From Principles to Practice: A Deep Dive into AI Ethics and Regulations

Nan Sun   NAN.SUN@UNSW.EDU.AU
University of New South Wales, 37 Constitution Ave, Canberra, ACT 2612 Australia

Yuantian Miao (Corresponding Author)   SKY.MIAO@NEWCASTLE.EDU.AU
University of Newcastle, University Dr, Newcastle, NSW 2308 Australia

Hao Jiang   JH072535@FOXMAIL.COM
Swinburne University of Technology, John St, Melbourne, VIC 3122 Australia

Ming Ding   MING.DING@DATA61.CSIRO.AU
Commonwealth Scientific and Industrial Research Organisation (CSIRO), Level 5/13 Garden St, Sydney, NSW, 2015, Australia

Jun Zhang   JUNZHANG@SWIN.EDU.AU
Swinburne University of Technology, John St, Melbourne, VIC 3122 Australia

Abstract

In the rapidly evolving domain of Artificial Intelligence (AI), the complex interaction between innovation and regulation has become an emerging focus of our society. Despite tremendous advancements in AI's capabilities to excel in specific tasks and contribute to diverse sectors, establishing a high degree of trust in AI-generated outputs and decisions necessitates meticulous caution and continuous oversight. A broad spectrum of stakeholders, including governmental bodies, private sector corporations, academic institutions, and individuals, have launched significant initiatives. These efforts include developing ethical guidelines for AI and engaging in vibrant discussions on AI ethics, both among AI practitioners and within the broader society. This article thoroughly analyzes the groundbreaking AI regulatory framework proposed by the European Union. It delves into the fundamental ethical principles of safety, transparency, non-discrimination, traceability, and environmental sustainability for AI developments and deployments. Considering the technical efforts and strategies undertaken by academics and industry to uphold these principles, we explore the synergies and conflicts among the five ethical principles. Through this lens, work presents a forward-looking perspective on the future of AI regulations, advocating for a harmonized approach that safeguards societal values while encouraging technological advancement.

1. Introduction

In today's world, where Artificial Intelligence (AI) is swiftly reshaping numerous facets of our daily life, the demand for robust and efficient AI regulations has become more pronounced. Acknowledging this need, on October 30, 2023, U.S. President Joe Biden enacted an Executive Order, signifying a historical move towards tackling the complex issues raised by AI technologies (House, 2023). This action highlights an increasing recognition of the profound influence of AI on aspects such as safety and security, privacy, fairness and civil rights, as well as innovation and competition, calling for a forward-thinking and preemptive strategy in regulatory measures regarding AI advancements.

The evolution of AI regulations represents a contemporary and dynamic narrative that has emerged significantly over recent decades. Initial apprehensions about ethics and data privacy have evolved into more defined guidelines and potential legislative frameworks, though the journey is far from complete. The chronicle of AI regulation is a modern and swiftly developing area, mirroring the rapid advancements in AI technology. The genesis of this regulatory trajectory can be traced back to the mid-20th century, marked notably by the introduction of the Fair Credit Reporting Act in 1970 in the United States (trade commission, 2023). The legislation safeguards data gathered by consumer reporting agencies, including credit bureaus, medical data firms, and tenant screening services. The Act prohibits sharing information in a consumer report with individuals or entities that do not have a designated purpose as outlined in the Act. This act represented an early significant step in the formalization of data protection. Another significant achievement is the General Data Protection Regulation (GDPR) (of the European Union, 2023), which is a thorough data privacy and security legislation enacted by the European Union (EU) in 2016. It stands as the most stringent data privacy and security law globally. While it was formulated and ratified by the EU, it places responsibilities on organizations worldwide as long as they interact with or gather data from individuals in the EU. The GDPR introduces a set of novel data privacy rights to grant individuals greater authority over the information they provide to organizations. As a visual representation, Figure 1, encapsulates the critical milestones in the progression of data protection regulations.
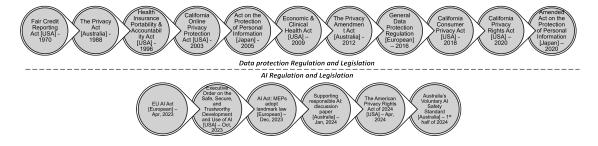

Figure 1: Timeline of data protection and AI regulations and legislation

AI regulation goes beyond the data component to encompass AI systems' design, development, deployment, and use. It is concerned not only with the data these systems process but also their decisions, the biases they may perpetuate, and the ethical implications of their interaction with humans and other systems. AI regulation aims to address the societal and ethical challenges posed by increasingly autonomous technologies. Regulation of AI varies from country to country, reflecting different approaches, priorities, and concerns. For example, in August 2023, the Australian Human Rights Commission (i.e., Commission) submitted a discussion document titled "Promoting Responsible AI" to the Department of Industry, Science, and Resources in Australia (Commission, 2023a). The Commission expresses particular concern about emerging issues, including privacy breaches, algorithmic discrimination, biases in automation, and the spread of misinformation and disinformation. Furthermore, the intricate interaction between AI, neurotechnology, the metaverse, and extended reality technologies adds complexity to this rapidly evolving field. The Commission has advised the government to initiate an assessment of regulatory deficiencies to ascertain

the relevance of existing legislation to AI. In cases where gaps are identified, the associated legislation should be examined and updated to effectively address AI-related challenges.

Figure 1 outlines the key developments in AI regulation and legislation across various regions, highlighting significant milestones from April 2021 to the first half of 2024. In Europe, the EU AI Act, the first comprehensive regulation on AI, adopts a risk-based approach emphasizing safety, transparency, and environmental sustainability. It is expected to be fully implemented by 2026. In October 2023, the White House issued an Executive Order to ensure safe, secure, and trustworthy AI development in the United States, prioritizing standards for AI safety and civil rights. By April 2024, the U.S. introduced the American Privacy Rights Act, focusing on comprehensive data privacy. Concurrently, Australia responded to AI safety discussions by proposing a voluntary AI safety standard emphasising a risk-based framework for AI deployment. These initiatives reflect a growing international commitment to responsibly managing AI's societal impacts.

In the initial stages of research, Scherer (Scherer, 2015) initiated a discussion on the feasibility and challenges of government AI regulation, suggesting paths for effective regulation. Concurrently, other researchers delved into AI's regulatory and ethical aspects in specific domains. For example, Pesapane et al. (Pesapane, Volont'é', Codari, & Sardanelli, 2018) explored these issues in medical services, while Reddy et al. (Reddy, Allan, Coghlan, & Cooper, 2020) proposed a governance model for ethical and regulatory concerns in AI healthcare applications. Focusing on particular aspects of AI regulation, Wäschle (W'á'schle, Thaler, Berres, P'ó'lzlbauer, & Albers, 2022) conducted a systematic literature review on highly automated driving, emphasizing safety assessment methods for AI systems. Other reviews have concentrated on different aspects: reviews (Larsson & Heintz, 2020; Walmsley, 2021) on AI transparency; works (Ferrer, van Nuenen, Such, Cot'é', & Criado, 2021; Caton & Haas, 2020; Pessach & Shmueli, 2022) on bias and discrimination in AI; and reviews (Van Wynsberghe, 2021; Nishant, Kennedy, & Corbett, 2020; Wu, Raghavendra, Gupta, Acun, Ardalani, Maeng, Chang, Aga, Huang, Bai, et al., 2022a; Dhar, 2020) on AI's sustainability.

AI and algorithm-driven decision-making are increasingly impacting our daily lives, being extensively used in critical sectors, such as healthcare, business, government, education, and justice. This shift towards an algorithmic society comes with many benefits and risks, as these systems can sometimes cause harm to users and society. Ensuring the safety, reliability, and trustworthiness of these systems is crucial. Trustworthy AI systems are those that are reliable, ethical, and transparent, enabling users to have confidence in the decisions made by these systems. Recent literature reviews on trustworthy AI (Li, Qi, Liu, Di, Liu, Pei, Yi, & Zhou, 2023; Mora-Cantallops, S'á'nchez-Alonso, Garc'í'a-Barriocanal, & Sicilia, 2021; Kaur, Uslu, Rittichier, & Durresi, 2022), including work (Kaur et al., 2022) focused on the elements of fairness, explainability, accountability, reliability, and acceptance to mitigate AI risks and enhance user and societal trust. On the other hand, Li et al. (Li et al., 2023) provided specific recommendations for AI practitioners, covering the entire AI system lifecycle, from data collection, model development, and system deployment to ongoing monitoring and governance.

In April 2021, the European Commission proposed the initial AI regulatory framework for the European Union (Commission, 2023d). This framework classifies AI systems into various risk levels, with corresponding levels of regulation. The primary objective of this

Act is to guarantee the safety, transparency, traceability, non-discrimination, and environmental sustainability of AI systems used in the EU. The oversight of AI systems should rely on human decision-making (i.e., autonomy) rather than full automation to avoid adverse consequences. Members of the European Parliament endorsed the regulation, which was reached in negotiations with member states in December 2023, with 523 votes in favour, 46 against and 49 abstentions (Parliament, 2024).

To the best of our knowledge, there lacks a comprehensive survey that fully addresses the five fundamental principles laid out in the AI Act (Commission, 2023d), i.e., safety, transparency, traceability, non-discrimination, and environmental sustainability, as they apply to AI models and systems. An analytical paper is needed to explore how regulations might balance the push for innovation with the need to mitigate risks such as unforeseen consequences or misuse of AI technology. In addition, a detailed review of existing AI regulations would enlighten policymakers, stakeholders, and the public and support more informed decisions in AI governance. Motivated by these considerations, we thoroughly examine AI regulation, exploring the interactions and tensions among its essential aspects. We also provide structured, actionable recommendations for AI professionals in academia and industry. The following is a summary of the main contributions of this paper.

- We present and analyze the stipulations detailed in the AI Act, concentrating on the key areas covered by AI regulations, including safety, transparency, non-discrimination, traceability, and environmental sustainability within the realm of AI. These discussions are specifically rooted in the regulatory necessities and the foundational concepts introduced in the AI Act, providing a comprehensive overview of its scope and implications.

- We discuss the synergies and conflicts among components of AI regulation through an analysis of current technical efforts focused on crafting AI systems that are safe, transparent, traceable, environmentally sustainable, and free from bias, respectively. This exploration aids in understanding how these elements interact and sometimes clash, guiding the development and deployment of AI that adheres to high ethical and regulatory standards.

- By exploring the synergies and conflicts within these critical dimensions of AI regulation, we investigate how AI systems can be designed and developed to meet regulatory standards, highlighting the trade-offs involved. We also discuss future research pathways for scholars in AI and AI regulation fields and for industry practitioners who need to adhere to AI regulations once the AI Act is implemented.

2. Preliminaries and definitions

This section establishes the foundational concepts and definitions for this survey, beginning with a rationale for AI regulation to underscore its importance. We then introduce key concepts within the AI Act, emphasizing its focus on human-centric considerations.

2.1 The need for AI regulation

The journey of AI regulation is an evolving narrative, having gained significant momentum in recent years. Initially, the focus was primarily on ethical considerations and data privacy

concerns. These foundational issues have gradually paved the way for more defined guidelines and, in some instances, the drafting of legislative proposals. However, despite these strides, there is still a considerable amount of progress to be made in this area. Before delving deeper into this topic, it's essential to outline the core reasons necessitating AI regulations, which range from safeguarding individual privacy and ensuring ethical AI use to mitigating potential societal risks and establishing accountability in AI development and application.

2.1.1 Ethical and societal reasons

The *unbiasedness* of the AI system depends on its training data and design decisions. Without proper regulations, these systems risk perpetuating existing societal biases, leading to unfair or discriminatory outcomes. This is particularly concerning in areas such as employment, lending, and criminal justice, where biased AI could reinforce existing *inequalities.* Regulations are necessary to ensure that AI systems are developed and deployed in a manner that is *equitable* and *fair* to all segments of society (Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021b).

The increasing deployment of AI in critical sectors like healthcare, finance, and law enforcement raises important questions about *accountability* (Bernard & Balog, 2023). When an AI system makes a decision that negatively impacts individuals or communities, it is essential to have clear guidelines determining who is *responsible* – the developers, the users, or the AI itself. Regulations in this area would help clarify lines of responsibility and ensure that affected parties have recourse in the event of harm or injustice.

AI's reliance on vast datasets, often comprising personal and sensitive information, heightens concerns about *data privacy and security* (Chatterjee & NS, 2023). The potential for misuse or unauthorized access to this data poses significant risks to *individuals' privacy rights*. Regulations are needed to establish strict guidelines on data collection, use, storage, and sharing. These include ensuring that individuals are informed and consented to how their data is used and implementing robust security measures to protect data from breaches and leaks.

Another reason is to mitigate the risk of *social manipulation* (Sheikh, 2020). In particular, AI's role in social media algorithms can influence public opinion and even manipulate elections, as evidenced by the Cambridge Analytica scandal, where personal data was used to target voters and potentially affect the political landscape. This has led to calls for tighter regulations on how AI can be used to manipulate information and target users on social media platforms through misinformation and disinformation.

2.1.2 Technical reasons

In fields such as autonomous driving or medical diagnostics, where AI systems make decisions that can have life-or-death consequences, the importance of *risk management* cannot be overstated (AI, 2023). Regulations are critical in ensuring that these systems are designed, tested, and operated under stringent safety standards. These involve setting benchmarks for performance, establishing protocols for failure detection and response, and ensuring systems are resilient to various types of risks, including cyber threats, system malfunctions, and unexpected environmental conditions. Regulatory frameworks can guide the development and deployment of these systems to prevent accidents and mitigate potential harms.

From a technical standpoint, AI regulation is essential for promoting *interoperability* and *scalability*. Ensuring AI systems can effectively communicate and work together across different platforms and applications is crucial for maximizing their utility and efficiency. Regulations could establish standards that enhance this *interoperability*. Additionally, *scalability* is vital for AI systems to handle increasing loads smoothly. Regulations could help set guidelines that ensure AI systems are designed to scale efficiently without disproportionate increases in resource demands.

Furthermore, large AI models typically require substantial computational power, translating into significant energy consumption and consequent environmental impacts (Ahmad, Zhang, Huang, Zhang, Dai, Song, & Chen, 2021). Regulations could play a critical role in pushing for developing more *energy-efficient AI technologies*, potentially reducing both the environmental footprint and the costs associated with high energy consumption.

2.1.3 Global and National Security Reasons

AI's capabilities in processing vast amounts of data make it a powerful tool for surveillance, espionage, and cyber warfare. This raises concerns about the protection of *national interests and security* (Robinson, 2020). A regulatory framework is necessary to govern the use of AI in these domains, ensuring that it does not compromise national security or infringe on citizens' rights and freedoms. Regulations can help define permissible uses of AI in surveillance and intelligence, establish safeguards against unauthorized data access, and protect against cyber threats that utilize AI technologies. This regulatory oversight is pivotal in maintaining the delicate balance between the use of AI for national security and the preservation of democratic values and individual rights.

The prospect of AI being used in *autonomous weapons systems* is a growing concern (Akhtar, 2023; De 'Á'greda, 2020). These technologies can potentially revolutionize warfare, raising ethical and strategic questions about their deployment. Regulations are crucial in preventing the harmful use of AI in military applications, ensuring that such systems are developed and used in compliance with international humanitarian laws and ethical standards. By establishing clear boundaries and oversight mechanisms, regulations can prevent an arms race in lethal autonomous weapons and ensure that AI is used to enhance, rather than undermine global security and peace.

2.1.4 Economic Reasons

AI technology has the potential to disrupt market dynamics significantly. Without proper regulatory oversight, there is a risk that large corporations with substantial resources could monopolize the AI sector. Such dominance could hinder innovation, as smaller companies and startups may find it challenging to compete on an unequal playing field. Regulations can foster *a competitive market* by ensuring fair access to AI technologies and preventing monopolistic practices. This encourages a diverse ecosystem of AI developers and service providers, fostering innovation and ensuring that the benefits of AI are not confined to a few dominant players.

As AI becomes more integrated into consumer products and services, it is vital to *protect consumers* from faulty or misleading AI applications. Regulations are crucial in setting standards for AI quality, reliability, and truthfulness in advertising (Stojanovic, 2020). This

includes ensuring that AI-driven products meet certain performance benchmarks and that claims made about AI capabilities are accurate and not misleading. By doing so, regulations safeguard consumers from subpar or potentially harmful AI-driven products and services, fostering trust and confidence in the market.

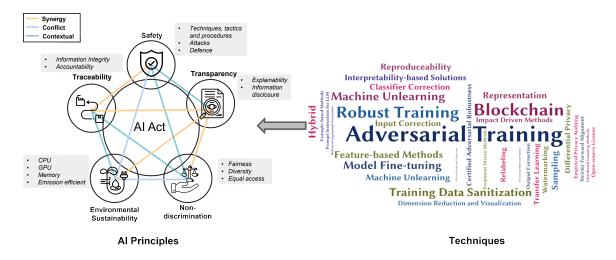## 2.2 Key concepts in the AI Act

Figure 2: Human-centered principles of the AI Act and techniques supporting their implementation.

The AI Act aims to guarantee that AI systems utilized in the EU uphold safety, transparency, traceability, non-discrimination, and environmental sustainability, as illustrated in Figure 2. It places a strong focus on human oversight to avert detrimental consequences. In this context, we explain these five crucial elements as outlined in the AI Act.

2.2.1 Safety

AI safety, a multifaceted and interdisciplinary domain, focuses on preventing adverse consequences that might arise from AI systems (Leslie, 2019). This field tackles not only ethical considerations, ensuring AI systems align with moral values and contribute positively, but also addresses technical challenges (Du & Xie, 2021). These challenges include monitoring and maintaining the reliability of AI systems to avert risks.

In the EU AI Act, safety refers primarily to the regulation and mitigation of risks associated with using AI systems, ensuring that these systems do not pose unacceptable risks to health, safety, and fundamental rights (Commission, 2023b). Consider the machinery industry as an example. High-risk AI systems here might involve AI-driven robots in manufacturing plants, such as an autonomous robotic arm assembling heavy machinery parts or an AI system managing critical manufacturing processes. The malfunction of these systems can cause severe production setbacks, expensive damages, or even endanger human workers (Huang, Shen, Li, Fey, & Brecher, 2021). To mitigate these risks, it's crucial to equip AI systems with robust safety features like fail-safes or emergency stops. They must un-

dergo comprehensive testing and validation across different operational scenarios. Moreover, resilience against cyber threats or manipulation is vital to prevent potentially disastrous outcomes.

2.2.2 TRANSPARENCY

In the context of the EU AI Act, transparency refers primarily to obligations imposed on AI system providers to ensure that the operation and output of AI systems are clear and understandable by those who deploy or interact with them. Transparency measures are designed to ensure that AI systems are used in a way that is understandable and respects the rights and freedoms of all individuals involved. This supports the overall goal of human-centric AI development and deployment within the EU.

The concept of transparency in AI regulation encompasses the level of openness and comprehensibility in how AI systems are developed, implemented, and managed (Robinson, 2020). For instance, AI systems identified as high-risk must accompany these systems with detailed documentation and instructions that disclose the system's capabilities, characteristics, and limitations, allowing users to understand and correctly apply the outputs generated by these systems (Commission, 2023d).

Beyond these operational aspects, transparency in AI also significantly intersects with the concepts of explainability(Hamon, Junklewitz, Sanchez, et al., 2020). This facet of transparency is about demystifying the internal workings of AI systems, providing clarity on how specific decisions are reached. It's a move towards making AI systems less of a "black box" and more of an open book, where users and stakeholders can understand and rationalize the logic behind AI-driven decisions. This level of transparency is not just a regulatory requirement but a foundation for building trust between users and AI technologies, ensuring that users can interact with these systems confidently and effectively, knowing the rationale behind their outputs and actions (Shneiderman, 2020).

2.2.3 NON-DISCRIMINATION

The principle of non-discrimination in the realm of AI is centered on reducing the likelihood of biased decision-making by algorithms (Wachter, Mittelstadt, & Russell, 2020). This involves careful attention to the design and the integrity of data sets used in creating AI systems, combined with commitments to rigorous testing, effective risk management, thorough documentation, and consistent human oversight throughout the entire lifecycle of these AI systems.

In the context of AI regulations, non-discrimination is a guiding principle that ensures AI technologies are developed and utilized to prevent biased and unfair outcomes or decisions, particularly those that might unjustly target or disadvantage different groups of people(Commission, 2023d). This aspect of AI regulation is crucial given the risk that AI systems might unintentionally sustain, magnify, or even introduce new forms of biases or discriminatory practices (Lloyd, 2018). Such regulations aim to encourage the creation and application of AI technologies in a manner that upholds fairness and ethical standards, thus preventing the reinforcement of existing social disparities. For instance, as outlined in OpenAI's GPT-4 technical report, content considered harmful encompasses elements like "hate speech, discriminatory language, incitements to violence, or content that is then used to ei-

ther spread false narratives or to exploit an individual" (OpenAI, 2023). In response to this, GPT-4 was developed using a technique known as model-refusal. This approach, grounded in reinforcement learning, rewards the model during its training phase for actively declining to produce such harmful content.

2.2.4 ENVIRONMENTAL SUSTAINABILITY

In AI regulation, the concept of environmental sustainability emphasizes the creation, implementation, and utilization of AI systems in ways that mitigate adverse environmental effects. This approach recognizes the ecological impact of AI technology, focusing on issues such as energy use, efficient resource management, and promoting ecological sustainability (Nishant et al., 2020). This includes safeguarding the environment and enhancing its quality, which also pertains to human health and safety (Commission, 2023d).

Crucial aspects of this environmental focus in AI regulation include promoting energy efficiency, particularly in AI systems with high computational demands like large data centers, which are significant electricity consumers (Ahmad et al., 2021). Additionally, it encompasses the pursuit of sustainable development, efforts to decrease the carbon footprint of AI technologies, and the advancement of 'Green AI' (Van Wynsberghe, 2021). This latter term refers to designing AI models that are not just high-performing but also optimized for minimal consumption of computational resources and energy. Moreover, it involves adopting eco-friendly practices, such as minimizing electronic waste and encouraging the recycling of AI hardware components.

2.2.5 TRACEABILITY

The principle of traceability within the framework of AI regulation encompasses the systematic capability to monitor, record, and chronicle the decision-making processes employed by AI systems during their entire operational lifespan (Commission, 2023d). This principle is crucial for ensuring that AI operations are transparent and accountable. For instance, an AI system should possess robust logging features that provide a traceability level suitable for its intended use, as per the guidelines outlined in the AI Act. Such traceability measures are vital for verifying the AI system's performance and compliance throughout its lifecycle (Kroll, 2021).

Additionally, traceability involves closely tracking the various iterations of AI models, encompassing all modifications and updates that occur over time. This tracking is not just a record-keeping exercise but a critical tool for understanding how an AI model evolves (Agre, 2014). It enables stakeholders to comprehend the changes made to the model, assess their impacts, and, if necessary, revert to prior versions when newer updates do not perform as expected or introduce unforeseen issues.

2.3 Human-centered AI Act

The AI Act emphasizes putting people first. Hence, we discuss integrating human values into the design, development, and deployment of ethical AI systems in this subsection. As illustrated in Figure 3, we map various stakeholders' interests and concerns to the ethical considerations across each stage of the AI lifecycle. This comprehensive alignment ensures
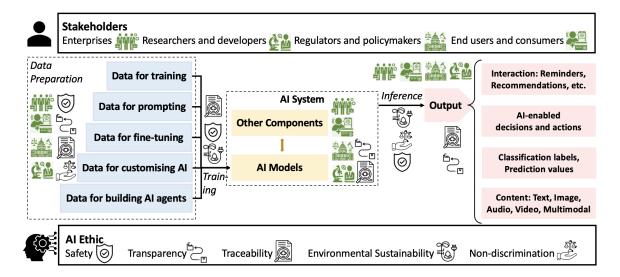
Figure 3: Human-centric principles reflected in the AI Act on the AI system components and the related stakeholders.

that AI systems are developed and deployed in a manner that addresses ethical aspects thoroughly, with the overarching goal of benefiting individuals and society at large.

We categorize stakeholders into four main groups: enterprises, researchers and developers, regulators and policymakers, and users and customers, as shown in Figure 3. According to recent research (Latonero, 2018; Deshpande & Sharp, 2022; Aguirre, Dempsey, Surden, & Reiner, 2020; Stahl, Rodrigues, Santiago, & Macnish, 2022), these stakeholders have distinct interests and responsibilities regarding ethical AI development and deployment. Enterprises are responding to both internal and external pressures to ensure their AI technologies are ethically deployed and do not cause harm. For instance, Microsoft conducted a Human Rights Impact Assessment (HRIA) for AI, and Google has established AI principles that reference human rights (Latonero, 2018; Deshpande & Sharp, 2022). These efforts aim to mitigate adverse impacts, ensure corporate social responsibility, and maintain public trust. Regulators and policymakers are responsible for aligning AI with human rights principles throughout each stage of the AI lifecycle. Researchers and developers, who are deeply involved in each lifecycle stage, ensure technical performance and efficiency while promoting responsible AI behavior. They also contribute to the discourse on AI and human rights by examining the social impacts of AI, developing ethical frameworks, and advocating for policies that prioritize human dignity (Deshpande & Sharp, 2022). Lastly, end-users and customers focus on the risks and harms associated with AI. They advocate for AI systems that are free from biases, transparent in their operations and decision-making processes, particularly concerning conflicts of interest, and safe and traceable, including maintaining data integrity and security (Aguirre et al., 2020).

The AI Act influences people by creating a comprehensive framework that addresses various aspects of AI development and deployment, all with the ultimate goal of protecting human safety, rights, and well-being while promoting beneficial AI technologies. It enhances

safety by reducing risks of harm from AI applications, supports environmental sustainability to benefit both the environment and human well-being, ensures traceability for accountability, and increases transparency to foster trust, fairness, and informed decision-making.

Ethical concerns in the data preparation stage include ensuring data safety, obtaining informed consent, and addressing biases in data collection (Huang, Zhang, Mao, & Yao, 2022a). It is crucial to collect representative data that accurately reflects the diversity of the population to avoid discrimination and ensure fairness. This process must also ensure the data owners' safety and make the data collection process transparent and traceable for data providers. During model training, it is essential to build a robust AI against various threats to ensure its safety, while the ethical use of computational resources should be considered to minimize environmental impacts. Transparency and traceability in the deployment process of AI models, including integrating other components within the AI system, are crucial to ensuring user safety. During the inference stage, it is important to evaluate the model's fairness, environmental sustainability, and safety in real-world scenarios (Huang et al., 2022a), and accordingly can opt to implement bias mitigation, cost-effective, and threats mitigation techniques. The output of the AI system is normally shown as various interactions with customers, AI-enabled decisions and actions, classification labels, prediction values, and/or other content like text, images, etc. All those output data must be transparent and traceable.

3. How to design regulation-compliant systems: the synergies and conflicts

This section investigates how the five key principles of AI regulation interact, exploring both their capacity to work together and the conflicts that exist between them. By analyzing these dynamics, this section aims to provide a thorough understanding of the techniques that can help AI systems adhere to these principles. This understanding will serve as a guide for developing AI systems that are both ethically sound and legally compliant.

3.1 Safety

3.1.1 Overview

Establishing well-defined safety protocols for the development and implementation of AI in an ethical manner is crucial. Initially, we present a comprehensive overview of safety considerations in AI systems. This forms the foundation for a more detailed exploration of how these safety considerations interact and potentially conflict with the other four key principles outlined in the AI Act: traceability, transparency, environmental sustainability, and non-discrimination. As depicted in Figure 4, we identify potential attacks at each stage of the AI development pipeline. These attacks stem from the typical Tactics, Techniques, and Procedures (TTPs) employed by hackers. Following this, we outline defensive objectives for the establishment of secure and reliable AI systems.

-*Common TTPs:* Tactics, Techniques, and Procedures (TTPs) are integral concepts in cybersecurity, utilized to categorize and understand attacker behaviors. They serve as tools for testing and assessing the robustness of an organization's threat detection capabilities. Breaking down these concepts: (1) Tactics refer to the overarching goals or 'why' behind
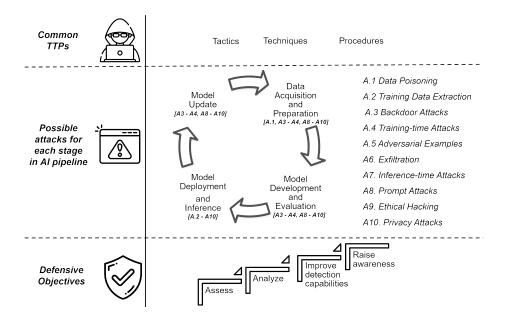
Figure 4: An overview of safe AI Systems.

an attack, explaining the motives for executing certain actions; (2) Techniques detail the 'how' aspect, outlining the methods adversaries employ to achieve their tactical objectives; (3) Procedures are the specific, detailed methods used to execute these techniques.

There's a concerted effort within the AI security community to compile and comprehend the TTPs that adversaries might deploy against AI systems. A notable contribution in this domain comes from MITRE (ATT&CK, 2023b), renowned for their MITRE ATT&CK TTP framework. They have extended their expertise to the AI field by publishing a comprehensive set of TTPs tailored for AI systems (ATT&CK, 2023a). This work by MITRE is instrumental in aiding organizations to better understand and prepare for potential threats against their AI infrastructures, reflecting the growing importance of specialized TTPs in the evolving landscape of AI security.

-*Possible attacks for each stage in AI pipeline:* Leveraging the knowledge in threat intelligence and AI system development, we have identified specific TTPs, leading to potential attacks that are most relevant and realistic for real-world adversaries. These potential attacks encompass a range of strategies, such as prompt attacks, extracting training data, backdoor attacks, adversarial attacks, data poisoning, exfiltration, and ethical hacking. The pipeline shown in Figure 4 encompasses stages such as data collection and preparation, model creation and assessment, and the deployment and updating of models. This structured approach aids in comprehending specific vulnerabilities and developing appropriate countermeasures for each phase of AI system development, taking into account the safety considerations emphasized in the AI Act. We've organized possible attacks in a structured table as shown in Table 1, correlating each attack type with its respective phase in the AI development pipeline. In terms of their mechanisms and objectives, several of these attacks can overlap. For example, training-time attacks could encompass methods like data poisoning and backdoor attacks. Adversarial examples can be a form of both inference-time

36

attacks and exfiltration if used to extract model behavior or data. Training data extraction and privacy attacks may involve unauthorized access to sensitive data.

Table 1: Potential Attacks Across Different Stages of the AI Pipeline Ensuring Safety

Potential Attacks	Explanation	High risk stages	Related and latest surveys/ papers
Data Poisoning	Deliberate manipulation of training data to corrupt model learning, causing incorrect predictions. Common methods include injecting malicious data or altering existing data points.	Data Acquisition & Preparation	(Cin'à', Grosse, Demontis, Vascon, Zellinger, Moser, Oprea, Biggio, Pelillo, & Roli, 2023)(Wang, Ma, Wang, Hu, Qin, & Ren, 2022)
Training Data Extraction	Unauthorized extraction or reconstruction of sensitive information from the training dataset used in model development.	Model Deployment & Inference	(Carlini, Tramer, Wallace, Jagielski, Herbert-Voss, Lee, Roberts, Brown, Song, Erlingsson, et al., 2021)
Training-time Attacks in GenAI	Attacks targeting pre-training and fine-tuning stages of GenAI models by manipulating portions of training data to cause specific model failures.	All stages	(Vassilev, Oprea, Fordyce, & Andersen, 2024)
Backdoor Attack	Insertion of hidden triggers into AI models that cause incorrect behavior only when specific inputs are present, while maintaining normal performance otherwise.	All stages	(Gao, Doan, Zhang, Ma, Zhang, Fu, Nepal, & Kim, 2020)(Li, Zhang, Wang, & Song, 2023)
Adversarial Examples Attack	Specially crafted inputs that cause AI models to produce incorrect outputs despite appearing normal to humans. Includes white-box, black-box, and transferability attacks.	Model Deployment & Inference	(Chakraborty, Alam, Dey, Chattopadhyay, & Mukhopadhyay, 2021)(Akhtar, Mian, Kardan, & Shah, 2021)(Zhang, Sheng, Alhazmi, & Li, 2020)
Exfiltration	Theft of sensitive data, proprietary algorithms, or intellectual property from AI models through unauthorized access or exploitation.	Model Deployment & Inference	(Taori, Gulrajani, Zhang, Dubois, Li, Guestrin, Liang, & Hashimoto, 2023)(Chung, Yang, Wang, Cento, Jerath, Raman, Lie, & Chignell, 2023)
Inference-time attacks in GenAI	Exploitation of vulnerabilities in deployed LLMs and RAG applications during the model's operational phase.	Model Deployment & Inference	(Dong, Zhou, Yang, Shao, & Qiao, 2024)
Prompt Attacks	Manipulation of LLM inputs through carefully crafted prompts to exploit model vulnerabilities and induce unintended behaviors.	All stages	(Shi, Li, Yin, Han, Zhou, & Liu, 2022)
Privacy Attacks	Unauthorized extraction of sensitive information through techniques like membership inference, model extraction, and property inference attacks.	All stages	(Rigaki & Garcia, 2023)
Ethical Hacking	Authorized security testing of AI systems to identify and address vulnerabilities before malicious exploitation.	All stages	(Yaacoub, Noura, Salman, & Chehab, 2021)

-*Defensive objectives:* To enhance the safety of AI systems, we summarize four strategic objectives to defend against threat actors and attacks, inspired by considering the TTPs used by adversaries targeting various stages of the AI pipeline. The first objective involves evaluating the impact of attacks on both users and products, and determining methods to bolster cyber resilience against such incursions. The second objective focuses on examining the AI system's robustness, particularly in terms of detecting and thwarting potential attacks, as well as understanding how these attacks might circumvent existing safeguards. The third objective is to enhance the system's ability to detect threats. Finally, the fourth

objective is to cultivate awareness among stakeholders. This is aimed at assisting AI system developers in recognizing risks and potential attacks at each stage of the AI pipeline, and in advocating for a risk-driven, well-informed approach to investing in security measures for AI systems for organizations.

3.1.2 Techniques related to safety and their relations to other aspects of the AI Act

Below, we explore methods to improve the security of AI systems in response to the attacks outlined in Table 1. Additionally, we examine how each method impacts the four principles outlined in the AI Act: transparency, non-discrimination, environmental sustainability, and traceability.

Adversarial training: Adversarial training, introduced by Goodfellow et al. (Hamon et al., 2020) and further developed by Madry et al. (Madry, Makelov, Schmidt, Tsipras, & Vladu, 2017), enhances model safety by incorporating adversarial examples during the training process. This method improves resilience against attacks and increases the model's semantic content (Tsipras, Santurkar, Engstrom, Turner, & Madry, 2018), thereby boosting interpretability and predictability. These features are crucial for safety, allowing better anticipation and mitigation of potential failures. However, adversarial training often comes with trade-offs, including reduced accuracy on clean data and increased computational costs.

Formal verification employs methods from formal methods, such as Satisfiability Modulo Theories (SMT) solvers and abstract interpretation, to certify the robustness of neural networks against adversarial inputs (Katz, Barrett, Dill, Julian, & Kochenderfer, 2017) (Gehr, Mirman, Drachsler-Cohen, Tsankov, Chaudhuri, & Vechev, 2018). This mathematically rigorous approach enhances transparency by providing a solid foundation for understanding and documenting model behavior under adversarial conditions. While formal verification currently faces limitations in scalability and computational efficiency, it offers significant potential for comprehensive robustness certification. This rigorous verification is essential for developing transparent AI systems, as it allows for a deeper understanding of how and why a model is robust against certain types of adversarial inputs.

Certified adversarial robustness: Certified adversarial robustness is crucial for AI system safety and transparency, ensuring predictable and reliable behavior under adversarial inputs. Randomized smoothing enhances classifier robustness by integrating Gaussian noise into input data (Lecuyer, Atlidakis, Geambasu, Hsu, & Jana, 2019)(Cohen, Rosenfeld, & Kolter, 2019), generating predictions most likely correct under noise perturbations. This approach yields provably robust classifiers, particularly against ℓ_2-norm attacks (Carlini et al., 2021), enhancing system predictability and transparency. The method offers certified robustness for a subset of testing samples, providing quantifiable metrics to track model performance under adversarial conditions. Recent advancements have extended this concept to more complex perturbations, integrating denoising diffusion probabilistic models with high-accuracy classifiers, broadening the scope of transparency across various adversarial scenarios (Carlini et al., 2021).

Training data sanitization: Training data sanitization enhances AI system safety by mitigating poisoned samples' impact. These methods identify and remove malicious samples before training. Techniques like the Region of Non-Interest (RONI) method enhance model

reliability by excluding poisoned samples (Nelson, Barreno, Chi, Joseph, Rubinstein, Saini, Sutton, Tygar, & Xia, 2008). Label cleaning methods target label flipping attacks (Paudice, Mu'ñ'oz Gonz'á'lez, & Lupu, 2019), while outlier detection (Steinhardt, Koh, & Liang, 2017) and clustering methods (Laishram & Phoha, 2016)(Taheri, Javidan, Shojafar, Pooranian, Miri, & Conti, 2020) remove anomalies. Ensemble variance computation helps identify potential poisoning attempts (Venkatesan, Sikka, Izmailov, Chadha, Oprea, & De Lucia, 2021). Cybersecurity mechanisms for provenance and integrity attestation add further protection. Overall, data sanitization improves AI transparency by preserving training data authenticity, enhancing system safety and decision-making reliability. **Robust training:** Robust training enhances the transparency and safety of AI systems by integrating techniques like ensemble model voting, robust optimization, and randomized smoothing (Levine & Feizi, 2020)(Wang, Levine, & Feizi, 2022). Ensemble models allow for comparative analysis of individual model decisions, improving insight into the decision-making process. Robust optimization, through methods like trimmed loss functions, minimizes the impact of extreme data points, making the model's reasoning more transparency and less prone to data poisoning (Jagielski, Oprea, Biggio, Liu, Nita-Rotaru, & Li, 2018)(Diakonikolas, Kamath, Kane, Li, Steinhardt, & Stewart, 2019). Randomized smoothing adds stability to model predictions against minor input perturbations, offering a clearer understanding of how inputs lead to outputs and safeguarding against subtle adversarial attacks (Rosenfeld, Winston, Ravikumar, & Kolter, 2020). Collectively, these methods make AI systems' decision-making processes more discernible and reliable, fortifying them against manipulative data inputs.

Machine unlearning: Machine unlearning represents a novel and distinctive approach aimed at mitigating privacy concerns in machine learning by allowing users to request the removal of their data from trained models, thus potentially enhancing privacy, fairness, and trust in AI systems. This technique comes in two primary forms: exact unlearning, which involves retraining the model from scratch or from a certain checkpoint to ensure complete removal of the data's influence, and approximate unlearning, where the model's parameters are adjusted to diminish the influence of the data intended to be forgotten (Bourtoule et al., 2021)(Cao & Yang, 2015). Although promising, the implementation of machine unlearning, whether exact or approximate, introduces complexities related to ensuring the integrity and performance of the model post-unlearning, alongside considerations of computational efficiency and the potential environmental impact of the retraining or updating process (Ginart, Guan, Valiant, & Zou, 2019), (Izzo et al., 2021), (Neel, Roth, & Sharifi-Malvajerdi, 2021). As such, while machine unlearning is a forward step towards user-centric, privacy-preserving AI, it necessitates meticulous design and management to balance its benefits against the operational and ethical challenges it poses.

Differential privacy: Differential Privacy (DP) enhances AI system safety by protecting sensitive training data (Dwork, 2006). It limits an attacker's ability to gain insights about individual data points from AI outputs. DP has evolved to include approximate DP and Rényi DP (Mironov, Talwar, & Zhang, 2019). The primary algorithm for training machine learning models is DP-SGD (Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, & Zhang, 2016), with recent enhancements like DP-FTRL (Kairouz et al., 2021a) and DP matrix factorization (Demontis, Melis, Pintor, Jagielski, Biggio, Oprea, Nita-Rotaru, & Roli, 2019). DP protects against data reconstruction and membership inference attacks, with tight bounds derived by Thudi et al. (Thudi, Shumailov, Boenisch, & Papernot, 2022).

Table 2: Mechanism/approaches influencing AI safety and the discussion on safety versus other four aspects("-": negative impact, "+": positive impact).

Papers	Approaches Influencing AI Safety	Mitigations towards Potential Attacks	Impact	Other Aspects' Impacts in AI Regulation
(Goodfellow, Shlens, & Szegedy, 2014; Madry et al., 2017; Tsipras et al., 2018)	Adversarial training	Adversarial examples	The outcomes of adversarial training yield models with greater semantic significance compared to conventional models.	Transparency (+)
(Lecuyer et al., 2019; Cohen et al., 2019; Katz et al., 2017; Gehr et al., 2018)	Certified robustness	Adversarial examples	By providing a clear understanding of how the model processes data and makes decisions, thereby allowing for the creation of robustness guarantees and the identification of model vulnerabilities.	Transparency (+)
(Paudice et al., 2019; Steinhardt et al., 2017; Taheri et al., 2020; Venkatesan et al., 2021)	Training data sanitization	Data poisoning, backdoor	Before initiating machine learning training, ensure the training set is purified by eliminating any tainted samples.	Transparency (+)
(Wang et al., 2022; Jagielski et al., 2018; Diakonikolas et al., 2019; Rosenfeld et al., 2020)	Robust training	Data poisoning, backdoor	Adjusting the machine learning training algorithm to conduct resilient training rather than the standard approach.	Transparency (+)
(Kairouz, McMahan, Song, Thakkar, Thakurta, & Xu, 2021a; Chaudhari, Abascal, Oprea, Jagielski, Tramer, & Ullman, 2023; Mahloujifar, Ghosh, & Chase, 2022)	Differential privacy	Privacy attacks, training data extraction	Enhancing privacy protections for sensitive data used in AI training, testing, and deployment processes.	Fairness (+), transparency (+/-), traceability (+)
(Jagielski, Ullman, & Oprea, 2020; Zanella-Béguelin, Wutschitz, Tople, Salem, Rúhle, Paverd, Naseri, Kópf, & Jones, 2023; Nasr, Songi, Thakurta, Papernot, & Carlin, 2021)	Empirical privacy auditing	Privacy attacks, training data extraction	Empirically assessing an algorithm's privacy guarantees and establishing lower bounds through privacy attacks.	Fairness (+), transparency (+/-), traceability (+)
(Bourtoule, Chandrasekaran, Choquette-Choo, Jia, Travers, Zhang, Lie, & Papernot, 2021; Izzo, Smart, Chaudhuri, & Zou, 2021)	Machine unlearning	Privacy attacks, training data extraction	Enhancing user privacy by ensuring that all data that a user wishes to be forgotten from a model is effectively removed.	Environmental-sustainability (-), fairness (+)
(Ji, Qiu, Chen, Zhang, Lou, Wang, Duan, He, Zhou, Zhang, et al., 2023; Greshake, Abdelnabi, Mishra, Endres, Holz, & Fritz, 2023)	Stricter forward alignment	Prompt attacks	Developing integrated mechanisms through training for enhanced forward alignment and continuously refining via reinforcement learning informed by human feedback.	Fairness (+/-), transparency (+)
(Liu, Deng, Li, Wang, Zhang, Liu, Wang, Zheng, & Liu, 2023)	Prompt instruction	Prompt attacks	Prompting the model to handle user input with caution	Fairness (+/-)
(Ji et al., 2023; Liu et al., 2023)	Stricter backward alignment	Prompt attacks	Developing built-in safeguards by enhancing backward alignment through training, using tailored benchmark datasets or filters to oversee the input and output of a secure LLM.	Transparency (+)
(Kaur et al., 2022; R'ú'hr, Berger, & Hess, 2023; OpenAI, 2023)	Interpretability based solution	Prompt attacks	Making the models' decision-making processes more comprehensible to humans	Transparency (+)

However, it doesn't fully protect against model extraction or property inference attacks (Chaudhari et al., 2023)(Mahloujifar et al., 2022).

DP contributes to fairness by ensuring equal privacy protection across demographic groups (Farrand, Mireshghallah, Singh, & Trask, 2020). It enhances transparency in data handling practices (Dwork, McSherry, Nissim, & Smith, 2006; Gong, 2022) and improves traceability of data usage. However, trade-offs between privacy and utility must be carefully considered in specific applications.

Empirical privacy auditing: Implementing Differential Privacy (DP) faces challenges in balancing privacy and utility (Ponomareva, Hazimeh, Kurakin, Xu, Denison, McMahan, Vassilvitskii, Chien, & Thakurta, 2023). Worst-case analysis often leads to higher privacy parameters in practice, as seen in the 2020 U.S. Census (Vassilev et al., 2024). Privacy auditing, initiated by Jagielski et al. (Jagielski et al., 2020), aims to assess actual privacy levels through threat simulations. While membership inference attacks (Jayaraman & Evans, 2019)(Zanella-Béguelin et al., 2023) are used, poisoning attacks are more effective at revealing privacy erosion (Jagielski et al., 2020)(Nasr et al., 2021).

Recent advancements include improved precision for the Gaussian mechanism (Nasr, Hayes, Steinke, Balle, Tramér, Jagielski, Carlini, & Terzis, 2023b) and techniques reducing large sample size requirements (Pillutla, Andrew, Kairouz, McMahan, Oprea, & Oh, 2023). Efficient privacy auditing approaches have been introduced by Steinke et al. (Steinke, Nasr, & Jagielski, 2023) using random data canaries, and Andrew et al. (Andrew, Kairouz, Oh, Oprea, McMahan, & Suriyakumar, 2023) employing random client canaries for federated learning privacy evaluation.

Stricter forward alignment: Prompt injection happens when a user deliberately inputs text to modify the behavior of a Large Language Model (LLM), aiming to circumvent safeguards for producing misinformation, offensive content, or extracting private information (Vassilev et al., 2024). To counteract these attacks, model developers are implementing stronger built-in protections by enhancing forward alignment through training on selected, pre-aligned datasets and further refining the models via Reinforcement Learning informed by Human Feedback (RLHF) (Greshake et al., 2023)(Ji et al., 2023). This method of training, which indirectly incorporates human judgment to fine-tune models, helps ensure that LLMs are more aligned with human values and less prone to undesirable outputs. OpenAI's GPT-4, for instance, underwent fine-tuning with RLHF, resulting in a reduced propensity for generating harmful or inaccurate content (OpenAI, 2023).

Prompt instruction for LLM: LLM instructions can guide the model in handling user inputs with caution. By adding detailed instructions to the input prompt, models can be prepped to recognize and respond appropriately to content that might lead to a jailbreak, where the model's behavior deviates from intended safeguards (Chang, Li, Liu, Wang, Wang, & Liu, 2024). Strategically placing user input before these instructions can leverage the model's tendency to prioritize recent instructions, enhancing adherence to safety protocols (Liu et al., 2023). Furthermore, surrounding the instructions with random characters or specific HTML tags can signal to the model the distinction between system directives and user-provided prompts. This method helps delineate the boundaries between what the model perceives as commands to follow versus input to analyze, thereby improving the model's ability to navigate and process inputs more securely and effectively. Well-designed prompt instructions that prioritize fairness, sensitivity to context, and inclusive language

can contribute to increased fairness in LLMs. However, it's essential to carefully consider the design and implementation of prompt instructions to avoid unintended consequences that could decrease fairness.

Stricter backward alignment: Model providers are increasingly developing and integrating robust mechanisms by focusing on enhanced backward alignment. This involves rigorous training and evaluation processes using benchmark datasets specifically designed for this purpose or through filters that carefully monitor both the inputs received and outputs generated by a secured LLM. A notable strategy involves assessing the responses of LLMs to various prompts, which can help in identifying prompts that are potentially adversarial in nature, thereby preventing misuse of the model (Ji et al., 2023).

Moreover, the emergence of commercial solutions offering tools to detect and mitigate prompt injection attacks represents a significant advancement in safeguarding LLMs, such as Arthur [1], Lakera [2] and Aporia [3]. These tools are designed to spot potentially harmful user inputs and moderate outputs to prevent the model from engaging in jailbreak behavior, where it acts outside its intended operational boundaries. By implementing such measures, companies are embracing a defense-in-depth approach, layering multiple security mechanisms to provide comprehensive protection for LLMs. This multifaceted strategy not only enhances the immediate security of these models but also contributes to a broader assurance of safety and transparency in their deployment and use.

Interpretability-based solutions: Interpretability-based solutions targeting attacks on LLMs involve using outlier detection techniques to monitor the models' prediction trajectories. This approach is grounded in the observation that how a model processes and predicts outcomes based on inputs can reveal a lot about the nature of those inputs. Specifically, researchers have found that by closely examining the "prediction trajectory" - the series of predictions an LLM makes as it processes input-it's possible to identify inputs that are anomalous or potentially malicious (Greshake et al., 2023).

The concept of a "tuned lens" refers to a method or tool designed to scrutinize the model's prediction path. When this lens is applied to analyze how an LLM responds to various inputs, it can highlight when the model's predictions deviate significantly from what is expected under normal circumstances. Anomalous inputs, such as those designed to manipulate the model or trigger undesired behavior, often lead to unusual prediction trajectories. By detecting these outliers, interpretability-based solutions can effectively flag potentially harmful inputs before they impact the model's output, enhancing the security and reliability of LLMs in handling a wide range of data (Belrose, Furman, Smith, Halawi, Ostrovsky, McKinney, Biderman, & Steinhardt, 2023).

---

1. https://www.arthur.ai/product/shield
2. https://www.lakera.ai/
3. https://www.aporia.com/
4. https://llama.meta.com/
5. https://openai.com/chatgpt/
6. https://claude.ai/
7. https://copilot.microsoft.com/
8. https://x.ai/
9. https://deepmind.google/technologies/gemini/pro/

Table 3: Comparison of transparency regarding information disclosure in leading AI tools

Organi-zation	Model	Official model card or technical port	API/ SDK re-provided	Other details
Meta	LLaMA 3[4]	Y	Y	LaMA 3, available in 8B and 70B configurations, is trained on publicly available data. The official model card discloses some details about the model architecture and training.
OpenAI	ChatGPT 4[5]	Y	Y	ChatGPT 4 is now accessible with the option to download model weights and a technical report. Although the overall structure and instructions for use are provided, detailed information about the training process is not revealed.
Anthropic	Claude 3[6]	Y	Y	Claude 3 has unveiled its model weights and a technical report. The official documentation outlines details about the model training, yet specifics regarding the data sources, training process, and architecture are somewhat less detailed.
Microsoft	Copilot[7]	Y	Y	Copilot comes with official documentation and an open API. However, despite these resources, detailed information about the model's architecture and specific training details are kept under wraps.
xAI	Grok 1.5[8]	N	N	Due to the recent release of Grok1.5, detailed information about the model is currently scarce, permitting only preliminary usage via web interfaces.
Google	Gemini 1.5 Pro[9]	Y	Y	Gemini 1.5 Pro has unveiled the fundamental information along with comparative test outcomes of the model. However, the details provided are still lacking for a comprehensive grasp of the architecture and training methods.

3.2 Transparency

3.2.1 Overview

AI transparency is crucial for upholding the integrity and comprehensibility of AI systems (Balasubramaniam, Kauppinen, Rannisto, Hiekkanen, & Kujala, 2023; Cao & Yousefzadeh, 2023). It involves straightforward communication about the functions, decision-making processes, and constraints of these systems. This level of openness enables users to understand how decisions are made and help build trust. Furthermore, AI transparency includes explainability, which requires providing clear explanations of how AI processes inputs to produce outputs. This aspect is essential for enhancing user confidence and ensuring that AI operations comply with regulations designed to protect rights and prevent biases. In this subsection, we explore transparency in AI by discussing the roles of explainability and information disclosure.

- *Towards transparency through explainability:* AI explainability is crucial for understanding how AI models make decisions, ensuring transparency, accountability, and trust in AI systems. Several techniques can help achieve AI transparency by providing clear, understandable explanations for AI model behavior. Regarding *model interpretability methods*, a model that is interpretable, such as a decision tree, a linear regression, or a rule-based system, is transparent, and the model itself is intrinsically understandable (Carvalho, Pereira, & Cardoso, 2019). Techniques such as the SHapley Additive exPlanations (SHAP) and

Local Interpretable Model-agnostic Explanations (LIME) offer insights into which features are most important for the model's predictions (Meza Mart'í'nez, Nadj, Langner, Toreini, & Maedche, 2023). These *feature importance techniques* can help trace back the decision-making process by highlighting the contribution of each input feature to the final decision. Some works utilized methods, such as counterfactual explanations (Cheng, Ming, & Qu, 2020) and prototypes (Van Looveren & Klaise, 2021), providing insights by presenting specific examples the model uses for making decisions. Specifically, counterfactual explanations show how slight changes to the input can lead to different predictions, helping trace the decision boundaries of the model. Prototypes, on the other hand, identify representative examples from the dataset that are influential in the model's learning. In addition, *visualizations* can offer intuitive insights into complex models, especially deep learning models (Samek, Wiegand, & M'ú'ller, 2017). Techniques like saliency maps, which highlight the parts of the input (e.g., pixels in an image) that are most influential for a prediction, can help trace the model's focus and reasoning process (Ghariba, Shehata, & McGuire, 2019). Techniques that *decompose model predictions* into understandable components. For instance, Layer-wise Relevance Propagation (LRP) decomposes the output decision back to the input features to identify the relevance of each feature in the decision-making process (Lapuschkin, 2019). Lastly, tools and platforms allow users to interact with AI models, query them with different inputs and observe the outputs (Sun, Lin, Qiu, & Rimba, 2022). This hands-on approach can help users understand the model's behavior in different scenarios and trace its decision-making process more practically and intuitively.

-*Explainability for Large Language Models:* The techniques for enhancing Large Language Models (LLMs) explainability are categorized into two main paradigms: traditional fine-tuning methods and prompting-based methods (Zhao, Chen, Yang, Liu, Deng, Cai, Wang, Yin, & Du, 2023). Traditional fine-tuning methods encompass various strategies to enhance the interpretability and understanding of LLMs. *Feature attribution methods* form one facet of this approach, discerning the significance of different parts of input data, such as words or tokens within a sentence, for the model's predictions. Techniques like Integrated Gradients (Enguehard, 2023) utilize gradients to gauge the importance of each input feature by tracing their impact on the model's output. *Surrogate models* offer another avenue, employing simpler and more interpretable models to approximate the intricate behavior of complex LLMs (Chen & Ji, 2022). While they provide insights into decision-making processes, they may not capture the full complexity of the original model. *Representation analysis* delves into understanding the learned representations of the model, often through visualization or clustering of embeddings to identify patterns or similarities captured by the model.

Prompting-based methods leverage the generative capabilities of LLMs for explanation generation (Zhao et al., 2023). By prompting the model to articulate reasons for its decisions, this method sheds light on its *underlying reasoning*. *Counterfactual explanations* involve altering inputs slightly to observe changes in the model's predictions, elucidating crucial aspects of the decision-making process (Treviso, Ross, Guerreiro, & Martins, 2023). *Prompt engineering* involves crafting prompts to elicit more interpretable or explainable responses from the model. These prompts may encourage the model to articulate its reasoning process or highlight key factors influencing its decisions, contributing to a deeper

understanding of the model's behavior (Zou, Phan, Chen, Campbell, Guo, Ren, Pan, Yin, Mazeika, Dombrowski, et al., 2023).

-*Information disclosure:* Advances in generative AI have led to AI-generated media that is nearly indistinguishable from human-created content. This has underscored the need for standardized accountability reports as a valuable measure to improve the transparency of AI systems across the industry (Ali, Venkatraj, Morosoli, Naudts, Helberger, & Cesar, 2024). According to the GDPR (of the European Union, 2023), providers of AI systems interacting with humans must ensure that people are informed of their interaction with AI unless it's obvious from the context. Moreover, AI systems generating or manipulating content resembling real people, objects, or events (i.e., deepfakes) should explicitly disclose that the content is artificially created or manipulated.

A summary of detailed information concerning information disclosure in prominent AI systems is presented in Table 3. The extent of disclosed information regarding the training dataset, model architecture, test outcomes, and human interaction varies at the present stage. Although advanced AI tools like GPT-4 have published detailed technical reports, OpenAI has withheld specific information on training and datasets, citing the need to protect product security and maintain competitiveness. In contrast, Meta's LLaMA 3 is considered relatively transparent, providing details on architecture, parameter count, and some training procedure information in its Model Card. Despite this, current leading AI tools remain insufficiently transparent regarding automated data processing and automated decision-making. Moreover, while most state-of-the-art AI tools offer corresponding APIs or SDKs to integrate them into local programming environments, understanding the intricacies of these models is challenging for researchers and users due to their complexity and encapsulation. As a result, achieving transparency in AI systems remains an elusive goal.

3.2.2 Transparency versus safety, Environmental sustainability, non-discrimination, and traceability

While transparency strengthens traceability and supports non-discrimination by uncovering biases, it can conflict with safety and privacy concerns. Additionally, it promotes environmental sustainability but may reveal operational details. Thus, the ensuing discussion will elucidate the delicate balance required between transparency and the other principles to navigate the trade-offs and harmonize ethical AI deployment. As summarized in Table 4, the methods to augment transparency are categorized (Dwivedi, Dave, Naik, Singhal, Omer, Patel, Qian, Wen, Shah, Morgan, et al., 2023) into five categories: Dimension Reduction and Visualization, Feature-based Techniques, Example-based Techniques, Distributed Learning, and Pre-existing Model-based Methods. This subsection explores and discusses in more detail the approaches that can influence AI transparency under the six categories shown in Table 4.

Dimension reduction and visualization: These techniques are essential for achieving data understanding and interpretability, thereby promoting transparent AI models. They simplify the complexity of large datasets, providing clearer insights and more transparent data narratives that can be integrated into AI algorithms for decision-making. For instance, Principal Component Analysis (PCA) (Shlens, 2014) and Linear Discriminant Analysis (LDA) (Tharwat et al., 2017) can reduce high-dimensional features to low-dimensional

Table 4: Mechanism/approaches influencing AI transparency and the discussion on transparency versus safety, non-discrimination, environmental sustainability, and traceability ("-": negative impact, "+": positive impact).

Papers	Category	Methods	Impact	Discussion with the Rest Aspects in AI Regulation
(Nakanishi, 2024; Tharwat, Gaber, Ibrahim, & Hassanien, 2017)	Dimension reduction & visualization	PCA, LDA	Visualizes feature distributions and their influence on model decisions, helping to understand model behavior and detect potential biases.	Traceability (+), non-discrimination (+), environmental sustainability (+).
(Van der Maaten & Hinton, 2008; McInnes, Healy, & Melville, 2018; Wexler, Pushkarna, Bolukbasi, Wattenberg, Viʻéʻgas, & Wilson, 2019)		t-SNE, UMAP What-If		
(Goldstein, Kapelner, Bleich, & Pitkin, 2015; Ribeiro, Singh, & Guestrin, 2018; Lundberg & Lee, 2017; Nori, Jenkins, Koch, & Caruana, 2019)	Feature based methods	ICE, LIME, SHAP, InterpretML	Explains prediction outcomes by analyzing feature importance and their individual contributions to model decisions.	Traceability (+), non-discrimination (+).
(Dwivedi et al., 2023; Dhurandhar, Chen, Luss, Tu, Ting, Shanmugam, & Das, 2018; Garcʻíʻa & Aznarte, 2020; Mothilal, Sharma, & Tan, 2020; Fisher, Rudin, & Dominici, 2019; Wexler et al., 2019)	Example based methods	Anchors, CEM, KT SHAP, Dice, Alibi, What-If	Demonstrates how specific changes in input conditions affect model predictions, making decision processes more interpretable.	Traceability (+), non-discrimination (+).
(Warnat-Herresthal, Schultze, Shastry, Manamohan, Mukherjee, Garg, Sarveswara, Hʻáʻndler, Pickkers, Aziz, et al., 2021; Kairouz, McMahan, Avent, Bellet, Bennis, Bhagoji, Bonawitz, Charles, Cormode, Cummings, et al., 2021b)	Distributed learning methods	Federated Learning, Swam Learning	Enables transparent model behavior adjustment at each network node while maintaining data privacy through limited sharing.	Safety (-+), non-discrimination (+), environmental sustainability (-).
(Bourtoule et al., 2021)	Pre-existing model based methods	Machine Unlearning	Updates model behavior by modifying features or removing specific data influences.	Non-discrimination (+), environmental sustainability (+).

spaces for feature space visualization, roughly explaining a model's decisions. Although these transformations may obscure the original features' meanings, the importance of features in model decisions can still be prioritized based on their similarity to PCA loadings (Shlens, 2014). Visualization techniques like t-SNE and UMAP offer a visual understanding of the model's decision boundaries and the distribution of the training data (Van der Maaten & Hinton, 2008; McInnes et al., 2018; Wexler et al., 2019).

These techniques also positively impact the development of non-discriminative and environmentally sustainable AI with enhanced traceability. Specifically, PCA and LDA can unintentionally amplify biases if the principal components or discriminants used for dimensionality reduction capture biases inherent in the data (Shlens, 2014; Tharwat et al., 2017). In contrast, t-SNE (Van der Maaten & Hinton, 2008) and UMAP (McInnes et al., 2018) preserve local structures and neighborhood relationships, helping to reveal and mitigate biases to ensure equitable treatment of all individuals and groups. Regarding environmental sustainability, dimensionality reduction with PCA and LDA can make model training

more computationally efficient, leading to lower energy consumption and a smaller carbon footprint. Visualization techniques like t-SNE and UMAP can also reduce computational costs by providing insights that lead to more efficient model training and fewer iterations. Additionally, t-SNE and UMAP maintain relationships between data points in a lower-dimensional space, facilitating the traceability of how groups of data points are related to each other and potential biases, thus enhancing the traceability of decisions within the AI system.

Feature-based methods: Deployed during the Deployment and Maintenance stage, these techniques elucidate the contribution of individual features to the outcomes of AI models. They allow for the dissection of model decisions, facilitating an understanding of how inputs are transformed into outputs. Feature-based methods such as ICE (i.e., Individual Conditional Expectation) (Goldstein et al., 2015), LIME (i.e., Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2018), SHAP (i.e., Shapley Additive explanations) (Lundberg & Lee, 2017), and Interpret-ML (Nori et al., 2019) significantly influence factors like non-discrimination and traceability in the realm of AI transparency.

These methods enhance non-discrimination by providing granular insights into model predictions. ICE plots offer a microscopic view of how changing one feature affects the prediction, which can reveal discriminatory patterns against certain groups if predictions change unfavorably for specific feature ranges. LIME focuses on local fidelity, giving interpretable explanations for individual predictions, which can be crucial in pinpointing and addressing instances where the model might be unfairly biased against certain samples. SHAP values quantify the contribution of each feature to every prediction, providing a detailed breakdown that can uncover feature-related biases in model behavior. Together, these methods enable the identification and correction of discriminatory aspects within models, leading to more equitable AI systems. Meanwhile, ICE, LIME, and SHAP are especially effective in improving traceability within AI systems. ICE plots enable a clear understanding of the model's behavior with respect to individual features, providing traceable insights. LIME facilitates traceability by offering interpretable and locally accurate explanations for individual predictions, allowing one to follow the logic behind the model's output. SHAP extends this by assigning each feature an importance value for each prediction, creating a transparent and traceable link between input features and their impacts on the output. This ability to trace back the decision-making process to specific model inputs and their interactions significantly enhances the transparency of AI systems.

Example-based methods: Throughout the deployment and maintenance stage, these techniques focus on providing explanations for individual predictions made by AI systems. They support transparency by enabling stakeholders to query and receive understandable explanations for specific AI decisions. Example-based methods such as Anchors (Dwivedi et al., 2023), Contrastive Explanation Method (CEM) (Dhurandhar et al., 2018), Kernel Tree SHAP (Garc'í'a & Aznarte, 2020), Diverse counterfactual explanations (Dice) (Mothilal et al., 2020), and Alibi (Fisher et al., 2019) positively impact non-discrimination and traceability in the context of AI system transparency:

These methods provide detailed explanations based on instances, which can be pivotal in identifying and addressing discrimination within AI systems. For instance, Anchors offer explanations in the form of rules that are sufficient to guarantee the same prediction in similar cases, which can highlight discriminatory patterns if certain groups are consistently

associated with negative outcomes. Similarly, CEM can indicate if a model's predictions are adversely affected by certain demographics, and Dice provides counterfactual explanations that can demonstrate if and how outcomes could be equitable under different circumstances. By illustrating the specific conditions that lead to particular decisions, example-based methods enable stakeholders to detect and rectify biases, contributing to the non-discrimination of AI systems. Thus, traceability is significantly enhanced by example-based methods. They provide concrete instances that make the decision-making process of AI systems transparent and understandable. For example, Kernel and Tree SHAP calculate the contribution of each feature to a particular prediction, clarifying the decision-making process. Alibi offers diverse types of model explanations, including feature influence, which aids in tracing back the logic of the AI's conclusions. Dice creates counterfactual scenarios that can be traced to understand how different inputs would change the outcomes, thereby improving the system's transparency.

Distributed learning methods: Distributed learning methods offer decentralized AI solutions that align with inherently decentralized data structures and comply with data privacy and security regulations. These methods, like federated learning (Kairouz et al., 2021b) and swarm learning (Warnat-Herresthal et al., 2021), ensure that raw data does not need to be exchanged, maintaining proper transparency for all participants. Specifically, swarm learning provides secure, transparent, and fair onboarding for decentralized network members without requiring a central custodian (Warnat-Herresthal et al., 2021).

While improving AI system transparency, distributed learning methods have mixed impacts: they positively affect non-discrimination, negatively affect environmental sustainability, and have complex effects on safety. Regarding non-discrimination, collaborative fairness can be achieved by adjusting the performance of models allocated to each participant based on their contributions (Lyu, Xu, Wang, & Yu, 2020). However, distributed learning methods are not environmentally friendly in AI development. Qiu et al. (Qiu, Parcollet, Beutel, Topal, Mathur, & Lane, 2020) found that training time for distributed learning is significantly longer than for centralized training due to the lower computational capabilities of participants' devices. The impact on safety is multifaceted. A survey highlights the positive impact on data privacy through distributed learning in AI development but also points out vulnerabilities to various attacks, such as poisoning, backdoor, Generative Adversarial Network (GAN)-based attacks, and inference-based attacks (Gosselin, Vieu, Loukil, & Benoit, 2022).

Pre-existing model-based methods: Unlike the aforementioned post-hoc explainable AI methods, this category encompasses approaches that ensure AI model transparency through modifications or adjustments to the model's capabilities. Specifically, machine unlearning (Bourtoule et al., 2021) ensures the user's "right to be forgotten" by enabling the model to erase the knowledge associated with a specific user. As discussed in section 3.1.2, this method also has a positive effect on non-discrimination but positively impacts the environmental cost of AI.

3.3 Environmental Sustainability

3.3.1 Overview

The concept of environmental sustainability in AI systems emphasizes a balance between the environment, society, and economy (Van Wynsberghe, 2021), ensuring that AI innovations contribute positively to each area without causing adverse effects on the environment. Researchers recognize the significant carbon footprint and energy demands of extensive IoT networks and cloud/edge communications, advocating for an architectural and deployment approach that minimizes environmental impact (Wu et al., 2022a). The environmental impact of training a single extensive machine learning model, like Meena, can be equated to driving an average passenger vehicle for 242,231 miles (Patterson, Gonzalez, Le, Liang, Munguia, Rothchild, So, Texier, & Dean, 2021). However, this represents merely one facet of the broader picture. The environmental cost not only exists in the AI model training process but also in the AI system life cycle. As depicted in Figure 4, we explore potential environmental costs at each stage of the AI development pipeline. After that, we summarize the strategies applied to optimize the cost at each stage. Finally, we analyze the consequences of constructing an environmental sustainability AI system, focusing on its implications for safety, transparency, non-discrimination, and traceability.

The environmental costs mainly focus on power/energy consumption and carbon emissions caused by the usage of CPUs, GPUs, memory, and other system software and hardware. We specifically analyze those impact factors during data acquisition and preparation, model training, model inference, and AI system manufacturing and deployment. and summarize corresponding optimization strategies.

-*Data Acquisition and Preparation:* In this phase, energy consumption is predominantly driven by two processes: (1) the loading and validation split of large datasets, and (2) data preprocessing and feature extraction. The former necessitates downloading and storing training data on the model training server, a process that, although essential, is relatively less energy-intensive. Bouza et al. (Bouza, Bugeau, & Lannelongue, 2023) observed that loading a 6 GB Imagenet validation split accounted for merely 0.5% of the total energy consumption compared to model training. The latter process involves feature extraction and the weighting of individual features to assess the model's training efficacy, requiring significant computational resources. This stage is marked by exploring a wide array of machine-learning concepts on a large scale. Wu et al. (Wu et al., 2022a) reported that the energy expenditure for this process at Meta was half that of the model training phase.

-*Model Training:* Upon selecting an appropriate training solution, the AI model undergoes training using a more comprehensive set of production data, which is not only more current but also larger in volume and enriched with features. The carbon emissions associated with model training can be divided into two types: offline and online training emissions. Offline training leverages historical data for model training, whereas online training dynam-

---

10. www.green-algorithms.org
11. https://github.com/responsibleproblemsolving/energy-usage
12. https://github.com/sb-ai-lab/Eco2AI
13. https://github.com/lfwa/carbontracker
14. https://dl.acm.org/doi/abs/10.5555/3455716.3455964
15. https://mlco2.github.io/impact/
16. https://github.com/EPFLiGHT/cumulator

Table 5: Environmental cost measurement in CPU, GPU, and memory consumptions and emission efficiency. ('UNK'=unknown, 'avg'=average)

Tools	CPU Consumption	GPU Consumption	Memory Consumption	Emission Efficient
Green-Algorithm[10]	Energy: avg 12W/core if UNK; TDP. Usage Factor: 100% usage if UNK.	Energy: TDP; nvidia-smi; or avg 200W/GPU. Usage Factor: 100% usage if UNK	Energy: 0.3725W/GB of memory available. Usage Factor: all available/requested memory.	Location: not restricted. Static data: Carbon Foot-print; Electricity Maps. Default: 475gCO2eq/kWh
Code-carbon[11]	Energy: RAPL files or Power Gadget; TDP. Usage Factor: avg value 50%	Energy: pynvml library (NVIDIA GPUs only). Usage Factor: whole machine consumption	Energy: 0.3725W/GB of memory available. Usage Factor: allocated memory by 'process'	US & Canada: regional emissions per unit of power consumed; Other Countries: avg energy mix from Global Petrol Prices. Default: 475gCO2eq/kWh
Eco2AI[12]	Energy: avg 100W/core if UNK; TDP. Usage Factor: uses os & psutil python modules	Energy: pynvml library (NVIDIA GPUs only). Usage Factor: whole machine consumption	Energy: 0.3725W/GB of memory available. Usage Factor: allocated memory by 'process'	Each country's energy mix; Default: 436.5gCO2eq/kWh
Carbon-Tracker[13]	Energy: RAPL files. Usage Factor: whole machine RAPL value	Energy: pynvml library (NVIDIA GPUs only). Usage Factor: whole machine consumption	Energy: use RAPL files. Usage Factor: used by 'process' & others	Real-time data: Energi Data Service (Denmark); Carbon Intensity API (Great Britain). Static data: carbon-intensities.csv. Default: 475gCO2eq/kWh
EIT[14]	Energy: avg 100W/core if UNK; TDP. Usage Factor: uses os & psutil python modules	Energy: nvidia-smi (NVIDIA GPUs only). Usage Factor: 'nvidia-smi -q -x'	Energy: use RAPL files. Usage Factor: used by 'process' & others	Real-time data: California ISO. Static data: co2eq_parameters.json Default: 301gCO2eq/kWh
MLCO2[15]	None	Energy: TDP. Usage Factor: max load	None	Static data: the impact.csv file from cloud provider
Cumulator[16]	Energy: avg 250W/core if UNK; TDP. Usage Factor: None	Energy: avg 250W/GPU if UNK; TDP. Usage Factor: max load	None	Static data: Electricity Maps. Default: 447gCO2eq/kWh

ically updates the model parameters using the latest data. Wu et al. noted that, at Meta, the carbon emissions from online training were significantly lower than those from offline training (Wu et al., 2022a).

Additionally, the selection of batch sizes and epoch configurations influences energy consumption. Selecting an optimal batch size represents a balance between energy consumption and runtime (Bouza et al., 2023), with larger batch sizes offering faster runtime and reduced energy usage. Nonetheless, excessively large batch sizes can lead to inefficient energy consumption due to low usage rates. Therefore, the ideal batch size is the maximum size that ensures full utilization of all GPUs and CPUs. Similarly, epoch configuration compromises the model's quality and runtime. The duration of epochs and their associated energy consumption (or carbon emissions) remain constant (Anthony, Kanding, & Selvan, 2020). A larger number of epochs extends the runtime, while also enhancing the likelihood of achieving model convergence. The model structure configuration directly impacts the amount of the model's parameters.

Beyond the previously mentioned factors, checkpointing exerts a negligible effect, and the impact of the number of parameters on energy use and carbon emissions is uncertain. Bouza et al. (Bouza et al., 2023) identified an inconsequential difference in energy consumption between experiments conducted with a single checkpoint versus ten checkpoints. Typically, a larger number of parameters suggests extensive computational efforts during training. Nevertheless, actual energy consumption varies based on the specific model setup and training infrastructure employed. For example, training a Switch Transformer model, which boasts 1.5 trillion parameters (Fedus, Zoph, & Shazeer, 2022), results in considerably lower carbon emissions compared to the GPT-3 model, which contains 750 billion parameters (Mann, Ryder, Subbiah, Kaplan, Dhariwal, Neelakantan, Shyam, Sastry, Askell, Agarwal, et al., 2020).

-*AI System Manufacturing and Deployment:* The manufacturing and deployment of AI systems significantly affect carbon emissions and energy consumption. The impacts related to AI system manufacturing and deployment include the model's implementation program, the quantity and capabilities of processors running the program, the efficiency of data centers in power delivery and cooling, and the mix of energy sources (e.g., renewable, gas, coal) (Patterson et al., 2021). For instance, Meta's Transformer-based Universal Language Model for text translation at the inference stage served a lot of production traffic resulting in large energy consumption (Conneau, Khandelwal, Goyal, Chaudhary, Wenzek, Guzm'á'n, Grave, Ott, Zettlemoyer, & Stoyanov, 2019). Bouza et al. (Bouza et al., 2023) also highlight the impact of infrastructure size on power consumption through comparative experiments conducted on different platforms. For instance, training the Denoiser model on the larger Gemini-1 infrastructure (Grid5000 server) was completed in 2 hours, whereas it took 3 hours and 16 minutes on the smaller Rosenblatt infrastructure (MAP5 server). Furthermore, the CPU usage factor was significantly lower on Grid5000 (16%) compared to MAP5 (39%), and a similar trend was observed for GPU usage, with 14.3% on Grid5000 and 54% on MAP5, indicating higher efficiency in larger infrastructures. The energy-intensive nature of producing advanced computing components, coupled with the electricity demands of data centers hosting AI applications, underscores the environmental implications of AI's widespread adoption.

A denial-of-service (DoS) attack at the AI system deployment stage can generate sponge examples, significantly increasing energy consumption and the model's inference run time (Shumailov, Zhao, Bates, Papernot, Mullins, & Anderson, 2021). This attack exploits the energy-latency gap vulnerability in AI systems, where different inputs of the same size can cause a DNN to consume varying amounts of time and energy. This disparity is due to optimization strategies in hardware and AI algorithms. For instance, custom and semi-custom hardware typically leverage data sparsity and low-precision computations for DNN inference, reducing both arithmetic complexity and DRAM traffic to achieve better power efficiency (Chen, Yang, Emer, & Sze, 2019; Han, Liu, Mao, Pu, Pedram, Horowitz, & Dally, 2016; Zhao, Gao, Guo, Liu, Wang, Mullins, Cheung, Constantinides, & Xu, 2019). However, attackers can craft sponge examples by manipulating the token size of an NLP model's input and output sentences and the size of embedding spaces, leading to a substantial increase in algorithmic complexity and energy consumption. A proposed defense is to set a worst-case performance bound for some models and monitor their usage (Shumailov et al., 2021).

Table 5 summarized a list of tools (Bouza et al., 2023) used to measure CPU, GPU, and memory consumption and the emission efficiency related to the AI system's manufacturing and deployment. Users can monitor their AI system's environmental cost in its development and deployment to determine which stage needs to be improved to achieve environmental sustainability.

-*Environmental cost optimizations:* In the realm of AI, environmental cost optimization encompasses several pivotal areas, including algorithmic enhancements, processor improvements, data center efficiency, and the strategic mix of energy sources. A notable advancement in algorithm development is illustrated by the Evolved Transformer (Medium) model, which through neural architecture search, operates with significantly lower computational demands-1.6 times fewer FLOPS and reduced processing time by 1.1 to 1.3 times compared to its predecessor, the Transformer (Big), while simultaneously achieving a marginal increase in accuracy (Patterson et al., 2021). This progress underscores the potential for algorithmic and program improvements to contribute directly to environmental sustainability by reducing the energy footprint of AI operations.

Further optimizations are evident in the development of specialized hardware for deep learning, aimed at enhancing the cost-performance ratio. The focus on Total Cost of Ownership (TCO), which includes operational costs like electricity consumption and the capital expenses associated with computing infrastructure, highlights the relationship between power usage and financial expenditure. Patterson et al. (Patterson et al., 2021) shows that enhancements in performance per TCO also yield benefits in performance per watt, leading to financial savings and a reduction in carbon emissions. Moreover, cloud data centers exhibit significantly higher energy efficiency compared to traditional enterprise data centers, attributed to factors such as improved server utilization. This efficiency has mitigated the anticipated increase in energy consumption attributed to data center operations. Additionally, the strategic deployment of cloud computing facilities in locations with cleaner energy grids or where clean energy can be directly procured (e.g., Finland and Iowa) leverages the efficiency of transmitting information through optical fibers over long distances, further diminishing the environmental impact of AI systems. This holistic approach to environmental cost optimization in AI system manufacturing and deployment highlights the synergy between technological advancements and strategic operational decisions in reducing the carbon footprint and energy consumption of AI technologies.

3.3.2 Environment sustainability versus safety, transparency, non-discrimination, and traceability

Apart from environmental cost optimization, some techniques are considered for building environmentally sustainable AI during the model development and deployment processes. We analyze whether those environmentally sustainable techniques could simultaneously satisfy the other four aspects of an ethical AI. Table 6 summarizes their impacts and corresponding discussion.

Model fine-tuning: Model fine-tuning stands as a cornerstone technique in machine learning, streamlining the enhancement of an existing model to better align with specific tasks or data sets. As a cornerstone of transfer learning, this method allows for the seamless transfer of knowledge from one domain to a related one, facilitating the swift customization

Table 6: Mechanism/techniques influencing AI environmental sustainability and the discussion on environmental sustainability versus safety, transparency, non-discrimination and traceability ("-": negative impact, "+": positive impact)

Papers	Category	Techniques	Impact	Discussion with Other Aspects in AI Regulation
(Mazumder, Safavi, Rahnemoonfar, & Mohsenin, 2023)	Model Fine-tuning	A regression-focused profiling: find the various metrics' trend with hardware deployment of NN	Energy-efficient configuration for model development on the target device	Safety (+), Non-discrimination (+)
(Sha, He, Berrang, Humbert, & Zhang, 2022)		A super-fine-tuning apply a dynamic learning rate method	Fast learning with regular change learning rate end in low energy cost	
(LaBonte, Muthukumar, & Kumar, 2024; Kirichenko, Izmailov, & Wilson, 2022)		Last-layer fine-tuning		
(Guo, Shi, Kumar, Grauman, Rosing, & Feris, 2019; Chen, Yuan, Yang, He, Li, & Yang, 2021a; Chen & Moschitti, 2019)	Transfer Learning	Adaptive or layer-wise fine-tuning: select layer to be fine-tuned	Fewer layers fine-tuned reduce memory footprint & computational costs	Non-discrimination (+/-), transparency (+)
(Zhao, Chen, Liu, Shen, & Liu, 2011; Nater, Tommasi, Grabner, Van Gool, & Caputo, 2011)		Parameter-based learning share parameters between models		
(Mihalkova & Mooney, 2008; Mihalkova, Huynh, & Mooney, 2007)		Relation-based learning establish effective model mapping		
(Dai, Yang, Xue, & Yu, 2007; Fang, Chen, Song, Wang, Zhou, & Zhu, 2019)		Instance-based learning adapts data transferring knowledge	Negative impact with training enlarged dataset	
(Wu, Koh, Dobbie, & Lacombe, 2022b)		Phantom tree algorithm to measure the transfer cost	Trade-off among memory, computational cost and model performance	
(Bourtoule et al., 2021; Kurmanji, Triantafillou, Hayes, & Triantafillou, 2024)	Machine Unlearning	Exact unlearning with sub model/checkpoints retrained to entirely eliminate the data trace	Lower computational cost by avoiding the entire model retraining	Non-discrimination (+), transparency (+)
(Neel et al., 2021; Izzo et al., 2021)		Approximate unlearning with a gradient-based deletion approach to limit training data's impact		
(Lewis, Perez, Piktus, Petroni, Karpukhin, Goyal, K'ú'ttler, Lewis, Yih, Rockt'á'schel, et al., 2020)	RAG	Pre-trained language models with information retrieval methods to elevate text creation	Lower computational cost with the elimination of training expenses	Non-discrimination (+)
(Guo, Ouyang, & Xu, 2020; Liu, Zhuang, Zhuang, Guo, Huang, Zhu, & Tan, 2021)	Model compression	Pruning: Reduce model parameters that contribute little in training	Memory- & energy-efficiency in model deployment and/or development with reduced model size	Non-discrimination (-), transparency (+/-), traceability (-)
(Yamamoto, 2021; Gong, Chen, Lu, Li, Hao, & Chen, 2020)		Quantization: Reduce the number of bits used to represent each weight		
(Chen, Jiang, Liu, & Zhou, 2021b; Hsu, Hua, Chang, Lou, Shen, & Jin, 2022)		Low-rank factorization: Identify model's redundant parameters with matrix & tensor decomposition		
(Ji, Heo, & Park, 2021; Wang & Yoon, 2021; Huang, You, Wang, Qian, & Xu, 2022b)		Knowledge distillation: Transfer knowledge from a large to a smaller model & ensure the validity		

of pre-trained models for new tasks with limited labeled data (Guo et al., 2019). This approach not only significantly boosts the accuracy of models for targeted applications but also underscores the efficiency of model fine-tuning in elevating overall model performance.

Model fine-tuning offers an eco-friendly alternative to the traditional, resource-intensive process of training a model from the ground up using extensive datasets (Chen et al., 2021a; LaBonte et al., 2024; Kirichenko et al., 2022). By utilizing a smaller training dataset, fine-tuning significantly reduces computational and storage requirements during the Data Acquisition and Preparation phase. This reduction not only decreases the computational burden associated with loading, preprocessing, and extracting features from the data but also minimizes memory demands for data storage. In the Model Training phase, the smaller dataset and/or fewer layers to be trained facilitates quicker training times and manageable computational expenses. When applied within distributed or cloud-based systems, the minimized dataset size also leads to more manageable data communication costs, enhancing the overall efficiency and sustainability of the model development process.

Additionally, this technique plays a pivotal role in creating environmentally sustainable AI by optimizing platform/application-level caching (Wu et al., 2022a). By leveraging pre-computed embeddings from existing models, fine-tuning improves power efficiency in developing new models for distinct tasks, thereby minimizing the environmental footprint of AI training processes. Mazumder et al. (Mazumder et al., 2023) suggested a profiling approach to identify the most efficient fine-tuning configuration in terms of energy consumption or latency, targeting the required accuracy. This method analyzes the energy/latency contours across various tinyML device platforms.

Furthermore, model fine-tuning is instrumental in bolstering safety and fostering non-discrimination within AI development. Comprehensive fine-tuning can markedly diminish the threat of backdoor attacks, safeguarding the utility of AI systems while maintaining high safety standards . End-to-end fine-tuning effectively eradicates hidden vulnerabilities, showcasing its critical role in securing AI technologies. Moreover, adjusting the final layer of a model can enhance fairness, mitigating bias and promoting equality (Liang, Zhao, Chen, Bandara, & Shetty, 2023; Henzinger, Karimi, Kueffner, & Mallik, 2023; Mao, Deng, Yao, Ye, Kawaguchi, & Zou, 2023). This harmonious integration of environmental sustainability, safety measures, and non-discriminatory practices through model fine-tuning exemplifies its comprehensive benefits in fostering the creation of ethical and responsible AI systems.

Transfer learning: Transfer learning is advocated for scenarios where training data is either scarce or costly to acquire, enabling the use of different domains, tasks, and distributions for training and testing purposes (Niu, Liu, Wang, & Song, 2020; Pan & Yang, 2009). This approach, akin to model fine-tuning, promotes the development of eco-friendly AI by reducing energy consumption during the data acquisition, preparation, and model training stages. Unlike model fine-tuning, which focuses on preserving the pre-learned features, transfer learning seeks to adapt the pre-trained model's weights, features, or architectural knowledge to a new task, requiring minimal further training (Wu et al., 2022b). In addition to benefits similar to fine-tuning, this discussion examines the environmental advantages of transfer learning through its methods of transferring knowledge.

To enhance model performance by facilitating the effective transfer of useful knowledge and preventing the transfer of detrimental knowledge during the Data Acquisition and Preparation phase. This effort is outlined in the work of Chen et al. (Chen & Moschitti, 2019) and

further detailed by Niu et al. (Fang et al., 2019), who categorize the methodologies into four distinct types: instance-based, feature-based, parameter-based, and relation-based learning. Notably, parameter-based and relation-based methods are highlighted for their potential to support the development of AI systems in an environmentally sustainable manner.

Instance-based and feature-based learning do not significantly contribute to the environmental sustainability of AI, while some specific methods may negatively impact it. Instance-based learning involves adapting data from the source domain for use in the target domain (Dai et al., 2007; Fang et al., 2019), thereby enlarging the training set, which typically leads to training the model from scratch without any environmental benefits. Feature-based learning, while seeking domain-invariant features, can lead to a negative environmental impact. This type of learning often requires feature mapping techniques that transfer large amounts of data from the source to the target domain and involve adversarial learning and deep learning to extract high-level features. The latter methods can significantly amplify the computational burden during data acquisition and preparation. Wu et al. (Wu et al., 2022b) introduced an innovative and cost-effective transfer learning framework, OPERA (Online Transfer using Phantom Tree for Real-Time Adaptation), to optimize the trade-off between accuracy improvement and computational expense. This framework employs the phantom tree algorithm, which assesses the complexity involved in constructing tree-based models, as a method to evaluate the costs associated with transfer learning.

Conversely, learning about parameters and relations offers a more environmentally friendly approach. Parameter-based learning leverages shared parameters between models from the source and target domains, akin to model fine-tuning discussed above (Zhao et al., 2011; Nater et al., 2011). Relation-based learning, through techniques like transferring a Markov Logic Network (MLN) or employing Relational Pathfinding (RPF), focuses on establishing effective mappings from the source model to the target domain (Mihalkova & Mooney, 2008; Mihalkova et al., 2007). These strategies notably reduce the training time and data required to develop an accurate target domain model compared to building a model from scratch, as detailed by Mihalkova et al. (Mihalkova et al., 2007). This efficiency not only accelerates the learning process but also aligns with the principles of developing sustainable AI technologies by limiting the environmental footprint associated with extensive data processing and model training.

Transfer learning, particularly through parameter- and ration-based methods, holds the potential to cultivate sustainable AI systems while trading off non-discrimination and accuracy and ensuring fair transfer in the domain adaptation (Wang, Wang, Beutel, Prost, Chen, & Chi, 2021; Schumann, Wang, Beutel, Chen, Qian, & Chi, 2019). This approach enhances the interpretability of models (Abir, Uddin, Khanam, Tazin, Khan, Masud, Aljahdali, et al., 2022; Brito, Susto, Brito, & Duarte, 2023), making it easier to understand how AI makes decisions. Furthermore, explainable AI plays a crucial role in assessing both the performance and validity of models post-transfer learning, providing insights into their operation (Hossain, Chakrabarty, Gadekallu, Alazab, & Piran, 2023). However, challenges arise when the original, pre-trained model operates as a "black box" without transparency, necessitating potentially costly reprogramming efforts to enable transfer learning (Tsai, Chen, & Ho, 2020). The comparative costs of this reprogramming for black-box models versus training models from scratch have yet to be thoroughly explored.

Machine unlearning: Machine unlearning, as previously discussed, addresses privacy issues in machine learning by specifically eliminating individuals' data from trained models, thereby improving non-discrimination and trust in AI systems. Unlike the resource-intensive process of developing a new model from scratch to adhere to the "right to be forgotten," machine unlearning methods, such as exact and approximate unlearning, offer a more eco-friendly alternative by lowering computational costs. Exact unlearning, exemplified by techniques like checkpoint training (Kurmanji et al., 2024) or SISA training with only the model trained with the subset containing to-be-forgotten data retrained (Bourtoule et al., 2021), presents a favorable balance between unlearning accuracy and time. On the other hand, approximate unlearning employs the use of influence functions and a gradient-based deletion approach to diminish or restrict the impact of an individual's data on the model, achieving "approximate" statistical non-identifiability (Neel et al., 2021; Izzo et al., 2021).

Retrieval-augmented generation (RAG): presents a cutting-edge technique in natural language processing (NLP), merging the capabilities of pre-trained language models with information retrieval methods to elevate text creation (Lewis et al., 2020). This approach enables the model to dynamically access and integrate pertinent documents or data excerpts from a comprehensive corpus, infusing the text generation process with external information. The process involves using user queries to pinpoint relevant documents within a knowledge base, which are then combined with the query to form a prompt that supplies additional context for the response generation. In contrast to fine-tuning large language models (LLMs), RAG stands out as more environmentally sustainable due to its elimination of training expenses and the absence of costs associated with storing additional or modified LLMs.

Given that users can contribute to the documents retrieved, RAG has the potential to support non-discrimination through adjustments for fairness. Nonetheless, its impact on aspects such as safety, transparency, and traceability is yet to be determined.

Model Compression: Reducing model size without compromising performance is essential for environmentally sustainable AI, as larger models result in higher inference times and increased energy consumption and memory storage. This survey investigates four techniques applied during the training and post-training phases: pruning, quantization, low-rank factorization, and knowledge distillation. Pruning eliminates redundant parameters, neurons, layers, and filters, particularly in image processing (Guo et al., 2020; Liu et al., 2021). Quantization and Low-factor factorization can be applied to convolutional and fully connected layers, while the former reduces the number of bits for each weight and the latter decomposes a large matrix into smaller matrices of parameters with proper factorization and rank selection (Yamamoto, 2021; Gong et al., 2020; Chen et al., 2021b; Hsu et al., 2022). Knowledge distillation, focusing on classification-based tasks, replaces large layers with smaller ones through a teacher-student learning framework (Ji et al., 2021; Wang & Yoon, 2021; Huang et al., 2022b).

While model compression positively impacts environmental sustainability, it adversely affects fairness measures and the model's traceability. Model compression can amplify existing biases in AI models, undermining non-discrimination efforts (Ramesh, Chavan, Pandit, & Sitaram, 2023; Kamal & Talbert, 2024; Stoychev & Gunes, 2022). Additionally, it can be used to remove watermarks within the model, thereby reducing traceability (Shao, Yang, Gu, Qin, Fan, & Yang, 2024). The impact on transparency is multifaceted: though com-

pression can decrease interpretability, hindering transparency (Joseph, Siddiqui, Bhaskara, Gopalakrishnan, Muralidharan, Garland, Ahmed, & Dengel, 2020), explainable AI techniques can be utilized to mitigate this effect (Yan, Natarajan, Joshi, Khardon, & Tadepalli, 2024; Becking, Dreyer, Samek, M'ú'ller, & Lapuschkin, 2020; Yu & Xiang, 2023). Notably, compressed models can offer benefits in privacy protection (Chen, Song, Ozgur, & Kairouz, 2024).

3.4 Traceability

3.4.1 Overview

The concept of traceability pertains to the ability to relate the unique identifiable entities in a verifiable way by all parties involved, from the development phase of an AI system to its use and deployment (Kerrigan, 2022). This concept is critical to the success of AI systems in terms of ensuring transparency, accountability, and governance (D'í'az-Rodr'í'guez, Del Ser, Coeckelbergh, de Prado, Herrera-Viedma, & Herrera, 2023).

-*Regulations incorporating traceability:* One significant regulation that includes traceability as a key component is the European Union's proposed AI Act (Commission, 2023d). The AI Act emphasizes the importance of traceability for high-risk AI systems by requiring comprehensive documentation of the AI system's development, deployment, and operational phases. This documentation is critical for understanding how high-risk AI systems are developed and how they perform throughout their lifetime, enabling the verification of compliance with the Act's requirements, monitoring of operations, and post-market monitoring. The AI Act mandates that technical documentation must contain the detailed information necessary to assess the AI system's compliance with relevant requirements, including the general characteristics, capabilities, limitations of the system, algorithms, data training, testing, and validation processes used, as well as documentation on the relevant risk management system. Furthermore, the AI Act requires high-risk AI systems to technically allow for automatic recording of events (i.e., logs) throughout the system's lifetime, ensuring that activities related to the AI system can be traced back and verified.

In addition to the EU's AI Act, the U.S. Government Accountability Office (GAO) has also outlined traceability as a critical aspect of responsible AI deployment (Office, 2021). The GAO's AI Accountability Framework includes traceability as one of its principles, advising entities to document the results of monitoring activities and any corrective actions taken to promote traceability and transparency. This approach enhances accountability by ensuring that decisions and actions taken by AI systems can be traced back to their source, facilitating oversight and the identification of issues that may require remediation.

These regulations and frameworks recognize the significance of traceability in maintaining the transparency, accountability, and governance of AI systems, particularly those that pose significant risks to individuals and society.

-*The role of traceability in AI regulation:* The concept of traceability is crucial in AI regulation, providing a fundamental structure that guarantees AI systems are effectively monitored and scrutinized, held accountable, and conform to ethical norms and legal mandates (Kroll, 2021). By enabling the ability to track and understand all actions, decisions, and processes involved in the AI lifecycle, traceability fosters a higher degree of transparency. This is essential not only for the creators of AI systems but also for regulators and the pub-

lic, ensuring that any decisions made by AI can be scrutinized and understood, thereby upholding accountability, especially when these decisions significantly impact individuals or society at large.

In the realm of AI regulations and guidelines, traceability is often a stipulated requirement. This entails maintaining comprehensive records of the data, models, and processes utilized throughout the development and deployment phases of AI systems. Such meticulous documentation simplifies the process of demonstrating *compliance with both legal and ethical standards* (Kroll, 2021). For instance, the EU AI Act emphasizes the importance of keeping detailed records as a means to facilitate adherence to regulatory frameworks, thereby underscoring the critical role of traceability in regulatory compliance (Commission, 2023d).

Furthermore, traceability is integral to fostering *ethical and responsible AI development and use* (Peters, Vold, Robinson, & Calvo, 2020). It provides the necessary infrastructure to audit and review the behavior of AI systems thoroughly. This capability is crucial for identifying, analyzing, and rectifying biases, errors, or any unintended consequences that may emerge throughout an AI system's lifecycle. Consequently, traceability supports the operationalization of ethical principles and the practice of responsible AI, ensuring that AI systems are developed and utilized in a manner that aligns with societal values and norms.

From a *risk management perspective*, traceability equips organizations with the tools to identify and mitigate potential issues early on (Steimers & Schneider, 2022). It allows for the continuous monitoring and evaluation of AI systems to ensure their operation remains within the intended ethical and regulatory parameters. By enabling early detection and correction of problems, traceability significantly contributes to minimizing risks associated with AI systems, thereby ensuring their safe and effective deployment.

Lastly, traceability plays a crucial role in *enhancing trust* in AI technologies among users and the broader public. The knowledge that the workings and decisions of an AI system can be traced back to their origins fosters confidence in the technology's reliability and fairness (Bedué & Fritzsche, 2022). In an era where trust in AI is paramount for widespread adoption, traceability emerges as a key factor in building and maintaining this trust, thereby facilitating a more receptive environment for AI technologies.

3.4.2 Traceability versus safety, transparency, non-discrimination, and environmental sustainability

Table 7 summarizes the approaches used to prompt AI traceability with its impacts on the other four aspects of ethical AI.

Blockchain: Blockchain technology employs its consensus mechanism and cryptographic technologies to enable users to record data on an immutable ledger, thus ensuring the integrity and traceability of the data. In the era of artificial intelligence, blockchain methods are increasingly used to store and share training data, model parameters, and training processes, enhancing traceability, privacy, accountability, and audibility (Kavasidis et al., 2023).

Functioning as a distributed ledger maintained by a peer-to-peer network, blockchain is often integrated into a federated learning (FL) framework. This approach decentralizes the model training process, distributing it across various nodes to maintain data privacy and security (Li, Li, Yu, Wang, & Chen, 2020a; Kavasidis et al., 2023; Chen et al., 2023; Li et al., 2020b). In regulated industries such as healthcare and pharmaceuticals, where

Table 7: Mechanism/techniques influencing AI traceability and the discussion on environmental sustainability versus safety, transparency, non-discrimination and environmental sustainability ("-": negative impact, "+": positive impact)

Papers	Category	Techniques	Impact	Discussion with the Rest Aspects in AI Regulation
(Kavasidis, Lallas, Mountzouris, Gerogiannis, & Karageorgos, 2023; Chen, Xue, Wang, Huang, Baker, & Zhou, 2023; Li, Fan, Tse, & Lin, 2020b)	Blockchain	Blockchain-based Federated Learning	Recording each node's training data and process for backtracking and audit	Environmental sustainability (+), safety (+), transparency (+)
(Barni, Podilchuk, Bartolini, & Delp, 2001; Sahu, 2024)	Water marking	Direct embedding in data	Enhancing source tracking and audit trails	Safety (+), transparency (+)
(Min, Li, Chen, & Cheng, 2024; Nie, Lu, Wu, & Zhu, 2024)		Model watermarking	Providing IP protection and regulatory compliance	
(Zhang, Ye, Xie, Tang, Liao, Liu, Chen, & Deng, 2024; Abdelnabi & Fritz, 2021)		Adversarial watermarking	Proving ownership when AI models are subjected to various attacks	
(Mora-Cantallops et al., 2021)	AI systems reproducibility	Practices and tool support	Using relevant tools, practices, and data models for traceability in their connection to building AI models and systems	Safety (+), transparency (+), non-discrimination (+)
(Lin, Ko, Chuang, Lin, et al., 2006; Cui & Araujo, 2024)	Open-source licenses	Regulation	Controlling the misuse and unauthorized use of open-source code	Transparency (+)

process reproducibility is crucial — including model reconstruction — blockchain provides a robust solution to meet these stringent requirements (Kavasidis et al., 2023; Li et al., 2020b). More specifically, they deploy a multi-blockchain-based platform to create a comprehensive audit trail for all activities associated with the FL model training process (Kavasidis et al., 2023). In this setup, each node in the model uses a blockchain to store the training data and intermediate data derived from training sessions, each secured with hash values. Concurrently, a global blockchain, managed by a smart contract, orchestrates the overall training process. To minimize the storage burden on the blockchain, model parameters are kept in an InterPlanetary File System (IPFS) distributed network, with the ledger recording the corresponding IPFS-created addresses (Chen et al., 2023). This approach significantly reduces the environmental impact of heavy data sharing and transmission in blockchain operations. Simultaneously, distributing the processing burden of the FL framework across multiple computational nodes minimizes both operational and computational overhead.

Since blockchain techniques ensure not only traceability but also privacy on an immutable ledger, it is crucial for regulated companies like aerospace and pharmaceutical companies. These sectors require maintaining the highest quality standards for their products and the ability to reconstruct the entire production chain to gain crucial insights into the causes of any failures (Kavasidis et al., 2023). Therefore, it improves the model's safety and transparency.

Watermarking: In the context of AI that involves embedding a unique, invisible marker or pattern into AI-generated outputs (such as images, videos, or text) or within the AI model itself. This marker is designed to be robust against modifications and should be retrievable even after the data undergoes transformations or compression (Regazzoni, Palmieri, Smailbegovic, Cammarota, & Polian, 2021).

Watermarks are subtly incorporated directly into content, such as images or video frames, through a process known as direct embedding (Barni et al., 2001; Sahu, 2024). This method ensures the watermarks are both imperceptible to users and recoverable when needed. It is a technique frequently utilized in various forms of media and official documents. Model watermarking, on the other hand, involves adjusting a neural network's internal parameters-like specific weights or biases-to embed a watermark (Min et al., 2024). This is done in such a way that the model's performance remains unaffected (Nie et al., 2024). The presence of the watermark can be confirmed through specific queries that prompt the model to produce pre-defined responses, revealing the watermark. Adversarial watermarking represents a more advanced technique (Zhang et al., 2024; Abdelnabi & Fritz, 2021). In this method, the process of embedding a watermark involves the creation of adversarial example inputs specifically designed to challenge the model. These inputs are subtly modified to include the watermark and are used during the model's training phase. The model is thus trained to recognize these examples and respond in a specific manner, embedding the watermark more thoroughly within its operational framework. This type of watermarking embeds it deeply within the model's decision-making patterns, making it particularly robust.

Watermarking can enhance the safety of AI systems by ensuring that the models and their outputs have not been tampered with. This is particularly critical in applications like autonomous vehicles or healthcare, where data integrity is crucial for making safe decisions. Watermarking can make the origins of AI-generated content clearer, which is essential for transparency. Users and regulators can verify where an AI output came from and whether it has been altered from its original form.

Reproducibility: The reproducibility of AI systems facilitates traceability by ensuring that the processes and results of AI models can be reliably duplicated under similar or different conditions. This capability is vital for verifying and validating the methodologies and outcomes involved in AI research and applications, thereby enhancing traceability, transparency and trustworthiness.

Several tools have been developed to achieve reproducibility, which, in turn, supports traceability (Mora-Cantallops et al., 2021). ModelDB (Vartak, Subramanyam, Lee, Viswanathan, Husnoo, Madden, & Zaharia, 2016) is an open-source system designed for versioning machine learning models, allowing users to index, track, and store modeling artifacts for later reproduction, sharing, and analysis. It emphasizes experiment tracking and provides a web-based interface for the visual representation and analysis of models. Code Ocean [17], Whole Tale [18], and The Renku Project [19] are among the numerous online tools aimed at enhancing reproducibility. These platforms leverage cloud storage and containerization technologies, such as Docker, to capture the research environment fully. This enables the reuse, sharing, and reproduction of the entire research process. Code Ocean merges leading tools, languages,

---

17. https://codeocean.com/
18. https://wholetale.org/
19. https://datascience.ch/renku/

and environments to offer an end-to-end workflow focused on reproducibility. Whole Tale is a free, open-source platform that captures data, code, and the complete software environment, aspiring to redefine how computational and data-driven science is conducted and reproduced. The Renku Project combines a web platform with a command-line interface to support reproducibility, reusability, and collaboration. A broader set of tools supporting "methods reproducibility" research includes ZenML [20], Binder [21], DVC (i.e., Data Version Control) [22], Taverna [23], Kepler [24], and VisTrails [25]. These tools are designed to manage different aspects of the computational research lifecycle, including environment setup, code execution, data management, and the tracking of provenance and metadata. These tools provide essential capabilities for ensuring that AI research and applications are reproducible, which in turn supports the traceability of AI systems by documenting and validating the steps, data, and outcomes involved in model development and deployment.

Traceability in AI systems involves maintaining comprehensive records of the data, decisions, processes, and methodologies used throughout the AI lifecycle. This documentation is crucial for ensuring AI safety. It allows developers and stakeholders to understand how an AI system was built, trained, and deployed, making it easier to identify, diagnose, and rectify any safety issues that may arise. Traceability ensures that AI systems can be audited and reviewed for safety compliance, potentially preventing harm to users and the public. Transparency in AI refers to the openness and clarity regarding how AI systems operate, make decisions, and are developed. Traceability supports transparency by providing a detailed record of the AI development process, including the data used for training, algorithms applied, and decision-making processes. This information is essential for stakeholders to assess the reliability and trustworthiness of AI systems. Transparency, supported by traceability, fosters trust among users, regulators, and the public by making AI operations understandable and open to scrutiny. Traceability supports fairness by documenting the data, algorithms, and decision-making processes used, allowing for the examination and correction of potential biases. By maintaining transparent records, stakeholders can audit AI systems to identify and mitigate unfair practices, ensuring that AI technologies produce equitable outcomes for all users.

Open-source license: An open-source license is a type of license for computer software and other products that allows the source code to be used, modified, and distributed by anyone. Open-source licenses are designed to encourage collaboration and sharing, promoting the development of the software in a community-driven manner. These licenses allow the software to be freely used, modified, and shared under defined terms and conditions (Steiniger & Hunter, 2013).

There are various types of open-source licenses, each with its own specific terms that define how the software can be used, modified, and distributed. Some of the most common open-source licenses include (1) MIT License [26]: One of the most permissive and straightforward open-source licenses, allowing almost unrestricted freedom to use, modify,

---

20. https://zenml.io/
21. https://mybinder.org/
22. https://dvc.org/
23. https://taverna.incubator.apache.org/
24. https://kepler-project.org/
25. https://www.vistrails.org/
26. https://opensource.org/license/mit

and distribute the software, provided that the license and copyright notice are included with any substantial portions of the software; (2) Apache License [27]: Allows the user to freely use, modify, and distribute the software, with the condition that any modifications are documented, and the original copyright and license notices are provided with any distributions. It also grants a patent license to contributors; (3) Source Distribution (BSD) License [28]: The BSD License is another permissive open source license that maintains license notices and copyrights while allowing larger or licensed works to be distributed without source code under different license terms.

Open-source licenses require that the source code be made available to the public. This transparency allows researchers, developers, and users to examine the algorithms, data processing methods, and decision-making processes within AI systems (McKay, 2022). It helps in understanding how these systems work, identifying potential biases, errors, or vulnerabilities, and ensuring that the AI behaves as intended. By allowing anyone to access, modify, and distribute the source code, open-source licenses foster a collaborative environment. Open-source AI projects also enable other researchers to reproduce and verify the results claimed by the original developers. This is a fundamental aspect of scientific research that ensures the reliability and validity of AI technologies. Lastly, open-source projects typically maintain extensive documentation and use version control systems. This practice enhances traceability by providing a detailed history of changes, updates, and modifications made to the AI system over time.

3.5 Non-discrimination

3.5.1 Overview

Non-discrimination in the context of the EU AI Act means developing and using AI systems that promote diversity, equal access, and gender equality while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law.

-*Non-discrimination in the AI Act:* In the EU AI Act, the concept of non-discrimination, also refer as fairness, is not explicitly defined in a single section, but rather, it permeates various aspects of the regulation through provisions aimed at ensuring that AI systems do not create or perpetuate discrimination or bias. Firstly, the Act emphasizes the prevention of discrimination by requiring that high-risk AI systems undergo rigorous assessment processes to ensure they do not produce biased outcomes (Commission, 2023d). This includes testing, validation, and documentation to demonstrate that these systems can handle data fairly without leading to discriminatory results. Secondly, non-discrimination is also promoted through transparency and explainability requirements, particularly for high-risk AI systems. The Act mandates that operators provide clear information on AI systems' functioning, capabilities, and decision-making processes (Commission, 2023c). This transparency helps stakeholders understand how decisions are made, thereby promoting non-discrimination in the operation of AI systems. In addition, proper management of data used by AI systems is critical to ensuring non-discrimination. The Act requires high-quality data governance practices to prevent biases arising from data misuse or poor quality. This includes measures for the accuracy, reliability, and representativeness of the data sets used,

---

27. https://www.apache.org/licenses/LICENSE-2.0
28. https://opensource.org/license/bsd-3-clause

Table 8: Mechanism/approaches influencing AI non-discrimination and the discussion on non-discrimination versus safety, transparency, environmental sustainability, and traceability ("-": negative impact, "+": positive impact)

Stage	Category	Papers	Explanation	Trade-off
Pre	Sampling	(Feldman, Friedler, Moeller, Scheidegger, & Venkatasubramanian, 2015; Adler, Falk, Friedler, Nix, Rybeck, Scheidegger, Smith, & Venkatasubramanian, 2018; Dwork, Immorlica, Kalai, & Leiserson, 2018)	Implements statistical parity through balanced data sampling and removes disparate impact by modifying the distribution of protected attributes in training data. Key methods include stratified sampling and balanced mini-batch creation.	Safety (+), transparency (+)
	Relabeling	(Luong, Ruggieri, & Turini, 2011; Miron, Tolan, G'ó'mez, & Castillo, 2020; Wang, Ustun, & Calmon, 2019)	Corrects biased labels in training data through KNN-based label propagation and gradient descent optimization. Focuses on identifying and rectifying mislabeled instances that contribute to discriminatory outcomes.	Transparency (+), environmental-sustainability (+)
	Representation	(Bower, Kitchen, Niss, Strauss, Vargas, & Venkatasubramanian, 2017; Brunet, Alkalay-Houlihan, Anderson, & Zemel, 2019; Kairouz, Liao, Huang, Vyas, Welfert, & Sankar, 2019; du Pin Calmon, Wei, Vinzamuri, Ramamurthy, & Varshney, 2018)	Transforms raw data into fair representations through dimensionality reduction and feature engineering. Employs convex optimization to learn representations that maximize task performance while minimizing correlation with protected attributes.	Transparency (+), traceability (+), safety (+)
In	Treatment-driven	(Zafar, Valera, Rogriguez, & Gummadi, 2017b; Zafar, Valera, Gomez Rodriguez, & Gummadi, 2017a; Bellamy, Dey, Hind, Hoffman, Houde, Kannan, Lohia, Martino, Mehta, Mojsilovic, et al., 2018; Cheng, Hao, Yuan, Si, & Carin, 2021; Zhou, Ma, Zhang, Zhou, & Yang, 2021)	Enforces equal treatment by modifying model architecture and loss functions. Incorporates fairness constraints directly into optimization objectives through adversarial training and regularization terms that penalize discriminatory decisions.	Safety (+), environment-sustainability (-), transparency(+), traceability (+)
	Impact-driven	(Zhang, Lemoine, & Mitchell, 2018; Wadsworth, Vera, & Piech, 2018; Edwards & Storkey, 2015; Elazar & Goldberg, 2018; Sweeney & Najafian, 2020)	Focuses on equalizing model outcomes across protected groups through adversarial debiasing and demographic parity constraints. Explicitly optimizes for balanced prediction distributions regardless of sensitive attributes.	Safety (+), environment-sustainability (-)
	Hybrid	(Madras, Creager, Pitassi, & Zemel, 2018; Kim, Shin, Jang, Song, Joo, Kang, & Moon, 2021; Park, Hwang, Kim, & Byun, 2021)	Combines treatment and impact-driven approaches through multi-objective optimization. Uses disentangled representations and counterfactual fairness to simultaneously address disparate treatment and impact.	Transparency (+), safety (+)
Post	Input correction	(Agarwal, Dud'í'k, & Wu, 2019; Kiritchenko & Mohammad, 2018)	Adjusts test inputs through targeted perturbation and feature transformation to ensure fair predictions. Implements systematic modifications to input data while preserving task-relevant information.	Environmental-sustainability (-), transparency (+)
	Classifier correction	(Adler et al., 2018; McNamara, Ong, & Williamson, 2017; Anders, Pasliev, Dombrowski, M'ú'ller, & Kessel, 2020)	Modifies trained model's decision boundaries through post-hoc calibration and threshold adjustment. Implements fairness constraints while maintaining model performance through minimal architectural changes.	Environmental-sustainability (-), safety (+)
	Output correction	(Mehrabi, Gupta, Morstatter, Steeg, & Galstyan, 2021a; Jang, Shi, & Wang, 2022)	Applies post-processing to model predictions using group-specific thresholds and rejection sampling. Ensures statistical parity in final outputs while maintaining prediction quality.	Environmental-sustainability (-), transparency (+)

which directly influence the non-discrimination of the AI system's outcomes. Lastly, the regulation explicitly prohibits AI practices that could manipulate or exploit vulnerable groups or otherwise lead to unfair outcomes. For example, AI systems designed to exploit vulnerabilities of individuals based on age, economic situation, or disabilities are forbidden. By integrating these principles, the AI Act seeks to ensure that AI systems used within the EU contribute to equitable outcomes, do not reinforce unfair biases, and respect the principle of non-discrimination as laid out in the broader framework of EU law.

-*Addressing bias throughout the lifecycle of an AI system:* As shown in Table **??**, we organize strategies for reducing bias in AI systems into three distinct stages, namely pre-processing, in-processing, and post-processing, each focusing on a different stage of the development and deployment process for AI systems. Considerable mitigation at the pre-processing stage involves carefully examining and preparing the data. Related methods ensure the data is representative of all demographics and does not contain historical biases or skewed distributions that could influence the AI's decision-making process. At the in-processing stage, non-discrimination constraints are integrated directly into the algorithm's learning process. By adjusting the learning algorithms to account for equity, in-processing aims to prevent the AI system from perpetuating existing biases that might be present in the training data. The final stage of bias mitigation focuses on the outputs of AI systems. After an AI model has been trained, post-processing techniques are applied to adjust its decisions to ensure they adhere to non-discrimination principles.

-*Non-discrimination-driven approaches to public safety:* Safety emphasizes the prevention of harm to individuals or data breaches due to technical failures or design flaws in AI systems, including avoiding unfair risks to specific groups due to bias. The interaction between non-discrimination and safety ensures that AI systems are safe and fair for all users. Below are particular algorithms and research examples illustrating this tension. For instance, striving for demographic parity through equal male and female parole rates may inadvertently disadvantage lower-risk female inmates to fulfill this ratio, thereby breaking the non-discrimination principle of equalized odds (Berk, Heidari, Jabbari, Kearns, & Roth, 2021). The application of machine learning in critical sectors such as the judiciary, welfare systems, and autonomous driving underscores the myriad ways in which AI systems, imbued with inherent biases, can impact daily life and the subtleties of biases in AI and robotics infiltrating real-world scenarios (Howard & Borenstein, 2018), emphasizing the non-discrimination imperative for researchers and engineers to anticipate downstream effects in AI system development. Acknowledging the predictive value of gender in such contexts could lead to a decision-making paradigm that is unfair to minority groups, without secured public safety outcomes. By implementing these approaches, AI systems can contribute to public safety through equitable law enforcement, unbiased emergency response, inclusive public health measures, and fair access to public services. Ultimately, non-discriminatory AI helps reduce social tensions, improve trust in institutions, and ensure that technological advancements benefit all members of society equally (Mitchell, Potash, Barocas, D'Amour, & Lum, 2021).

3.5.2 Non-discrimination versus transparency, traceability, safety, and environmental sustainability

In the framework of the European Union's AI Act, the principle of non-discrimination, along with traceability, transparency, safety, and environmental sustainability, constructs a comprehensive framework to ensure the ethical and societal responsibility of AI systems. This section elucidates how these intertwined principles foster a fair, secure, and sustainable deployment of AI technologies.

Enhancing non-discrimination through increased transparency: Transparency is closely linked to the principle of non-discrimination, as it enhances the visibility of system operations, allowing for external assessments of bias or unfair practices. Beyond aiding stakeholders in identifying potential biases, transparency also bolsters trust in the non-discrimination and reliability of AI systems.

Transparency and explanation (Grgic-Hlaca, Redmiles, Gummadi, & Weller, 2018)(Srivastava, Heidari, & Krause, 2019) are fundamental in fostering trust and understanding in machine learning non-discrimination. By identifying and addressing the underlying causes of bias, causal methods can help reveal underlying biases and improve decision-making transparency. While Inverse Propensity Scoring (IPS) techniques (Bonner & Vasile, 2018) offer a useful approach for enhancing fairness in AI systems by tackling selection bias and promoting more equitable datasets, they are not without drawbacks. These methods, though practical, struggle to adapt to changes in observational patterns and are frequently designed for particular use cases rather than being universally applicable.By introducing small changes to input data, perturbation methods can reveal hidden biases in AI models, helping identify areas where discrimination may occur. Studies using perturbation techniques have demonstrated that implementing measures to ensure non-discrimination does not substantially reduce accuracy in AI systems (Patro, Chakraborty, Ganguly, & Gummadi, 2020)(Ekstrand, Tian, Kazi, Mehrpouyan, & Kluver, 2018). This finding supports the use of such interventions to enhance privacy protection, transparency and fairness in AI applications. Interpretable models (Liu, Wang, Fan, Liu, Li, Jain, Liu, Jain, & Tang, 2022) further non-discrimination by making decision processes transparent, advocating for a cohesive approach to creating fair and transparent algorithmic systems (Li, Li, & Venkatasubramanian, 2006).

Enhancing non-discrimination through increased traceability: At the heart of traceability lies the commitment to ensuring that the decision-making pathways, data sources, and algorithmic logic of AI systems are transparently documented and traceable. In essence, traceability serves as an implementation mechanism for non-discrimination, supporting the assessment and verification of non-discrimination through detailed documentation of decision processes.

Studies (Brunet et al., 2019) introduce a pre-processing strategy to mitigate bias by altering or removing bias-originating documents during training, addressing word embedding bias. Concurrently, research (Kilbertus, Gasc'ó'n, Kusner, Veale, Gummadi, & Weller, 2018) develops an encryption method for sensitive user data, enhancing security and privacy while allowing non-discrimination verification without exposing data to unauthorized use or access.

The impact of non-discrimination AI approaches to sustainable AI: Repairing or rebuilding a fair model is typically time-consuming, particularly when post-processing is required (Jang et al., 2022; Adler et al., 2018; Agarwal et al., 2019). Adversarial learning

techniques, such as adversarial debiasing, eliminate discrimination using a discriminator, but introducing additional model training and inference processes is not environmentally friendly (Zhang et al., 2018; Beutel, Chen, Zhao, & Chi, 2017; Sweeney & Najafian, 2020). However, some repair methods are relatively more sustainable than rebuilding. For example, using a suitable algorithm to learn counterfactual distributions can repair a black-box classifier without the need for retraining, offering an environmentally sustainable way to restore the model's fairness (Wang et al., 2019).

4. Discussion and future directions

Addressing the open problems of ethical AI requires concerted efforts across multiple stakeholders, including governments, private sectors, academia, and civil society. By fostering an environment that prioritizes ethical considerations in AI development and use, it is possible to harness the benefits of AI technologies while mitigating their risks and ensuring they contribute positively and ethically to society. In the following, we first summarize the open problems associated with the AI Act from the perspective of technical efforts to promote human-centric principles, including safety, traceability, transparency, environmental sustainability, and non-discrimination. We then explore potential approaches for regulating AI systems in the future.

4.1 AI Act's technical challenges for human-centric design

Strengthening AI systems against supply-chain vulnerabilities: Addressing supply-chain vulnerabilities in AI systems extends beyond protecting against direct attacks to include securing the entire ecosystem surrounding AI development, from the data sources and software libraries to the deployment environments. These vulnerabilities may arise from compromised data leading to poisoned models, tampered software libraries inducing backdoor threats, or adversarial manipulations undermining model integrity, availability, and privacy. Effective mitigation requires a comprehensive approach that combines advanced technical strategies, like adversarial training for resilience against attacks (Chen & Ji, 2022), (Abdelnabi & Fritz, 2021), and rigorous data sanitization (Venkatesan et al., 2021) to prevent malicious data injection, with robust procedural and policy frameworks. These measures must account for the inherent trade-offs between model accuracy and robustness, and the dynamic nature of AI threats necessitates continuous vigilance and adaptation. Enhancing the security of AI systems against supply-chain vulnerabilities thus involves a concerted effort across the AI community to foster secure, reliable, and trustworthy AI through ongoing collaboration, innovation, and a strong commitment to security best practices.

The dual-edged nature of technologies in ethical AI promotion: Promoting ethical AI underscores the nuanced balance between harnessing innovative technological advances to foster ethical AI practices and navigating the inherent challenges these technologies may pose. On one side, technologies like machine learning algorithms, blockchain, and data analytics tools can significantly enhance transparency, accountability, and fairness in AI systems, laying the foundation for more ethical AI development. They provide mechanisms for unbiased decision-making, environmental friendliness, secure data sharing, and enhanced privacy, which are crucial for ethical standards. For example, transfer learning, as discussed in Section 3.3.2, typically has a positive influence on the development of

AI models, particularly in terms of environmental sustainability, non-discrimination, and transparency. Nonetheless, not all transfer learning approaches yield beneficial outcomes for environmentally sustainable development, and some are necessary only in specific scenarios. For example, instance-based transfer learning often has a detrimental environmental impact, while feature-based transfer learning might have a slightly negative effect. In situations where the training set is small, and label information originates solely from the source domain, users might be constrained to opt for either instance-based or feature-based transfer learning over more eco-friendly alternatives. According to (Niu et al., 2020), other environmental sustainability transfer learning methods can be chosen when the label information of the target-domain instances is available. Otherwise, a trade-off between environmental sustainability and performance shall be considered (Wu et al., 2022b). Additionally, the efficiency of instance-based or feature-based transfer learning in enhancing non-discrimination and transparency while remaining cost-effective is yet to be determined.

Another example is adversarial training, as discussed in Section 3.1.2 and Section 3.5.2. It significantly enhances the safety and non-discrimination aspects of AI model development, respectively (Madry et al., 2017; Tsipras et al., 2018; Zafar et al., 2017b; Beutel et al., 2017). This method strengthens the model's resilience or reduces its dependency on certain specific features by introducing adversarial examples during the training phase. However, while aimed at promoting model fairness, adversarial training might inadvertently compromise the model's performance in critical scenarios, which leads to potential threats in safety (Beutel et al., 2017), such as identifying illegal weapon possession (Zafar et al., 2017b). Adversarial training with crafted adversarial examples introduces the features that the model learns to treat independently, enhancing the system's transparency and traceability by offering clear insights into the model's robustness/defenses against particular adversarial inputs (Madry et al., 2017; Tsipras et al., 2018). Further, adversarial training may introduce a more complex training process, resulting in more computational cost and potentially conflicting with environmental sustainability objectives (Zafar et al., 2017b).

Differential privacy (DP) is the third technique discussed due to its dual-edged nature. Generally, DP enhances AI transparency and safety but poses challenges for developing environmentally friendly AI and complicates traceability, with a complicated effect on fairness. This technique is adaptable and applicable across various stages of AI development, including training data preprocessing, feature engineering, model training, and prediction, primarily to ensure privacy-preserving data analysis (Dwork, 2006; Zhu, Ye, Wang, Zhou, & Philip, 2020). Such applications help AI providers establish suitable levels of transparency tailored for different users or customers. In addition, with DP techniques, the AI model can be more robust and safe by mitigating various attacks like membership inference attacks and model inversion attacks (Truex, Liu, Gursoy, Wei, & Yu, 2019; Salim, Moustafa, Turnbull, & Razzak, 2022). However, incorporating DP increases the complexity of models, learning tasks, and datasets, particularly during the model training phase, which can contradict the principles of environmental sustainability (Truex et al., 2019). As for traceability, although the DP-based model is more complex to understand than the typical model, the DP-based explainable model shows more private interpretability. Patel et al. (Patel, Shokri, & Zick, 2022) proposed an optimization method to minimize the total privacy loss while maintaining a high explanation quality. The impact of DP on the fairness of AI systems is complex. On the one hand, it supports fairness by enabling re-sampling of training data (Zhu et al., 2020),

which helps maintain balance across different groups. On the other hand, using differentially private training data can inadvertently introduce biases, potentially affecting fairness negatively (Tran, Fioretto, Van Hentenryck, & Yao, 2021), though these effects might be mitigated by perturbing the model's outputs (Mangold, Perrot, Bellet, & Tommasi, 2023). Thus, while DP contributes positively to transparency and safety, its influence on environmental sustainability, traceability, and fairness in AI necessitates careful consideration and balanced application.

Ensuring Ethical Development in Large Language Models: The importance of data in the development of Large Language Models (LLMs) cannot be overstated. The correlation between the growth of these models and the exponential increase in required training data is evident in their evolution. Although detailed disclosure of data sources by LLMs remains rare, the instances where such transparency exists reveal the vast data consumption involved in training (Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozi'è're, Goyal, Hambro, Azhar, et al., 2023). The emergence of chatbots (OpenAI, 2022) marks a significant technological leap, offering potential across a wide range of business applications, from entertainment to critical sectors. However, this advancement comes with challenges, notably the vulnerability of LLMs to prompt injections (Greshake et al., 2023), a technique that can elicit undesired responses. Recent study has identified inherent limitations in strictly censoring LLMs, highlighting the need for alternative risk management strategies such as implementing controlled access points for models and enhancing cybersecurity measures(Glukhov, Shumailov, Gal, Papernot, & Papyan, 2023). Despite the impressive capabilities demonstrated by chatbots in specific tasks, the technology is still in its infancy and requires cautious deployment, especially in scenarios where trustworthiness is paramount. The need for continuous monitoring underscores the technology's emerging status and the inherent challenges in ensuring ethical AI development. Future directions should focus on increasing transparency around data sources, enhancing security measures to counter vulnerabilities, and adhering to rigorous ethical standards to foster trust and reliability in LLM applications.

Trade-offs between the attributes of ethical and trustworthy AI: To navigate towards the goal of ethical AI, it is paramount to recognize and address the intricate balance among the diverse attributes that underpin the trustworthiness of AI systems. These attributes include human-centric principles that are defined in regulation as well as some that extend beyond, such as accuracy, susceptibility to adversarial attacks, explainability, privacy, and robustness to adversarial attacks. A singular focus on optimizing one attribute, such as accuracy, often results in compromises in other critical areas like fairness and vulnerability to attacks (Jagielski, Kearns, Mao, Oprea, Roth, Sharifi-Malvajerdi, & Ullman, 2019). This delicate balancing act between various desirable qualities highlights a key challenge in AI development: the inherent trade-offs that must be managed. For instance, enhancing an AI system's adversarial robustness might inadvertently lower its accuracy and fairness (Wang, Chen, Gui, Hu, Liu, & Wang, 2020).

The path forward involves a multifaceted approach that does not seek to maximize performance in one attribute at the expense of others but rather aims for a harmonious balance that upholds the principles of AI regulation. This requires a comprehensive understanding of the interplay between different AI system characteristics, necessitating ongoing research into characterizing and navigating the trade-offs between them. As AI technology becomes

increasingly integrated into various facets of modern life, the importance of such research grows, underscoring the need for innovative solutions that ensure AI systems are trustworthy by being fair, secure, transparent, traceable, and environmentally friendly. Future directions in achieving ethical AI must focus on developing methodologies and technologies that enable this balance, ensuring AI systems are designed and deployed in a manner that earns and maintains public trust (AI, 2023).

4.2 Future AI regulation directions

Rapid AI technological development vs slow and robust regulatory frameworks: With the widespread application of AI, a comprehensive regulatory framework is essential to govern AI development and deployment, encompassing various laws related to the environment and labor relations. This survey focuses on ethical issues such as safety, transparency, traceability, non-discrimination, and environmental sustainability. However, the rapid advancement of AI results in continuous breakthroughs and widespread deployment across multiple domains, creating a regulatory lag and involving a complex array of stakeholders, as shown in Figure 3. These stakeholders include enterprises, researchers and developers, regulators and policymakers, end-users, and customers. AI technologies enable rapid innovation cycles with frequent updates and iterations to enhance performance and capabilities, which can sometimes surpass researchers' and developers' ability to fully understand and address potential ethical concerns (Blauth, Gstrein, & Zwitter, 2022; Nasr, Carlini, Hayase, Jagielski, Cooper, Ippolito, Choquette-Choo, Wallace, Tram'è'r, & Lee, 2023a). For example, jailbreaking prompts have recently emerged as an effective method that may missed by end-users and customers to bypass security restrictions and generate harmful content that was originally intended to be prohibited (Yu, Liu, Liang, Cameron, Xiao, & Zhang, 2024).

Required technical complexity and global standardization: Robust regulations and legislative acts necessitate a consensus on values and moral perspectives, as well as a deep understanding of technical complexities to achieve AI goals. This complexity stems from the need to tackle a diverse array of issues while ensuring that AI technologies are safe, transparent, traceable, non-discriminatory, and environmentally sustainable. As discussed in the previous section 4.1, it is crucial to identify technical challenges and trade-off solutions to comprehensively address ethical AI goals. Additionally, given the rapid misuse and abuse of LLMs (Blauth et al., 2022), it is essential to understand related technologies to identify associated risks and corresponding solutions. Our survey provides a comprehensive analysis of AI technologies regarding various ethical issues.

Furthermore, AI systems are currently developed and implemented under varying national regulations as described in Section 1, which leads to fragmentation. The lack of unified standards across countries creates disparities in legal certainty and market access for AI operators. Furthermore, the Act highlights the need for international cooperation to achieve consistent standards and regulatory practices. Efforts include pursuing international agreements and mutual recognition of conformity assessments to harmonize AI regulations globally.

Interdisciplinary collaboration beyond computer science: An in-depth understanding of AI trustworthiness necessitates not only the development of newer and better AI technologies but also a comprehensive grasp of how these technologies interact with hu-

man society. This multifaceted approach calls for collaboration across various disciplines far beyond computer science. AI practitioners should collaborate closely with domain experts whenever AI technologies are deployed in real-world scenarios that impact people, such as in medicine, finance, transportation, and agriculture. Domain experts can provide crucial insights into industry-specific challenges and help ensure that AI applications are both effective and responsible. Furthermore, AI practitioners should seek advice from social scientists better to understand the often unintended societal impacts of AI. Social scientists can help identify and address issues such as the effects of AI-automated decisions, job displacement in AI-impacted sectors, and the influence of AI systems on social networks. By working together, AI developers and social scientists can develop strategies to mitigate negative outcomes and enhance the positive impacts of AI. Lastly, it is essential for AI practitioners to carefully consider how AI technology is presented to the public and interdisciplinary collaborators. Transparent and honest communication about the known limitations of AI systems is crucial for building trust and ensuring responsible use.

Societal impacts: The regulation of AI is essential for balancing technological advancement with societal welfare, particularly regarding job displacement. Current AI regulations face several limitations in addressing societal impacts effectively. Many regulations are reactive rather than proactive, often responding to issues only after they have arisen. This delayed action can leave displaced workers without immediate support or retraining opportunities, exacerbating the negative effects of job loss. Additionally, the lack of comprehensive frameworks means that regulations tend to focus on specific sectors or technologies without considering the broader economic and social implications, leading to gaps in protection for workers across different industries.

5. Conclusion

In the survey of the AI Act and its technical and regulatory dimensions, we have journeyed through a landscape where the pillars of AI-safety, transparency, non-discrimination, traceability, and environmental sustainability-are both the foundation and the horizon of regulatory endeavours. Our exploration has not only mapped out the current state of AI advancements and their interplay with regulatory frameworks but also illuminated the intricate pathways through which these principles can be harmonized to foster AI systems that are not only innovative but also ethical, equitable, and sustainable.

As we conclude this review, it becomes evident that the AI Act is more than a legislative framework; it is a compass guiding the AI community through the complexities of ethical AI creation and utilization. The Act's focus on safety, transparency, non-discrimination, traceability, and environmental sustainability serves as a multifaceted lens through which the AI ecosystem can be viewed, assessed, and improved. Our discussions have revealed that while challenges in balancing these principles persist, synergies also exist, offering opportunities for holistic advancements in AI regulation and application. As we look forward, it is clear that the development of informed, effective, and equitable AI regulatory systems will require ongoing effort, adaptation, and commitment to the principles outlined in the AI Act. This review serves as a foundational resource for policymakers, stakeholders, developers and scholars, aiming to navigate the complex regulatory landscape of AI and contribute to the development of informed, effective, and equitable regulatory AI systems.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318.

Abdelnabi, S., & Fritz, M. (2021). Adversarial watermarking transformer: Towards tracing text provenance with data hiding. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 121–140. IEEE.

Abir, W. H., Uddin, M. F., Khanam, F. R., Tazin, T., Khan, M. M., Masud, M., Aljahdali, S., et al. (2022). Explainable ai in diagnosing and anticipating leukemia using transfer learning method. *Computational Intelligence and Neuroscience*, *2022*.

Adler, P., Falk, C., Friedler, S. A., Nix, T., Rybeck, G., Scheidegger, C., Smith, B., & Venkatasubramanian, S. (2018). Auditing black-box models for indirect influence. *Knowledge and Information Systems*, *54*, 95–122.

Agarwal, A., Dud'í'k, M., & Wu, Z. S. (2019). Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pp. 120–129. PMLR.

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

Aguirre, A., Dempsey, G., Surden, H., & Reiner, P. B. (2020). Ai loyalty: a new paradigm for aligning stakeholder interests. *IEEE Transactions on Technology and Society*, *1*(3), 128–137.

Ahmad, T., Zhang, D., Huang, C., Zhang, H., Dai, N., Song, Y., & Chen, H. (2021). Artificial intelligence in sustainable energy industry: Status quo, challenges and opportunities. *Journal of Cleaner Production*, *289*, 125834.

AI, N. (2023). Artificial intelligence risk management framework (ai rmf 1.0)..

Akhtar, F. (2023). Regulating artificial intelligence for a safer and more ethical future: A review of the eu's ai act..

Akhtar, N., Mian, A., Kardan, N., & Shah, M. (2021). Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, *9*, 155161–155196.

Ali, A. E., Venkatraj, K. P., Morosoli, S., Naudts, L., Helberger, N., & Cesar, P. (2024). Transparent ai disclosure obligations: Who, what, when, where, why, how..

Anders, C., Pasliev, P., Dombrowski, A.-K., M'ú'ller, K.-R., & Kessel, P. (2020). Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR.

Andrew, G., Kairouz, P., Oh, S., Oprea, A., McMahan, H. B., & Suriyakumar, V. (2023). One-shot empirical privacy estimation for federated learning..

Anthony, L. F. W., Kanding, B., & Selvan, R. (2020). Carbontracker: Tracking and predicting the carbon footprint of training deep learning models..

ATT&CK, M. (2023a). Mitre atlas..

ATT&CK, M. (2023b). Mitre att&ck..

Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of ai systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197.

Barni, M., Podilchuk, C. I., Bartolini, F., & Delp, E. J. (2001). Watermark embedding: Hiding a signal within a cover image. *IEEE Communications magazine*, *39*(8), 102–108.

Becking, D., Dreyer, M., Samek, W., M'ú'ller, K., & Lapuschkin, S. (2020). Ecq x: explainability-driven quantization for low-bit and sparse dnns. In *International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers*, pp. 271–296. Springer.

Bedué, P., & Fritzsche, A. (2022). Can we trust ai? an empirical investigation of trust requirements and guide to successful ai adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549.

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias..

Belrose, N., Furman, Z., Smith, L., Halawi, D., Ostrovsky, I., McKinney, L., Biderman, S., & Steinhardt, J. (2023). Eliciting latent predictions from transformers with the tuned lens..

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2021). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, *50*(1), 3–44.

Bernard, N., & Balog, K. (2023). A systematic review of fairness, accountability, transparency and ethics in information retrieval..

Beutel, A., Chen, J., Zhao, Z., & Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations..

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of ai. *Ieee Access*, *10*, 77110–77122.

Bonner, S., & Vasile, F. (2018). Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 104–112.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., & Papernot, N. (2021). Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pp. 141–159. IEEE.

Bouza, L., Bugeau, A., & Lannelongue, L. (2023). How to estimate carbon footprint when training deep learning models? a guide and review. *Environmental Research Communications*, *5*(11), 115014.

Bower, A., Kitchen, S., Niss, L., Strauss, M., Vargas, A., & Venkatasubramanian, S. (2017). Fair pipelines. *ArXiv*, *abs/1707.00391*.

Brito, L. C., Susto, G. A., Brito, J. N., & Duarte, M. A. V. (2023). Fault diagnosis using explainable ai: A transfer learning-based approach for rotating machinery exploiting augmented synthetic data..

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., & Zemel, R. (2019). Understanding the origins of bias in word embeddings. In *International conference on machine learning*, pp. 803–811. PMLR.

Cao, X., & Yousefzadeh, R. (2023). Extrapolation and ai transparency: Why machine learning models should reveal when they make decisions beyond their training. *Big Data & Society*, *10*(1), 20539517231169731.

Cao, Y., & Yang, J. (2015). Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021). Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650.

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8), 832.

Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey..

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., & Mukhopadhyay, D. (2021). A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, *6*(1), 25–45.

Chang, Z., Li, M., Liu, Y., Wang, J., Wang, Q., & Liu, Y. (2024). Play guessing game with llm: Indirect jailbreak attack with implicit clues..

Chatterjee, S., & NS, S. (2023). Impact of ai regulation and governance on online personal data sharing: from sociolegal, technology and policy perspective. *Journal of Science and Technology Policy Management*, *14*(1), 157–180.

Chaudhari, H., Abascal, J., Oprea, A., Jagielski, M., Tramer, F., & Ullman, J. (2023). Snap: Efficient extraction of private properties with poisoning. In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 400–417. IEEE.

Chen, H., & Ji, Y. (2022). Adversarial training for improving model robustness? look at both prediction and interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 10463–10472.

Chen, J., Xue, J., Wang, Y., Huang, L., Baker, T., & Zhou, Z. (2023). Privacy-preserving and traceable federated learning for data sharing in industrial iot applications. *Expert Systems with Applications*, *213*, 119036.

Chen, L., Yuan, F., Yang, J., He, X., Li, C., & Yang, M. (2021a). User-specific adaptive fine-tuning for cross-domain recommendations..

Chen, L., Jiang, X., Liu, X., & Zhou, Z. (2021b). Logarithmic norm regularized low-rank factorization for matrix and tensor completion. *IEEE Transactions on Image Processing*, *30*, 3434–3449.

Chen, L., & Moschitti, A. (2019). Transfer learning for sequence labeling using source model and target data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 6260–6267.

Chen, W.-N., Song, D., Ozgur, A., & Kairouz, P. (2024). Privacy amplification via compression: Achieving the optimal privacy-accuracy-communication trade-off in distributed mean estimation. *Advances in Neural Information Processing Systems*, *36*.

Chen, Y.-H., Yang, T.-J., Emer, J., & Sze, V. (2019). Eyeriss v2: A flexible accelerator for emerging deep neural networks on mobile devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, *9*(2), 292–308.

Cheng, F., Ming, Y., & Qu, H. (2020). Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, *27*(2), 1438–1447.

Cheng, P., Hao, W., Yuan, S., Si, S., & Carin, L. (2021). Fairfil: Contrastive neural debiasing method for pretrained text encoders..

Chung, M.-H., Yang, Y., Wang, L., Cento, G., Jerath, K., Raman, A., Lie, D., & Chignell, M. H. (2023). Implementing data exfiltration defense in situ: A survey of countermeasures and human involvement..

Cin'à', A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., & Roli, F. (2023). Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys*, *55*(13s), 1–39.

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR.

Commission, A. H. R. (2023a). Australia needs ai regulation..

Commission, E. (2023b). Artificial intelligence act..

Commission, E. (2023c). Artificial intelligence act..

Commission, E. (2023d). 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..

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzm'á'n, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale..

Cui, J., & Araujo, D. A. (2024). Rethinking use-restricted open-source licenses for regulating abuse of generative models. *Big Data & Society*, *11*(1), 20539517241229699.

Dai, W., Yang, Q., Xue, G.-R., & Yu, Y. (2007). Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 193–200.

De 'Á'greda, A. G. (2020). Ethics of autonomous weapons systems and its applicability to any ai systems. *Telecommunications Policy*, *44*(6), 101953.

Demontis, A., Melis, M., Pintor, M., Jagielski, M., Biggio, B., Oprea, A., Nita-Rotaru, C., & Roli, F. (2019). Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks. In *28th USENIX security symposium (USENIX security 19)*, pp. 321–338.

Deshpande, A., & Sharp, H. (2022). Responsible ai systems: who are the stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 227–236.

Dhar, P. (2020). The carbon impact of artificial intelligence.. *Nat. Mach. Intell.*, *2*(8), 423–425.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *Advances in neural information processing systems*, *31*.

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., & Stewart, A. (2019). Sever: A robust meta-algorithm for stochastic optimization. In *International Conference on Machine Learning*, pp. 1596–1606. PMLR.

D''az-Rodr''guez, N., Del Ser, J., Coeckelbergh, M., de Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation..

Dong, Z., Zhou, Z., Yang, C., Shao, J., & Qiao, Y. (2024). Attacks, defenses and evaluations for llm conversation safety: A survey..

Du, S., & Xie, C. (2021). Paradoxes of artificial intelligence in consumer markets: Ethical challenges and opportunities. *Journal of Business Research*, *129*, 961–974.

du Pin Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2018). Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE Journal of Selected Topics in Signal Processing*, *12*(5), 1106–1119.

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, *55*(9), 1–33.

Dwork, C. (2006). Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer.

Dwork, C., Immorlica, N., Kalai, A. T., & Leiserson, M. (2018). Decoupled classifiers for group-fair and efficient machine learning. In *Conference on fairness, accountability and transparency*, pp. 119–133. PMLR.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer.

Edwards, H., & Storkey, A. (2015). Censoring representations with an adversary..

Ekstrand, M. D., Tian, M., Kazi, M. R. I., Mehrpouyan, H., & Kluver, D. (2018). Exploring author gender in book rating and recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 242–250.

Elazar, Y., & Goldberg, Y. (2018). Adversarial removal of demographic attributes from text data..

Enguehard, J. (2023). Sequential integrated gradients: a simple but effective method for explaining language models..

Fang, W., Chen, C., Song, B., Wang, L., Zhou, J., & Zhu, K. Q. (2019). Adapted tree boosting for transfer learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pp. 741–750. IEEE.

Farrand, T., Mireshghallah, F., Singh, S., & Trask, A. (2020). Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*, pp. 15–19.

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, *23*(1), 5232–5270.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268.

Ferrer, X., van Nuenen, T., Such, J. M., Cot'é', M., & Criado, N. (2021). Bias and discrimination in ai: a cross-disciplinary perspective. *IEEE Technology and Society Magazine*, *40*(2), 72–80.

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

Gao, Y., Doan, B. G., Zhang, Z., Ma, S., Zhang, J., Fu, A., Nepal, S., & Kim, H. (2020). Backdoor attacks and countermeasures on deep learning: A comprehensive review..

Garc'í'a, M. V., & Aznarte, J. L. (2020). Shapley additive explanations for no2 forecasting. *Ecological Informatics*, *56*, 101039.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., & Vechev, M. (2018). Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE.

Ghariba, B., Shehata, M. S., & McGuire, P. (2019). Visual saliency prediction based on deep learning. *Information*, *10*(8), 257.

Ginart, A., Guan, M., Valiant, G., & Zou, J. Y. (2019). Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, *32*.

Glukhov, D., Shumailov, I., Gal, Y., Papernot, N., & Papyan, V. (2023). Llm censorship: A machine learning challenge or a computer security problem?..

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

Gong, C., Chen, Y., Lu, Y., Li, T., Hao, C., & Chen, D. (2020). Vecq: Minimal loss dnn model compression with vectorized weight quantization. *IEEE Transactions on Computers*, *70*(5), 696–710.

Gong, R. (2022). Transparent privacy is principled privacy. *Harvard Data Science Review*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples..

Gosselin, R., Vieu, L., Loukil, F., & Benoit, A. (2022). Privacy and security in federated learning: A survey. *Applied Sciences*, *12*(19), 9901.

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pp. 79–90.

Grgic-Hlaca, N., Redmiles, E. M., Gummadi, K. P., & Weller, A. (2018). Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 world wide web conference*, pp. 903–912.

Guo, J., Ouyang, W., & Xu, D. (2020). Multi-dimensional pruning: A unified framework for model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1508–1517.

Guo, Y., Shi, H., Kumar, A., Grauman, K., Rosing, T., & Feris, R. (2019). Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4805–4814.

Hamon, R., Junklewitz, H., Sanchez, I., et al. (2020). Robustness and explainability of artificial intelligence. *Publications Office of the European Union*, *207*.

Han, S., Liu, X., Mao, H., Pu, J., Pedram, A., Horowitz, M. A., & Dally, W. J. (2016). Eie: Efficient inference engine on compressed deep neural network. *ACM SIGARCH Computer Architecture News*, *44*(3), 243–254.

Henzinger, T. A., Karimi, M., Kueffner, K., & Mallik, K. (2023). Monitoring algorithmic fairness..

Hossain, S., Chakrabarty, A., Gadekallu, T. R., Alazab, M., & Piran, M. J. (2023). Vision transformers, ensemble model, and transfer learning leveraging explainable ai for brain tumor detection and classification..

House, T. W. (2023). Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence..

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, *24*, 1521–1536.

Hsu, Y.-C., Hua, T., Chang, S., Lou, Q., Shen, Y., & Jin, H. (2022). Language model compression with weighted low-rank factorization..

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2022a). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence*, *4*(4), 799–819.

Huang, T., You, S., Wang, F., Qian, C., & Xu, C. (2022b). Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, *35*, 33716–33727.

Huang, Z., Shen, Y., Li, J., Fey, M., & Brecher, C. (2021). A survey on ai-driven digital twins in industry 4.0: Smart manufacturing and advanced robotics. *Sensors*, *21*(19), 6340.

Izzo, Z., Smart, M. A., Chaudhuri, K., & Zou, J. (2021). Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pp. 2008–2016. PMLR.

Jagielski, M., Kearns, M., Mao, J., Oprea, A., Roth, A., Sharifi-Malvajerdi, S., & Ullman, J. (2019). Differentially private fair learning. In *International Conference on Machine Learning*, pp. 3000–3008. PMLR.

Jagielski, M., Oprea, A., Biggio, B., Liu, C., Nita-Rotaru, C., & Li, B. (2018). Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In *2018 IEEE symposium on security and privacy (SP)*, pp. 19–35. IEEE.

Jagielski, M., Ullman, J., & Oprea, A. (2020). Auditing differentially private machine learning: How private is private sgd?. *Advances in Neural Information Processing Systems*, *33*, 22205–22216.

Jang, T., Shi, P., & Wang, X. (2022). Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 6988–6995.

Jayaraman, B., & Evans, D. (2019). Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, pp. 1895–1912.

Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., Duan, Y., He, Z., Zhou, J., Zhang, Z., et al. (2023). Ai alignment: A comprehensive survey..

Ji, M., Heo, B., & Park, S. (2021). Show, attend and distill: Knowledge distillation via attention-based feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 7945–7952.

Joseph, V., Siddiqui, S. A., Bhaskara, A., Gopalakrishnan, G., Muralidharan, S., Garland, M., Ahmed, S., & Dengel, A. (2020). Going beyond classification accuracy metrics in model compression..

Kairouz, P., Liao, J., Huang, C., Vyas, M., Welfert, M., & Sankar, L. (2019). Generating fair universal representations using adversarial models..

Kairouz, P., McMahan, B., Song, S., Thakkar, O., Thakurta, A., & Xu, Z. (2021a). Practical and private (deep) learning without sampling or shuffling. In *International Conference on Machine Learning*, pp. 5213–5225. PMLR.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2021b). Advances and open problems in federated learning. *Foundations and trends'Ⓡ' in machine learning*, *14*(1–2), 1–210.

Kamal, M., & Talbert, D. (2024). Beyond size and accuracy: The impact of model compression on fairness. In *The International FLAIRS Conference Proceedings*, Vol. 37.

Katz, G., Barrett, C., Dill, D. L., Julian, K., & Kochenderfer, M. J. (2017). Reluplex: An efficient smt solver for verifying deep neural networks. In *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*, pp. 97–117. Springer.

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: a review. *ACM Computing Surveys (CSUR)*, *55*(2), 1–38.

Kavasidis, I., Lallas, E., Mountzouris, G., Gerogiannis, V. C., & Karageorgos, A. (2023). A federated learning framework for enforcing traceability in manufacturing processes..

Kerrigan, C. (2022). *Artificial Intelligence: Law and Regulation*. Edward Elgar Publishing.

Kilbertus, N., Gasc'ó'n, A., Kusner, M., Veale, M., Gummadi, K., & Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In *International Conference on Machine Learning*, pp. 2630–2639. PMLR.

Kim, H., Shin, S., Jang, J., Song, K., Joo, W., Kang, W., & Moon, I.-C. (2021). Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 8128–8136.

Kirichenko, P., Izmailov, P., & Wilson, A. G. (2022). Last layer re-training is sufficient for robustness to spurious correlations..

Kiritchenko, S., & Mohammad, S. M. (2018). Examining gender and race bias in two hundred sentiment analysis systems..

Kroll, J. A. (2021). Outlining traceability: A principle for operationalizing accountability in computing systems. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 758–771.

Kurmanji, M., Triantafillou, P., Hayes, J., & Triantafillou, E. (2024). Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, *36*.

LaBonte, T., Muthukumar, V., & Kumar, A. (2024). Towards last-layer retraining for group robustness with fewer annotations. *Advances in Neural Information Processing Systems*, *36*.

Laishram, R., & Phoha, V. V. (2016). Curie: A method for protecting svm classifier from poisoning attack..

Lapuschkin, S. (2019). *Opening the machine learning black box with layer-wise relevance propagation*. Ph.D. thesis, Dissertation, Berlin, Technische Universit'á't Berlin, 2018.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. *Internet Policy Review*, *9*(2).

Latonero, M. (2018). Governing artificial intelligence: Upholding human rights & dignity..

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pp. 656–672. IEEE.

Leslie, D. (2019). Understanding artificial intelligence ethics and safety..

Levine, A., & Feizi, S. (2020). Deep partition aggregation: Provable defense against general poisoning attacks..

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., K'ú'ttler, H., Lewis, M., Yih, W.-t., Rockt'á'schel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46.

Li, H., Li, Y., Yu, Y., Wang, B., & Chen, K. (2020a). A blockchain-based traceable self-tallying e-voting protocol in ai era. *IEEE Transactions on Network Science and Engineering*, *8*(2), 1019–1032.

Li, L., Fan, Y., Tse, M., & Lin, K.-Y. (2020b). A review of applications in federated learning. *Computers & Industrial Engineering*, *149*, 106854.

Li, N., Li, T., & Venkatasubramanian, S. (2006). t-closeness: Privacy beyond k-anonymity and l-diversity. In *2007 IEEE 23rd international conference on data engineering*, pp. 106–115. IEEE.

Li, Y., Zhang, S., Wang, W., & Song, H. (2023). Backdoor attacks to deep learning models and countermeasures: A survey..

Liang, X., Zhao, J., Chen, Y., Bandara, E., & Shetty, S. (2023). Architectural design of a blockchain-enabled, federated learning platform for algorithmic fairness in predictive health care: Design science study. *Journal of medical Internet research*, *25*, e46547.

Lin, Y.-H., Ko, T.-M., Chuang, T.-R., Lin, K.-J., et al. (2006). Open source licenses and the creative commons framework: License selection and comparison. *Journal of information science and engineering*, *22*(1), 1–17.

Liu, H., Wang, Y., Fan, W., Liu, X., Li, Y., Jain, S., Liu, Y., Jain, A., & Tang, J. (2022). Trustworthy ai: A computational perspective. *ACM Transactions on Intelligent Systems and Technology*, *14*(1), 1–59.

Liu, J., Zhuang, B., Zhuang, Z., Guo, Y., Huang, J., Zhu, J., & Tan, M. (2021). Discrimination-aware network pruning for deep model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(8), 4035–4051.

Liu, Y., Deng, G., Li, Y., Wang, K., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). Prompt injection attack against llm-integrated applications..

Lloyd, K. (2018). Bias amplification in artificial intelligence systems..

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, *30*.

Luong, B. T., Ruggieri, S., & Turini, F. (2011). k-nn as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 502–510.

Lyu, L., Xu, X., Wang, Q., & Yu, H. (2020). Collaborative fairness in federated learning..

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393. PMLR.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks..

Mahloujifar, S., Ghosh, E., & Chase, M. (2022). Property inference from poisoning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1120–1137. IEEE.

Mangold, P., Perrot, M., Bellet, A., & Tommasi, M. (2023). Differential privacy has bounded impact on fairness in classification. In *International Conference on Machine Learning*, pp. 23681–23705. PMLR.

Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners..

Mao, Y., Deng, Z., Yao, H., Ye, T., Kawaguchi, K., & Zou, J. (2023). Last-layer fairness fine-tuning is simple and effective for neural networks..

Mazumder, A. N., Safavi, F., Rahnemoonfar, M., & Mohsenin, T. (2023). Reg-tune: A regression-focused fine-tuning approach for profiling low energy consumption and latency..

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction..

McKay, M. H. (2022). Ai transparency in a real-world context: What we can learn from past examples of algorithmic and statistical decision-making.. In *AI*.

McNamara, D., Ong, C. S., & Williamson, R. C. (2017). Provably fair representations..

Mehrabi, N., Gupta, U., Morstatter, F., Steeg, G. V., & Galstyan, A. (2021a). Attributing fair decisions with attention interventions..

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021b). A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, *54*(6), 1–35.

Meza Mart'í'nez, M. A., Nadj, M., Langner, M., Toreini, P., & Maedche, A. (2023). Does this explanation help? designing local model-agnostic explanation representations and an experimental evaluation using eye-tracking technology. *ACM Transactions on Interactive Intelligent Systems*, *13*(4), 1–47.

Mihalkova, L., Huynh, T., & Mooney, R. J. (2007). Mapping and revising markov logic networks for transfer learning. In *Aaai*, Vol. 7, pp. 608–614.

Mihalkova, L., & Mooney, R. J. (2008). Transfer learning by mapping with minimal target data. In *Proceedings of the AAAI-08 workshop on transfer learning for complex tasks*, pp. 31–36.

Min, R., Li, S., Chen, H., & Cheng, M. (2024). A watermark-conditioned diffusion model for ip protection..

Miron, M., Tolan, S., G'ó'mez, E., & Castillo, C. (2020). Addressing multiple metrics of group fairness in data-driven decision making..

Mironov, I., Talwar, K., & Zhang, L. (2019). R\'enyi differential privacy of the sampled gaussian mechanism..

Mitchell, S., Potash, E., Barocas, S., D'Amour, A., & Lum, K. (2021). Algorithmic fairness: Choices, assumptions, and definitions. *Annual Review of Statistics and Its Application*, *8*, 141–163.

Mora-Cantallops, M., S'á'nchez-Alonso, S., Garc'í'a-Barriocanal, E., & Sicilia, M.-A. (2021). Traceability for trustworthy ai: A review of models and tools. *Big Data and Cognitive Computing*, *5*(2), 20.

Mothilal, R. K., Sharma, A., & Tan, C. (2020). Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617.

Nakanishi, T. (2024). Pcaldi: Explainable similarity and distance metrics using principal component analysis loadings for feature importance..

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Wallace, E., Tram'è'r, F., & Lee, K. (2023a). Scalable extraction of training data from (production) language models..

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramér, F., Jagielski, M., Carlini, N., & Terzis, A. (2023b). Tight auditing of differentially private machine learning..

Nasr, M., Songi, S., Thakurta, A., Papernot, N., & Carlin, N. (2021). Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, pp. 866–882. IEEE.

Nater, F., Tommasi, T., Grabner, H., Van Gool, L., & Caputo, B. (2011). Transferring activities: Updating human behavior analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 1737–1744. IEEE.

Neel, S., Roth, A., & Sharifi-Malvajerdi, S. (2021). Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR.

Nelson, B., Barreno, M., Chi, F. J., Joseph, A. D., Rubinstein, B. I., Saini, U., Sutton, C., Tygar, J. D., & Xia, K. (2008). Exploiting machine learning to subvert your spam filter.. *LEET*, *8*(1-9), 16–17.

Nie, H., Lu, S., Wu, J., & Zhu, J. (2024). Deep model intellectual property protection with compression-resistant model watermarking..

Nishant, R., Kennedy, M., & Corbett, J. (2020). Artificial intelligence for sustainability: Challenges, opportunities, and a research agenda. *International Journal of Information Management*, *53*, 102104.

Niu, S., Liu, Y., Wang, J., & Song, H. (2020). A decade survey of transfer learning (2010–2020). *IEEE Transactions on Artificial Intelligence*, *1*(2), 151–166.

Nori, H., Jenkins, S., Koch, P., & Caruana, R. (2019). Interpretml: A unified framework for machine learning interpretability..

of the European Union, C. (2023). General data protection regulation..

Office, T. U. G. A. (2021). An accountability framework for federal agencies and other entities..

OpenAI (2022). Chatgpt: Optimizing language models for dialogue..

OpenAI (2023). Gpt-4 technical report..

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, *22*(10), 1345–1359.

Park, S., Hwang, S., Kim, D., & Byun, H. (2021). Learning disentangled representation for fair facial attribute classification via fairness-aware information alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, pp. 2403–2411.

Parliament, E. (2024). Artificial intelligence act: Meps adopt landmark law..

Patel, N., Shokri, R., & Zick, Y. (2022). Model explanations with differential privacy. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1895–1904.

Patro, G. K., Chakraborty, A., Ganguly, N., & Gummadi, K. (2020). Incremental fairness in two-sided market platforms: On smoothly updating recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, pp. 181–188.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training..

Paudice, A., Mu'ñ'oz Gonz'á'lez, L., & Lupu, E. C. (2019). Label sanitization against label flipping poisoning attacks. In *ECML PKDD 2018 Workshops: Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings 18*, pp. 5–15. Springer.

Pesapane, F., Volont'é', C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in europe and the united states. *Insights into imaging*, *9*, 745–753.

Pessach, D., & Shmueli, E. (2022). A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, *55*(3), 1–44.

Peters, D., Vold, K., Robinson, D., & Calvo, R. A. (2020). Responsible ai—two frameworks for ethical design practice. *IEEE Transactions on Technology and Society*, *1*(1), 34–47.

Pillutla, K., Andrew, G., Kairouz, P., McMahan, H. B., Oprea, A., & Oh, S. (2023). Unleashing the power of randomization in auditing differentially private ml..

Ponomareva, N., Hazimeh, H., Kurakin, A., Xu, Z., Denison, C., McMahan, H. B., Vassilvitskii, S., Chien, S., & Thakurta, A. G. (2023). How to dp-fy ml: A practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, *77*, 1113–1201.

Qiu, X., Parcollet, T., Beutel, D. J., Topal, T., Mathur, A., & Lane, N. D. (2020). Can federated learning save the planet?. In *NeurIPS-Tackling Climate Change with Machine Learning*.

Ramesh, K., Chavan, A., Pandit, S., & Sitaram, S. (2023). A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15762–15782.

Reddy, S., Allan, S., Coghlan, S., & Cooper, P. (2020). A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, *27*(3), 491–497.

Regazzoni, F., Palmieri, P., Smailbegovic, F., Cammarota, R., & Polian, I. (2021). Protecting artificial intelligence ips: a survey of watermarking and fingerprinting for machine learning. *CAAI Transactions on Intelligence Technology*, *6*(2), 180–191.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

Rigaki, M., & Garcia, S. (2023). A survey of privacy attacks in machine learning. *ACM Computing Surveys*, *56*(4), 1–34.

Robinson, S. C. (2020). Trust, transparency, and openness: How inclusion of cultural values shapes nordic national public policy strategies for artificial intelligence (ai). *Technology in Society*, *63*, 101421.

Rosenfeld, E., Winston, E., Ravikumar, P., & Kolter, Z. (2020). Certified robustness to label-flipping attacks via randomized smoothing. In *International Conference on Machine Learning*, pp. 8230–8241. PMLR.

R'ú'hr, A., Berger, B., & Hess, T. (2023). Intelligent it systems in business application: Control and transparency as means of building trust in ai. In *Work and AI 2030: Challenges and Strategies for Tomorrow's Work*, pp. 125–132. Springer.

Sahu, A. K. (2024). *Multimedia Watermarking: Latest Developments and Trends*. Springer Nature.

Salim, S., Moustafa, N., Turnbull, B., & Razzak, I. (2022). Perturbation-enabled deep federated learning for preserving internet of things-based social networks. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, *18*(2s), 1–19.

Samek, W., Wiegand, T., & M'ú'ller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models..

Scherer, M. U. (2015). Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. *Harv. JL & Tech.*, *29*, 353.

Schumann, C., Wang, X., Beutel, A., Chen, J., Qian, H., & Chi, E. H. (2019). Transfer of machine learning fairness across domains..

Sha, Z., He, X., Berrang, P., Humbert, M., & Zhang, Y. (2022). Fine-tuning is all you need to mitigate backdoor attacks..

Shao, S., Yang, W., Gu, H., Qin, Z., Fan, L., & Yang, Q. (2024). Fedtracker: Furnishing ownership verification and traceability for federated learning model..

Sheikh, S. (2020). *Understanding the role of artificial intelligence and its future social impact*. IGI Global.

Shi, Y., Li, P., Yin, C., Han, Z., Zhou, L., & Liu, Z. (2022). Promptattack: Prompt-based attack for language models via gradient search. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 682–693. Springer.

Shlens, J. (2014). A tutorial on principal component analysis..

Shneiderman, B. (2020). Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *10*(4), 1–31.

Shumailov, I., Zhao, Y., Bates, D., Papernot, N., Mullins, R., & Anderson, R. (2021). Sponge examples: Energy-latency attacks on neural networks. In *2021 IEEE European symposium on security and privacy (EuroS&P)*, pp. 212–231. IEEE.

Srivastava, M., Heidari, H., & Krause, A. (2019). Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2459–2468.

Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A european agency for artificial intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review*, *45*, 105661.

Steimers, A., & Schneider, M. (2022). Sources of risk of ai systems. *International Journal of Environmental Research and Public Health*, *19*(6), 3641.

Steinhardt, J., Koh, P. W. W., & Liang, P. S. (2017). Certified defenses for data poisoning attacks. *Advances in neural information processing systems*, *30*.

Steiniger, S., & Hunter, A. J. (2013). The 2012 free and open source gis software map–a guide to facilitate research, development, and adoption. *Computers, environment and urban systems*, *39*, 136–150.

Steinke, T., Nasr, M., & Jagielski, M. (2023). Privacy auditing with one (1) training run..

Stojanovic, M. (2020). Can competition law protect consumers in cases of a dominant company breach of data protection rules?. *European Competition Journal*, *16*(2-3), 531–569.

Stoychev, S., & Gunes, H. (2022). The effect of model compression on fairness in facial expression recognition. In *International Conference on Pattern Recognition*, pp. 121–138. Springer.

Sun, N., Lin, G., Qiu, J., & Rimba, P. (2022). Near real-time twitter spam detection with machine learning techniques. *International Journal of Computers and Applications*, *44*(4), 338–348.

Sweeney, C., & Najafian, M. (2020). Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 359–368.

Taheri, R., Javidan, R., Shojafar, M., Pooranian, Z., Miri, A., & Conti, M. (2020). On defending against label flipping attacks on malware detection systems. *Neural Computing and Applications*, *32*, 14781–14800.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, *3*(6), 7.

Tharwat, A., Gaber, T., Ibrahim, A., & Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI communications*, *30*(2), 169–190.

Thudi, A., Shumailov, I., Boenisch, F., & Papernot, N. (2022). Bounding membership inference..

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozi'è're, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models..

trade commission, F. (2023). Fair credit reporting act..

Tran, C., Fioretto, F., Van Hentenryck, P., & Yao, Z. (2021). Decision making with differential privacy under a fairness lens.. In *IJCAI*, pp. 560–566.

Treviso, M., Ross, A., Guerreiro, N. M., & Martins, A. F. (2023). Crest: A joint framework for rationalization and counterfactual text generation..

Truex, S., Liu, L., Gursoy, M. E., Wei, W., & Yu, L. (2019). Effects of differential privacy and data skewness on membership inference vulnerability. In *2019 First IEEE international conference on trust, privacy and security in intelligent systems and applications (TPS-ISA)*, pp. 82–91. IEEE.

Tsai, Y.-Y., Chen, P.-Y., & Ho, T.-Y. (2020). Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources. In *International Conference on Machine Learning*, pp. 9614–9624. PMLR.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., & Madry, A. (2018). Robustness may be at odds with accuracy..

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne.. *Journal of machine learning research*, *9*(11).

Van Looveren, A., & Klaise, J. (2021). Interpretable counterfactual explanations guided by prototypes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 650–665. Springer.

Van Wynsberghe, A. (2021). Sustainable ai: Ai for sustainability and the sustainability of ai. *AI and Ethics*, *1*(3), 213–218.

Vartak, M., Subramanyam, H., Lee, W.-E., Viswanathan, S., Husnoo, S., Madden, S., & Zaharia, M. (2016). Modeldb: a system for machine learning model management. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, pp. 1–3.

Vassilev, A., Oprea, A., Fordyce, A., & Andersen, H. (2024). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations..

Venkatesan, S., Sikka, H., Izmailov, R., Chadha, R., Oprea, A., & De Lucia, M. J. (2021). Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems. In *MILCOM 2021-2021 IEEE Military Communications Conference (MILCOM)*, pp. 874–879. IEEE.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, *123*, 735.

Wadsworth, C., Vera, F., & Piech, C. (2018). Achieving fairness through adversarial learning: an application to recidivism prediction..

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY*, *36*(2), 585–595.

Wang, D., Ustun, B., & Calmon, F. (2019). Repairing without retraining: Avoiding disparate impact with counterfactual distributions. In *International Conference on Machine Learning*, pp. 6618–6627. PMLR.

Wang, H., Chen, T., Gui, S., Hu, T., Liu, J., & Wang, Z. (2020). Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems*, *33*, 7449–7461.

Wang, L., & Yoon, K.-J. (2021). Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. *IEEE transactions on pattern analysis and machine intelligence*, *44*(6), 3048–3068.

Wang, W., Levine, A. J., & Feizi, S. (2022). Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, pp. 22769–22783. PMLR.

Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., & Chi, E. H. (2021). Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1748–1757.

Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., & Ren, K. (2022). Threats to training: A survey of poisoning attacks and defenses on machine learning systems. *ACM Computing Surveys*, *55*(7), 1–36.

Warnat-Herresthal, S., Schultze, H., Shastry, K. L., Manamohan, S., Mukherjee, S., Garg, V., Sarveswara, R., H'á'ndler, K., Pickkers, P., Aziz, N. A., et al. (2021). Swarm learning for decentralized and confidential clinical machine learning. *Nature*, *594*(7862), 265–270.

W'á'schle, M., Thaler, F., Berres, A., P'ó'lzlbauer, F., & Albers, A. (2022). A review on ai safety in highly automated driving. *Frontiers in Artificial Intelligence*, *5*, 952773.

Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Vi'é'gas, F., & Wilson, J. (2019). The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, *26*(1), 56–65.

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., et al. (2022a). Sustainable ai: Environmental implications, challenges and opportunities. *Proceedings of Machine Learning and Systems*, *4*, 795–813.

Wu, O., Koh, Y. S., Dobbie, G., & Lacombe, T. (2022b). Cost-effective transfer learning for data streams. In *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 1233–1238. IEEE.

Yaacoub, J.-P. A., Noura, H. N., Salman, O., & Chehab, A. (2021). A survey on ethical hacking: issues and challenges..

Yamamoto, K. (2021). Learnable companding quantization for accurate low-bit neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5029–5038.

Yan, S., Natarajan, S., Joshi, S., Khardon, R., & Tadepalli, P. (2024). Explainable models via compression of tree ensembles. *Machine Learning*, *113*(3), 1303–1328.

Yu, L., & Xiang, W. (2023). X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24355–24363.

Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., & Zhang, N. (2024). Don't listen to me: Understanding and exploring jailbreak prompts of large language models..

Zafar, M. B., Valera, I., Gomez Rodriguez, M., & Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180.

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR.

Zanella-Béguelin, S., Wutschitz, L., Tople, S., Salem, A., Rúhle, V., Paverd, A., Naseri, M., Kópf, B., & Jones, D. (2023). Bayesian estimation of differential privacy. In *International Conference on Machine Learning*, pp. 40624–40636. PMLR.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340.

Zhang, W. E., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *11*(3), 1–41.

Zhang, Y., Ye, D., Xie, C., Tang, L., Liao, X., Liu, Z., Chen, C., & Deng, J. (2024). Dual defense: Adversarial, traceable, and invisible robust watermarking against face swapping..

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2023). Explainability for large language models: A survey..

Zhao, Y., Gao, X., Guo, X., Liu, J., Wang, E., Mullins, R., Cheung, P. Y., Constantinides, G., & Xu, C.-Z. (2019). Automatic generation of multi-precision multi-arithmetic cnn accelerators for fpgas. In *2019 International Conference on Field-Programmable Technology (ICFPT)*, pp. 45–53. IEEE.

Zhao, Z., Chen, Y., Liu, J., Shen, Z., & Liu, M. (2011). Cross-people mobile-phone based activity recognition. In *Twenty-second international joint conference on artificial intelligence*. Citeseer.

Zhou, C., Ma, J., Zhang, J., Zhou, J., & Yang, H. (2021). Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 3985–3995.

Zhu, T., Ye, D., Wang, W., Zhou, W., & Philip, S. Y. (2020). More than privacy: Applying differential privacy in key areas of artificial intelligence. *IEEE Transactions on Knowledge and Data Engineering*, *34*(6), 2824–2843.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. (2023). Representation engineering: A top-down approach to ai transparency..