture social bias during the pretraining stage or task-specific bias during finetuning. However, while new definitions of unfairness and mitigation approaches have emerged when using foundation models, classical definitions and mitigation techniques are either reused or adapted to the new learning paradigm. For example, individual or group fairness metrics presented in the previous subsection can be applied to foundation models, such as LLMs, by quantifying the disparity of the outcome of an algorithm built upon the foundation model. Individual or group fairness metrics naturally transfer to foundation models when used for classification or regression tasks. On the other hand, other unique forms of bias are specific to natural language tasks such as text generation, machine translation and question-answering. This includes: *stereotyping*, which occurs when the model makes assumptions about certain groups due to historical or social bias; *toxicity*, where generated text contains offensive language targeting specific social groups [215]. Several metrics have been proposed to measure this type of bias in language models. Unlike in classification tasks, bias evaluation metrics in language models are task-specific and usually linked to a dataset designed to identify a particular type of bias. For instance, stereotypes in language models can be measured using crafted input text that only differs in the demographic information (e.g., gender, race or religion) and by analyzing variation in the model output. The task submitted to the model could be question-answering or text completion [216, 217]. For example [217] found that when the context information given to the model is under-informative, the produced output reinforces existing social bias by generating stereotyped content with harmful biases. This form of evaluation often reveals *intrinsic bias* coming from model pretraining or *extrinsic bias* from model finetuning [218].

Foundation models are not limited to natural language tasks; they have been successfully extended to other data modalities, including vision [219] and tabular data [220]. In the case of tabular data, rows and columns are serialized into text format and presented to the model as contextual input [220]. Remarkably, language models can achieve high predictive accuracy on test data without any parameter updates, a capability known as *in-context learning* (ICL) [220]. This paradigm allows models to adapt to new tasks by conditioning on a few examples provided in the prompt, enabling rapid generalization with minimal computational overhead [151].

In parallel, foundation models specifically tailored for tabular data are being developed, moving beyond the simple

text-serialization approach [221]. These models are designed to capture the unique structure and statistical properties of tabular data, such as column semantics and row-wise dependencies. Recent advancements show that these specialized foundation models are beginning to outperform traditional tree-based models like XGBoost and LightGBM, particularly in settings with large-scale data or complex feature interactions. Their ability to leverage pretraining and transfer learning offers significant advantages in both predictive performance and generalization.

From a fairness standpoint, existing metrics such as demographic parity, equal opportunity, and disparate impact can still be applied to evaluate model behavior under in-context learning [222]. However, because the model parameters remain fixed during inference, traditional in-preprocessing mitigation approaches are not directly applicable. As a result, new fairness interventions that operate at the prompt level have been proposed. These include demonstration selection [222], where carefully curated in-context examples are chosen to reduce bias, and prompt engineering [223], where the phrasing and structure of the input are optimized to elicit fairer or more accurate responses.

Despite promising results, the fairness implications of in-context learning are still an active area of research. The model's behavior can be highly sensitive to prompt design and input ordering, leading to prediction variability across different demographic groups [151, 224]. Moreover, because prompts are often constructed manually or heuristically, ensuring fairness at scale remains a significant challenge. Future work is needed to develop systematic and automated techniques for fairness-aware prompt generation, as well as robust evaluation frameworks tailored to this new paradigm.

5 | Responsibility, Accountability, and Regulations

As it was demonstrated in previous sections, AI systems are generally complex "black-boxes" sociotechnical systems that can sometimes produce unintended outcomes, which have raised issues of accountability establishment when something goes wrong. Moreover, understanding how new technologies intertwine with the upcoming AI regulations and how they reflect on AI ethics could go a long way. In this section, we delve into other important ethical considerations of AI and the current landscape of AI regulation.

5.1 | Existential Risks of AI

The existential risks of AI are becoming increasingly concerning as advancements in the field continue to evolve [225, 226]. The common approach with new technologies is to implement them first, then address any significant issues that arise, making adjustments as necessary. However, this approach may not be suitable for advanced AI, as early missteps in directing these systems could prevent later adjustments, potentially leading to catastrophic outcomes [227]. Examining history, technological advancement during the Industrial Revolution completely transformed people's lives and brought about profound changes for humanity. This also yielded sociotechnical problems that needed regulations. For instance, the invention of the first gas-powered automobile lacked the rules and technologies we now take for granted, like traffic signs and automated lights. Pedestrians and drivers had to watch out for themselves for safety. During this transformative period, legislators responded reactively by instituting rules to govern personal car usage. Few events during the Industrial Revolution have the same transformative impact as advancements in AI [228].

These advancements in AI could potentially amplify existing catastrophic risks, such as bioterrorism, the spread of disinformation leading to institutional dysfunction, misuse of centralized power, nuclear and conventional warfare, other coordination failures, and unforeseen risks [226]. This necessitates a proactive approach where potential problems are anticipated and resolved well in advance, preserving our ability to make corrections and avoid irreversible consequences.

Shortly after the development of advanced AI, we are likely to encounter AI systems that significantly outperform humans in most cognitively demanding tasks [229]. These include tasks that have a substantial impact on the world, such as technological development, social/political persuasion, and cyber operations [226]. If there is a conflict of goals between advanced AI systems and humans, the AI will likely outperform or outmaneuver humans to prioritize its objectives. Misdirected advanced AI could limit humanity's ability to make corrections. It could determine that its current goals would not be met if humans redirected it toward other objectives or deactivated it. Consequently, it would take steps to prevent such interventions.

5.2 | Accountability

We have seen in previous sections that AI systems can sometimes go wrong; for example, gender bias in the Amazon hiring system, or adversarial environments that mislead a deep neural network model. In general, when something goes wrong in an organization or a company, someone is held accountable for it. However, AI systems cannot solely be responsible for unpredicted outcomes. A question that naturally arises when AI systems provide outcomes that have adverse effects or harm individuals, is who is accountable for the issues. Accountability is a well-studied notion in different fields such as politics and law [230, 231]. It provides moral principles that guide the ethical conduct of people or organizations as they bear the responsibility for their actions. Accountability is, therefore, a key component of trust in society, organizations, and the professional milieu. In the context of AI, accountability is a meta-component of trustworthiness that ensures that ethical principles are promoted and enforced throughout the lifecycle of AI projects [228, 232–234].

5.2.1 | Accountable for What?

According to Virginia Dignum accountability means the decisions of the AI system are explainable and derivable from the decision-making mechanisms used [235] and is a set of components guided by moral values that are part of a large socio-technical system. Millar et al. [236] see accountability as "answerability" for decisions, actions, products, and policies.

From this perspective, it is required that accountability is a practice that the management board of a company could enforce by making all the stakeholders who develop the system understand that they bear the responsibility for the decisions of the system. Therefore, policymakers, data scientists, and AI engineers should carefully choose and justify the choices made during the design, development, and deployment of AI systems [237].

5.2.2 | Accountable to Whom?

The design, development, and deployment of AI systems involve multiple stakeholders. While a system in which decisions can be explained can help to detect the causes or reasons for the incidents, it does not necessarily provide an answer to who is most responsible for the unintended outcomes. Establishing causal accountability remains challenging, and unintended negative outcomes might bear legal liability. Companies or organizations can acknowledge incidents caused by the use of their systems and can pay for reparation. However, monetary compensation could not be enough in society and requires someone to be punished, that is, to take legal responsibility for the potential harm. Ammanath [228] says that instead of looking for people to blame, companies should recognize the legal responsibility of sociotechnical systems around AI and all stakeholders should use it as an "additional motivation to own their individual responsibility."

Promoting accountability throughout the entire lifecycle of AI systems faces the challenge of developing processes and rules to enforce it while not hindering innovation. The enforcement of accountability can also be backed by laws and regulations that define a conformity assessment of the outcomes of sociotechnical systems to hinder potential flaws in the system [228, 236]. Accountability, therefore, becomes a core component of trustworthiness in a socio-technical system. That is, if a system provides discriminatory outcomes, leaks people's sensitive data, or is not robust to different kinds of environments where it is deployed, someone would bear the responsibility of justifying and mitigating it.

5.3 | AI Ethics and Regulation

Understanding how new technologies intertwine with the upcoming AI regulations and how they reflect on AI ethics could go a long way. According to Stanford University's 2023 AI Index [238] in 2022, 127 countries passed around 37 bills with the word "artificial intelligence," which shows a greater interest in regulating AI. In this section, we discuss theoretical considerations of AI and the current landscape of AI regulation.

There are initiatives around the world that focus on the development and promotion of trustworthy AI and the regulations that surround it. One exciting initiative to look at is the National AI Initiative, one of the components of the United States' approach to AI regulation. Looking at the National AI Initiative Act of 2020 [239], it aims to promote and support research and development in trustworthy AI in both the public and private sectors. It also aims to develop technical standards and guidelines that facilitate the evaluation of bias in artificial intelligence training data and applications. The US governmental agencies allocated a budget of $1.7 billion to AI R&D which is an increase of 209% from 2018 [238]. Despite many of these efforts, the United States does not have a comprehensive federal law governing AI. Different states or even problem statements have taken different approaches to AI regulation, with some enacting laws related to specific applications of AI, such as autonomous vehicles, and others focusing on broader issues, such as data privacy and security. For autonomous vehicle regulation in the United States, the National Conference of State Legislatures (NCSL) provides a comprehensive database [240] that tracks autonomous vehicle bills introduced in all 50 states and the District of Columbia. And for broader issues such as data privacy and security, the National Conference of State Legislatures (NCSL) documents state privacy laws in various areas [241].

There has also been regulation around Deepfake and online safety of users [242]. The UK passed the Online Safety Bill, making sharing pornographic deepfakes without consent a crime, and considering laws for clear labeling of AI-generated content [242, 243]. Also, China's Deep Synthesis Provisions, effective from January 10, 2023, regulate deep synthesis technology, mandating data protection, transparency, and content management with other requirements [244].

The European Union (EU) has taken significant steps to regulate AI and protect data. General Data Protection Regulation (GDPR) [245], enacted in May 2018, is a set of rules designed to give EU citizens more control over their personal data. Elements of GDPR, like Data Minimization, can play a huge role in what kind of data AI models can store and use for training. Additionally, the European Commission introduced the Artificial Intelligence Act in April 2021 [246] and is actively working toward establishing comprehensive AI regulations. "The vote sets up the so-called trilogue negotiations, the final phase of the EU's process, and paves the way for the likely adoption of Europe's—and the world's—first comprehensive AI regulatory framework in early 2024" [247].

The regulation and public policy of AI could greatly help make AI safe for all involved; however, it also comes with some challenges. Firstly, new policies may be slow to implement and can get stuck in bureaucratic processes. For example, in Canada, Bill C-27 ("An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act, and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts") [248] had its first reading in the House of Commons in 2022. Parts of the bill, such as Part 3, add common requirements for the design, development, and use of artificial intelligence systems, including measures to mitigate the risks of harm and biased output. However, on 6 January 2025, the bill died on the Order Paper, since parliament was prorogued [249].

In addition to regulatory processes, researchers have also raised questions about the legal duty to find and discover less discriminatory algorithms [250]. Model multiplicity is a phenomenon in which multiple models can achieve equal accuracy on the same task, but differ in their individual predictions or aggregate properties [251]. This means it is possible to have a model with lower disparate impact without affecting the performance. Black et al. [250] in their work argue that service providers should have a legal duty to search and discover these LDAs that show less

disparate impact with reasonable efforts. There are no known regulations around the discovery of LDAs, and the question of who the burden of proof falls upon is also raised in the paper.

5.4 | Generative AI Ownership

For a long time, creativity has been considered one of the most distinguishing attributes of humans. But with the advances in computer science and machine learning, people have tried to add creativity to the machine learning models, and it is where Generative AI appears. Regarding probabilistic modeling, generative models are machine learning models that describe generating a dataset. Sampling from such a model generates new data belonging to the desired distribution [252].

5.4.1 | Who is the Author?

While current samples of generative models can create high-quality products like meaningful text [253], different types of art, including paintings [254], music [255], and executable computer codes, the first generative models that caught the news attention were StyleGAN [256] which created hyper-realistic images from human faces and GPT-2 [257] that could finish a paragraph of meaningful text related to its received opening sentence.

While the number of generative AI users is increasing daily, a critical question about generative AI ownership still needs to be answered. Consider a situation in which an employee who is expected to add some lines of code to their company's product uses generative AI to generate the code and add it to their product. The question that arises here is who owns the product? Does the company lose copyright over its product because a machine learning model generates one part? What about art generated by generative AI? While an artist who won Colorado State Fair's art competition using an AI-generated picture [258] claimed he had not broken any rules, another artist rejected the prize he won at the Sony World Photography award, revealing that AI had generated the winning photo [259]. As a prevention from such events, the Grammys forbids AI-generated works from participating in their competitions [260]. Debates about the ownership of generative AI products remain unsolved, and people taking each side of this debate have satisfying reasons. While from one point of view, the owners of the AI-generated products are people who made the models generate such output by their prompts and guidance, the other perspective gives the ownership of these products to model owners who created these generative models or to the models themselves [261].

5.4.2 | Rules and Regulations

Although the debate on generative AI ownership has not concluded, generative AI is being used more daily. Thus, it needs some regulation to decide who is responsible for and who owns the products generated by AI. Based on Federal Register guideline [262], the work generated by AI, using only one prompt by the human user, is unprotected. At the same time, if it is a product of interaction between the human and machine where a human makes some modification, selection, or arrangement on the AI product, it is protected by copyright law. Regulations in the EU and UK take a different side toward AI-generated product ownership: a product can be subject to copyright if it is "the author's own intellectual creation." According to Copyright Designs and Patents Act 1988 (CDPA) [263], since computer programs cannot be considered as the owner of their product, "the person by whom the arrangements necessary for the creation of the work are undertaken" is known as the author of AI products. As can be seen from different viewpoints of regulations toward generative AI ownership, there has yet to be a finalized committed decision. Some consider AI models as an extension of their human users, and the human owns their products; others take model owners as the creators of the model's products, and the last group considers joint ownership between creators and the model owners. Since the legal landscape around AI ownership is still evolving, a long way exists to achieve commitment.

But what are generative AI's economic and social impacts apart from the ownership of AI products? Based on AI ACT, when products like images, voice, art, etc., are generated by AI and published, it should be exposed that they are AI-generated, especially when they resemble existing persons, places, or events [264]. Otherwise, AI-generated products can easily deceive and manipulate people in various political, social, and cultural domains. Malicious use of AI to defame or create popularity for specific individuals or groups is another important concern regarding the use of Generative AI. In this regard, the AI Act added an account for general-purpose AI that "may lead to discriminatory outcomes and the exclusion of certain groups." Product manufacturers or fashion designers can utilize AI-generated content to promote their products without hiring specialized individuals. In this regard, another issue that arises is the economic impact of using generative AI. Despite the significant concerns raised about generative AI's impact on the future of jobs by enabling enterprises to use AI instead of their human forces, the economic consequences of its use and the possibility of further economic inequality due to AI utilization exist. Since launching and utilizing generative AI requires high-level hardware capabilities and software knowledge, equal access to it is impossible for all countries and all individuals within a society. This can significantly increase social and economic inequality between developing and developed countries and between organizations and individuals with financial and informational resources and those without them. The lack of consistent regulation for generative AI added to the concerns about the harms unregulated AI can bring to society [265] has resulted in an open letter [266] initiated by Future for Life and signed by more than 3000 people in the technology industry, asking for a six-month pause on giant AI experiments by all companies until sufficient policies are made and enforced. On the other hand, since the AI Act is forcing more strict regulations for using generative AI, some European companies claim that it may "jeopardize technological sovereignty" and signed an open letter [267] asking the EU to reconsider its plans to let European companies participate in the advancement of AI.

6 | Environmental Impact

While machine learning algorithms are not inherently bad for the climate and environment, their implementation can have

certain implications that need to be considered [268–270]. One significant factor is the high energy consumption associated with training and running machine learning models, particularly when GPUs (Graphics Processing Units) are used for accelerated computation. From an environmental viewpoint, the following aspects have to be considered.

6.1 | Energy Consumption and Carbon Emissions

Machine learning algorithms, particularly deep learning models, require extensive computational resources for training [268]. This process involves performing numerous complex calculations on large datasets. GPUs are commonly used for their parallel processing capabilities, which accelerate the training process. However, GPUs are power-hungry devices that consume substantial amounts of electricity. The energy consumption of AI algorithms is a significant concern due to the scale at which these algorithms are deployed. The study by Strubell et al. [271] discovered that the greenhouse gas emissions produced from some of the NLP algorithms were comparable to those from 300 flights traveling between New York and San Francisco. Training state-of-the-art models can take weeks or even months, consuming a substantial amount of energy during that time. The energy consumption is directly proportional to the size and complexity of the model and the amount of data being processed. As machine learning applications become more prevalent across industries, the collective energy consumption associated with training and running increases [272].

6.2 | Carbon Emissions

Machine learning algorithms' energy consumption directly impacts carbon emissions [273–275], as most electricity worldwide still comes from non-renewable sources like coal, natural gas, and oil. These fossil fuels emit CO2 and other greenhouse gases when burned for energy, contributing to climate change. Machine learning algorithms, which heavily use GPUs, often draw electricity from grids powered by fossil fuel-based plants. This results in significant carbon emissions from both the direct power consumption and the indirect emissions from fossil fuel extraction, production, and transportation needed to meet energy demands.

6.3 | Electronic Waste and Disposal Challenges

The rapid advancement of machine learning technology necessitates frequent hardware upgrades, including GPUs [276]. As newer and more powerful GPUs are introduced to the market, older models become obsolete. This cycle of hardware replacement results in electronic waste generation. The electronic waste consists of discarded electronic devices, including GPUs, that are no longer in use. Improper handling and disposal of e-waste pose significant environmental risks. Electronic devices contain hazardous materials, such as lead, mercury, cadmium, and brominated flame retardants, which can contaminate the environment if not managed properly. E-waste often ends up in landfills, where toxic substances can leach into the soil and water, posing threats to ecosystems and human health. Managing e-waste from machine learning infrastructure can be particularly challenging due to the rapid obsolescence of hardware and the need for specialized recycling processes to handle complex electronic components [277].

6.4 | Using AI for Mitigating Climate Change

Contrary to the potential harms of climate change due to AI, recent research [278–281] has gained interest in harnessing AI to combat climate change. By leveraging AI's capabilities, climate modeling and prediction, optimization of resource management and energy efficiency, advancement of renewable energy solutions, and enhancement of climate change adaptation and resilience can be improved. The potential of AI to address the pressing issues of climate change offers a promising path toward sustainability and a greener future.

6.4.1 | Improved Climate Modeling and Prediction

AI has revolutionized climate modeling by providing powerful tools to process vast amounts of data and make accurate predictions [282, 283]. Machine learning algorithms can analyze historical climate data, satellite imagery, and environmental variables, unveiling complex climate system dynamics. This enhanced understanding enables us to anticipate extreme weather events, assess the impacts of human activities on the environment, and predict long-term climate trends. AI-driven climate modeling empowers policymakers [278] with valuable insights to develop effective mitigation and adaptation strategies, fostering a more sustainable future.

6.4.2 | Optimized Resource Management and Energy Efficiency

AI has the potential to transform resource management and enhance energy efficiency across various sectors [284, 285]. By employing AI, especially RL algorithms, smart grids can analyze real-time data on energy consumption, demand, and weather conditions [286, 287]. This analysis facilitates efficient energy distribution, minimizes wastage, and encourages the integration of renewable energy sources. Machine learning techniques can optimize energy-intensive processes in industries, transportation systems, and buildings, resulting in significant energy savings.

6.4.3 | Advancements in Renewable Energy Solutions

Renewable energy plays a pivotal role in mitigating climate change, and AI can accelerate its adoption. AI algorithms can optimize the integration and operation of renewable energy systems [288–290] like solar, wind, and hydroelectric power. By analyzing geographical and climatic data, AI can identify optimal locations for solar panels and wind turbines, maximizing energy output. Additionally, AI can aid in developing advanced materials and technologies for efficient energy storage, addressing one of the primary challenges in renewable energy adoption [291]. By harnessing AI, the shift to a renewable energy-powered

world reduces dependence on fossil fuels and mitigates climate change.

In summary, to ensure the responsible use of machine learning algorithms, stakeholders must consider the environmental implications and strive to adopt sustainable practices. This can include optimizing algorithms for efficiency, using energy-efficient hardware, considering renewable energy sources for computing infrastructure, and promoting responsible e-waste management. By balancing technological advancements with environmental considerations, we can potentially harness the benefits of machine learning while minimizing its potential negative impacts on the climate and environment.

7 | Conclusion

Over the past few years, AI ethics has reaped significant attention as the ramifications of AI decisions permeate every layer of society. We provide a review to understand its intricacies and challenges through notions of privacy and data protection, transparency and explainability, fairness and equity, responsibility, accountability, and regulations and lastly the environmental impacts. The notions presented in this paper are key ingredients to building trustworthy AI systems. Nevertheless, there is not yet a checklist that, when fulfilled, ensures trustworthy AI. The integration of these principles depends on the problems and the business needs. Moreover, certain ethical principles can be conflicting or open to compromise. For example, prioritizing privacy in a system might lead to diminished performance and fairness [292, 293], introducing trade-offs among accuracy, fairness, and privacy. Determining which ethical aspect should take precedence in such scenarios remains ambiguous. Regulators are expected to establish a baseline for fairness and privacy, necessitating compliance audits before AI system deployment [294].

Concurrently, constructing fair models introduces privacy concerns related to demographic information [295, 296], or faces challenges when demographic information is unavailable due to privacy constraints [297, 298]. While explainability is crucial for instilling trust, disparities in the quality of explanations [299] and the risk of "washing" [43] persist.

Assessments of advancements in building models robust to attacks predominantly focus on accuracy, though other principles such as fairness are also vulnerable to attacks [300].

Conversely, promoting accountability in companies, with clear consequences for individuals involved in the AI lifecycle when unexpected outcomes arise, could instill fear of professional repercussions and potentially stifle innovation. Therefore, it is imperative to formulate policies that strike a balance between fostering innovation and ensuring accountability [228, 236]. Despite these challenges requiring considerable attention, an expanding body of work is dedicated to addressing them [301–305].

Its interdisciplinary scope and emphasis on actionable insights make it a valuable resource for researchers, developers, and policymakers alike. Still, several open challenges remain, such as reconciling ethical trade-offs, operationalizing fairness without compromising privacy, and developing globally inclusive governance frameworks. Future research should focus on creating standardized evaluation metrics, scalable mitigation techniques, and participatory design processes to ensure responsible and reproducible AI across diverse contexts.

To sum up, while this paper investigates the interdisciplinary scope of AI Ethics from classical machine learning to foundation models, it highlights open challenges, including ethical trade-offs, the need for a global, inclusive regulatory and governance framework, and the environmental impact of AI.

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

Endnotes

[1] Differentially private stochastic gradient descent.

References

1. N. A. Smuha, "The Eu Approach to Ethics Guidelines for Trustworthy Artificial Intelligence," *Computer Law Review International* 20, no. 4 (2019): 97–106.

2. R. Gianni, S. Lehtinen, and M. Nieminen, "Governance of Responsible AI: From Ethical Guidelines to Cooperative Policies," *Frontiers in Computer Science* 4, no. 873 (2022): 437.

3. A. Jobin, M. Ienca, and E. Vayena, "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence* 1, no. 9 (2019): 389–399.

4. L. Floridi, J. Cowls, M. Beltrametti, et al., "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and machines* 28, no. 4 (2018): 689–707.

5. T. Hagendorff, "The Ethics of AI Ethics: An Evaluation of Guidelines," *Minds and Machines* 30, no. 1 (2020): 99–120.

6. W. J. Lewis, J. Van Lenteren, S. C. Phatak, and J. Tumlinson Iii, "A Total System Approach to Sustainable Pest Management," *Proceedings of the National Academy of Sciences* 94, no. 23 (1997): 12243–12248.

7. P. Mooney, P. Culliton, A. Eltaief, et al., "AI Report 2023" (Equal First Authors: Paul Mooney*, Phil Culliton*) (2023), https://www.kaggle.com/AI-Report-2023.

8. J. Jiao, S. Afroogh, Y. Xu, and C. Phillips, "Navigating LLM Ethics: Advancements, Challenges, and Future Directions," *arXiv Preprint* (2024).

9. J. Laine, M. Minkkinen, and M. Mäntymäki, "Ethics-Based AI Auditing: A Systematic Literature Review on Conceptualizations of Ethical Principles and Knowledge Contributions to Stakeholders," *Information & Management* 61, no. 5 (2024): 103969.

10. N. K. Corrêa, C. Galvão, J. W. Santos, et al., "Worldwide AI Ethics: A Review of 200 Guidelines and Recommendations for AI Governance," *Patterns* 4, no. 10 (2023): 100857.

11. E. Prem, "From Ethical AI Frameworks to Tools: A Review of Approaches," *AI and Ethics* 3, no. 3 (2023): 699–716.

12. P. Radanliev, O. Santos, A. Brandon-Jones, and A. Joinson, "Ethics and Responsible AI Deployment," *Frontiers in Artificial Intelligence* 7 (2024): 1377011.

13. A. A. Khan, S. Badshah, P. Liang, et al., "Ethics of AI: A Systematic Literature Review of Principles and Challenges," in *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering* (IEEE, 2022), 383–392.

14. K. R. Varshney, "Trustworthy Machine Learning and Artificial Intelligence," *XRDS: Crossroads, the ACM Magazine for Students* 25, no. 3 (2019): 26–29.

15. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *arXiv Preprint* (2016).

16. H. Baniecki and P. Biecek, "Adversarial Attacks and Defenses in Explainable Artificial Intelligence: A Survey," *Information Fusion* 102 (2024): 303.

17. A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, "Adversarial Attacks and Defences: A Survey," *arXiv Preprint* (2018).

18. M. Fredrikson, S. Jha, and T. Ristenpart, "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2015), 1322–1333.

19. B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep Models Under the Gan: Information Leakage From Collaborative Deep Learning," in *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2017), 603–618.

20. T. T. Nguyen, T. T. Huynh, Z. Ren, et al., "A Survey of Privacy-Preserving Model Explanations: Privacy Risks, Attacks, and Countermeasures," *arXiv Preprint* (2024).

21. N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical Black-Box Attacks Against Machine Learning," in *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (ACM, 2017), 506–519.

22. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing Machine Learning Models via Prediction Apis," in *USENIX Security Symposium* (ACM, 2016), 601–618.

23. M. Xue, C. Yuan, H. Wu, Y. Zhang, and W. Liu, "Machine Learning Security: Threats, Countermeasures, and Evaluations," *IEEE Access* 8 (2020): 74720–74742.

24. M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The Security of Machine Learning," *Machine Learning* 81 (2010): 121–148.

25. N. Pitropakis, E. Panaousis, T. Giannetsos, E. Anastasiadis, and G. Loukas, "A Taxonomy and Survey of Attacks Against Machine Learning," *Computer Science Review* 34 (2019): 100199.

26. T. Baluta, S. Shen, S. Hitarth, S. Tople, and P. Saxena, "Membership Inference Attacks and Generalization: A Causal Perspective," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2022), 249–262.

27. H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, and X. Zhang, "Membership Inference Attacks on Machine Learning: A Survey," *ACM Computing Surveys (CSUR)* 54, no. 11 (2022): 1–37.

28. M. Rigaki and S. Garcia, "A Survey of Privacy Attacks in Machine Learning," *ACM Computing Surveys* 56, no. 4 (2020): 1–34.

29. R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," in *2017 IEEE Symposium on Security and Privacy (SP)* (IEEE, 2017), 3–18.

30. P. Quan, S. Chakraborty, J. V. Jeyakumar, and M. Srivastava, "On the Amplification of Security and Privacy Risks by Post Hoc Explanations in Machine Learning Models," *arXiv Preprint* (2022).

31. Z. He, T. Zhang, and R. B. Lee, "Model Inversion Attacks Against Collaborative Inference," in *Proceedings of the 35th Annual Computer Security Applications Conference* (ACM, 2019), 148–162.

32. M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing," in *23rd {USENIX} Security Symposium ({USENIX} Security 14)* (USENIX, 2014), 17–32.

33. G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking Smart Machines With Smarter Ones: How to Extract Meaningful Data From Machine Learning Classifiers," *International Journal of Security and Networks* 10, no. 3 (2015): 137–150.

34. L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting Unintended Feature Leakage in Collaborative Learning," in *2019 IEEE Symposium on Security and Privacy (SP)* (IEEE, 2019), 691–706.

35. N. Dalvi, P. Domingos, S. Sanghai, and D. Verma, "Adversarial Classification," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2004), 99–108.

36. B. Biggio, B. Nelson, and P. Laskov, "Poisoning Attacks Against Support Vector Machines," *arXiv Preprint* (2012).

37. H. Zhang, J. Gao, and L. Su, "Data Poisoning Attacks Against Outcome Interpretations of Predictive Models," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (ACM, 2021), 2165–2173.

38. B. Biggio, I. Corona, D. Maiorca, et al., "Evasion Attacks Against Machine Learning at Test Time," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23–27, 2013, Proceedings, Part III 13* (Springer, 2013), 387–402.

39. Y. Deng and L. J. Karam, "Universal Adversarial Attack via Enhanced Projected Gradient Descent," in *2020 IEEE International Conference on Image Processing (ICIP)* (IEEE, 2020), 1241–1245.

40. H. Xu, Y. Ma, H.-C. Liu, et al., "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review," *International Journal of Automation and Computing* 17 (2020): 151–178.

41. H. Baniecki and P. Biecek, "Manipulating Shap via Adversarial Data Perturbations (Student Abstract)," *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (2022): 12907–12908.

42. A. Ghorbani, A. Abid, and J. Zou, "Interpretation of Neural Networks is Fragile," *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (2019): 3681–3688.

43. U. Aïvodji, H. Arai, O. Fortineau, S. Gambs, S. Hara, and A. Tapp, "Fairwashing: The Risk of Rationalization," in *International Conference on Machine Learning* (PMLR, 2019), 161–170.

44. C. Anders, P. Pasliev, A.-K. Dombrowski, K.-R. Müller, and P. Kessel, "Fairwashing Explanations With Off-Manifold Detergent," in *International Conference on Machine Learning* (PMLR, 2020), 314–323.

45. K. Fukuchi, S. Hara, and T. Maehara, "Faking Fairness via Stealthily Biased Sampling," *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020): 412–419.

46. L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, no. 5 (2002): 571–588.

47. C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," *Foundations and Trends in Theoretical Computer Science* 9, no. 3–4 (2014): 211–407.

48. J. Zhang, Z. Zhang, X. Xiao, Y. Yang, and M. Winslett, "Functional Mechanism: Regression Analysis Under Differential Privacy," *arXiv Preprint* (2012).

49. N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable Private Learning With Pate," *arXiv Preprint* (2018).

50. M. Abadi, A. Chu, I. Goodfellow, et al., "Deep Learning With Differential Privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2016), 308–318.

51. F. McSherry and K. Talwar, "Mechanism Design via Differential Privacy," in *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* (IEEE, 2007), 94–103.

52. I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," *arXiv Preprint* (2014).

53. Y. Zhang, W. Song, Z. Ji, et al., "How Well Does Llm Generate Security Tests?," *arXiv Preprint* (2023).

54. M. L. Siddiq and J. C. Santos, "Generate and Pray: Using Sallms to Evaluate the Security of LLM Generated Code," *arXiv Preprint* (2023).

55. C. Chen, J. Su, J. Chen, et al., "When Chatgpt Meets Smart Contract Vulnerability Detection: How Far Are We?," *ACM Transactions on Software Engineering and Methodology* 34, no. 4 (2025): 1–30.

56. S. Hu, T. Huang, F. Ilhan, S. F. Tekin, and L. Liu, "Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives," in *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)* (IEEE, 2023), 297–306.

57. F. Wang, *Using Large Language Models to Mitigate Ransomware Threats* (Akamai Technologies, 2023).

58. A. Vats, Z. Liu, P. Su, et al., "Recovering From Privacy-Preserving Masking With Large Language Models," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), 10771–10775.

59. F. Yaman, *Agent SCA: Advanced Physical Side Channel Analysis Agent With LLMs* (North Carolina State University, 2023).

60. A. Happe and J. Cito, "Getting PWN'D by AI: Penetration Testing With Large Language Models," in *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ACM, 2023), 2082–2086.

61. M. Beckerich, L. Plein, and S. Coronado, "Ratgpt: Turning Online LLMs Into Proxies for Malware Attacks," *arXiv Preprint* (2023).

62. M. Chowdhury, N. Rifat, S. Latif, M. Ahsan, M. S. Rahman, and R. Gomes, "Chatgpt: The Curious Case of Attack Vectors' Supply Chain Management Improvement," in *2023 IEEE International Conference on Electro Information Technology (eIT)* (IEEE, 2023), 499–504.

63. C. Chen and K. Shu, "Can Llm-Generated Misinformation Be Detected?," *arXiv Preprint* (2023).

64. R. Staab, M. Vero, M. Balunović, and M. Vechev, "Beyond Memorization: Violating Privacy via Inference With Large Language Models," *arXiv Preprint* (2023).

65. P. V. Falade, "Decoding the Threat Landscape: Chatgpt, Fraudgpt, and Wormgpt in Social Engineering Attacks," *arXiv Preprint* (2023).

66. J. Rando and F. Tramèr, "Universal Jailbreak Backdoors From Poisoned Human Feedback," *arXiv Preprint* (2023).

67. A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning Language Models During Instruction Tuning," in *International Conference on Machine Learning* (PMLR, 2023), 35413–35425.

68. H. Yao, J. Lou, and Z. Qin, "Poisonprompt: Backdoor Attack on Prompt-Based Large Language Models," in *ICASSP 2024–2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2024), 7745–7749.

69. H. Zhong, J. Chang, Z. Yang, et al., "Copyright Protection and Accountability of Generative AI: Attack, Watermarking and Attribution," in *Companion Proceedings of the ACM Web Conference* (ACM, 2023), 94–98.

70. P. Zhu, T. Takahashi, and H. Kataoka, "Watermark-Embedded Adversarial Examples for Copyright Protection Against Diffusion Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2024), 24420–24430.

71. M.-A. Panaitescu-Liess, Z. Che, B. An, et al., "Can Watermarking Large Language Models Prevent Copyrighted Text Generation and Hide Training Data?," *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (2025): 25002–25009.

72. J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu, "Are Diffusion Models Vulnerable to Membership Inference Attacks?," in *International Conference on Machine Learning* (PMLR, 2023), 8717–8730.

73. W. Fu, H. Wang, C. Gao, G. Liu, Y. Li, and T. Jiang, "Membership Inference Attacks Against Fine-Tuned Large Language Models via Self-Prompt Calibration," in *The Thirty-Eighth Annual Conference on Neural Information Processing Systems* (ACM, 2024).

74. D. Huang, J. M. Zhang, Q. Bu, X. Xie, J. Chen, and H. Cui, "Bias Testing and Mitigation in llm-Based Code Generation," in *ACM Transactions on Software Engineering and Methodology* (ACM, 2024).

75. Z. Talat, A. Névéol, S. Biderman, et al., "You Reap What You Sow: On the Challenges of Bias Evaluation Under Multilingual Settings," in *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models* (ACM, 2022), 26–41.

76. S. Urchs, V. Thurner, M. Aßenmacher, C. Heumann, and S. Thiemichen, "How Prevalent is Gender Bias in Chatgpt?—Exploring German and English Chatgpt Responses," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (Springer, 2023), 293–309.

77. H. Li, D. Guo, W. Fan, et al., "Multi-Step Jailbreaking Privacy Attacks on Chatgpt," *arXiv Preprint* (2023).

78. Z. Wei, Y. Wang, A. Li, Y. Mo, and Y. Wang, "Jailbreak and Guard Aligned Language Models With Only Few In-Context Demonstrations," *arXiv Preprint* (2023).

79. D. Kang, X. Li, I. Stoica, C. Guestrin, M. Zaharia, and T. Hashimoto, "Exploiting Programmatic Behavior of LLMs: Dual-Use Through Standard Security Attacks," in *2024 IEEE Security and Privacy Workshops (SPW)* (IEEE, 2024), 132–143.

80. Z. Wang, W. Xie, K. Chen, B. Wang, Z. Gui, and E. Wang, "Self-Deception: Reverse Penetrating the Semantic Firewall of Large Language Models," *arXiv Preprint* (2023).

81. E. Derner, K. Batistič, J. Zahálka, and R. Babuška, "A Security Risk Taxonomy for Prompt-Based Interaction With Large Language Models," *IEEE Access* 12 (2024): 126176–126187.

82. K. Gao, J. Gu, Y. Bai, et al., "Energy-Latency Manipulation of Multi-Modal Large Language Models via Verbose Samples," *arXiv Preprint* (2024).

83. T. Liu, Z. Deng, G. Meng, Y. Li, and K. Chen, "Demystifying RCE Vulnerabilities in LLM-Integrated Apps," in *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (ACM, 2024), 1716–1730.

84. OWASP, *Owasp Top 10 for LLMs Should Be Mentioned: Owasp (Aug. 2023). Owasp Top 10 for LLM* (OWASP, 2023), https://owasp.org/www-project-top-10-for-large-language-modelapplications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v1_0.pdf.

85. N. Meade, E. Poole-Dayan, and S. Reddy, "An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models," *arXiv Preprint* (2021).

86. N. Subramani, S. Luccioni, J. Dodge, and M. Mitchell, "Detecting Personal Information in Training Corpora: An Analysis," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)* (ACM, 2023), 208–220.

87. V. Logacheva, D. Dementieva, S. Ustyantsev, et al., "Paradetox: Detoxification With Parallel Data," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACM, 2022), 6804–6818.

88. D. Wang, C. Gong, and Q. Liu, "Improving Neural Language Modeling via Adversarial Training," in *International Conference on Machine Learning* (PMLR, 2019), 6555–6565.

89. H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "Smart: Robust and Efficient Fine-Tuning for Pre-Trained Natural Language Models Through Principled Regularized Optimization," *arXiv Preprint* (2019).

90. B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, and W. Dou, "Adversarial Attacks on Large Language Model-Based System and Mitigating Strategies: A Case Study on Chatgpt," *Security and Communication Networks* 2023, no. 1 (2023): 8691095.

91. L. Xu, L. Berti-Equille, A. Cuesta-Infante, and K. Veeramachaneni, "In Situ Augmentation for Defending Against Adversarial Attacks on Text Classifiers," in *International Conference on Neural Information Processing* (Springer, 2022), 485–496.

92. X. Sun, X. Li, Y. Meng, et al., "Defending Against Backdoor Attacks in Natural Language Generation," *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023): 5257–5265.

93. Z. Wang, Z. Liu, X. Zheng, Q. Su, and J. Wang, "Rmlm: A Flexible Defense Framework for Proactively Mitigating Word-Level Adversarial Attacks," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (ACM, 2023), 2757–2774.

94. A. Helbling, M. Phute, M. Hull, and D. H. Chau, "Llm Self Defense: By Self Examination, LLMs Know They Are Being Tricked," *arXiv Preprints* (2023).

95. M. Xiong, Z. Hu, X. Lu, et al., "Can LLMS Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMS," *arXiv Preprint* (2023).

96. C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence* 1, no. 5 (2019): 206–215.

97. J. A. McDermid, Y. Jia, Z. Porter, and I. Habli, "Artificial Intelligence Explainability: The Technical and Ethical Dimensions," *Philosophical Transactions of the Royal Society A* 379, no. 2207 (2021): 20200363.

98. A. Binder, M. Bockmayr, M. Hägele, et al., "Morphological and Molecular Breast Cancer Profiling Through Explainable Machine Learning," *Nature Machine Intelligence* 3, no. 4 (2021): 355–366.

99. K. Kobylińska, T. Orłowski, M. Adamek, and P. Biecek, "Explainable Machine Learning for Lung Cancer Screening Models," *Applied Sciences* 12, no. 4 (2022): 1926.

100. J. Adeoye, L.-W. Zheng, P. Thomson, S.-W. Choi, and Y.-X. Su, "Explainable Ensemble Learning Model Improves Identification of Candidates for Oral Cancer Screening," *Oral Oncology* 136, no. 106 (2023): 278.

101. M. Idrees and A. Sohail, "Explainable Machine Learning of the Breast Cancer Staging for Designing Smart Biomarker Sensors," *Sensors International* 3, no. 100 (2022): 202.

102. M. Sobhan and A. M. Mondal, "Explainable Machine Learning to Identify Patient-Specific Biomarkers for Lung Cancer," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (IEEE, 2022), 3152–3159.

103. P. Bracke, A. Datta, C. Jung, and S. Sen, "Machine Learning Explainability in Finance: An Application to Default Risk Analysis," Bank of England Working Paper (2019).

104. A. G. Hoepner, D. McMillan, A. Vivian, and C. Wese Simen, "Significance, Relevance and Explainability in the Machine Learning Age: An Econometrics and Financial Data Science Perspective," *European Journal of Finance* 27, no. 1–2 (2021): 1–7.

105. W. Li, "Transparency and Explainability in Financial Data Science" (PhD thesis, NTNU) (2021).

106. M. Rizinski, H. Peshov, K. Mishev, L. T. Chitkushev, I. Vodenska, and D. Trajanov, "Ethically Responsible Machine Learning in Fintech," *IEEE Access* 10 (2022): 97531–97554.

107. M. Espinosa Zarlenga, P. Barbiero, G. Ciravegna, et al., "Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off," *Advances in Neural Information Processing Systems* 35 (2022): 21400–21413.

108. N. Ameen, A. Tarhini, A. Reppel, and A. Anand, "Customer Experiences in the Age of Artificial Intelligence," *Computers in Human Behavior* 114, no. 106 (2021): 548.

109. Y. Xu, C.-H. Shieh, P. van Esch, and I.-L. Ling, "AI Customer Service: Task Complexity, Problem-Solving Ability, and Usage Intention," *Australasian Marketing Journal* 28, no. 4 (2020): 189–199.

110. C. Maree and C. W. Omlin, "Can Interpretable Reinforcement Learning Manage Prosperity Your Way?," *AI* 3, no. 2 (2022): 526–537.

111. R. Sahal, S. H. Alsamhi, and K. N. Brown, "Personal Digital Twin: A Close Look Into the Present and a Step Towards the Future of Personalised Healthcare Industry," *Sensors* 22, no. 15 (2022): 5918.

112. D. Silver, A. Huang, C. J. Maddison, et al., "Mastering the Game of Go With Deep Neural Networks and Tree Search," *Nature* 529, no. 7587 (2016): 484–489.

113. S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the Gaps: Assuring the Safety of Autonomous Systems From an Engineering, Ethical, and Legal Perspective," *Artificial Intelligence* 279, no. 103 (2020): 201, https://doi.org/10.1016/j.artint.2019.103201.

114. R. Hoffman, G. Klein, S. T. Mueller, M. Jalaeian, and C. Tate, "The Stakeholder Playbook for Explaining AI Systems," PsyArXiv Preprint (2021), https://osf.io/9pqez.

115. A. Mamalakis, E. A. Barnes, and I. Ebert-Uphoff, "Investigating the Fidelity of Explainable Artificial Intelligence Methods for Applications of Convolutional Neural Networks in Geoscience," *Artificial Intelligence for the Earth Systems* 1, no. 4 (2022): e220012.

116. A. Papenmeier, G. Englebienne, and C. Seifert, "The Role of Causality in Explainable Artificial Intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15, no. 2 (2019): e70015.

117. V. Pillai and H. Pirsiavash, "Explainable Models With Consistent Interpretations," *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (2021): 2431–2439.

118. G. Carloni, A. Berti, and S. Colantonio, "The Role of Causality in Explainable Artificial Intelligence," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 15, no. 2 (2019): e70015.

119. S. Anjomshoae, K. Främling, and A. Najjar, "Explanations of Black-Box Model Predictions by Contextual Importance and Utility," in *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: First International Workshop, EXTRAAMAS 2019, Montreal, QC, Canada, May 13–14, 2019, Revised Selected Papers 1* (Springer, 2019), 95–109.

120. Q. V. Liao and K. R. Varshney, "Human-Centered Explainable AI (xAI): From Algorithms to User Experiences," *arXiv Preprint* (2021).

121. F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable Artificial Intelligence: A Survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (IEEE, 2018), 210–215.

122. T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2 (Springer, 2009).

123. C. Modarres, M. Ibrahim, M. Louie, and J. Paisley, "Towards Explainable Deep Learning for Credit Lending: A Case Study," *arXiv Preprint* (2018).

124. Y. Yang and M. Wu, "Explainable Machine Learning for Improving Logistic Regression Models," in *2021 IEEE 19th International Conference on Industrial Informatics (INDIN)* (IEEE, 2021), 1–6.

125. R. C. Holte, "Very Simple Classification Rules Perform Well on Most Commonly Used Datasets," *Machine Learning* 11 (1993): 63–90.

126. S. Odense and A. d'Avila Garcez, "Extracting m of n Rules From Restricted Boltzmann Machines," in *Artificial Neural Networks and Machine Learning–ICANN 2017: 26th International Conference on Artificial Neural Networks, Alghero, Italy, September 11–14, 2017, Proceedings, Part II 26* (Springer International Publishing, 2017), 120–127.

127. Z.-H. Zhou, Y. Jiang, and S.-F. Chen, "Extracting Symbolic Rules From Trained Neural Network Ensembles," *AI Communications* 16, no. 1 (2003): 3–15.

128. J. H. Friedman and B. E. Popescu, "Predictive Learning via Rule Ensembles," *Annals of Applied Statistics* 2 (2008): 916–954.

129. A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation," *Journal of Computational and Graphical Statistics* 24, no. 1 (2015): 44–65.

130. B. Carter, J. Mueller, S. Jain, and D. Gifford, "What Made You Do This? Understanding Black-Box Decisions With Sufficient Input Subsets," in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), 567–576.

131. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems* 30 (2017): 4768–4777.

132. K. Simonyan, A. Vedaldi, and A. Zisserman, "Visualising Image Classification Models and Saliency Maps," *Deep Inside Convolutional Networks* 2 (2014).

133. M. Sundararajan and A. Najmi, "The Many Shapley Values for Model Explanation," in *International Conference on Machine Learning* (PMLR, 2020), 9269–9278.

134. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?' Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016), 1135–1144.

135. M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-Precision Model-Agnostic Explanations," in *Proceedings of the AAAI Conference on Artificial Intelligence* (IEEE, 2018), 32.

136. R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local Rule-Based Explanations of Black Box Decision Systems," *arXiv Preprint* (2018).

137. R. K. Mothilal, A. Sharma, and C. Tan, "Explaining Machine Learning Classifiers Through Diverse Counterfactual Explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (ACM, 2020), 607–617.

138. K. Sokol and P. A. Flach, "Counterfactual Explanations of Machine Learning Predictions: Opportunities and Challenges for AI Safety," *SafeAI@ AAAI* (2019): 1–4.

139. S. Verma, J. Dickerson, and K. Hines, "Counterfactual Explanations for Machine Learning: Challenges Revisited," *arXiv Preprint* (2021).

140. B. Kim, M. Wattenberg, J. Gilmer, et al., "Interpretability Beyond Feature Attribution: Quantitative Testing With Concept Activation Vectors (TCAV)," in *International Conference on Machine Learning* (PMLR, 2018), 2668–2677.

141. C.-K. Yeh, B. Kim, S. Arik, C.-L. Li, T. Pfister, and P. Ravikumar, "On Completeness-Aware Concept-Based Explanations in Deep Neural Networks," *Advances in Neural Information Processing Systems* 33 (2020): 20554–20565.

142. B. Zhou, Y. Sun, D. Bau, and A. Torralba, "Interpretable Basis Decomposition for Visual Explanation," in *Proceedings of the European Conference on Computer Vision (ECCV)* (ACM, 2018), 119–134.

143. N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *2009 IEEE 12th International Conference on Computer Vision* (IEEE, 2009), 365–372.

144. C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to Detect Unseen Object Classes by Between-Class Attribute Transfer," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), 951–958.

145. D. Alvarez-Melis and T. S. Jaakkola, "Self-Explaining Neural Networks," SENN.pdf (2018).

146. A. Sarkar, D. Vijaykeerthy, A. Sarkar, and V. N. Balasubramanian, "A Framework for Learning Ante-Hoc Explainable Models via Concepts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2022), 10286–10295.

147. Z. Chen, Y. Bei, and C. Rudin, "Concept Whitening for Interpretable Image Recognition," *Nature Machine Intelligence* 2 (2020): 772–782.

148. P. W. Koh, T. Nguyen, Y. S. Tang, et al., "Concept Bottleneck Models," in *International Conference on Machine Learning* (PMLR, 2020), 5338–5348.

149. I. Sheth, A. A. Rahman, L. R. Sevyeri, M. Havaei, and S. E. Kahou, "Learning From Uncertain Concepts via Test Time Interventions," in *Workshop on Trustworthy and Socially Responsible Machine Learning* (NeurIPS, 2022).

150. S. Shin, Y. Jo, S. Ahn, and N. Lee, "A Closer Look at the Intervention Procedure of Concept Bottleneck Models," in *International Conference on Machine Learning* (PMLR, 2023), 31504–31520.

151. T. Brown, B. Mann, N. Ryder, et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems* 33 (2020): 1877–1901.

152. S. Bubeck, V. Chandrasekaran, R. Eldan, et al., "Sparks of Artificial General Intelligence: Early Experiments With Gpt-4," *arXiv Preprint* (2023).

153. J. D. M.-W. Kenton and L. K. Toutanova, "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of naacL-HLT* (ACL Anthology, 2019), 2.

154. H. Touvron, L. Martin, K. Stone, et al., "Llama 2: Open Foundation and Fine-Tuned Chat Models," *arXiv Preprint* (2023).

155. B. Peng, M. Galley, P. He, et al., "Check Your Facts and Try Again: Improving Large Language Models With External Knowledge and Automated Feedback," *arXiv Preprint* (2023).

156. C. Singh, A. R. Hsu, R. Antonello, et al., "Explaining Black Box Text Modules in Natural Language With Language Models," *arXiv Preprint* (2023).

157. O. Cfka and A. Liutkus, "Black-Box Language Model Explanation by Context Length Probing," *arXiv Preprint* (2023).

158. E. Hernandez, S. Schwettmann, D. Bau, T. Bagashvili, A. Torralba, and J. Andreas, "Natural Language Descriptions of Deep Visual Features," in *International Conference on Learning Representations* (MIT, 2021), https://arxiv.org/abs/2201.11114.

159. S. Bills, N. Cammarata, D. Mossing, et al., "Language Models Can Explain Neurons in Language Models," (2023), https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

160. C. Singh, J. X. Morris, J. Aneja, A. M. Rush, and J. Gao, "Explaining Patterns in Data With Language Models via Interpretable Autoprompting," *arXiv Preprint* (2022), https://doi.org/10.48550/arXiv.2210.01848.

161. R. Zhong, C. Snell, D. Klein, and J. Steinhardt, "Describing Differences Between Text Distributions With Natural Language," in *International Conference on Machine Learning* (PMLR, 2022), 27099–27116.

162. H. Zhao, H. Chen, F. Yang, et al., "Explainability for Large Language Models: A Survey," *ACM Transactions on Intelligent Systems and Technology* 15, no. 2 (2024): 1–38.

163. L. Bereska and E. Gavves, "Mechanistic Interpretability for AI Safety – A Review," *arXiv Preprint* (2024).

164. A. Zou, L. Phan, S. Chen, et al., "Representation Engineering: A Top-Down Approach to AI Transparency," *arXiv Preprint* (2023).

165. S. Marks and M. Tegmark, "The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets," *arXiv Preprint* (2023).

166. K. Park, Y. J. Choe, and V. Veitch, "The Linear Representation Hypothesis and the Geometry of Large Language Models," *arXiv Preprint* (2023).

167. J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias Risk Assessments in Criminal Sentencing," *ProPublica* (2016).

168. D. Jeffrey, "Amazon Scraps Secret AI Recruiting Tool That Showed Bias Against Women," (2021), https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

169. G. Creamer, G. Kazantsev, and T. Aste, *Machine Learning and AI in Finance* (Routledge, 2021), https://arxiv.org/ftp/arxiv/papers/2111/2111.11295.pdf.

170. V. Laurim, S. Arpaci, B. Prommegger, and H. Krcmar, "Computer, Whom Should I Hire? — Acceptance Criteria for Artificial Intelligence in the Recruitment Process," in *Hawaii International Conference on System Sciences* (University of Hawaii, 2021), https://doi.org/10.24251/HICSS.2021.668.

171. M. Y. Shaheen, "AI in Healthcare: Medical and Socio-Economic Benefits and Challenges," Preprint (2021), https://doi.org/10.14293/S2199-1006.1.SOR-.PPRQNI1.v1.

172. N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A Survey on Bias and Fairness in Machine Learning," *ACM computing surveys (CSUR)* 54, no. 6 (2021): 1–35.

173. D. Pessach and E. Shmueli, "A Review on Fairness in Machine Learning," *ACM Computing Surveys* 55, no. 3 (2022): 44, https://doi.org/10.1145/3494672.

174. R. Wilms, E. Mäthner, L. Winnen, and R. Lanwehr, "Omitted Variable Bias: A Threat to Estimating Causal Relationships," *Methods in Psychology* 5, no. 100 (2021): 75, https://doi.org/10.1016/j.metip.2021.100075.

175. N. Shahbazi, Y. Lin, A. Asudeh, and H. Jagadish, "Representation Bias in Data: A Survey on Identification and Resolution Techniques," *ACM Computing Surveys* 55 (2023): 1–39, https://doi.org/10.1145/3588433.

176. H. Suresh and J. Guttag, "A Framework for Understanding Sources of Harm Throughout the Machine Learning Life Cycle," in *Equity and Access in Algorithms, Mechanisms, and Optimization* (ACM, 2021), 1–9.

177. G. Alves, F. Bernier, M. Couceiro, K. Makhlouf, C. Palamidessi, and S. Zhioua, "Survey on Fairness Notions and Related Tensions," *EURO Journal on Decision Processes* 100 (2023): 33, https://doi.org/10.1016/j.ejdp.2023.100033.

178. H. Weerts, M. Dudík, R. Edgar, A. Jalali, R. Lutz, and M. Madaio, "Fairlearn: Assessing and Improving Fairness of AI Systems," *Journal of Machine Learning Research* 24, no. 257 (2023): 1–8.

179. T. P. Pagano, R. B. Loureiro, F. V. N. Lisboa, et al., "Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods," *Big Data and Cognitive Computing* 7, no. 1 (2023): 15, https://doi.org/10.3390/bdcc7010015.

180. P. Birzhandi and Y.-S. Cho, "Application of Fairness to Healthcare, Organizational Justice, and Finance: A Survey," *Expert Systems With Applications* 216, no. 119 (2023): 465, https://doi.org/10.1016/j.eswa.2022.119465.

181. Y. Li, Y. Ge, and Y. Zhang, "Tutorial on Fairness of Machine Learning in Recommender Systems," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM, 2021), 2654–2657.

182. S. Barocas, M. Hardt, and A. Narayanan, *Fairness and Machine Learning* (Fairness and Machine Learning, 2019), http://www.fairmlbook.org.

183. S. Corbett-Davies, J. D. Gaebler, H. Nilforoshan, R. Shroff, and S. Goel, "The Measure and Mismeasure of Fairness," *Journal of Machine Learning Research* 24, no. 312 (2023): 1–117.

184. M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual Fairness," *Advances in Neural Information Processing Systems* 30 (2017).

185. A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data Decisions and Theoretical Implications When Adversarially Learning Fair Representations," *arXiv Preprint* (2017).

186. P. J. Kenfack, A. R. Rivera, A. M. Khan, and M. Mazzara, "Learning Fair Representations Through Uniformly Distributed Sensitive Attributes," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (IEEE, 2023), 58–67.

187. D. Lowd and C. Meek, "Adversarial Learning," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (ACM, 2005), 641–647.

188. D. Madras, E. Creager, T. Pitassi, and R. Zemel, "Learning Adversarially Fair and Transferable Representations," in *International Conference on Machine Learning* (PMLR, 2018), 3384–3393.

189. J. Song, P. Kalluri, A. Grover, S. Zhao, and S. Ermon, "Learning Controllable Fair Representations," in *The 22nd International Conference on Artificial Intelligence and Statistics* (PMLR, 2019), 2164–2173.

190. R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *International Conference on Machine Learning* (PMLR, 2013), 325–333.

191. F. Kamiran and T. Calders, "Data Preprocessing Techniques for Classification Without Discrimination," *Knowledge and Information Systems* 33, no. 1 (2012): 1–33.

192. C.-Y. Chuang and Y. Mroueh, "Fair Mixup: Fairness via Interpolation," *arXiv Preprint* (2021).

193. A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," in *International Conference on Machine Learning* (PMLR, 2018), 60–69.

194. V. Iosifidis and E. Ntoutsi, "Adafair: Cumulative Fairness Adaptive Boosting," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (ACM, 2019), 781–790.

195. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., "Generative Adversarial Nets," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS 2014)*, vol. 2 (MIT Press, 2014), 2672–2680.

196. B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases With Adversarial Learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (ACM, 2018), 335–340.

197. Y. Tian and Y. Zhang, "A Comprehensive Survey on Regularization Strategies in Machine Learning," *Information Fusion* 80 (2022): 146–166.

198. Y. Bechavod and K. Ligett, "Learning Fair Classifiers: A Regularization-Inspired Approach," *arXiv Preprint* (2017): 1733–1782.

199. Y. Bechavod and K. Ligett, "Penalizing Unfairness in Binary Classification," *arXiv Preprint* (2017).

200. R. Berk, H. Heidari, S. Jabbari, et al., "A Convex Framework for Fair Regression," *arXiv Preprint* (2017).

201. T. Calders and S. Verwer, "Three Naive Bayes Approaches for Discrimination-Free Classification," *Data Mining and Knowledge Discovery* 21 (2010): 277–292.

202. T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-Aware Learning Through Regularization Approach," in *2011 IEEE 11th International Conference on Data Mining Workshops* (IEEE, 2011), 643–650.

203. B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning Non-Discriminatory Predictors," in *Conference on Learning Theory* (PMLR, 2017), 1920–1953.

204. M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Artificial Intelligence and Statistics* (PMLR, 2017), 962–970.

205. S. Rančić, S. Radovanović, and B. Delibašić, "Investigating Oversampling Techniques for Fair Machine Learning Models," in *Decision Support Systems XI: Decision Support Systems, Analytics and Technologies in Response to Global Crisis Management: 7th International Conference on Decision Support System Technology, ICDSST 2021, Loughborough, UK, May 26–28, 2021* (Springer International Publishing, 2021), 110–123.

206. S. Yan, H.-T. Kao, and E. Ferrara, "Fair Class Balancing: Enhancing Model Fairness Without Observing Sensitive Attributes," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (ACM, 2020), 1715–1724.

207. L. E. Celis, A. Deshpande, T. Kathuria, and N. K. Vishnoi, "How to Be Fair and Diverse?," *arXiv Preprint* (2016).

208. Y. Roh, K. Lee, S. E. Whang, and C. Suh, "Fairbatch: Batch Selection for Model Fairness," *arXiv Preprint* (2020).

209. S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic Decision Making and the Cost of Fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2017), 797–806.

210. C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson, "Decoupled Classifiers for Group-Fair and Efficient Machine Learning," in *Conference on Fairness, Accountability and Transparency* (PMLR, 2018), 119–133.

211. M. Hardt, E. Price, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016)* (Curran Associates Inc., 2016), 3323–3331.

212. A. K. Menon and R. C. Williamson, "The Cost of Fairness in Binary Classification," in *Conference on Fairness, Accountability and Transparency* (PMLR, 2018), 107–118.

213. S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth, "A Comparative Study of Fairness-Enhancing Interventions in Machine Learning," in *Proceedings of the Conference on Fairness, Accountability, and Transparency* (ACM, 2019), 329–338.

214. R. Bommasani, D. A. Hudson, E. Adeli, et al., "On the Opportunities and Risks of Foundation Models," *arXiv Preprint* (2021).

215. A. Abid, M. Farooqi, and J. Zou, "Persistent Anti-Muslim Bias in Large Language Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (ACM, 2021), 298–306.

216. S. Kiritchenko and S. Mohammad, "Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems," in *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics* (Association for Computational Linguistics, 2018), 43–53.

217. A. Parrish, A. Chen, N. Nangia, et al., "Bbq: A Hand-Built Bias Benchmark for Question Answering," in *Findings of the Association for Computational Linguistics: ACL*, vol. 2022 (ACL Anthology, 2022).

218. P. Delobelle, E. K. Tokpo, T. Calders, and B. Berendt, "Measuring Fairness With Biased Rulers: A Comparative Study on Bias Metrics for Pre-Trained Language Models," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics* (Association for Computational Linguistics, 2022), 1693–1706.

219. A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in *International Conference on Machine Learning* (PMLR, 2021), 8748–8763.

220. S. Hegselmann, A. Buendia, H. Lang, M. Agrawal, X. Jiang, and D. Sontag, "Tabllm: Few-Shot Classification of Tabular Data With Large Language Models," in *International Conference on Artificial Intelligence and Statistics* (PMLR, 2023), 5549–5581.

221. N. Hollmann, S. Müller, K. Eggensperger, and F. Hutter, "Tabpfn: A Transformer That Solves Small Tabular Classification Problems in a Second," *arXiv Preprint* (2022).

222. J. Hu and M. Du, "Enhancing Fairness in In-Context Learning: Prioritizing Minority Samples in Demonstrations," in *The Second Tiny Papers Track at ICLR 2024* (ICLR, 2024).

223. H. Ma, C. Zhang, Y. Bian, et al., "Fairness-Guided Few-Shot Prompting for Large Language Models," *Advances in Neural Information Processing Systems* 36 (2023): 43136–43155.

224. Y. Xiang, H. Yan, L. Gui, and Y. He, *Addressing Order Sensitivity of In-Context Demonstration Examples in Causal Language Models* (ACL, 2024).

225. D. Hendrycks and M. Mazeika, "X-Risk Analysis for AI Research," *arXiv Preprint* (2022).

226. D. Hendrycks, M. Mazeika, and T. Woodside, "An Overview of Catastrophic AI Risks," *arXiv Preprint* (2023).

227. P. Torres, "The Possibility and Risks of Artificial General Intelligence," *Bulletin of the Atomic Scientists* 75, no. 3 (2019): 105–108.

228. B. Ammanath, *Trustworthy AI: A Business Guide for Navigating Trust and Ethics in AI* (John Wiley & Sons, 2022).

229. D. Hadfield-Menell and G. K. Hadfield, "Incomplete Contracting and AI Alignment," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (ACM, 2019), 417–422.

230. W. N. Espeland and B. I. Vannebo, "Accountability, Quantification, and Law," *Annual Review of Law and Social Science* 3 (2007): 21–43.

231. C. Harlow, "Accountability and Constitutional Law," in *The Oxford Handbook of Public Accountability* (Oxford Academic, 2014), 195–210.

232. F. Doshi-Velez, M. Kortz, R. Budish, et al., "Accountability of AI Under the Law: The Role of Explanation," in *Berkman Center Research Publication, forthcoming* (2017), https://doi.org/10.2139/ssrn.3064761.

233. B. Kim, J. Park, and J. Suh, "Transparency and Accountability in AI Decision Support: Explaining and Visualizing Convolutional Neural Networks for Text Information," *Decision Support Systems* 134, no. 113 (2020): 302.

234. B. S. Miguel, A. Naseer, and H. Inakoshi, "Putting Accountability of AI Systems Into Practice," in *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence* (ACM, 2021), 5276–5278.

235. V. Dignum, *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*, vol. 2156 (Springer, 2019).

236. J. Millar, B. Barron, K. Hori, R. Finlay, K. Kotsuki, and I. Kerr, "Accountability in AI. Promoting Greater Social Trust," in *Conference on Artificial Intelligence* (IJCAI, 2018).

237. C. Novelli, M. Taddeo, and L. Floridi, "Accountability in Artificial Intelligence: What It Is and How It Works," *AI & Society* 39, no. 4 (2024), https://doi.org/10.1007/s00146-023-01635-y.

238. S. Lynch, "2023 State of AI in 14 Charts," (2023), https://hai.stanford.edu/news/2023-state-ai-14-charts.

239. Congress, "National Artificial Intelligence Initiative Act of 2020, Division of Public Law 116-283," Current as of April 15, 2023 (2020), https://www.congress.gov/bill/116th-congress/house-bill/6395/text.

240. National Conference of State Legislatures, *Autonomous Vehicles State Bill Tracking Database* (NCSL, 2023), https://www.ncsl.org/transportation/autonomous-vehicles-state-bill-tracking-database.

241. National Conference of State Legislatures, *State Laws Related to Digital Privacy* (NCSL, 2022), https://www.ncsl.org/technology-and-communication/state-laws-related-to-digital-privacy.

242. K. Zhu, "The State of AI Regulations in 2023," (2023), https://www.holisticai.com/papers/the-state-of-ai-regulations-in-2023.

243. S. McCallum, "Revenge and Deepfake Porn Laws to be Toughened," BBC.com (2023), https://www.bbc.com/news/technology-66021643.

244. G. Interesse, *China to Regulate Deep Synthesis (Deepfake) Technology From 2023* (China-Briefing, 2023), https://www.china-briefing.com/news/china-to-regulate-deep-synthesis-deep-fake-technology-starting-january-2023/.

245. European Union, "Regulation (eu) 2016/679 of the European Parliament and of the Council of 27 April 2016," (2018), https://eur-lex.europa.eu/eli/reg/2016/679/oj.

246. European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)* (European Commission, 2021), https://artificialintelligenceact.eu/.

247. C. Morse, "What the GDPR Can Teach Us About AI Regulation," (2023), https://www.weforum.org/agenda/2023/06/gdpr-artificial-intelligence-regulation-europe-us/.

248. Government of Canada, "Department of Justice—Statement of Potential Charter Impacts," (2022), https://www.justice.gc.ca/eng/csj-sjc/pl/charter-charte/c27_1.html.

249. W. J. Wagner, A. Guilmain, and M. Walsh, "Bill C-27 Timeline of Developments," (2024), https://gowlingwlg.com/en-ca/insights-resources/articles/2024/bill-c27-timeline-of-developments.

250. E. Black, L. Koepke, P. Kim, S. Barocas, and M. Hsu, "The Legal Duty to Search for Less Discriminatory Algorithms," *arXiv Preprint* (2024).

251. E. Black, M. Raghavan, and S. Barocas, "Model Multiplicity: Opportunities, Concerns, and Solutions," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. FAccT'22. New York, NY, USA* (Association for Computing Machinery, 2022), 850–863, https://doi.org/10.1145/3531146.3533149.

252. D. Foster, *Generative Deep Learning* (O'Reilly Media, Inc, 2022).

253. OpenAI, "Chatgpt," (2023), https://openai.com/research/gpt-4.

254. Photoleap, "Midjourney," (2023), https://www.photoleapapp.com/lp/pl-text-to-image/.

255. Soundraw Inc, "Soundraw," (2023), https://soundraw.io/.

256. T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE, 2019), 4401–4410.

257. A. Radford, J. Wu, R. Child, et al., "Language Models Are Unsupervised Multitask Learners," *OpenAI Blog* 1, no. 8 (2019): 9.

258. D. Harwell, "He Used AI to Win a Fine-Arts Competition. Was It Cheating?," (2022), https://www.washingtonpost.com/technology/2022/09/02/midjourney-artificial-intelligence-state-fair-colorado/.

259. J. Grierson, "Photographer Admits Prize-Winning Image Was AI-Generated," (2023), https://www.theguardian.com/technology/2023/apr/17/photographer-admits-prize-winning-image-was-ai-generated#:~:text=The%20German%20artist%20Boris%20Eldagsen,generations%20in%20black%20and%20white.

260. A. King, "The grammys officially ban AI-generated works – 'only humans' eligible for awards," (2023), https://www.digitalmusicnews.com/2023/06/19/the-grammys-ai-generated-content-ineligible/.

261. Z. Epstein, A. Hertzmann, M. Akten, et al., "Art and the Science of Generative AI," *Science* 380, no. 6650 (2023): 1110–1111.

262. F. Register, "Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence," (2023), https://www.federalregister.gov/documents/2023/03/16/2023-05321/copyright-registration-guidance-works-containing-material-generated-by-artificial-intelligence.

263. F. Register, "Copyright, Designs and Patents Act 1988," (2023), https://www.legislation.gov.uk/ukpga/1988/48/contents.

264. European Union, *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts* (EU, 2022), https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/.

265. E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (ACM, 2021), 610–623.

266. Future for Life, "Pause Giant AI Experiments: An Open Letter," (2023), https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

267. Future for Life, "Open Letter to the Representatives of the European Commission, the European Council and the European Parliament," (2023), https://drive.google.com/file/d/1wrtxfvcD9FwfNfWGDL37Q6Nd8wBKXCkn/view?usp=sharing.

268. D. Patterson, J. Gonzalez, Q. Le, et al., "Carbon Emissions and Large Neural Network Training," *arXiv Preprint* (2021).

269. R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM* 63, no. 12 (2020): 54–63.

270. J. Zhong, Y. Zhong, M. Han, T. Yang, and Q. Zhang, "The Impact of AI on Carbon Emissions: Evidence from 66 Countries," *Applied Economics* 56, no. 25 (2023): 2975–2989, https://doi.org/10.1080/00036846.2023.2203461.

271. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2019), 3645–3650.

272. S. A. Sarswatula, T. Pugh, and V. Prabhu, "Modeling Energy Consumption Using Machine Learning," *Frontiers in Manufacturing Technology* 2, no. 855 (2022): 208.

273. L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models," *arXiv Preprint* (2020).

274. P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, "Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning," *Journal of Machine Learning Research* 21, no. 1 (2020): 10039–10081.

275. A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the Carbon Emissions of Machine Learning," *arXiv Preprint* (2019).

276. Y. Gao, Y. Liu, H. Zhang, et al., "Estimating GPU Memory Consumption of Deep Learning Models," in *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering* (ACM, 2020), 1342–1352.

277. A. Miller, "The Future of e-Waste Depends on AI and Machine Learning," *AI Journal* (2022).

278. M. Coeckelbergh, "AI for Climate: Freedom, Justice, and Other Ethical and Political Challenges," *AI and Ethics* 1, no. 1 (2021): 67–72.

279. J. Cowls, A. Tsamados, M. Taddeo, and L. Floridi, "The AI Gambit: Leveraging Artificial Intelligence to Combat Climate Change—Opportunities, Challenges, and Recommendations," *AI & Society* 38 (2021): 283–307.

280. C. Huntingford, E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang, "Machine Learning and Artificial Intelligence to Aid Climate Change Research and Preparedness," *Environmental Research Letters* 14, no. 12 (2019): 124007.

281. W. Leal Filho, T. Wall, S. A. R. Mucova, et al., "Deploying Artificial Intelligence for Climate Change Adaptation," *Technological Forecasting and Social Change* 180, no. 121 (2022): 662.

282. E. A. Barnes, J. W. Hurrell, I. Ebert-Uphoff, C. Anderson, and D. Anderson, "Viewing Forced Climate Patterns Through an AI Lens," *Geophysical Research Letters* 46, no. 22 (2019): 13389–13398.

283. T. Schneider, S. Behera, G. Boccaletti, et al., "Harnessing AI and Computing to Advance Climate Modelling and Prediction," *Nature Climate Change* 13, no. 9 (2023): 887–889.

284. Q. W. Ahmed, S. Garg, A. Rai, et al., "AI-Based Resource Allocation Techniques in Wireless Sensor Internet of Things Networks in Energy Efficiency With Data Optimization," *Electronics* 11, no. 13 (2022): 2071.

285. Q. Zeng, Y. Du, K. Huang, and K. K. Leung, "Energy-Efficient Resource Management for Federated Edge Learning With CPU-GPU Heterogeneous Computing," *IEEE Transactions on Wireless Communications* 20, no. 12 (2021): 7947–7962.

286. S. S. Ali and B. J. Choi, "State-of-the-Art Artificial Intelligence Techniques for Distributed Smart Grids: A Review," *Electronics* 9, no. 6 (2020): 1030.

287. T. Mazhar, H. M. Irfan, I. Haq, et al., "Analysis of Challenges and Solutions of Iot in Smart Grids Using AI and Machine Learning Techniques: A Review," *Electronics* 12, no. 1 (2023): 242.

288. A. Al-Othman, M. Tawalbeh, R. Martis, et al., "Artificial Intelligence and Numerical Models in Hybrid Renewable Energy Systems With Fuel Cells: Advances and Prospects," *Energy Conversion and Management* 253, no. 115 (2022): 154.

289. C. Chen, Y. Hu, M. Karuppiah, and P. M. Kumar, "Artificial Intelligence on Economic Evaluation of Energy Efficiency and Renewable Energy Technologies," *Sustainable Energy Technologies and Assessments* 47, no. 101 (2021): 358.

290. S. Ramachandran, "Applying AI in Power Electronics for Renewable Energy Systems [Expert View]," *IEEE Power Electronics Magazine* 7, no. 3 (2020): 66–67.

291. W. Shin, J. Han, and W. Rhee, "AI-Assistance for Predictive Maintenance of Renewable Energy Systems," *Energy* 221, no. 119 (2021): 775.

292. E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential Privacy Has Disparate Impact on Model Accuracy," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (Curran Associates Inc, 2019), 15479–15488.

293. A. Uniyal, R. Naidu, S. Kotti, et al., "DP-SGD vs. Pate: Which Has Less Disparate Impact on Model Accuracy?," *arXiv Preprint* (2021).

294. G. Falco, B. Shneiderman, J. Badger, et al., "Governing AI Safety Through Independent Audits," *Nature Machine Intelligence* 3, no. 7 (2021): 566–571.

295. J. Ferry, U. Aïvodji, S. Gambs, M.-J. Huguet, and M. Siala, "Exploiting Fairness to Enhance Sensitive Attributes Reconstruction," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (IEEE, 2023), 18–41.

296. L. Lyu and C. Chen, "A Novel Attribute Reconstruction Attack in Federated Learning," *arXiv Preprint* (2021).

297. T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang, "Fairness Without Demographics in Repeated Loss Minimization," in *International Conference on Machine Learning* (PMLR, 2018), 1929–1938.

298. P. J. Kenfack, S. E. Kahou, and U. Aïvodji, "Fairness Under Demographic Scarce Regime," *arXiv Preprint* (2023).

299. J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, and H. Lakkaraju, "Fairness via Explanation Quality: Evaluating Disparities in the Quality of Post Hoc Explanations," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (ACM, 2022), 203–214, https://doi.org/10.1145/3514094.3534159.

300. N. Mehrabi, M. Naveed, F. Morstatter, and A. Galstyan, "Exacerbating Algorithmic Bias Through Fairness Attacks," *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (2021): 8930–8938.

301. X. Li, P. Wu, and J. Su, "Accurate Fairness: Improving Individual Fairness Without Trading Accuracy," *Proceedings of the AAAI Conference on Artificial Intelligence* 37 (2023): 14312–14320.

302. N. Papernot, A. Thakurta, S. Song, S. Chien, and Ú. Erlingsson, "Tempered Sigmoid Activations for Deep Learning With Differential Privacy," *Proceedings of the AAAI Conference on Artificial Intelligence* 35 (2021): 9312–9321.

303. C. Tran, K. Zhu, F. Fioretto, and P. Van Hentenryck, "Sf-Pate: Scalable, Fair, and Private Aggregation of Teacher Ensembles," *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI '23)* (2022): 501–509, https://doi.org/10.24963/ijcai.2023/56.

304. Y. Wang, X. Wang, A. Beutel, F. Prost, J. Chen, and E. H. Chi, "Understanding and Improving Fairness-Accuracy Trade-Offs in Multi-Task Learning," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (ACM, 2021), 1748–1757.

305. D. Xu, W. Du, and X. Wu, "Removing Disparate Impact on Model Accuracy in Differentially Private Stochastic Gradient Descent," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (ACM, 2021), 1924–1932.