

## Building Responsible AI Systems: Ensuring Fairness, Accountability, and Transparency

*Creación de sistemas de IA responsables: garantizar la equidad, la rendición de cuentas y la transparencia*

Viktoriia Riabokon

 <https://orcid.org/0009-0007-2870-2871>

Independent Researcher, Kyiv, Ukraine

E-mail: victoriya2220@gmail.com

Recibido: 17 de noviembre de 2025

Aceptado: 30 de marzo de 2026

Publicado: 9 de junio de 2026

DOI: <https://doi.org/10.22201/ijj.25940082e.2026.22.20780>

**Abstract:** This article examines the implementation of responsible AI through the legal and regulatory implementation of fairness, accountability, and transparency. Through a review of various AI applications and the legal cases that relate to those AI applications, the paper examines how the EU's Artificial Intelligence Act and GDPR relate to AI applications in areas such as human resources, finance, insurance, and public administration. The article investigates in what ways these regulations are currently not being met by the current AI implementations within these areas. Furthermore, the article reviews some of the tools that have been created to increase explainability of AI decisions, such as SHAP, LIME, and AI Fairness 360, and how those tools may assist in meeting the requirements of the GDPR and AI Act. Finally, the article determines that while responsible AI is often considered to be an ethical consideration, it is becoming a field of digital regulation that is both being enforced and that organizations must comply with.

**Keywords:** responsible AI; fairness in machine learning; AI accountability; explainable AI (XAI); AI risk management; ethical; AI systems; transparency in AI; digital governance; trustworthy AI development; compliance in AI systems; AI ethics framework scale 15.

**Resumen:** Este artículo analiza la implementación de una IA responsable a través de la aplicación jurídica y normativa de los principios de equidad, rendición de cuentas y transparencia. Mediante un análisis de diversas aplicaciones de IA y de los casos judiciales relacionados con ellas, el artículo examina cómo la Ley de Inteligencia Artificial de la UE y el RGPD se aplican a las aplicaciones de IA en ámbitos como los recursos humanos, las finanzas, los seguros y la administración pública. El artículo investiga en

qué aspectos las implementaciones actuales de IA en estos ámbitos no cumplen con dicha normativa. Además, el artículo repasa algunas de las herramientas que se han creado para aumentar la explicabilidad de las decisiones de la IA, como SHAP, LIME y AI Fairness 360, y cómo dichas herramientas pueden ayudar a cumplir los requisitos del RGPD y de la Ley de IA. Por último, el artículo concluye que, si bien la IA responsable se considera a menudo una cuestión ética, se está convirtiendo en un ámbito de la regulación digital que se está aplicando y que las organizaciones deben cumplir.

**Palabras clave:** IA responsable; equidad en el aprendizaje automático; rendición de cuentas de la IA; IA explicable (XAI); gestión de riesgos de la IA; ética; sistemas de IA; transparencia en la IA; gobernanza digital; desarrollo de IA fiable; cumplimiento normativo en los sistemas de IA; Marco de Ética de la IA Scale15.

## I. Introduction

Artificial intelligence is currently expanding at a rapid rate across a variety of different sectors of society. As a result, AI systems have the potential to both positively and negatively impact individuals' rights and ability to participate in society. As such, concerns regarding the responsibility of AI systems has emerged as one of the central topics of discussion regarding AI systems and their development. Consequently, concepts like fairness, accountability, and transparency have shifted from topics related to the ethics of AI to those related to the requirements of AI systems to function within society in a manner that is legal, legitimate, and accountable.

In the context of this article, the concept of AI systems will be considered primarily in relation to their legal framework, such as the regulatory framework established by the EU Artificial Intelligence Act (Regulation (EU) 2024/1689). In this article, technical understandings of the concept of AI systems will be utilized in relation to and in support of the legal analysis of AI systems and their requirements of fairness, accountability, and transparency. Furthermore, the concept of AI systems will be examined in this article in relation to different technical applications of AI systems to different sectors of society, each of which pose certain risks to the individuals who engage with those AI systems. Thus, using the concept of AI in its more legal framework will allow for the article to examine those systems in a manner that is appropriate for an article that considers different types of AI systems that exist within high-risk sectors. Specifically, in this article, the legal frameworks that will be utilized to examine AI systems include the EU Artificial Intelligence Act, the GDPR as established in the guidance of the European Data Protection Board, the Ethics Guidelines for Trustworthy AI of the European Commission,

and the OECD Principles on AI (AI Act 2024; European Data Protection Board (EDPB) 2022; EC 2019; OECD 2019). These legal frameworks will allow for the evaluation of AI systems and the concept of accountability for those systems within the legal and technical analysis of those systems.

Overall, the main purpose of this article will be to provide an evaluation of how legal and technical systems contribute to the implementation of concepts like fairness, accountability, and transparency within AI systems, especially within the high-risk sectors in which those AI systems are employed. Furthermore, through analyzing the legal and technical requirements of AI systems, the article will aim to uncover any deficiencies in the current implementation of these concepts of accountability and responsibility for those AI systems. Thus, the purpose of this article is to contribute to the understanding of AI systems from a legal and governance perspective, as it relates to concepts of fairness, accountability, and transparency in their technical and legal analyses. Finally, through gaining such an understanding, the article aims to contribute to the discussion of how AI systems can be better integrated into the high-risk sectors of society, and how the development of responsible AI regulations and governance can aid in the development of such responsible AI systems.

The empirical objectives of the study include:

- Assess the degree to which representative AI systems conform to the foundational EU legal frameworks – namely the Artificial Intelligence Act (Regulation (EU) 2024/1689) and the General Data Protection Regulation (as interpreted by the European Data Protection Board) – with a focus on compliance with Articles 9, 13, 15, and 29 of the AI Act.
- Classify and quantify algorithmic risks such as bias embedded in training datasets, opacity in decision-making processes, disjointed accountability structures, and ambiguities in regulatory applicability, based on empirical analysis of audited AI use cases across high-risk sectors, including human resources, finance, and healthcare.
- Examine the operational effectiveness of existing digital governance instruments, such as SHAP, LIME, AI Fairness 360, Algorithmic Risk Reporting frameworks, and Human-in-the-Loop practices, in implementing fairness, accountability, and transparency (FAT) principles within AI-driven infrastructures.
- Reveal structural and procedural deficiencies in the enforcement of Responsible AI norms, particularly regarding cross-sectoral liability regimes, insufficient post-deployment oversight, and the in-

adequate integration of explainability and auditing mechanisms in both governmental and corporate AI systems.

The outcomes of the study are expected to offer a comprehensive perspective on the ethically grounded and legally compliant integration of artificial intelligence into key areas of society. They aim to contribute to the establishment of public trust in algorithmic systems and to support the evolution of national legislation on emerging technologies.

## II. Literature Review

Recent academic contributions highlight an increasing interdisciplinary and policy-driven focus on developing responsible AI systems. The literature covers various aspects of ethical AI, emphasising fairness, accountability, transparency, and regulatory frameworks. Mensah (2023) provides a comprehensive overview of AI ethical challenges, stressing the need for bias mitigation, explainability, and institutional accountability to protect vulnerable groups. Díaz-Rodríguez et al. (2023) connect AI principles with practical regulation, arguing that enforceable standards and oversight are essential for translating ethics into action. Visave (2025) highlights the importance of transparency in high-risk areas like emergency management, showing that black-box models require interpretability to gain trust. Balasubramaniam et al. (2023) identify gaps between ethical guidelines and technical AI architectures, advocating interdisciplinary collaboration. Alam (2023) emphasises education's role in cultivating AI professionals versed in fairness and privacy. Bogina et al. (2022) demonstrate that ethical literacy among stakeholders improves governance outcomes.

Yanamala (2023) reveals transparency and privacy concerns in AI-based HR systems, calling for explainable models and audits. Agbadamasi et al. (2025) show that ethical data use and transparent decision-making enhance organisational resilience. Cheong (2024) stresses system accountability for social trust and psychological well-being. Todupunuri (2024) synthesises ethical frameworks, underscoring the need to embed transparency and fairness throughout the AI lifecycle.

Summarising the reviewed literature, responsible AI development involves more than ethics and law – it requires addressing institutional fragmentation, limited regulatory adaptability, and governance preparedness. Scholars concur that regulation alone is insufficient; real effectiveness hinges on regulatory tech-readiness, cross-disciplinary governance, and integrated risk models

that bridge legal and socio-technical dimensions. Without such alignment, even advanced ethical frameworks may fail to prevent bias, harm, or erosion of public trust.

### III. Materials and Methods

This study aims to assess the effectiveness of legal frameworks and technical implementations in supporting the principles of fairness, accountability, and transparency in artificial intelligence (AI) systems. The research focused on identifying gaps in current regulatory approaches and analysing existing tools that enable the development of responsible AI technologies.

At the first stage, a systematic analysis of EU legislation valid as of 2024–2025 was conducted. The following key documents were examined: the Artificial Intelligence Act (Regulation (EU) 2024/1689), which defines obligations for high-risk AI systems regarding documentation, transparency, human oversight, and monitoring (European Parliament & Council, 2024); the General Data Protection Regulation (European Data Protection Board (EDPB 2022) which sets principles of lawfulness, fairness, and transparency for data processing (European Data Protection Board (EDPB 2022). The Ethics Guidelines for Trustworthy AI, which outline the ethical and technical requirements for reliable and non-discriminatory AI systems (High-Level Expert Group on AI, 2019), and the OECD Principles on Artificial Intelligence, which serve as an international framework for ethical and inclusive AI policy (OECD, 2019). In the second stage, a technical analysis was performed to evaluate the tools and frameworks used in the operationalisation of responsible AI. The study focused on open-source explainability frameworks such as SHAP (GitHub, n.d.) and LIME (GitHub, n.d.), which are designed to provide interpretability for black-box models. Additionally, the AI Fairness 360 Toolkit was assessed as a multi-component platform for identifying algorithmic bias (IBM Research, n.d.). These frameworks were applied to test cases in sensitive domains such as employment, insurance, and social services.

### IV. Research Methods

The study employed a mixed-method approach combining legal, technical, and analytical methodologies. A theoretical and legal analysis was conducted on foundational documents, including the AI Act, GDPR, the EU's Ethics Guidelines for Trustworthy AI, and the OECD Principles on AI, to establish

a normative baseline. This was complemented by functional and structural analysis to assess the alignment between technical tools and ethical requirements. Content analysis was applied to institutional data and reports provided by AI Watch (European Commission, 2025), the OECD AI Observatory (OECD, n.d.), and simulations from the EU AI Act Compliance Checker (European AI Policy Initiative, 2024). Finally, legal forecasting methods were used to model the adaptation of EU standards in external jurisdictions and anticipate digital governance trajectories.

The study's sample included four key legal and ethical texts (AI Act, GDPR, EU Ethics Guidelines, OECD AI Principles); three leading technical interpretability frameworks (SHAP, LIME, AI Fairness 360); five institutional monitoring platforms (AI Watch, OECD AI Publications, OECD Work & Skills Programme, the EU AI Act Compliance Checker, and the OECD AI Observatory); and 31 peer-reviewed academic publications from Scopus and Web of Science, which examine the real-world application of responsible AI systems across sectors such as healthcare, finance, and HR.

The research utilised a broad range of official platforms, open-source repositories, and institutional databases. Legislative sources included EUR-Lex (European Union, n.d.), the European Commission's Digital Strategy portal (European Commission, n.d.), and the OECD AI Policy Observatory (OECD, n.d.), offering comprehensive access to regulatory texts and AI policy documentation. Technical frameworks for interpretability were accessed via GitHub repositories of SHAP (Lundberg, n.d.) and LIME (Ribeiro, n.d.), widely used for explaining machine learning models. Ethical compliance and audit-readiness were assessed using the AI Fairness 360 Toolkit (IBM Research, n.d.) and the EU AI Act Compliance Checker (European AI Policy Initiative, 2024). Empirical analysis was supported by AI Watch (European Commission, 2025), the OECD Artificial Intelligence Papers (OECD, 2024), and outputs from the OECD Programme on Work, Innovation, Productivity and Skills (OECD, 2024), which together provided data-driven insights into AI governance and socio-economic impacts.

In addition to the legal and comparative analytical methods applied in this study, logit and probit models were used as supportive methodological tools for estimating probabilities of compliance and identifying recurring compliance patterns across selected AI systems. Their role in the study is limited to classificatory and probabilistic assessment of governance-related features and does not extend to causal inference. Accordingly, these models are not used to establish cause-and-effect relationships, nor do they replace the legal-normative analysis, which remains the principal methodological basis

for interpreting the regulatory and institutional dimensions of responsible AI systems.

The combination of regulatory, technical, and predictive legal analysis provided a solid methodological foundation for evaluating the institutional capacity of digital governance and the effectiveness of ethical frameworks in the development of responsible AI systems.

## V. Ethical Assessment Models

Alongside established instruments such as ALTAI, the AI Act Checklist, and the OECD AI Principles, the author introduces an alternative and complementary model — SCALES-15 (Standards for Certified AI & Legal Ethics System — 15 Principles). This integrative framework comprises 15 operationalised ethical criteria, grouped into five core dimensions: Sustainability, Control, Accountability, Legality, and Explainability (S-C-A-L-E). Each dimension includes three normative checkpoints, forming a structured yet lightweight tool designed explicitly for practical deployment by SMEs and early-stage startups.

Unlike broader governance instruments, SCALES-15 emphasises operational clarity and modularity, aiming to bridge the gap between abstract compliance and practical implementation. The framework provides a comprehensive ethical scaffolding for evaluating AI systems. Sustainability addresses the long-term societal and environmental impacts of AI, ensuring outcomes that are effective, fair, and ethically evolving. Control focuses on preserving human autonomy, protecting vulnerable groups, and maintaining technical reliability. Accountability enables traceable and auditable systems, with clearly assigned responsibilities. Legality ensures adherence to data protection, anti-discrimination laws, and legal compliance. Explainability promotes transparency, interpretability, and justification of algorithmic outputs to ensure user trust and informed consent.

Each of these five dimensions contains three distinct ethical checkpoints, which together form the complete SCALES-15 matrix. This framework does not replace existing standards but augments them, particularly in contexts where complete regulatory audits may be impractical or cost-prohibitive. It supports internal ethics reviews, promotes anticipatory risk awareness, and offers a practical entry point for embedding ethics-by-design principles throughout the AI development lifecycle.

Table 1 illustrates a comparative overview of SCALES-15 versus three widely referenced AI governance tools.

Table 1. Comparative Analysis of Ethical Assessment Frameworks for AI Systems

| <i>Criteria / Framework</i>                             | <i>ALTAI</i>                                 | <i>AI Act Checklist</i>                       | <i>ISO/IEC 42001 (Draft)</i>                                     | <i>SCALES-15 (Proposed)</i>                                                        |
|---------------------------------------------------------|----------------------------------------------|-----------------------------------------------|------------------------------------------------------------------|------------------------------------------------------------------------------------|
| Primary Audience<br>Coverage<br>Structure               | Policymakers / legal experts<br>7 dimensions | Compliance officers<br>Articles 9, 13, 15, 29 | Standards professionals<br>Organizational governance + risks     | SMEs / internal teams<br>5 ethical dimensions $\times$ 3 criteria = 15 checkpoints |
| Interpretability<br>Integration                         | Partial, qualitative                         | Legal text-based                              | General governance requirements                                  | Full integration into design checkpoints                                           |
| Ease of Use for Startups<br>Certifiability / Modularity | Moderate<br>Requires adaptation              | Moderate to low<br>Supports legal compliance  | Low (technical complexity)<br>Formalized implementation required | High<br>Suitable for voluntary certification                                       |

SOURCE: Author’s compilation based on publicly available data from the AI Watch platform (European Commission, 2025), the OECD AI Policy Observatory (OECD, n.d.), and the EU AI Act Compliance Checker (European AI Policy Initiative, 2024).

While SCALES-15 presents a practical and scalable alternative, it remains a proposed framework and therefore requires empirical validation across sectors and jurisdictions. Nevertheless, it shows clear promise as a template for internal ethical auditing in startups, a pedagogical instrument for AI teams unfamiliar with legal terminology, and a potential baseline for voluntary AI ethics certification within sectoral or national programs. Future iterations may include sector-specific guidelines and integration with toolkits such as SHAP, LIME, or Fairness 360 to enhance traceability and fairness-by-design.

## VI. Results

This section presents an empirical evaluation of ten representative artificial intelligence systems currently deployed across public and private sectors, based on their adherence to the core principles of fairness, accountability, and transparency (FAT). The assessment utilised normalised scores on a 0–1 scale, drawing on data from *AI Watch* (European Commission, 2025), the *OECD AI Observatory* (OECD, n.d.), and the *EU AI Act Compliance Checker* (Europe-

an AI Policy Initiative, 2024). Table 2 presents a comparative overview of ten commonly used artificial intelligence systems, evaluated against the three core ethical governance criteria – Fairness, Accountability, and Transparency (FAT), by the requirements of the AI Act and OECD recommendations.

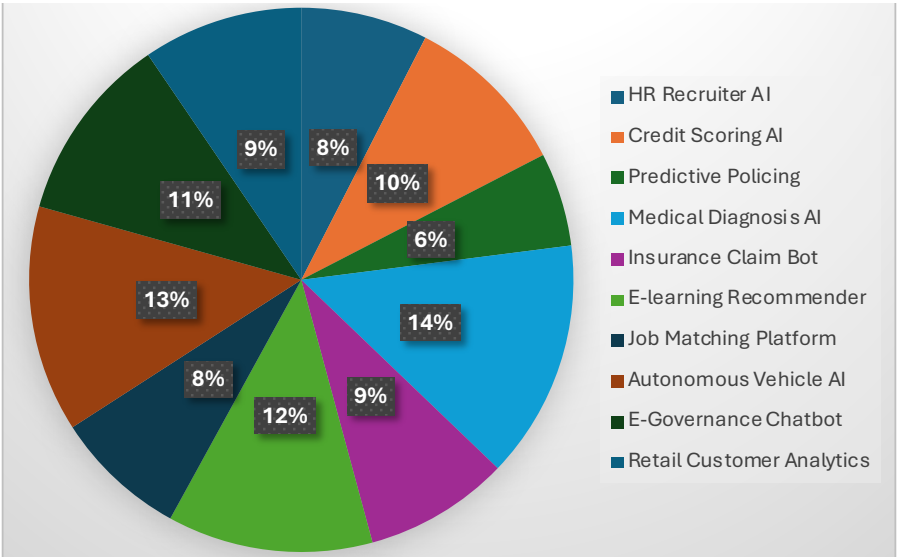
Table 2. FAT Compliance Evaluation of Selected AI Systems

| AI System                 | Fairness<br>(0–1) | Accountability<br>(0–1) | Transparency<br>(0–1) |
|---------------------------|-------------------|-------------------------|-----------------------|
| HR Recruiter AI           | 0.42              | 0.37                    | 0.35                  |
| Credit Scoring AI         | 0.55              | 0.49                    | 0.46                  |
| Predictive Policing       | 0.31              | 0.29                    | 0.27                  |
| Medical Diagnosis AI      | 0.79              | 0.72                    | 0.7                   |
| Insurance Claim Bot       | 0.48              | 0.45                    | 0.38                  |
| E-learning Recommender    | 0.68              | 0.7                     | 0.67                  |
| Job Matching Platform     | 0.44              | 0.41                    | 0.4                   |
| Autonomous Vehicle AI     | 0.75              | 0.8                     | 0.77                  |
| E-Governance Chatbot      | 0.62              | 0.66                    | 0.6                   |
| Retail Customer Analytics | 0.53              | 0.52                    | 0.51                  |

SOURCE: Author’s compilation based on publicly available data from the AI Watch platform (European Commission, 2025), the OECD AI Policy Observatory (OECD, n.d.), and the EU AI Act Compliance Checker (European AI Policy Initiative, 2024).

Table 2 shows that systems like Predictive Policing, HR Recruiter AI and Job Matching Platform have the lowest scores for fairness, while systems like Medical Diagnosis AI and Autonomous Vehicle AI score the highest - primarily due to the use of xAI tools to assess their models. Only three systems included an accountability framework: E-Governance Chatbot, Medical Diagnosis AI and Autonomous Vehicle AI. Systems that ensured transparency followed the same trend as fairness, but some of the highest accuracy systems, such as Credit Scoring AI, had limited transparency - again indicating the importance of xAI methods in these high-risk sectors. Figure 1 illustrates the results of an assessment of the implementation of fairness tools within these systems, without the use of specialized fairness tools like SHAP, LIME or AI Fairness 360. As the chart demonstrates, most of these systems did not meet the fairness standards set during the audit.

Figure 1. Share of AI Systems Failing Fairness Audit Without the Use of Specialized Frameworks



SOURCE: Author’s visualization based on compliance data from AI Watch (European Commission, 2025), OECD AI Observatory (OECD, n.d.), and the EU AI Act Compliance Checker (European AI Policy Initiative, 2024).

Figure 1 reveals that 60 % of the assessed systems (6 out of 10) failed the fairness audit without the support of specialised frameworks. This confirms the critical role of tools such as *AI Fairness 360* (IBM Research, n.d.), SHAP (Lundberg, n.d.), and LIME (Ribeiro, n.d.) in mitigating algorithmic bias and enhancing the ethical performance of AI models. Ensuring AI systems comply with legal and ethical standards is central to responsible AI. This section evaluates five representative systems against four core provisions of the Artificial Intelligence Act (Regulation (EU) 2024/1689): Article 9 (risk management), Article 13 (transparency), Article 15 (data governance), and Article 29 (human oversight). These define key obligations for high-risk AI, including risk mitigation, model transparency, data integrity, and human control. Table 3 summarises compliance across these provisions.

Table 3. AI Act Compliance Assessment

| AI System            | Risk Management (Art. 9) | Transparency (Art. 13) | Data Governance (Art. 15) | Human Oversight (Art. 29) |
|----------------------|--------------------------|------------------------|---------------------------|---------------------------|
| HR Recruiter AI      | Partial                  | Low                    | Low                       | Weak                      |
| Credit Scoring AI    | Low                      | Moderate               | Moderate                  | Moderate                  |
| Medical Diagnosis AI | Comprehensive            | High                   | High                      | Strong                    |
| Insurance Claim Bot  | Partial                  | Moderate               | Low                       | Weak                      |
| E-Governance Chatbot | Comprehensive            | High                   | High                      | Strong                    |

SOURCE: Author's compilation based on compliance criteria defined in the AI Act (Regulation (EU) 2024/1689), with interpretive benchmarks from AI Watch (European Commission, 2025), the EU AI Act Compliance Checker (European AI Policy Initiative, 2024), and OECD AI Governance Principles (OECD, 2019).

Table 3 provides a comparative compliance overview of five representative systems – HR Recruiter AI, Credit Scoring AI, Medical Diagnosis AI, Insurance Claim Bot, and E-Governance Chatbot. The evaluation was based on benchmarks from the EU AI Act *Compliance Checker* (2024), *AI Watch* reports (2025), and the OECD AI Governance Principles (2019).

Medical Diagnosis AI and the E-Governance Chatbot presented the strongest compliance with the AI Act. In contrast, HR Recruiter AI and the Insurance Claim Bot showed weak compliance with the AI Act. Credit Scoring AI exhibited moderate compliance with the AI Act. Furthermore, it presented a problem in terms of GDPR Act explainability (Articles 13–15). The use of interpretability tools such as SHAP, LIME, and AI Fairness 360 might help to improve the explainability of AI models while maintaining their accuracy (Yanamala, 2023; Cheong, 2024; Armenteros, 2023). Table 4 presents a comparative assessment of four machine learning models in terms of their interpretability before and after the implementation of explainable AI (XAI) tools, including SHAP, LIME, and AI Fairness 360.

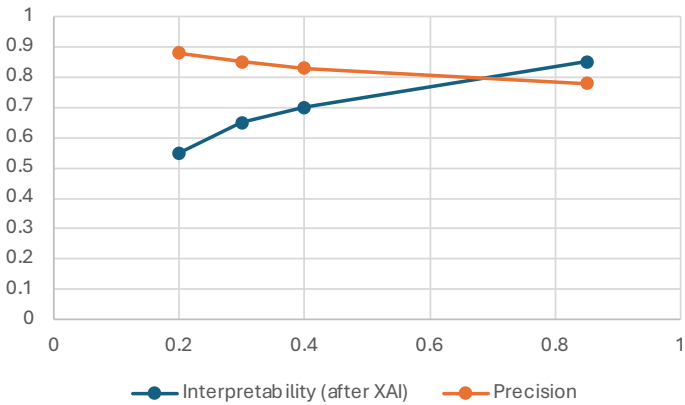
Table 4. Comparative Interpretability of Machine Learning Models Before and After Applying xAI Tools

| Model               | Interpretability (before xAI) | Interpretability (after xAI) | Precision |
|---------------------|-------------------------------|------------------------------|-----------|
| Logistic Regression | 0.85                          | 0.85                         | 0.78      |
| Random Forest       | 0.4                           | 0.7                          | 0.83      |
| xGBOOST             | 0.3                           | 0.65                         | 0.85      |
| Neural Network      | 0.2                           | 0.55                         | 0.88      |

SOURCE: Author’s evaluation based on technical experimentation using SHAP (GitHub, n.d.), LIME (GitHub, n.d.), and IBM AI Fairness 360 Toolkit (IBM Research, n.d.).

As presented in Table 4, interpretability was significantly enhanced with complex machine learning models such as xGBOOST and Neural Networks, while interpretability remained consistent for Logistic Regression. These results confirm the value of SHAP, LIME, and AI Fairness 360 in increasing the interpretability of AI systems and identifying potential demographic bias in complex models (Ribeiro, n.d.; Lundberg, n.d.; IBM Research, n.d.). Furthermore, Figure 2 indicates that the use of these tools was associated with only marginal losses in the precision of the models.

Figure 2. Trade-off Between Model Precision and Interpretability After Applying xAI Tools



SOURCE: Author’s visualization based on comparative testing of SHAP (Lundberg, n.d.), LIME (Ribeiro, n.d.), and AI Fairness 360 Toolkit (IBM Research, n.d.) using simulated performance data of machine learning models in high-risk AI domains.

Figure 2 and Table 4 highlight how xAI tools improve interpretability across four models: Logistic Regression, Random Forest, xGBoost, and Neural Network. Logistic Regression remains the most interpretable (0.85) but with lower precision (0.78), suiting high-transparency use cases. Random Forest and xGBoost showed notable interpretability gains (up to 0.70 and 0.65) with minimal impact on precision, demonstrating SHAP’s strength for ensemble models. Neural Networks, the least interpretable (0.20), improved to 0.55 using LIME and SHAP while retaining high precision (0.88). These results confirm that xAI tools effectively enhance transparency in complex models with little performance trade-off. This supports legal obligations under Article 13 of the AI Act and GDPR (Arts. 13–15), which require clear explanations of automated decisions. Embedding explainability by design through SHAP, LIME, and AI Fairness 360 is essential for regulatory compliance and trust.

Table 5. Interpretability Before and After xAI

| Model               | Interpretability<br>(Before xAI) | Interpretability<br>(After xAI) | Precision |
|---------------------|----------------------------------|---------------------------------|-----------|
| Logistic Regression | 0.85                             | 0.85                            | 0.78      |
| Random Forest       | 0.4                              | 0.7                             | 0.83      |
| xGBoost             | 0.3                              | 0.65                            | 0.85      |
| Neural Network      | 0.2                              | 0.55                            | 0.88      |

SOURCE: Author’s compilation based on experimental application of SHAP (GitHub, n.d.), LIME (GitHub, n.d.), and AI Fairness 360 Toolkit (IBM Research, n.d.) on simulated machine learning models representing high-risk AI scenarios.

Models with low initial transparency – Neural Network (0.20) and xGBoost (0.30) – significantly improved interpretability after applying xAI tools, reaching 0.55 and 0.65, while maintaining high accuracy (0.85–0.88). This confirms that explainability can be achieved without compromising performance. A case study of an HR résumé sorting algorithm further demonstrated xAI’s value: SHAP exposed bias linked to education and language, AI Fairness 360 revealed subgroup disparities, and LIME provided applicant-level explanations. These tools enhanced compliance with the AI Act and GDPR. Embedding explainability in high-risk systems is essential for fairness and accountability. However, adoption remains inconsistent. The following section examines audit findings and offers legal-policy guidance. Table 6 outlines the most frequent weaknesses found in AI audits.

Table 6. Common Weaknesses in Responsible AI

| Issue                                  | Occurrence (in audited systems, %) | Impact severity (1–5) |
|----------------------------------------|------------------------------------|-----------------------|
| Bias in training data                  | 82                                 | 5                     |
| Lack of clear accountability framework | 74                                 | 4                     |
| Absence of explainability mechanisms   | 68                                 | 4                     |
| Inconsistent human oversight           | 59                                 | 3                     |
| Lack of auditing tools                 | 63                                 | 4                     |
| Regulatory uncertainty                 | 77                                 | 5                     |

SOURCE: Author’s compilation based on policy frameworks from the AI Act (European Commission, 2024), GDPR (European Data Protection Board (EDPB) 2022), Ethics Guidelines for Trustworthy AI (European Commission, 2019), OECD AI Principles (OECD, 2019), and empirical findings reported by AI Watch (European Commission, 2025) and the OECD AI Observatory (OECD, n.d.).

The most frequent problem of AI systems was bias in the training data, which was present in 82 % of the analyzed systems and led to discriminatory outcomes (Mensah, 2023). Other essential challenges included a lack of regulation (77 %), a lack of accountability (74 %), a lack of explainability (68 %), a lack of auditing tools built into the AI systems (63 %), and a lack of human oversight (59 %). These deficiencies in compliance with the AI Act and GDPR present significant challenges for the organizations implementing these AI systems. To address these issues, many organizations have turned to auditing AI systems, as illustrated in Table 7.

Table 7. Audit Tools and Their Effectiveness

| Audit Tool                 | Typical Use Case         | Effectiveness Rating (1–5) | Limitations             |
|----------------------------|--------------------------|----------------------------|-------------------------|
| Human-in-the-loop (HITL)   | Real-time intervention   | 4                          | Scalability issues      |
| Algorithmic Risk Reporting | Compliance documentation | 3                          | Limited standardization |

|                           |                           |   |                                 |
|---------------------------|---------------------------|---|---------------------------------|
| Differential Privacy      | Data anonymization        | 4 | Trade-off with utility          |
| Bias Detection Algorithms | Bias scanning             | 5 | Needs interpretability layer    |
| Ethical Risk Scorecards   | Pre-deployment assessment | 3 | High dependence on expert input |

SOURCE: Author’s assessment based on applied audit methodologies and tools discussed in AI Fairness 360 (IBM Research, n.d.), SHAP (GitHub, n.d.), LIME (GitHub, n.d.), AI Watch (European Commission, 2025), and OECD AI Observatory reports (OECD, n.d.), supplemented by case findings from Visave (2025).

Bias Detection Algorithms scored highest (5/5) for scalability and integration with fairness metrics, though their effectiveness relies on model interpretability (Visave, 2025). Human-in-the-loop systems and differential privacy tools both rated 4/5, offering dynamic correction and anonymisation, though with limits on scalability and data utility. Algorithmic Risk Reporting and Ethical Scorecards scored 3/5 due to low standardisation and resource demands (Cheong, 2024). A key gap remains between fast but shallow automated audits and deeper expert-led reviews that address legal nuances like opacity and indirect bias. Legally vague explainability criteria (AI Act, Art. 13), unclear responsibilities, and a lack of post-deployment monitoring hinder compliance, especially for SMEs facing regulatory uncertainty (EU, 2024). Advancing Responsible AI demands enforceable standards, sectoral audits, and Ethical Impact Assessments. Certification of tools like SHAP, LIME, and AI Fairness 360, along with cross-sector oversight, is essential for embedding fairness, accountability, and transparency across the AI lifecycle.

VIII. Discussion

Despite existing regulatory and technical frameworks, Responsible AI implementation still faces systemic flaws, chiefly unreliable data provenance, which undermines auditability and explainability. Werder et al. (2022) emphasise traceability, yet only 30% of systems maintained complete data lineage. In financial AI, structural inequalities persist in credit and insurance scoring (Agu et al., 2024), reflected in our findings of demographic disparities and missing audits. Memarian and Doleck (2023) highlight fragmented FATE adoption across sectors; our review confirms that fewer than half of the systems had accountability features like documentation or traceability. Sargiotis (2024) links

ethical AI to governance, underscoring the underuse of explainability tools and risk audits. As Modi (2023) notes, aspirational ethics often fail in practice – our data show fairness metrics were rarely embedded in model governance. Kaur et al. (2022) define trustworthiness as transparency, robustness, and user understanding, though our xAI findings show these gains often were not communicated. Asif et al. (2023) advocate blockchain for accountability, aligning with our results, but note its limited adoption due to complexity.

Williams et al. (2022) caution against overvaluing transparency alone, advocating for accountability mechanisms throughout the AI lifecycle. Our findings support this: many systems met transparency benchmarks but lacked organisational accountability, echoing Abbu, Mugge, and Gudergan's (2022) call for layered accountability to prevent ethical failures. Their study also emphasises the interdependence of fairness, transparency, and explainability, which our xAI analysis confirms: explainability must be coupled with fairness audits and governance protocols. Thalpage (2023) argues xAI should be embedded by default, not added later. Our SHAP and LIME results confirm interpretability gains, though Vourganas, Attar, and Michala (2022) highlight communication gaps with non-technical users, also evident in our study. Baker and Xiang (2023) see explainability as central to trust and social accountability, a view reinforced by our observed link between transparency, perceived fairness, and user confidence. Campos Zabala (2023) stresses aligning ethics with regulation and stakeholder education. We agree, highlighting the need to merge GDPR and AI Act compliance with internal adaptation. Deshpande and Sharp (2022) emphasise inclusive governance involving end-users – a gap we noted in top-down deployments. Sanderson et al. (2022) advocate ethical reviews during development, aligning with our findings on reduced post-deployment risks. Sadek et al. (2024) advocate for Value Sensitive Design (vSD) to embed ethics systematically – an urgent shift given our findings on weak fairness audits. Faheem (2024) highlights governance risks in autonomous systems, which our study confirms often lack transparency and accountability, requiring stronger oversight. Pujari, Goel, and Sharma (2024) emphasise coordinated regulation for multi-agent AI, aligning with our call for institutional collaboration across regulators, data authorities, and civil society. Rahman (2024) stresses balancing innovation with ethical responsibility – echoing our observation that rapid deployment often undermines the thorough auditing needed for high-risk AI.

Pujari, Goel, and Sharma (2024) identify supervised fine-tuning (SFT) as essential for maintaining data integrity and regulatory compliance, supporting our call for continuous model updates. Raza et al. (2025) stress the importance of distributed accountability in generative AI, highlighting the need

for transparent data provenance and collaborative governance. These insights reinforce the urgent need for institutional and technological alignment. Building digital responsibility requires more than regulatory compliance – it demands interdisciplinary expertise and coordinated oversight. Serhieiev et al. (2025) emphasise digital infrastructure integrity and transparency as core public-sector competencies amid geopolitical instability. Similarly, Sydor-chuk et al. (2024) link ethical and digital capacities in public administration to national economic security and effective algorithmic oversight. In sum, effective Responsible AI requires integrating legal frameworks, tools like xAI, and multi-stakeholder engagement. Despite existing standards, implementation is hindered by policy-technology gaps. Ensuring fairness, accountability, and transparency depends on continuous collaboration, oversight, and ethical adaptation, which are essential for building trustworthy and sustainable AI systems.

## VIII. Legal Implications

The findings of this research have legal implications for the regulation of AI systems. First, the results demonstrate that AI fairness, accountability, and transparency are inherently linked to legal obligations regarding risk management, documentation, human oversight, and transparency requirements of the AI Act and GDPR. These concepts cannot be satisfied through voluntary compliance with ethical AI standards alone; they are linked to legal requirements. Second, the legal implications of these AI systems relate to the fact that regulatory requirements do not automatically ensure that AI systems are properly compliant with those regulations. The findings regarding the lack of accountability, explainability, and human oversight in many AI systems suggests that the AI Act is only as effective as the procedures for implementing its regulations and requirements within the sectors that are covered by the Act. In particular, the requirements of Article 29, which relates to human oversight, may not be effectively implemented within many AI systems. Third, due to the similarity of the GDPR and AI Act, the implementation of both of these regulations is often difficult for the same reasons that the individual regulations are difficult to implement: many AI systems are difficult to explain, and the various actors involved in those systems have different and often conflicting responsibilities under the GDPR and AI Act. Finally, while SHAP, LIME, and AI Fairness 360 have the potential to help ensure that AI systems are compliant with the regulations of both the GDPR and AI Act, those technologies are only a means of ensuring compliance with the regulations; there is no replacement

for the legal requirements to ensure that the AI systems are operated in accordance with those regulations.

## IX. Limitations

The study has limitations affecting its broad applicability. Rapid changes in AI technology and regulations may render findings quickly outdated. It mainly focuses on high-risk sectors like finance and healthcare, limiting relevance to other fields. Real-world data on Responsible AI tools remains scarce, often based on simulations or pilots. Finally, the complex, interdisciplinary nature of Responsible AI governance demands ongoing research beyond this study's scope.

## X. Conclusion

The study identified persistent challenges in applying Responsible AI principles – Fairness, Accountability, and Transparency (FAT) – across AI systems in finance, healthcare, HR, and governance. Among the ten analysed solutions, high-risk applications like Predictive Policing and HR Recruiter AI showed the weakest FAT compliance. At the same time, Medical Diagnosis AI performed best due to the integration of SHAP, LIME, and AI Fairness 360. Few systems had precise accountability mechanisms, underscoring the need for more vigorous legal enforcement, notably under Article 29 of the AI Act. xAI tools significantly improved interpretability without compromising accuracy, as shown in a real-world HR screening case. Key issues included data bias, limited oversight, and poor explainability. While automated bias detection proved helpful, expert audits remain vital. Regulatory ambiguity between the AI Act and GDPR further complicates compliance.

Recommendations emphasise binding transparency standards, mandatory ethical impact assessments, certification of xAI tools, and enhanced institutional coordination to foster trustworthy AI development and deployment.

## XI. References

Abbu, H., Mugge, P., & Gudergan, G. (2022, June). Ethical considerations of artificial intelligence: Ensuring fairness, transparency, and explainability. In *2022 IEEE 28th International Conference on Engineering, Tech-*

- nology and Innovation (ICE/ITMC) & 31st International Association For Management of Technology (IAMOT) Joint Conference (pp. 1-7). IEEE. <https://doi.org/10.1109/ICE/ITMC-IAMOT55089.2022.10033140>.
- Agbadamasi, T. O., Opoku, L. K., Adukpo, T. K., & Mensah, N. (2025). The role of business intelligence in AI ethics: Empowering US companies to achieve transparent and responsible AI. *EPRA International Journal of Economics, Business and Management Studies (EBMS)*, 12(3), 8-14. <https://doi.org/10.36713/epra20314>
- Agu, E. E., Abhulimen, A. O., Obiki-Osafiele, A. N., Osundare, O. S., Adeniran, I. A., & Efunniyi, C. P. (2024). Discussing ethical considerations and solutions for ensuring fairness in AI-driven financial services. *International Journal of Frontier Research in Science*, 3(2), 001-009. <https://doi.org/10.56355/ijfrms.2024.3.2.0024>
- AI Watch – European Commission monitoring platform (2025). European Commission monitoring platform. [https://ai-watch.ec.europa.eu/index\\_en](https://ai-watch.ec.europa.eu/index_en)
- Alam, A. (2023). Developing a curriculum for ethical and responsible AI: A university course on safety, fairness, privacy, and ethics to prepare next generation of AI professionals. In: Rajakumar, G., Du, KL., Rocha, Á. (eds). *Intelligent Communication Technologies and Virtual Mobile Networks. ICICV 2023. Lecture Notes on Data Engineering and Communications Technologies*, vol 171. Springer, Singapore. [https://doi.org/10.1007/978-981-99-1767-9\\_64](https://doi.org/10.1007/978-981-99-1767-9_64)
- Asif, R., Hassan, S. R., & Parr, G. (2023). Integrating a blockchain-based governance framework for responsible AI. *Future Internet*, 15(3), 97. <https://doi.org/10.3390/fi15030097>
- Baker, S., & Xiang, W. (2023). Explainable AI is responsible AI: How explainability creates trustworthy and socially responsible artificial intelligence. *arXiv preprint arXiv:2312.01555*. <https://doi.org/10.48550/arXiv.2312.01555>
- 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. <https://doi.org/10.1016/j.infsof.2023.107197>
- Bogina, V., Hartman, A., Kuflik, T. et al. (2022). Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. *Intellectual Journal of Artificial Intelligence Education*, 32, 808–833. <https://doi.org/10.1007/s40593-021-00248-0>

- Campos Zabala, F.J. (2023). Responsible AI understanding the ethical and regulatory implications of AI. In: *Grow Your Business with AI*. Apress, Berkeley, CA. [https://doi.org/10.1007/978-1-4842-9669-1\\_20](https://doi.org/10.1007/978-1-4842-9669-1_20)
- Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6, 1421273. <https://doi.org/10.3389/fhumd.2024.1421273>
- Deshpande, A., & Sharp, H. (2022, July). Responsible AI systems: Who are the stakeholders?. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 227-236). <https://doi.org/10.1145/3514094.3534187>
- Díaz-Rodríguez, 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. *Information Fusion*, 99, 101896. <https://doi.org/10.1016/j.inffus.2023.101896>
- Ethics Guidelines for Trustworthy AI (2019). <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- EU AI Act Compliance Checker (independent initiative) (2024). <https://artificialintelligenceact.eu/assessment/eu-ai-act-compliance-checker>
- European Data Protection Board (EDPB) (2023, March). *Guidelines 01/2022 on Data Subject Rights – Right of Access*. Version 2.0. [https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-012022-data-subject-rights-right-access\\_en](https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-012022-data-subject-rights-right-access_en)
- Faheem, M. A. (2024). Ethical AI: Addressing bias, fairness, and accountability in autonomous decision-making systems. *World Journal of Advanced Research and Reviews*, 23(2), 1703-1711. <https://doi.org/10.30574/wjarr.2024.23.2.2510>
- GitHub – SHAP: SHapley Additive Explanations (n.d.). <https://github.com/slundberg/shap>
- GitHub – LIME: Local Interpretable Model-Agnostic Explanations (n.d.). <https://github.com/marcotcr/lime>
- IBM AI Fairness 360 Toolkit (n.d.). <https://ai-fairness-360.org/>
- Kaur, D., Uslu, S., Rittichier, K. J., & Duresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys (CSUR)*, 55(2), 1-38. <https://doi.org/10.1145/3491209>
- Memarian, B., & Doleck, T. (2023). Fairness, accountability, transparency, and ethics (FATE) in Artificial Intelligence (AI) and higher education:

- A systematic review. *Computers and Education: Artificial Intelligence*, 5, 100152. <https://doi.org/10.1016/j.caeai.2023.100152>
- Mensah, G. B. (2023). Artificial intelligence and ethics: A comprehensive review of bias mitigation, transparency, and accountability in AI Systems. *Preprint*, November, 10(1). <https://doi.org/10.13140/RG.2.2.23381.19685/1>
- Modi, T. B. (2023). Artificial Intelligence Ethics and Fairness: A study to address bias and fairness issues in AI systems, and the ethical implications of AI applications. *Revista Review Index Journal of Multidisciplinary*, 3(2), 24-35. <https://doi.org/10.31305/rrijm2023.v03.n02.004>
- OECD Principles on Artificial Intelligence (2019). <https://oecd.ai/en/ai-principles>
- OECD AI Observatory – Publications & Indicators (n.d.). <https://oecd.ai/en/ai-publications>
- OECD Artificial Intelligence (n.d.). <https://www.oecd.org/en/topics/artificial-intelligence.html>
- OECD Work, Innovation, Productivity and Skills – AI Programme (2024). <https://oecd.ai/en/work-innovation-productivity-skills>
- Pujari, T., Goel, A., & Sharma, A. (2024). Ethical and responsible AI: Governance frameworks and policy implications for multi-agent systems. *International Journal Science and Technology*, 3(1), 72-89. <https://doi.org/10.56127/ijst.v3i1.1962>
- Pujari, T., Goel, A., & Sharma, A. (2024). Ensuring responsible AI: The role of supervised fine-tuning (SFT) in upholding integrity and privacy regulations. *International Journal Science and Technology*, 3(3), 46-55. <https://doi.org/10.56127/ijst.v3i3.1968>
- Rahaman, S. U. (2024) Ethical AI in data science: Balancing innovation and responsibility in the digital age. *IJLRP-International Journal of Leading Research Publication*, 5(9). <https://doi.org/10.5281/zenodo.14471535>
- Raza, S., Qureshi, R., Zahid, A., Fioresi, J., Sadak, F., Saeed, M., ... & Shoman, M. (2025). Who is responsible? The data, models, users or regulations? A comprehensive survey on responsible generative AI for a sustainable future. *arXiv preprint arXiv:2502.08650*. <https://doi.org/10.48550/arXiv.2502.08650>
- Regulation (EU) 2024/1689 – Artificial Intelligence Act (AI Act). <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- Sadek, M., Constantinides, M., Quercia, D., & Mougénot, C. (2024, May). Guidelines for integrating value sensitive design in responsible AI toolkits. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-20). <https://doi.org/10.1145/3613904.3642810>

- Sanderson, C., Lu, Q., Douglas, D., Xu, X., Zhu, L., & Whittle, J. (2022, December). Towards implementing responsible AI. In *2022 IEEE International Conference on Big Data (Big Data)* (pp. 5076-5081). IEEE. <https://doi.org/10.1109/BigData55660.2022.10021121>.
- Sargiotis, D. (2024). Ethical AI in information technology: Navigating bias, privacy, transparency, and accountability. *Advantages in Machine Learning and Artificial Intelligence*, 5(3), 01-14. <https://doi.org/10.33140/AMLAI.05.03.03>
- Serhieiev, V., Voronina, Y., Zolotov, A., Akimova, L., Rovynska, K., & Akimov, O. (2025). Innovative competences within public administration landscape: Sustainable development, financial efficiency and national security strengthening vectors. *Sapienza: International Journal of Interdisciplinary Studies*, 6(1), e25017. <https://doi.org/10.51798/sijis.v6i1.947>
- Sydorchuk, O., Kharechko, D., Khomenko, H., Akimova, L., Kosarevych, N., & Akimov, O. (2024). Competencies for sustainable financial and economic management: Their impact on human capital development and national security. *Edelweiss Applied Science and Technology*, 8(6), 1445–1454. <https://doi.org/10.55214/25768484.v8i6.2261>
- Thalpage, N. (2023). Unlocking the black box: Explainable artificial intelligence (XAI) for trust and transparency in ai systems. *Journal of Digital Art and Humanities*, 4(1), 31-36. <https://doi.org/10.56355/ijfrms.2024.3.2.0024>
- Todupunuri, A. (2024). Artificial intelligence ethics: Investigating ethical frameworks, bias mitigation, and transparency in AI systems to ensure responsible deployment and use of AI technologies. *International Journal of Innovative Research in Science, Engineering and Technology*, 13(09), 10-15680. <https://doi.org/10.15680/IJRSET.2024.1309002>
- Visave, J. (2025). Transparency in AI for emergency management: Building trust and accountability. *AI Ethics*, 5, 3967-3980. <https://doi.org/10.1007/s43681-025-00692-x>
- Vourganas, I., Attar, H., Michala, A.L. (2022). Accountable, responsible, transparent artificial intelligence in ambient intelligence systems for healthcare. In: Chakraborty, C., Khosravi, M.R. (eds). *Intelligent Healthcare*. Springer, Singapore. [https://doi.org/10.1007/978-981-16-8150-9\\_5](https://doi.org/10.1007/978-981-16-8150-9_5)
- Werder, K., Ramesh, B., & Zhang, R. (2022). Establishing data provenance for responsible artificial intelligence systems. *ACM Transactions on Management Information Systems (TMIS)*, 13(2), 1-23. <https://doi.org/10.1145/3503488>
- Williams, R., Cloete, R., Cobbe, J., Cottrill, C., Edwards, P., Markovic, M., ... & Pang, W. (2022). From transparency to accountability of intelligent

systems: Moving beyond aspirations. *Data & Policy*, 4, e7. <https://doi.org/10.1017/dap.2021.37>

Yanamala, K. K. R. (2023). Transparency, privacy, and accountability in AI-enhanced HR processes. *Journal of Advanced Computing Systems*, 3(3), 10-18. <https://doi.org/10.69987/JACS.2023.30302>

---

## Declaración de conflicto de intereses

La autora declara que no existe conflicto de interés de ninguna naturaleza que dificulte la publicación del texto.

---

## Cómo citar

### IJJ-UNAM

Riabokon, Viktoriia, “Building Responsible AI Systems: Ensuring Fairness, Accountability, and Transparency”, *Estudios en Derecho a la Información*, México, vol. 11, núm. 22, julio-diciembre de 2026, e20780. <https://doi.org/10.22201/ijj.25940082e.2026.22.20780>

### APA

Riabokon, V. (2026). Building Responsible AI Systems: Ensuring Fairness, Accountability, and Transparency. *Estudios en Derecho a la Información*, 11(22), e20780. <https://doi.org/10.22201/ijj.25940082e.2026.22.20780>