Transparency and Explainable AI: Bridging Privacy, Fairness, and Accountability

Dragos VIERU[1]
Renée-Maria SCHMITT[2]

Abstract

This paper examines whether Explainable Artificial Intelligence (XAI) can address AI ethics, specifically, privacy, fairness, and accountability, by linking technical transparency with interpretability to facilitate ethical oversight. It combines review and normative analysis from multiple fields. Three case studies - COMPAS, Uber, and Clearview AI - illustrate XAI's role in revealing biases, tracing accountability, and identifying privacy risks. We argue that transparency alone is insufficient; understanding is hindered by overload, misinterpretation, and lack of context. Effective explainability requires coupling transparency with interpretability within legal, social, and organizational frameworks. XAI highlights ethical issues but needs human judgment and safeguards. The paper connects XAI to ethical principles, demonstrating that explainability is necessary but insufficient, and emphasizes the need for interdisciplinary and regulatory collaboration, such as the EU AI Act, to guide the development of responsible AI.

Keywords: *XAI, Privacy, Fairness, Accountability, Understandability, Transparency, Explainability, Interpretability.*

JEL classification: C88, D83, K23, O33

DOI: 10.24818/RMCI.2025.5.994

1. Introduction

Artificial Intelligence (AI) and machine learning (ML) have recently garnered significant attention, particularly with the emergence of generative AI tools like ChatGPT. AI covers areas such as natural language processing, speech, vision, robotics, and ML (Chazette et al., 2019). Due to their complexity, powerful AI systems often lack transparency, making it hard to understand their inner workings and decision-making (Rosenberger et al., 2025). Transparency and understanding are crucial to ensure their ethical behavior and alignment with human values, avoiding unethical decisions related to privacy, fairness, and accountability (von Eschenbach, 2021). AI systems are used across various domains like education, law, healthcare, finance, and transportation (Agarwal et al., 2022), creating a shift toward a more algorithmic society (Adadi and Berrada, 2018). The impact of AI on organizations, human lives, and society remains a topic of debate (Floridi et al., 2018). While some popular AI-driven tools make less problematic decisions, such

[1] Dragoș Vieru, TELUQ University, Montreal, Canada, e-mail: dragos.vieru@teluq.ca
[2] Renée-Maria Schmitt, Technical University of Munich, Germany, e-mail: reneemaria@web.de

as music recommendations, they are also employed in critical areas like medicine and autonomous transportation (Arrieta et al., 2020). Challenges and risks from harmful AI decisions highlight the importance of understanding their reasoning, requiring humans to reliably interpret AI outputs and how they are produced (Chazette et al., 2019; Coeckelbergh, 2020).

Explainable Artificial Intelligence (XAI) represents a set of processes and methods that aim to produce more transparent, understandable, and explainable AI systems without sacrificing performance or accuracy (Arrieta et al., 2020; Silva et al., 2023). It addresses digital responsibility and social, ethical, and ecological aspects of information system usage (Sovrano et al., 2022). Many stakeholders, including algorithm experts, regulators, lawyers, philosophers, and futurologists, agree on the relevance of XAI today (Waardenburg & Huysman, 2022). It encompasses much more than a few individual technological methods. It is considered a movement and part of the “third-wave AI,” the next generation of AI development (Waardenburg et al., 2022). XAI provides methods and practices that make model behavior interpretable and justifiable without sacrificing performance (Arrieta et al., 2020; Guidotti et al., 2018). In human-centric AI ethics, explanations serve as socio-technical interfaces that link algorithmic operations with human reasoning, trust, and contestation (Coeckelbergh, 2020; Miller, 2019). Policymakers also prioritize transparency and explainability (Jobin et al., 2019), with proposals like the EU’s AI Act requiring explainability for high-risk systems (EU AI Act, 2025). However, scholarship warns that transparency alone isn’t enough; disclosures can overwhelm or mislead if not suited to human cognition and roles (Lipton, 2018; Sovrano et al., 2021).

This paper examines how XAI can bridge the gap between technical transparency and ethical practices to address key issues, including privacy, fairness, and accountability - issues that are central to global AI guidelines and shape public trust (Jobin et al., 2019). We argue that XAI can serve as a socio-technical bridge between technical transparency and the ethical imperatives of privacy, fairness, and accountability, thereby offering both theoretical insights and practical guidance for more responsible AI governance. This is essential for addressing privacy, fairness, and accountability (Miller, 2019; Coeckelbergh, 2020). Guided by the question *Can XAI help mitigate the ethical issues of privacy, fairness, and accountability in AI systems?*, our analysis suggests that XAI can identify privacy violations, surface unfair biases, and clarify loci of responsibility, thereby enabling more ethical outcomes; nevertheless, human and institutional action remains essential to act on these insights.

2. XAI Main Concepts

Before we analyze how XAI might alleviate AI’s ethical issues, it is essential that we define the core terminology. The literature on AI has introduced a cluster of related terms – transparency, explainability, interpretability, and understandability – which are sometimes used inconsistently or interchangeably (Lipton, 2018; Guidotti et al., 2018). Clear definitions are needed to avoid confusion (cf. Floridi & Sanders,

2002) and to ensure we evaluate XAI on a coherent basis (Arrieta et al., 2020; Vainio-Pekka et al., 2023).

Transparency in AI refers to making information about how an AI system operates visible, serving as a “pro-ethical condition” that supports accountability and fairness. A pro-ethical condition is a feature of a system or technology that facilitates ethical behavior or decision-making, even though it is not ethical in itself (Turilli & Floridi, 2009). In AI, this involves access to the algorithm’s source code, model, or training data, allowing experts or the public to understand the decision-making process (Jobin et al., 2019). However, transparency can vary, often referring to explainability or interpretability. While necessary, open information doesn't always ensure AI behavior is understandable or outcomes are justified (Rawal et al., 2022).

Explainability involves actively providing reasons for an AI’s behavior, focusing on communicating decisions in understandable terms for stakeholders (Lipton, 2018; Arrieta et al., 2020). It often uses post hoc methods, like highlighting influential features or giving natural-language rationales (Ribeiro et al., 2016). Explanations bridge humans and AI, clarifying how and why outputs are generated (Miller, 2019; Mittelstadt et al., 2019). For instance, a loan system might show that income and credit score were key factors, making decisions contestable and understandable (Wachter et al., 2017). Explainability also builds trust; users are more likely to accept decisions when provided with justifications (Kizilcec, 2016).

Interpretability refers to the extent to which a human can understand a model's decisions. It measures how easily an observer can internalize and use the system's information. In this context, the effectiveness of the model depends on how well the audience can understand it and its explanations (Gilpin et al., 2018). Lipton (2018) states that the term is somewhat vague but generally relates to simplicity and the ability to simulate the model, such as decision trees or linear regression, where reasoning can be traced.

Understandability, also known as comprehensibility, is the quality or state of something being understandable, meaning that the mind can comprehend or grasp it. In AI, understanding pertains to knowledge about, for instance, how an AI model functions internally and how it predicts outcomes. Understandability reflects a model's capacity to allow humans to understand it (Arrieta et al., 2020).

It must be noted that in literature and practice, these terms sometimes overlap or are used differently. Many AI ethics guidelines and research papers use “transparency” and “explainability” almost synonymously, or bundle interpretability under the umbrella of explainability (Jobin et al., 2019; Gerlings et al., 2021; Rawal et al., 2022). Vainio-Pekka et al. (2023) observe that the fields of AI ethics and XAI lack a common framework and conceptualization, resulting in vagueness in the usage of these terms. Our definitions above align with a consensus view that can be summarized as follows:

- AI Transparency: The AI’s workings are visible or accessible (the information is available).
- AI Explainability: The AI actively provides reasons or explanations (the communication of information).

- AI Interpretability: A human's ability to make sense of the AI with the given information (the comprehension achieved).
- AI Understandability: The overall extent to which the AI is understandable to humans (the quality of being understood).

3. Ethics in AI: Privacy, Fairness, and Accountability

The rapid AI advancement has raised ethical concerns across sectors like health, law, finance, and transportation (Jobin et al., 2019). AI ethics aims to ensure AI aligns with moral values and avoids harm (Floridi et al., 2018; Hagendorff, 2020). Many principles exist for "trustworthy AI," but core ones like privacy, fairness, and accountability are most common (Jobin et al., 2019; Fjeld et al., 2020). Studies highlight these principles as especially prominent and challenging for XAI to address (Khan et al., 2022).

3.1 Privacy

Privacy is an individual's right to control personal data (Brunotte et al., 2023). In AI, this is at risk because models rely on large datasets, often containing personal information, to learn and decide. As AI integrates into daily life via smart speakers, wearables, and online platforms, personal data is collected, shared, and analyzed on an unprecedented scale (Floridi et al., 2018; Hagendorff, 2020). This raises concerns about misuse, surveillance, loss of anonymity, and breaches (Stinson, 2022). AI threatens privacy by exposing personal data (e.g., facial recognition without consent), inferring sensitive info, or accumulating data beyond control (Jobin et al., 2019). Lack of transparency worsens these risks, as opaque AI prevents users from knowing what data is collected, how it's used, or shared (Doshi-Velez & Kim, 2017).

The Clearview case. Clearview AI, a U.S. startup, developed a facial recognition tool by scraping billions of images from the internet and social media without consent, raising privacy and ethical concerns. Law enforcement secretly used it before it became known, sparking legal issues due to a lack of consent and user control. Using personal images from social platforms violated data ownership principles, and a lack of transparency eroded trust (Porter, 2020). The database's scale threatened free expression through mass surveillance. Regulators responded: Canadian privacy commissioners ruled practices unlawful (Thompson, 2021); the UK fined £7.5 million and ordered data deletion; U.S. lawsuits under BIPA led to settlements limiting law enforcement sales (Hern, 2022). Companies like Twitter, Google, IBM, Microsoft, and Amazon halted facial recognition sales due to privacy concerns. This case highlights how AI surveillance can erode privacy through overreach, lack of transparency, and disproportionate monitoring. Privacy in AI requires security, transparency, proportionality, and control, encompassing data minimization, consent, and an explanation of data use, as mandated by laws such as the EU's GDPR. Explainable AI reveals data flows and decision logic, making privacy-related info transparent. Privacy explanations show what data was accessed and its influence, boosting trust and accountability (Brunotte et al., 2023).

XAI may improve privacy by clarifying what the AI knows about individuals and how data is used. Clear disclosures empower users and policymakers to protect rights; without explainability, privacy breaches may go unnoticed until harm occurs.

3.1 Fairness

Fairness in AI involves ensuring decisions are impartial and free from bias, avoiding systematic disadvantaging based on characteristics like race, gender, or ethnicity (Mehrabi et al., 2022). It includes concepts like equality of opportunity, outcomes, or treatment, but can conflict across definitions (Verma & Rubin, 2018). Biased data and algorithms can lead to discriminatory results, making fairness a key challenge in AI ethics (Barocas et al., 2023). Biases can occur during data collection, model design, or deployment—for example, facial recognition errors for darker-skinned faces (Buolamwini & Gebru, 2018; Chakraborty et al., 2021) or hiring algorithms mirroring gender bias (Dastin, 2018). Formal fairness criteria often conflict, and no single metric suits all contexts (Franklin et al., 2022).

XAI may reveal which features influence decisions, helping detect unfair reasoning. For example, if a credit score system cites a ZIP code for denial, it may proxy race or socioeconomic status (Ribeiro et al., 2016). While explainability can't guarantee fairness, it aids in auditing, detecting discrimination, and explaining decisions (Dodge et al., 2022; Zhou et al., 2022).

The COMPAS case. A concrete example of fairness challenges in AI is the COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) algorithm, a risk assessment tool used in parts of the United States to predict the likelihood of recidivism. This algorithm is used in U.S. courtrooms to assign a risk score. Developed by Northpointe (now Equivant), COMPAS assigns defendants to a risk category (low, medium, or high) to predict their likelihood of reoffending. The 2016 ProPublica investigation highlighted racial bias in the COMPAS algorithm, which judges used for pre-trial and sentencing decisions. Analysis of 7,000 Broward County arrestees showed Black defendants were nearly twice as likely to be falsely labeled "high risk," and white defendants were often misclassified as "low risk" despite reoffending. These biases challenged claims of neutrality. Transparency issues arose because factors such as age and prior convictions had unknown weights, making scores difficult to challenge. The Wisconsin Supreme Court, in State v. Loomis (2016), upheld COMPAS but warned that it shouldn't be the sole basis for sentencing due to its limitations.

The COMPAS case illustrates how opaque tools can perpetuate inequalities in the justice system. Without clear explanations, individuals cannot contest errors or contextualize results. Miller (2019) notes explanations must align with human needs, and COMPAS's opacity deprived defendants of such justifications. Lack of transparency undermines fairness and trust in the justice system. Using XAI methods, risk assessment tools could become more open, clearer, and ethically justified, ensuring individuals are judged transparently rather than by mysterious numbers. However, improving fairness might reduce accuracy, and adding explainability could violate privacy or other values.

3.3 Accountability

Ethical AI guidelines state that only humans can be held accountable for harm (Abrassart et al., 2018; Loi and Spielkamp, 2021). Loi and Spielkamp (2021) stress the clarification of accountability's forms and dimensions. Bovens (2007) defines accountability as responsible parties explaining and justifying actions to affected parties, involving questions, judgments, and consequences. This relational view includes five elements: actor, forum, relationship, content, account criteria, and consequences (Wieringa, 2020). In AI, actors include the AI system and stakeholders; decisions are the content and criteria, called "algorithmic accountability." Effective accountability has three phases: information, explanation/justification, and consequences (Bovens, 2007). Actors share info, justify conduct, answer questions, and are willing to justify actions to the forum. Consequences are imposed or possible (Busuioc, 2021). Loi and Spielkamp (2021) highlight responsibility, answerability, and sanctionability.

Uber's 2018 fatality case. The literature suggests that autonomous systems, such as vehicles or medical diagnosis applications, can make decisions that require human oversight (cf., Santosh and Wall, 2022). For example, an Uber autonomous car was involved in a fatal accident in 2018 (Wakabayashi, 2018). Although a driver was in the car, it was driving autonomously at the time of the accident. This example alone underscores the need to establish mechanisms that ensure accountability (Henriksen et al., 2021; Cooper et al., 2022). Nevertheless, with the increasing prevalence of AI systems, it becomes increasingly problematic to locate and assign accountability (Goodall, 2018; Langer et al., 2021).

Nissenbaum (1996) identifies four obstacles that complicate the allocation of accountability with the advent of computerized systems: 1. the problem of "many hands" – addresses the fact that the number of parties involved changes from only a few towards a complex system of parties; 2. the occurrence of computer bugs; 3. blaming computer systems for their decisions and using them as "scapegoats"; and 4. the challenge of "ownership without liability" refers to the problem of property rights of the systems and their components.

The 2018 Uber self-driving car fatality highlights accountability issues. On March 18, 2018, an Uber prototype struck and killed pedestrian Elaine Herzberg in Tempe, Arizona. NTSB investigations revealed the AI detected Herzberg 6 seconds before impact, but kept changing her classification, initially unknown, then vehicle, then bicycle, causing uncertainty about her trajectory. This raised questions about responsibility: the safety driver, Uber engineers, the manufacturer, or the AI itself. This case exemplifies Nissenbaum's "many hands" problem, with AI logs crucial in understanding the failure. Uber was not criminally charged; the safety driver was.

Ethically, this is unsatisfying if AI flaws were central. Cooper et al. (2022) note incidents like this show the need for accountability mechanisms. Better transparency might have allowed early detection and correction of classification issues, potentially preventing the crash. Without clear explanations of AI decisions, assigning accountability remains difficult, risking repeated mistakes. More importantly, the propensity of opaque AI systems to continuously learn from data, rather than having explicit written code, aggravates the allocation of accountability (Bovens, 2007; Henriksen et al., 2021). A “responsibility gap” emerges (Lima et al., 2022), and a possible direct result of this is the autonomous decisions with adverse outcomes that no one directly accounts for, as too many people may have (untraceably) contributed to the harm (cf., Cooper et al., 2022).

XAI enhances accountability by making AI decision-making traceable and justifiable, allowing deployment parties to justify outcomes. Explanations identify “accountable junctures,” such as bias origins in data or model logic, essential for accountability. Accountability involves answerability and responsibility, with explainability making AI actions visible and comprehensible. Transparency is necessary for accountability. Thus, XAI may enhance accountability by providing clarity for oversight, redress, and the allocation of responsibility.

4. Discussion: AI Transparency vs. AI Understandability

4.1 The Cognitive Science of Explanations

Research shows humans prefer minimal, contrastive explanations like “Why X rather than Y?” (Lombrozo, 2006; Miller, 2019). For example, when a loan is denied, the key question is “Why was it denied instead of approved?” An effective explanation highlights the crucial factor, such as insufficient income, rather than overwhelming the reader with details. Raw transparency (like revealing every neural network weight) doesn't always lead to understanding. Cognitive science reveals that people have a limited information-processing capacity. Miller (1956) argued that working memory can hold about “7±2” items, and too much data hinders understanding, creating a transparency paradox: more disclosure may reduce comprehension. Showing a full decision tree is less helpful than highlighting key rules (Lipton, 2018). Communication studies emphasize that explanations must be tailored to the specific audience. Regulators may need raw logs, but end-users benefit from simple reasons (Mittelstadt et al., 2019). Miller (2019) and others stress explanations are social, interactive, and personalized. Evidence shows that explanations boost trust only when meaningful, like counterfactuals, not just confidence scores (Kizilcec, 2016). A true AI should reason about its beliefs and actions, enabling trust (Selvaraju et al., 2017). These insights indicate transparency alone isn’t enough; it needs abstraction and context. Sovrano et al. (2021) distinguish 'explainable' from truly explained systems. Providing model code offers transparency but doesn't ensure understanding. Effective explanations must be clear and meaningful, aligning with Floridi and Sanders’s (2002) view that ethical understanding requires access to information and interpretation.

4.2 Transparency Alone: The Illusion of Understanding

Equating transparency with explainability risks information overload, confusion, and misinterpretation. Selective explanations are often more effective (Ribeiro et al., 2016). Misinterpretation is a risk, especially with complex models, until simplified tools are used (Rudin, 2019). Transparency without guidance can harm fairness; partial explanations like saliency highlights may give false security and lead to overtrust (Köhl et al., 2019; Jacovi & Goldberg, 2020). User diversity means one explanation can't suit all. Technical reports suit engineers but may exclude laypeople. Effective communication needs multiple formats, like probability statements for experts and frequency-based comparisons for the public (Gigerenzer & Edwards, 2003). To address these challenges, Páez (2019) proposes a pragmatic turn in XAI, focusing less on exposing internal model logic and more on the communicative context.

This aligns with Floridi and Sanders's (2002) early emphasis on intelligibility: explanations must make sense to human agents in their roles. A useful analogy is software: having source code ensures transparency, but documentation ensures understanding. Similarly, AI systems should be accompanied by user manuals that explain their purpose, limitations, and use cases in plain language. Recent policy proposals, such as the EU AI Act, suggest requiring such documentation for high-risk AI, bridging the gap between formal transparency and practical explainability (EU AI Act, 2025).

4.3 Bridging the Gap: Toward Understandable Transparency

To harness the ethical benefits of XAI, explanations must move beyond transparency to understandability. We identified several strategies from the literature that point toward actionable design.

Simplified models and post-hoc explainers. One approach is to use interpretable models or approximate complex models with simpler rule-based surrogates. Research shows users prefer slightly less accurate but more interpretable models in high-stakes contexts (Rudin, 2019). However, simplifications must stay faithful to the underlying system to prevent misleading users (Guidotti et al., 2018). Thus, this highlights the interpretability–accuracy trade-off: in industries like finance or criminal justice, clarity can outweigh small accuracy gains.

Interactive explanations. Explanations are more effective when dialogic, allowing users to ask “Why not X?” or “What if Y?” (Miller, 2019). Counterfactual and contrastive explanations provide users with a concrete understanding of how input changes would alter outcomes (Byrne, 2019). Studies suggest that such interactivity enhances trust calibration, letting users judge when to rely on or override AI recommendations (Wachter et al., 2017).

Human-centered design. Explanations should be tested with actual users. Think-aloud protocols and user studies reveal whether explanations enhance mental models or, instead, create confusion (Doshi-Velez & Kim, 2017). Explanations succeed only if they foster appropriate trust and actionable understanding (Jacovi & Goldberg, 2020). Experiments confirm that some explanation formats reduce overreliance, while others unintentionally encourage blind trust (Kizilcec, 2016). Thus, we argue that empirical evaluation is crucial in the design of XAI.

Janus-faced transparency. Turilli and Floridi (2009) view transparency as a proethical value: valuable when supporting goals such as accountability and fairness, but absolute openness can be harmful, overwhelming users, or exposing sensitive information. XAI should aim for appropriate transparency – sufficient for oversight, but not at the expense of privacy or security. For example, a bank might not reveal proprietary model weights, but it can explain the reasons for loan denials. Hence, transparency is 'Janus-faced": it must promote accountability while safeguarding against harm.

These strategies show that true explainability combines transparency and interpretability, with transparency alone often insufficient. The Uber self-driving crash illustrates that the release of source code didn't prevent harm without clear warnings to recognize and act on in real-time. The COMPAS case reveals that even if disclosures were made, fairness issues persist without accessible explanations of bias. Similarly, the Clearview AI controversy shows that revealing methods, like scraping billions of images, don't adequately protect privacy. In all cases, explanations that are actionable, clear, and user-focused transform transparency into accountability, fairness, and privacy. Miller (2019) notes that explanations are most effective when they are selective, contrastive, and socially meaningful. For example, a denied bank loan stating, "Your loan was denied because your income is below $50,000; if it were higher, the outcome would differ," is human-centered, unlike raw formulas. Experiments by Westphal et al. (2023) show that showing model probabilities alone doesn't improve performance, but explanations help users calibrate trust. This supports that explainability needs interpretability layered onto transparency.

Based on the previous argumentation, we suggest that effective XAI must be designed with users in mind. Explanations should be understandable, actionable, and proportional, drawing on cognitive psychology, human–computer interaction, and ethics. Transparency may reveal how a model works, but only interpretation ensures that the message sent by the AI is the message received by the human.

5. Conclusions

This paper explores whether Explainable AI (XAI) can address privacy, fairness, and accountability, linking transparency and ethics. Our analysis, which includes theory and case studies, demonstrates that XAI helps reveal and mitigate ethical issues, but it isn't a complete solution; human judgment remains essential (Mittelstadt et al., 2019; Rudin, 2019). XAI can improve privacy by revealing data influence (Wang et al., 2024), exposing bias, like in the COMPAS case (Miller, 2019), and tracing responsibility, as with Uber and Clearview AI. Overall, XAI connects technical processes with human oversight, translating them into ethics and rights (Floridi & Sanders, 2002). But it doesn't automatically fix ethical issues. Turilli & Floridi (2009) note transparency often supports ethics, but isn't enough; human intervention is needed to adjust data or decisions (Páez, 2019). In Uber's case, explainability wouldn't have prevented harm without a strong safety culture and effective regulations. In COMPAS, transparency allowed critique, but reform required political action. With Clearview AI, knowing about image scraping didn't protect privacy without legal rulings. So, XAI highlights problems, but fixing them depends on human and institutional action (Coeckelbergh, 2020).

Our analysis suggests that effective XAI is about human communication, and explanations must be selective, contrastive, and socially meaningful to be actionable (Lombrozo, 2006; Miller, 2019). We argue that showing raw data like code or heatmaps isn't enough; explanations should match the audience's understanding. Success is determined by whether stakeholders understand and can act, rather than the volume of data (Hafermalz & Huysman, 2021; Jacovi & Goldberg, 2020). The EU's AI Act, which requires explanations for high-risk AI, supports this, as a lack of explanations hinders contesting decisions, causing injustice and eroding trust (Wachter et al., 2017). Goals include meaningful explainability, such as plain language or user-testing explanations (Sovrano et al., 2021).

This paper contributes to the AI ethics research domain by framing explainable AI (XAI) as a socio-technical link between technical transparency and ethical principles. We hope that it clarifies the distinctions among transparency, explainability, interpretability, and understandability, which are often used interchangeably, providing a clearer theoretical base. Drawing from computer ethics, cognitive science, and information systems, it contextualizes XAI as more than a technical tool, but part of normative debates.

This approach suggests that XAI can shed light on, but not resolve, ethical issues such as privacy, bias, and accountability. We demonstrate how XAI identifies and mitigates ethical risks through case studies, such as Uber, COMPAS, and Clearview AI, illustrating how explainability exposes bias, clarifies responsibility, and enhances transparency in high-stakes decisions. These lessons can help practitioners, software developers, and policymakers utilize XAI for accountability, contestability, and trust, emphasizing XAI as a key component of ethical AI governance, rather than a panacea.

References

1. Abrassart, C., Bengio, Y., Chicoisne, G., de Marcellis-Warin, N., Dilhac, M.-A., Gambs, S., … & Voarino, N. (2018). *The Montreal Declaration for a Responsible Development of Artificial Intelligence*, Université de Montréal, available at https://montrealdeclaration-responsibleai.com/the-declaration/ (accessed 10 November 2025).
2. Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), *IEEE Access,* 6, 52138-52160.
3. Agarwal, A., Agarwal, H., & Agarwal, N. (2022). Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems, *AI and Ethics*, 3, 267-279.

4. Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Information Fusion*, 58, 82-115.
5. Bovens, M. (2007). Analysing and Assessing Accountability: A Conceptual Framework", *European Law Journal*, 13(4), 447-468.
6. Brunotte, W., Specht, A., Chazette, L., &Schneider, K. (2023). Privacy Explanations - A Means to End-User Trust, *Journal of Systems and Software*, 195, 111545.
7. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification, *Proceedings of Machine Learning Research,* 81, 1-15.
8. Busuioc, M. (2021). Accountable Artificial Intelligence: Holding Algorithms to Account, *Public Administration Review*, 81(5), 825-836.
9. Byrne, R. M. J. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning, *IJCAI*, 6276-6282.
10. Chakraborty, J., Majumder, S., & Menzies, T. (2021). Bias in machine learning software: Why? How? What to do?, *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, Athens, Greece, 429-440.
11. Chazette, L., Brunotte, W., & Speith, T. (2021). Exploring Explainability: A Definition, a Model, and a Knowledge Catalogue, P*roceedings of the IEEE 29th International Requirements Engineering Conference*, Notre Dame, IN, USA, 197-208.
12. Chazette, L., Karras, O., & Schneider, K. (2019). Do End-Users Want Explanations? Analyzing the Role of Explainability as an Emerging Aspect of Non-Functional Requirements, *Proceedings of the IEEE 27th International Requirements Engineering Conference*, Jeju, S. Korea, 223-233.
13. Coeckelbergh, M. (2020). *AI Ethics*. MIT Press.
14. Cooper, A. F., Moss, E., Laufer, B., and Nissenbaum, H. (2022). Accountability in an Algorithmic Society: Relationality, Responsibility, and Robustness in Machine Learning, *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Seoul, S. Korea, 864-876.
15. Cortese, J. F. N. B., Cozman, F. G., Lucca-Silveira, M. P., & Bechara, A. F. (2022). Should explainability be a fifth ethical principle in AI ethics?, *AI and Ethics*, 3, 23-134.
16. Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women*, Reuters, available at: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G (accessed 7 September 2025)
17. Dexe, J., Franke, U., and Söderström, F. (2021), "Explainable artificial intelligence for accountability: A policy perspective," AI & Society, Vol. 36, No. 2, pp. 431-444.
18. Dodge, J., Hilderbrand, C., Newman, J., Leiser, N., Prabhakaran, V., Carter, L., & Gebru, T. (2022). Can AI explain unfairness? A framework for evaluating bias in machine learning with explainability, *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, Association for Computing Machinery, 848-859.
19. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning, arXiv preprint arXiv:1702.08608.
20. EU AI Act. (2025). *Future of Life Institute*, available at: https://artificialintelligenceact.eu/about/ (accessed 7 September 2025)
21. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Cemter Research Publication, (2020-1), available at: http://nrs.harvard.edu/urn-3:HUL.InstRepos:42160420 (accessed 14 september 2025).

22. Floridi, L.,& Sanders, J. W. (2002). Mapping the foundationalist debate in computer ethics, *Ethics and Information Technology*, 4(1), 1-9.
23. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, … & Vayena, E. (2018). AI4People-An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations, *Minds and Machines,* 28(4), 689-707.
24. Franklin, J. S., Bhanot, K., Ghalwash, M., Bennett, K. P., McCusker, J., & McGuinness, D. L. (2022). An Ontology for Fairness Metrics, P*roceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, Oxford, United Kingdom, 265-275.
25. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining Explanations: An Overview of Interpretability of Machine Learning, *Proceedings of the IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, Turin, Italy, 80-89.
26. Gigerenzer, G., & Edwards, A. (2003). Simple tools for understanding risks: From innumeracy to insight," *BMJ*, 327(7417), 741-744.
27. Goodall, N. J. (2018). Machine ethics and automated vehicles, In M. Maurer, J. C. Gerdes, B. Lenz, and H. Winner (Eds.), *Autonomous driving*, Springer, 93-112.
28. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models, *ACM Computing Surveys*, 51(5), 1-42.
29. Hafermalz, E., & Huysman, M. (2021). Please Explain: Key Questions for Explainable AI research from an Organizational perspective, *Morals & Machines*, 1(2), 10-23.
30. Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines, *Minds and Machines*, 30(1), 99-120.
31. Hern, A. (2022). TechScape: Clearview AI fined £7.5m for brazenly harvesting your data, *The Guardian*. available at: https://www.theguardian.com/technology/2022/may/25/techscape-clearview-ai-facial-recognition-fine (accessed 7 September, 2025)
32. Henriksen, A., Enni, S., & Bechmann, A. (2021). Situated Accountability: Ethical Principles, Certification Standards, and Explanation Methods in Applied AI, *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, Association for Computing Machinery, 574-585.
33. Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?, *ACL Anthology*, 4198-4205.
34. Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface, *Proceedings of the 34th Annual ACM Conference on Human Factors in Computing Systems (CHI'16)*, 2390-2395.
35. Khan, A. A., Badshah, S., Liang, P., Waseem, M., Khan, B., Ahmad, A., & Azeem Akbar, M. (2022). Ethics of AI: A Systematic Literature Review of Principles and Challenges, *Proceedings of the International Conference on Evaluation and Assessment in Software Engineering 2022*, Gothenburg, Sweden, 383-392.
36. Köhl, M. A., Baum, K., Langer, M., Oster, D., Speith, T., & Bohlender, D. (2019). Explainability as a Non-Functional Requirement, *Proceedings of the IEEE 27th International Requirements Engineering Conference*, Jeju, S. Korea, 363-368.
37. Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., & Baum, K. (2021) What Do We Want From Explainable Artificial Intelligence (XAI)? - A Stakeholder Perspective on XAI and a Conceptual Model Guiding Interdisciplinary XAI Research, *Artificial Intelligence*, 296, 103473.

38. Lima, G., Grgić-Hlača, N., Jeong, J. K., & Cha, M. (2022). The Conflict Between Explainable and Accountable Decision-Making Algorithms, *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, Seoul, S. Korea, 2103-2113.
39. Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery, *Queue*, 16(2), 31-57.
40. Loi, M., & Spielkamp, M. (2021). Towards Accountability in the Use of Artificial Intelligence for Public Administrations, *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 757-766.
41. Lombrozo, T. (2006). The structure and function of explanations, *Trends in Cognitive Sciences*, 10(10), 464-470.
42. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). Survey on Bias and Fairness in Machine Learning, *ACM Computing Surveys*, 54(6), 1-35.
43. Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information, *Psychological Review*, 63(2), 81-97.
44. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences, *Artificial Intelligence*, 267, 1-38.
45. Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, USA, 279-288.
46. Nissenbaum, H. (1996). Accountability in a computerized society, *Science and Engineering Ethics,* 2(1), 25-42.
47. National Transportation Safety Board (NTSB). (2019). *Collision between vehicle controlled by developmental automated driving system and pedestrian, Tempe, Arizona, March 18, 2018*, (Report No. NTSB/HAR-19/03). available at:
https://www.ntsb.gov/investigations/accidentreports/pages/hwy18mh010.aspx (accessed 7 September 2025)
48. Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI), *Minds and Machines*, 29(3), 441-459.
49. Porter, J. (2020, January 20). Go read this NYT expose on a creepy new facial recognition database used by US police, *The Verge*, available at: https://www.theverge.com/2020/1/20/21073718/clearview-ai-facial-recognition-database-new-york-times-investigation (accessed 8 September, 2025)
50. Possati, L. M. (2022). The ontological transparency of algorithms: Accountability and opacity in AI, *AI & Society*, Vol. 37, No. 4, pp. 1531-1542.
51. Rawal, A., McCoy, J., Rawat, D. B., Sadler, B. M., & Amant, R. St. (2022). Recent Advances in Trustworthy Explainable Artificial Intelligence: Status, Challenges, and Perspectives, *IEEE Transactions on Artificial Intelligence*, 3(6), 852-866.
52. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier, *KDD 2016 Proceedings*, 1135-1144.
53. Rosenberger, P., Zeng, Y., and Timan, T. (2025). From transparency to intelligibility: Rethinking explainability in AI governance, *AI and Ethics*, 5(2), 211-225.
54. Rudin, C. (2019). Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, 1(5), 206-215.
55. Santosh, K., & Wall, C. (2022). *AI, Ethical Issues and Explainability - Applied Biometrics*, Springer Nature Singapore, 1-46.
56. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization, *Proceedings of the IEEE International Conference on Computer Vision (ICCV 2017)*, IEEE, 618-626.

57. Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable artificial intelligence: Evaluating the objective and subjective impacts of XAI on human-agent interaction. *International Journal of Human-Computer Interaction*, 39(7), 1390-1404.
58. Sovrano, F., Vitali, F., & Palmirani, M. (2021). Making Things Explainable vs Explaining: Requirements and Challenges Under the GDPR, in *AI Approaches to the Complexity of Legal Systems XI-XII*, V. Rodríguez-Doncel, M. Palmirani, M. Araszkiewicz, P. Casanovas, U. Pagallo and G. Sartor (eds.), Springer, 169-182.
59. Stinson, C. (2022). Algorithms are not neutral: Bias in collaborative filtering, *AI and Ethics*, 2(4), 763-770.
60. Thompson, E. (2021, February 4). U.S. technology company Clearview AI violated Canadian privacy law: report, *CBC/Radio-Canada*, available at: https://www.cbc.ca/news/politics/technology-clearview-facial-recognition-1.5899008 (accessed 1 September 2025).
61. Turilli, M., & Floridi, L. (2009). The Ethics of Information Transparency, *Ethics and Information Technology*, 11(2), 105-112.
62. Vainio-Pekka, H., Agbese, M. O. O., Jantunen, M., Vakkuri, V., Mikkonen, T., Rousi, R., & Abrahamsson, P. (2023). The role of explainable AI in the research field of AI ethics, *ACM Transactions on Interactive Intelligent Systems*, 13(4), 1-39.
63. Verma, S., & Rubin, J. (2018). Fairness definitions explained, *Proceedings of the International Workshop on Software Fairness*, ACM, 1-7.
64. von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI, *Philosophy & Technology,* 34(4), 1607-1622.
65. Wakabayashi, D. (2018). Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam, *The New York Times*, available at: https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html (accessed 7 September 2025).
66. Waardenburg, L., Huysman, M., & Sergeeva, A. V. (2022). In the land of the blind, the one-eyed man is king: Knowledge brokerage in the age of learning algorithms. *Organization Science,* 33(1), 59-82.
67. Waardenburg, L., & Huysman, M. (2022). From coexistence to co-creation: Blurring boundaries in the age of AI. *Information and Organization*, 32(4), 100432.
68. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box, *Harvard Journal of Law & Technology*, 31(2), 841-887.
69. Wang, X., Wu, Y. C., Zhou, M., & Fu, H. (2024). Beyond surveillance: Privacy, ethics, and regulations in face recognition technology, *Frontiers in Big Data*, 7, 1337465.
70. Westphal, M., Vössing, M., Satzger, G., Yom-Tov, G. B., & Rafaeli, A. (2023). Decision control and explanations in human-AI collaboration: Improving user perceptions and compliance, *Computers in Human Behavior*, 144, 107714.
71. Wieringa, M. (2020). What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain, 1-18.
72. Zhou, J., Chen, F., & Holzinger, A. (2022). Towards Explainability for AI Fairness, in *xxAI - Beyond Explainable AI,* A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller & W. Samek (eds.), Springer, 375-386.