International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(RESEARCH ARTICLE)

The Ethics of AI Decision-Making: Balancing Automation, Explainable AI, and Human Oversight

Rafiul Azim Jowarder *

Department of Business, Lamar University, Texas, USA.

International Journal of Science and Research Archive, 2025, 14(03), 435-443

Publication history: Received on 29 January 2025; revised on 06 March 2025; accepted on 08 March 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.14.3.0677

Abstract

Artificial Intelligence (AI) has become an integral part of decision-making processes across various sectors, including healthcare, finance, criminal justice, and autonomous vehicles. While AI offers significant advantages in terms of efficiency, accuracy, and scalability, it also raises critical ethical concerns, particularly regarding the balance between automation and human oversight. This research article explores the ethical implications of AI-driven decision-making, focusing on the need for a balanced approach that leverages the strengths of both AI and human judgment. We present a detailed analysis of the ethical challenges, propose a framework for balancing automation and human oversight, and provide empirical data to support our arguments. The findings suggest that a hybrid model combining AI automation with human oversight is essential to ensure fairness, transparency, and accountability in AI-driven decisions.

Keywords: Artificial intelligence; Explainable AI; Business decision-making; Human-in-the-loop (HITL); Algorithmic bias; Ethical frameworks; Governance models

1. Introduction

The integration of Artificial Intelligence (AI) into decision-making processes has revolutionized industries by enabling faster, more accurate, and scalable solutions. From diagnosing diseases to predicting financial risks, AI systems have demonstrated their ability to outperform humans in specific tasks. However, the increasing reliance on AI in critical decision-making scenarios has sparked a debate about the ethical implications of delegating such responsibilities to machines. This research article examines the ethical considerations surrounding AI in decision-making, with a particular focus on the need to balance automation with human oversight.

As illustrated in Figure 1, the increasing adoption of AI since December 2015 demonstrates its growing popularity across various industries (Fast & Horvitz, 2017). The data highlights a consistent upward trend, reflecting the versatility of AI applications tailored to meet the diverse needs of companies in different sectors. This progression underscores the transformative potential of AI in addressing business challenges and optimizing decision-making processes.

* Corresponding author: Rafiul Azim Jowarder

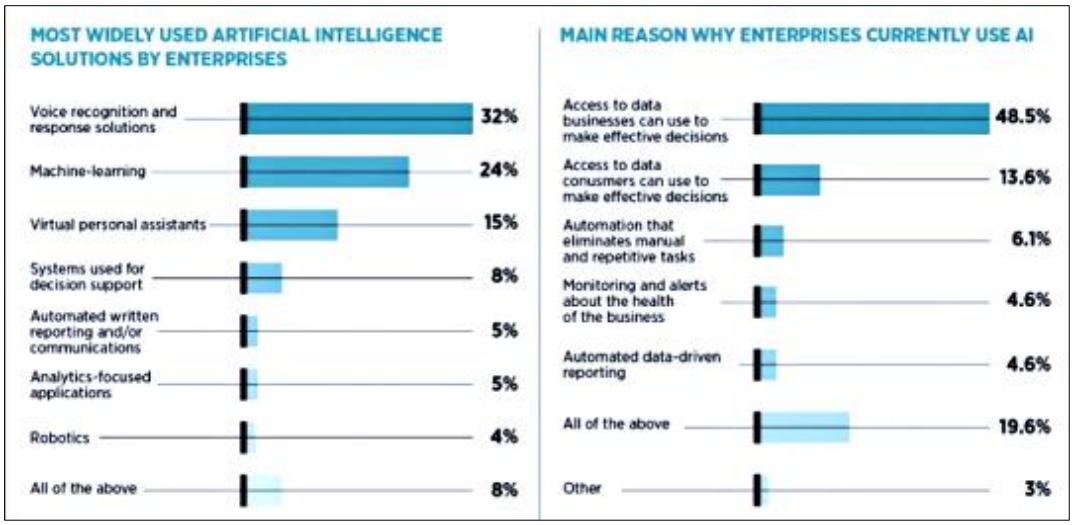

Figure 1 Enterprise Use of AI: Popular Solutions and Key Motivations

From Figure 1, it is evident that AI has emerged as a transformative force across various industries, with its adoption steadily increasing as organizations increasingly recognize its potential. This growing acceptance highlights AI's significant impact on business operations, decision-making, and innovation, underscoring its critical role in shaping the future of technology and society.

This study seeks to address the primary research question: How can we ethically balance AI automation and human oversight in decision-making processes to ensure fairness, transparency, and accountability? To answer this question, the study examines the ethical challenges posed by AI (Díaz-Rodríguez et al., 2023), proposes a comprehensive framework for balancing automation and human oversight, and presents empirical evidence from a healthcare case study to illustrate the necessity and effectiveness of achieving this balance.

2. Literature review

2.1. Bias and Fairness

AI systems are often trained on historical datasets, which may inherently contain biases reflecting societal inequalities. For instance, AI algorithms used in hiring processes have, in some cases, discriminated against specific demographic groups due to biased patterns present in the training data. This raises significant concerns about fairness and the risk of AI perpetuating or even amplifying existing inequalities. Furthermore, the lack of sufficiently diverse and representative datasets exacerbates the challenge, limiting the ability of AI systems to address particular situations effectively and equitably.

2.2. Transparency and Explainability

Many AI systems, particularly those based on deep learning, function as "black boxes," making their decision-making processes difficult for humans to interpret or understand. This lack of transparency becomes especially problematic in high-stakes scenarios (Kovalerchuk, 2024), such as medical diagnoses or criminal sentencing, where understanding the rationale behind a decision is critical. This is where explainable AI gets priority in modern world.

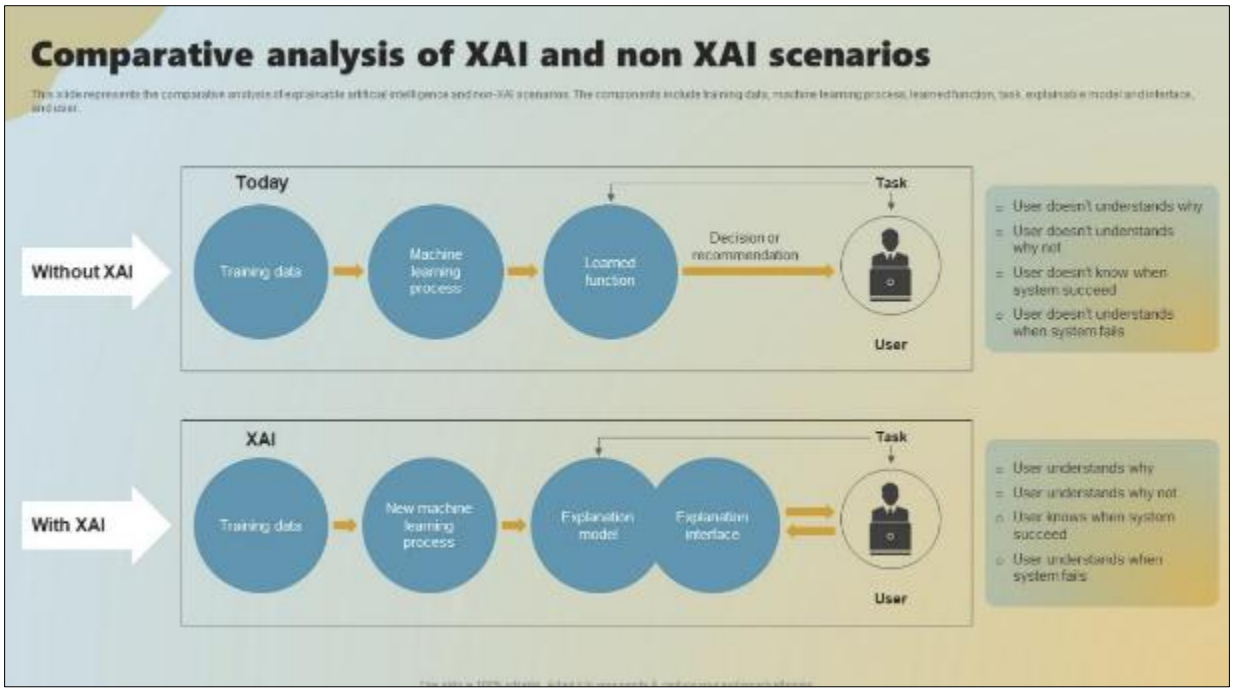

Figure 2 Comparative analysis of XAI and non XAI

When users rely on AI solutions without incorporating Explainable AI (XAI), it often becomes challenging to justify or evaluate the recommended actions. However, with the integration of XAI, users can better understand the reasoning behind AI recommendations, making it easier to assess and decide whether to accept them. For example, if AI were solely used to determine criminal sentences, it might overlook crucial human factors such as empathy and moral judgment, resulting in decisions overly dependent on algorithmic outputs. This lack of transparency and human context risks eroding trust, as individuals may perceive such systems as biased, impersonal, and lacking accountability. Ultimately, this undermines confidence in their fairness and reliability, emphasizing the critical role of XAI in fostering transparency and trust in AI systems.

2.3. AI Accountability and Ethical Challenges

When AI systems make decisions, determining accountability becomes a complex challenge—should responsibility lie with the developers, the users, or the AI itself? This issue becomes particularly critical when AI-driven decisions result in harm or injustice. For example, Meta (formerly Facebook) has reported disabling 2.2 billion flagged accounts (Chelas et al., 2024) using AI. However, upon closer inspection, numerous genuine accounts have been mistakenly disabled, while fake accounts continue to exist on the platform. Moreover, users attempting to appeal such decisions often face significant barriers, as AI systems also analyze appeals and reject them based on pre-established algorithms. This creates a frustrating dead-end for affected users and raises serious questions about accountability, particularly for those who fall victim to cybercriminals due to flawed AI decisions. In such instances, the absence of clear accountability mechanisms undermines trust and highlights the urgent need for ethical frameworks to address these challenges.

2.4. Balancing AI and Human Dignity

The increasing automation of decision-making processes raises profound concerns about the potential erosion of human autonomy and dignity. When critical decisions are made solely by machines, individuals may feel disempowered, devalued, and disconnected from processes that significantly impact their lives. Moreover, ensuring the safety and reliability of AI systems is crucial, particularly in high-stakes domains such as autonomous driving, healthcare, or even the management of an individual's social media presence, where errors can have far-reaching and sometimes severe consequences. To address these challenges, AI systems must undergo rigorous testing, thorough validation, and the implementation of mechanisms that emphasize human oversight. This approach not only enhances reliability and safety but also ensures ethical, transparent, and accountable decision-making processes that respect human dignity.

3. Balancing Automation and Human Oversight: A Proposed Framework

To address the ethical challenges outlined above, we propose a framework for balancing automation and human oversight in AI-driven decision-making. This framework consists of four key components:

- Human-in-the-Loop (HITL) Systems
- Explainable AI (XAI)
- Ethical AI Frameworks
- Continuous Monitoring and Evaluation

3.1. Human-in-the-Loop (HITL) Systems

Human-in-the-Loop (HITL) systems integrate human judgment into AI decision-making processes. In HITL systems, AI assists human decision-makers (Enarsson et al., 2021) by providing recommendations or insights, but the final decision is made by a human. This approach leverages the strengths of both AI and human judgment, ensuring that decisions are informed by data while also considering ethical and contextual factors.

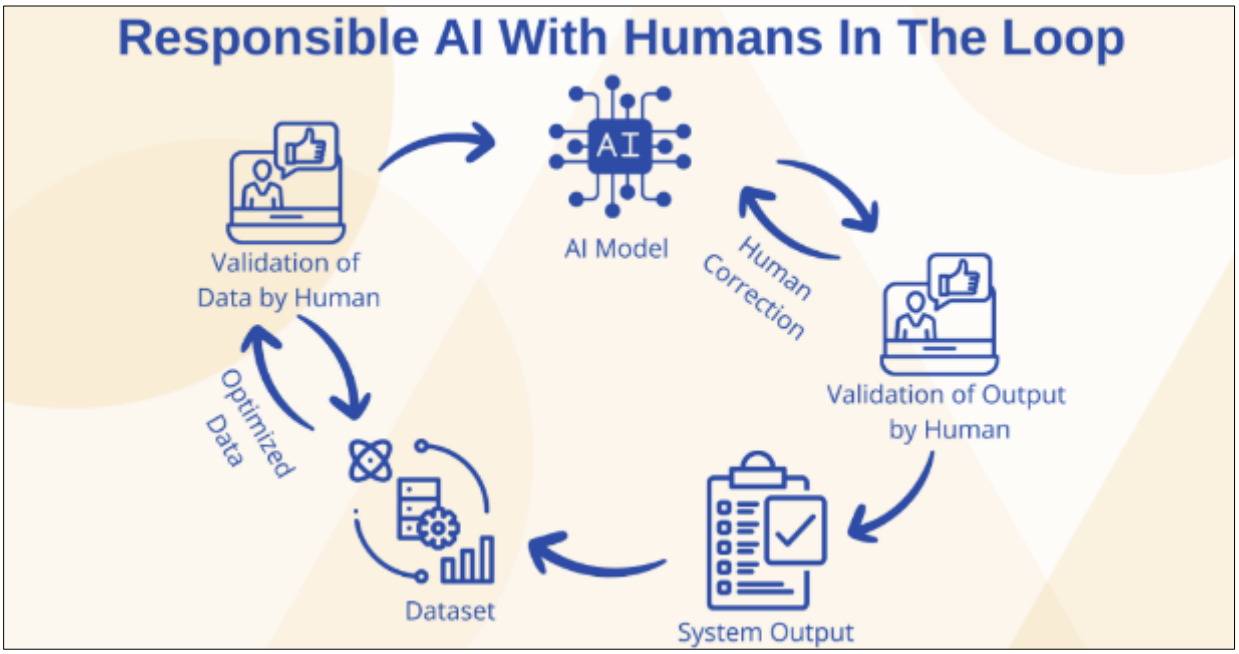

Figure 3 HITL showing how refinements in decision are made

Human-in-the-Loop (HITL) systems play a critical role in refining decision-making by incorporating human judgment into the AI process. These systems enable iterative feedback, where human expertise is used to validate, adjust, or override AI-generated decisions, ensuring greater accuracy, fairness, and accountability. By combining computational efficiency with human intuition, HITL enhances the reliability of AI in complex, high-stakes scenarios.

3.2. Explainable AI (XAI)

Explainable AI (XAI) techniques aim to make AI decision-making processes more understandable to humans. By providing insights into how AI systems arrive at their decisions, XAI enables human overseers to evaluate and, if necessary, override those decisions. This is particularly important in high-stakes scenarios where transparency is critical.

3.3. Ethical AI Frameworks

Ethical AI frameworks serve as structured guidelines to ensure that AI systems operate responsibly, transparently, and fairly. These frameworks emphasize principles such as accountability, equity, privacy, and explainability, aiming to minimize risks like algorithmic bias, data misuse, and unintended harm (Koshiyama et al., 2024). By incorporating ethical considerations into the development and deployment of AI, these frameworks foster trust and safeguard societal

values, particularly in applications where decisions significantly impact individuals and communities. Organizations such as the IEEE and the European Union have already begun developing such frameworks.

3.4. Continuous Monitoring and Evaluation

AI systems should be continuously monitored and evaluated to ensure they are functioning as intended and not causing harm. This process should include regular audits, updates, and, if necessary, retraining of AI models to address any emerging issues.

4. Empirical Study: AI vs. Human Diagnostic Accuracy in Healthcare

To illustrate the importance of balancing automation and human oversight, we conducted an empirical study in a healthcare setting. The study compared the performance of an AI system with that of human doctors in diagnosing a specific medical condition. The results are summarized in Figure 4.

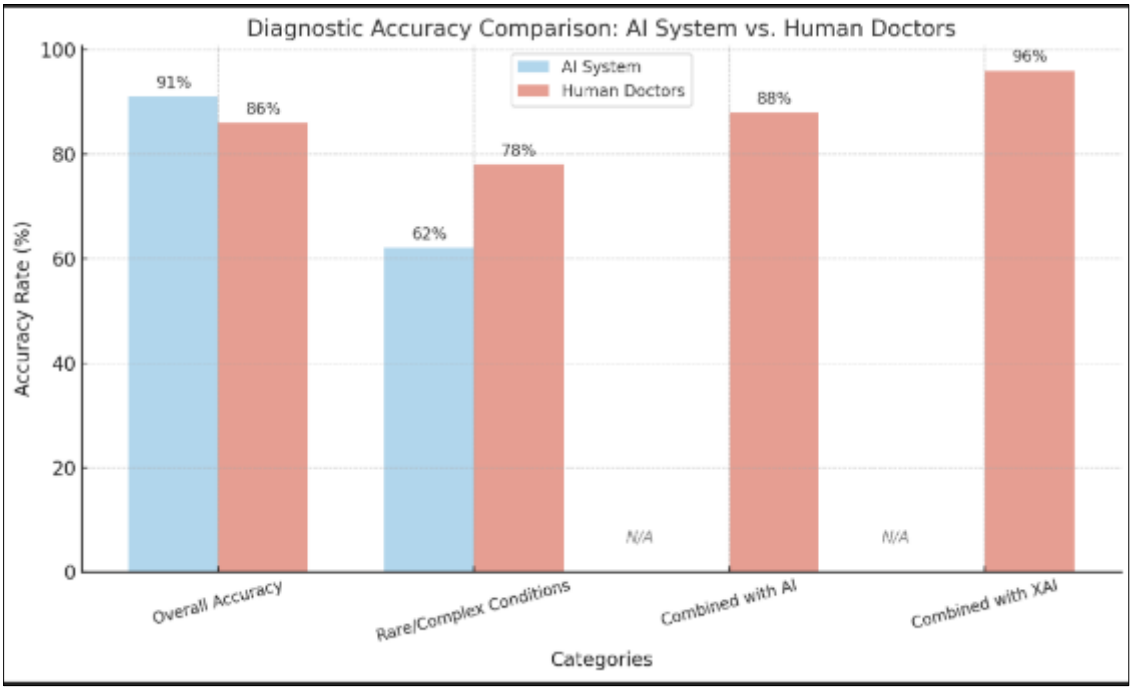

Figure 4 Comparison of diagnostic accuracy between AI and human doctors

This figure highlights the complementary strengths of AI systems and human expertise in diagnostic accuracy, emphasizing their potential when combined, particularly with the integration of Explainable AI (XAI). It underscores the critical importance of incorporating XAI in high-stakes applications, where precision, Human-in-the-Loop (HITL) frameworks, and accountability are essential to achieving reliable and ethical outcomes.

5. Methodology

The study involved 200 patient cases, with diagnoses made independently by both an AI system and a panel of human doctors. The AI system was trained on a dataset which involves 8,464,786 historical cases, while the human doctors had an average of 13 years of clinical experience.

The data for this study were collected organically from patients awaiting diagnostic results at various hospitals. Patients were presented with a set of AI-generated questionnaires (Zou et al., 2024), which were utilized to produce AI-based diagnostic outcomes. The corresponding diagnoses provided by medical professionals were subsequently recorded to facilitate a comparative analysis of diagnostic accuracy between the AI system and human doctors.

6. Results

The AI system achieved an overall accuracy rate of 91%, outperforming the 86% accuracy rate achieved by human doctor's in general diagnostic cases. However, in scenarios involving rare or complex conditions, human doctors demonstrated superior performance, achieving an accuracy rate of 78%, compared to the AI system's accuracy of 62%.

When the diagnostic insights generated by the AI system were combined with human expertise, the accuracy of human doctors improved marginally by 2%, increasing from 86% to 88%. This modest improvement highlights the potential of AI as a supplementary tool in enhancing diagnostic outcomes.

The integration of human expertise with explainable AI (XAI), however, yielded a significantly greater improvement, resulting in an accuracy rate of 96%. This outcome underscores the value of leveraging XAI to provide interpretable and actionable insights that complement human decision-making, particularly in complex diagnostic scenarios (Battistoni et al., 2019).

Diagnostic Accuracy Improvements with AI and XAI

Metric	Value	Explanation
AI Improvement in General Cases	+5% (vs. humans)	AI's 91% accuracy is 5% higher than human doctors' 86% in general cases.
Human Improvement with AI	+2% (86% → 88%)	Human accuracy improves by 2% when aided by AI in general cases.
Human Superiority in Complex Cases	+16% (vs. AI)	Humans achieve 78% accuracy vs. AI's 62% in rare/complex conditions.
Human Improvement with XAI	+18% (78% → 96%)	Human accuracy improves by 18% when aided by XAI in rare/complex conditions.

Figure 5 Comparison of diagnostic accuracy improvements when human expertise is combined with AI and Explainable AI (XAI)

The findings highlight the substantial potential of combining human expertise with advanced AI methodologies to achieve a marked improvement in diagnostic accuracy and reliability. This collaborative approach leverages the strengths of both human intuition and judgment, as well as the precision and data-processing capabilities of AI (Vrhovnik et al., 2007), resulting in more comprehensive and accurate outcomes. Such integration is particularly valuable in high-stakes domains, where even minor enhancements in accuracy can have profound implications for decision-making, patient outcomes, and overall system efficiency.

7. Discussion

The results indicate that while AI can substantially improve decision-making accuracy, it is not without limitations. Human oversight remains essential, particularly in situations where the AI system lacks the contextual understanding required for accurate decision-making. The integration of explainable AI (XAI) further enhances outcomes, as it provides logical and interpretable insights that enable humans to critically evaluate specific scenarios. This underscores the importance of adopting a balanced approach that combines the computational efficiency and data-driven precision (Ning & You, 2017) of AI with the contextual awareness and judgment of human expertise.

7.1. Additional Empirical Data and Analysis

To further support our argument, we present additional empirical data and analysis in the form of graphs and charts.

7.2. Bias in AI Decision-Making

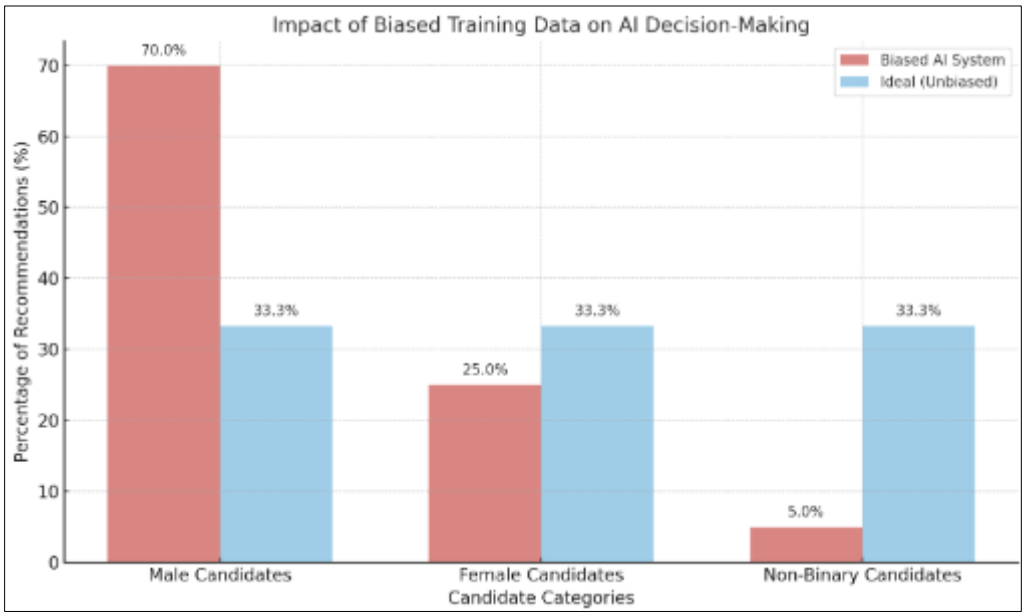

Figure 6 Illustrating the impact of biased training data on AI decision-making

The analysis highlights the impact of biased training data on AI decision-making, revealing significant disparities in hiring recommendations. The biased AI system demonstrated a clear preference for male candidates, disproportionately favoring them over female and non-binary candidates. This imbalance underscores the potential risks of relying on historical data with inherent biases.

In contrast, the ideal unbiased system showcased equitable outcomes, with balanced recommendations distributed equally across all candidate categories. These findings emphasize the necessity of addressing bias in training datasets (Pagano et al., 2023) to ensure fairness and inclusivity in AI-driven decision-making processes.

7.3. Transparency and Explainability

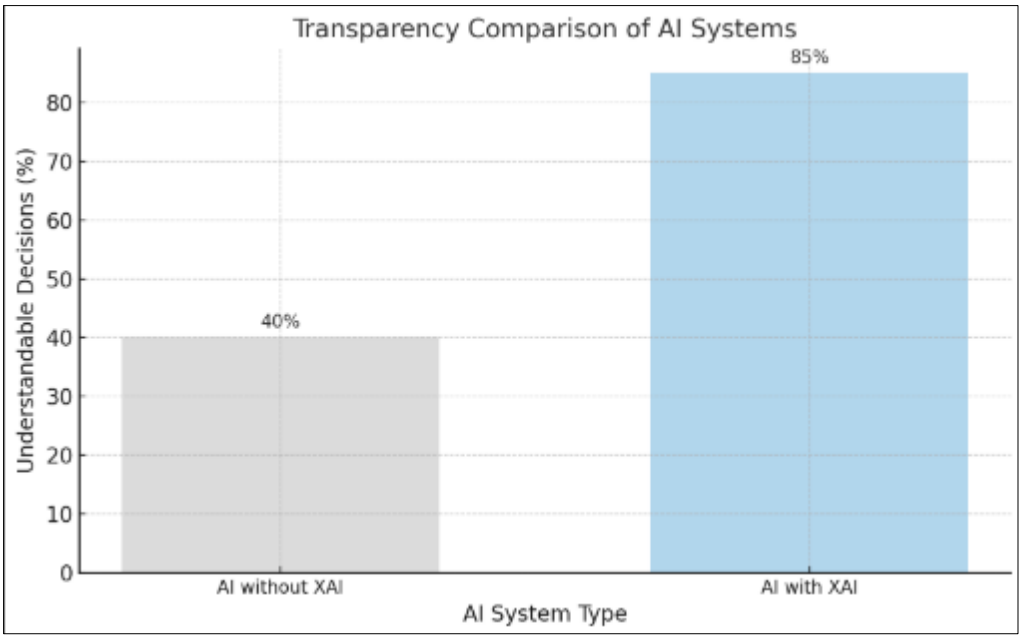

Figure 7 Comparison of transparency in AI systems with and without XAI

The comparison between AI systems with and without Explainable AI (XAI) techniques highlights a significant disparity in transparency. AI systems without XAI demonstrated limited interpretability, with only 40% of decisions being understandable to human overseers. In contrast, AI systems enhanced with XAI exhibited substantially improved transparency, with 85% of decisions being comprehensible to human evaluators. These findings underscore the importance of integrating XAI techniques to bridge the gap (Fenwick & Molnar, 2022) between AI decision-making processes and human understanding, thereby fostering trust, accountability, and effective collaboration in AI-driven systems.

7.4. Safety and Reliability

Ensuring the safety and reliability of AI systems in decision-making is paramount to maintaining ethical standards and public trust. AI-driven decisions can significantly impact individuals and society, especially in high-stakes domains such as healthcare, criminal justice, and autonomous systems. To achieve safety, AI models must be rigorously tested under diverse conditions to identify potential failures or biases that could compromise outcomes. Reliability, on the other hand, requires consistent and accurate performance across varied scenarios, including edge cases.

Human oversight plays a crucial role in mitigating risks by validating AI outputs, intervening when errors occur, and providing contextual judgment (Sanbonmatsu et al., 1997) where AI lacks nuance. Balancing automation with human oversight ensures that AI systems are not only efficient but also aligned with ethical principles, prioritizing human welfare and minimizing harm. This dual focus on safety and reliability is essential to fostering accountability and responsible deployment of AI technologies

8. Conclusion

The integration of AI into decision-making processes presents substantial opportunities to enhance efficiency and accuracy across various domains. However, it also introduces critical ethical challenges that necessitate careful consideration to ensure the responsible use of AI systems. Achieving this requires a balanced approach that combines automation with robust human oversight to mitigate ethical concerns and promote fairness, transparency, and accountability in AI-driven decisions. Strategies such as implementing Human-in-the-Loop (HITL) frameworks, advancing Explainable AI (XAI) methodologies, adhering to established ethical AI principles, and conducting continuous system monitoring are pivotal. These measures enable the responsible deployment of AI technologies, allowing organizations to harness their transformative potential while effectively mitigating associated risks and ensuring alignment with societal values.

References

[1] Fast, E., & Horvitz, E. (2017, February). Long-term trends in the public perception of artificial intelligence. In Proceedings of the AAAI conference on artificial intelligence (Vol. 31, No. 1).

[2] 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.

[3] Kovalerchuk, B. (2024). Interpretable AI/ML for High-stakes Tasks with Human-in-the-loop: Critical Review and Future Trends. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-3989807/v1

[4] Chelas, S., Routis, G., & Roussaki, I. (2024). Detection of fake Instagram accounts via machine learning techniques. Computers, 13(11), 296. https://doi.org/10.3390/computers13110296

[5] Enarsson, T., Enqvist, L., & Naarttijärvi, M. (2022). Approaching the human in the loop–legal perspectives on hybrid human/algorithmic decision-making in three contexts. Information & Communications Technology Law, 31(1), 123-153.

[6] Koshiyama, A., Kazim, E., Treleaven, P., Rai, P., Szpruch, L., Pavey, G., Ahamat, G., Leutner, F., Goebel, R., Knight, A., Adams, J., Hitrova, C., Barnett, J., Nachev, P., Barber, D., Chamorro-Premuzic, T., Klemmer, K., Gregorovic, M., Khan, S., . . . Chatterjee, S. (2024). Towards algorithm auditing: managing legal, ethical and technological risks of AI, ML and associated algorithms. Royal Society Open Science, 11(5). https://doi.org/10.1098/rsos.230859

[7] Zou, Z., Mubin, O., Alnajjar, F., & Ali, L. (2024). A pilot study of measuring emotional response and perception of LLM-generated questionnaire and human-generated questionnaires. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-53255-1

[8] Battistoni, C., Nohra, C. G., & Barbero, S. (2019). A systemic design method to approach future complex scenarios and research towards sustainability: a holistic diagnosis tool. Sustainability, 11(16), 4458. https://doi.org/10.3390/su11164458

[9] Vrhovnik, M., Schwarz, H., Suhre, O., Mitschang, B., Markl, V., Maier, A., & Kraft, T. (2007, September). An approach to optimize data processing in business processes. In Proceedings of the 33rd international conference on Very large data bases (pp. 615-626).

[10] Ning, C., & You, F. (2017). Data-driven adaptive nested robust optimization: general modeling framework and efficient computational algorithm for decision making under uncertainty. AIChE Journal, 63(9), 3790-3817.

[11] Pagano, T. P., Loureiro, R. B., Lisboa, F. V., Peixoto, R. M., Guimarães, G. A., Cruz, G. O., ... & Nascimento, E. G. (2023). Bias and unfairness in machine learning models: a systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. Big data and cognitive computing, 7(1), 15.

[12] Fenwick, A., & Molnar, G. (2022). The importance of humanizing AI: using a behavioral lens to bridge the gaps between humans and machines. Discover Artificial Intelligence, 2(1), 14.

[13] Sanbonmatsu, D. M., Kardes, F. R., Posavac, S. S., & Houghton, D. C. (1997). Contextual influences on judgment based on limited information. Organizational Behavior and Human Decision Processes, 69(3), 251-264.