Clear-Mind AI: Pioneering Transparent Machine Decisions

Pranav Thombre[1], Anish Niravane[2]

[1,2]Department of Computer Science, PVG's College of Science and Commerce

Abstract

The integration of Artificial Intelligence (AI) into critical decision-making processes across various sectors, including healthcare and criminal justice, underscores the necessity of Explainable Artificial Intelligence (XAI) systems. This paper delves into the theoretical foundations of human decision-making and their implications for designing user centric XAI systems. Our findings emphasize the importance of aligning XAI features with human reasoning processes to improve decision-making outcomes, ensuring greater transparency, trust, and accountability in AI-driven systems.

Keywords: Explainable Artificial Intelligence (XAI), Transparency, Trust, Accountability, Human-AI collaboration, Decision-making processes, User-centric design, Ethical implications, Fairness, Visualizations, Methodology.

1. Introduction:

The rapid integration of artificial intelligence (AI) into decision-making processes across various sectors has transformed how choices are made, enhancing efficiency and outcomes [1]. However, this advancement raises critical ethical considerations, particularly regarding transparency, accountability, and fairness [2]. The opacity of AI algorithms often leads to distrust among users, as the complexity of these systems can obscure their reasoning processes [3]. Explainable AI (XAI) has emerged as a potential solution to address these concerns, aiming to provide clarity on how decisions are made [4]. Despite its promise, the implementation of XAI is complicated by the wicked nature of the challenges faced in governance and decision-making, where ambiguity and a lack of consensus on ethical standards prevail [2].

As AI continues to evolve, it is essential to develop frameworks that ensure ethical practices, mitigate biases, and foster trust in AI-driven decisions [5].

In particular, the increasing use of AI technologies and assistants for decision-making in public affairs has sparked a lively debate on the benefits and potential harms of self-learning technologies.[6] This has spurred research interest in explainable AI (XAI)[7]. The effectiveness of these systems will be limited by the machine's inability to explain its thoughts and actions to human users in these critical situations[8].

This paper seeks to strengthen empirical application-specific investigations of XAI by exploring theoretical underpinnings of human decision making, drawing from the fields of philosophy and psychology[8].By articulating a detailed design space of technical features of XAI and connecting them with requirements of human reasoning, we aim to help developers build more user-centric explainable AI-based systems[7].In this context, the importance of transparency in AI decision-making is paramount, as it can significantly affect public perceptions of the legitimacy of AI decisions and decision-makers[6].

2. Abbreviations and Acronyms:

- AI - Artificial Intelligence
- XAI - Explainable AI
- HELM - Holistic Evaluation of Language Models
- MASS - Maritime Autonomous Surface Ships
- ANOVA - Analysis of Variance
- Tukey’s post-hoc test - Tukey’s Honestly Significant Difference (HSD) Test

3. XAI Architecture

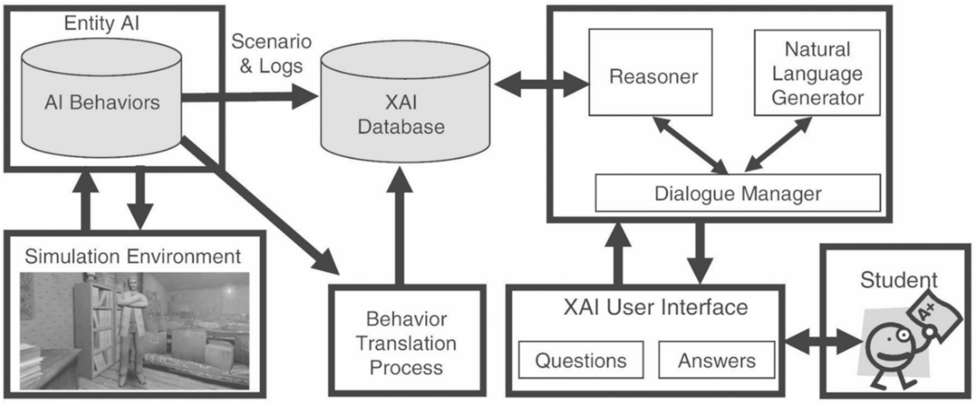

Fig.1 XAI Architecture[9]

4. Gaps and Challenges:

The integration of artificial intelligence (AI) into decision-making processes presents several gaps and challenges that need to be addressed to ensure effective and ethical use. One significant gap is the lack of transparency in AI algorithms, which often operate as "black boxes," making it difficult for users to understand how decisions are made[2][3][5]. This opacity can lead to distrust among users, particularly in high-stakes environments where decisions significantly impact individuals or groups [1].Another challenge is the potential for bias in AI systems, which can perpetuate existing inequalities if the training data is flawed or unrepresentative[5]. The ethical implications of such biases necessitate ongoing monitoring and evaluation of AI systems to ensure fairness and accountability [2][1].

Furthermore, the complexity of AI models complicates the establishment of clear accountability frameworks, making it challenging to determine who is responsible for decisions made by AI systems[1][4].Additionally, the wicked nature of the challenges faced in governance and decision-making complicates the implementation of explainable AI (XAI) solutions[4]. These challenges arise from the ambiguity of problems and the lack of consensus on ethical standards, which can hinder the effective deployment of AI technologies [2][3]. The need for transparency, accountability, and fairness in AI decision-making processes is also highlighted by the papers[1][5].

Moreover, the papers emphasize the importance of developing strategies for explainability that consider the context and potential biases of AI decision-making[2][4]. The need for ongoing monitoring and evaluation of AI systems to ensure fairness and accountability is also stressed [1][5].Despite the growing interest in explainable AI[10], there are significant gaps and challenges that remain unaddressed.

One major issue is that while many algorithmic approaches have been developed, little justification is provided for choosing different explanation types or representations, making it unclear why these explanations will be feasibly useful to actual users[8].Additionally, while XAI techniques have implicitly been designed to support rational reasoning, there is a need for more empirical research regarding how justifications should be designed and presented to gain public acceptance[8].

The challenge lies in bridging the gap between algorithm-generated explanations and human decision-making theories, as well as ensuring that explanations are tailored to meet the diverse needs of users across various domains[7].

5. Techniques for Transparent AI:

The integration of advanced AI systems in medical practice raises significant concerns regarding transparency, particularly in the context of patient-centred decision-making.

The authors argue that the complexity of these AI systems often renders them as "black boxes," where the underlying processes and decision-making criteria are not comprehensible to human practitioners. This lack of transparency conflicts with the core ideals of patient-centred medicine, which emphasizes informed decision-making based on shared information and deliberation between practitioners and patients[11].Transparent AI involves several key techniques to enhance understanding and trust in AI systems. Algorithmic transparency provides clear documentation of algorithms, while Explainable AI (XAI) enables systems to articulate their reasoning in human-friendly terms. Data transparency discloses sources and methods of data collection to mitigate biases. User education fosters informed decision-making, and regulatory frameworks establish legal guidelines for transparency.

Stakeholder engagement incorporates diverse perspectives, while regular audits ensure fairness and compliance. Feedback loops allow users to voice concerns, enhancing system performance. Open-source initiatives promote scrutiny, and interdisciplinary approaches develop comprehensive transparency strategies, ultimately benefiting users and society[12].One effective approach is to implement modular and generic architectures for explaining the behaviour of AI systems. This allows for the integration of explanation capabilities into various applications without being tied to a specific AI system[7].

Techniques such as feature attribution, where the importance of input features is highlighted, can help users grasp how specific data points influence the AI's predictions [3]. Additionally, employing counterfactual explanations allows users to explore alternative scenarios, enhancing their understanding of the decision-making process[8].Furthermore, integrating intelligibility queries into AI systems can enhance user interaction and understanding. By allowing users to ask specific questions about the AI's reasoning, such as "Why did this decision occur?" or "What factors influenced this outcome?", the system can provide tailored explanations that align with the user's inquiry[6].

6. Methodology:

This systematic review involved a thematic analysis of existing research on traffic alerts and collision avoidance systems related to decision transparency for maritime autonomous surface ships (MASS). The authors followed the PRISMA guidelines to extract and assess articles published between January 2012 and September 2023. They identified 22 relevant publications and categorized their findings into three main groups: strategies, visualization, and technology, with respective subgroups[13].The research adopted an interpretivist philosophy to gain deeper insights into artificial intelligence, utilizing an inductive approach to collect data on AI's ethical framework, transparency, and bias mitigation. An archival strategy was employed, focusing on qualitative studies to explore ethical behaviour in organizations utilizing AI. Secondary data collection was conducted using databases such as ResearchGate, IEEE, and Google, with a search strategy that included keywords like "artificial intelligence," "AI technologies," "ethical framework in AI," and "transparency within AI systems." Articles published from 2020 onwards were included, while those providing only abstract sections were excluded. Qualitative thematic analysis was performed to identify themes related to AI ethics, bias mitigation, and transparency. The study also adhered to the Copyright, Designs, and Patents Act 1988 in its data collection from secondary sources[14].The research design comprises three experimental studies focused on perceptions of justice in algorithmic decision-making across five application contexts: personal financial loans, workplace promotions, car insurance pricing, airline flight re-routing, and the freezing of bank accounts.

Data collection involved a lab study with in-person semi-structured interviews of 19 participants, who reflected on their agreement with various justice measures based on fictionalized cases. Additionally, two online studies were conducted to gather quantitative data, examining the effects of different explanation styles—Input Influence, Sensitivity, Case-based, and Demographic—on perceived justice levels. Participants rated their agreement with statements related to informational, procedural, and distributive justice using a 5-point Likert scale. Data analysis included qualitative thematic analysis for the lab study and statistical analysis (ANOVA and Tukey's post-hoc tests) for the online studies to evaluate the impact of explanation styles on perceptions of justice[15].

The study investigates how AI explanations and fairness influence human-AI trust and perceived fairness in AI-informed decision-making scenarios. A user study was conducted online, simulating AI-assisted decision-making in two contexts: health insurance and medical treatment, involving 25 participants due to pandemic restrictions. Participants were presented with different types of AI explanations—example-based and feature importance-based—alongside varying levels of introduced fairness (low, high, and control).

Statistical analyses, including two-way ANOVA tests, were employed to assess the interactions between explanation types and fairness levels on trust and perceived fairness, measured through self-report scales. The results indicated that low levels of introduced fairness diminished user trust, while high levels did not significantly impact trust. Additionally, AI explanations enhanced user trust, with no notable differences observed between the explanation types, and perceived fairness was positively influenced by high levels of introduced fairness[2].

7. Literature Review:

The ethical implications of AI transparency are critically examined, particularly in relation to the "AI knowledge gap." The authors argue that organizations utilizing AI decision systems must fulfil their obligations to stakeholders, which can be jeopardized by the opacity of AI models.

They stress that managers bear ethical responsibilities to justify their decisions, and the AI knowledge gap arises when these systems fail to provide adequate information, impeding managers' ability to meet their obligations. The complexity of opaque AI systems can result in strategic and ethical missteps, such as unfair predictions and a lack of accountability, underscoring the necessity for sufficient knowledge about AI models as both a strategic and moral imperative. The paper details the types of knowledge required for managers to justify their decisions, emphasizing the need for transparency, explainability, and interpretability in AI systems. The authors argue that employing black box AI models is unethical if it obstructs managers from fulfilling their responsibilities and advocate for a shift from merely generating knowledge about AI to understanding the knowledge demands of stakeholders.

They propose that stakeholder obligations should inform the design and development of AI models, ensuring these systems provide the necessary information to uphold ethical responsibilities. In conclusion, the authors urge firms to acknowledge their obligations to stakeholders when adopting AI systems, promoting a more responsible approach to AI procurement that prioritizes transparency and accountability in decision-making processes[16].This issue brief emphasizes the critical need for transparency in AI language models for policymakers, researchers, and the public, introducing the Holistic Evaluation of Language Models (HELM) framework as a benchmarking tool to enhance transparency.

The authors argue that understanding AI systems and their impacts requires more than traditional evaluation methods that focus solely on model accuracy, given the diverse applications of language models. HELM is proposed as a comprehensive framework that assesses language models across various metrics, including fairness, efficiency, robustness, and toxicity, facilitating a more holistic understanding of their performance. The brief also highlights the challenges in the current landscape of language model evaluation, where differing benchmarks hinder comparisons, and positions HELM as a standardizing tool for head-to-head model comparisons. Additionally, HELM acts as a public reporting mechanism for AI models, particularly those that are closed-access or widely deployed, empowering decision-makers to grasp their functions and impacts while ensuring alignment with human-centred values.

The authors stress that transparency is vital for effective policymaking and advocate for enhanced evaluation efforts to mitigate risks associated with AI deployment, such as bias and misinformation. Finally, the brief calls for further research to expand the HELM framework and address evaluation gaps, particularly for languages beyond English[17].The review focuses on cutting-edge research regarding AI decision transparency in the context of maritime autonomous surface ships (MASSs), highlighting the shift in navigational roles due to AI integration and the essential need for transparency in AI decision-making processes. The authors define AI decision transparency as the clarity with which AI systems communicate their decision-making processes to human operators, allowing them to understand and verify AI decisions. Following a systematic review methodology in line with PRISMA guidelines, the authors analysed 111 articles, narrowing their focus to 22 relevant publications that specifically address AI transparency in MASS contexts. The findings are categorized into three main themes: strategies, visualization, and technology, each exploring different aspects of decision transparency, including human factors, risk assessment, design principles, and visualization techniques. The necessity of integrating human cognitive processes into AI system designs is emphasized to ensure interpretability and accountability. Various visualization methods, such as colour coding and bounding boxes, are discussed for their potential to enhance operators' situational awareness and understanding of AI decisions.

Furthermore, the authors examine technological frameworks that can support decision transparency, including situational awareness models and route exchange mechanisms between autonomous and manned vessels. In conclusion, the authors call for further research to explore how the identified strategies and technologies can enhance decision transparency and improve human-AI collaboration in maritime environments[13].The first phase of our methodology involves a comprehensive literature review to explore the current state of explainable AI (XAI). This review examines various frameworks and techniques that have been proposed to enhance transparency and interpretability in AI systems.

We analyse the theoretical foundations of XAI, focusing on how different explanation types, such as local and global explanations, impact user understanding and trust[8]. Additionally, we investigate the role of cognitive biases in human decision-making processes and how these biases can affect the acceptance of AI-generated explanations[6]. This literature review serves as a foundation for developing a conceptual framework that aligns XAI techniques with user needs and decision-making strategies.

8. Application and Evaluation:

The insights from this paper can be applied to the development of autonomous shipping systems by ensuring that AI decision-making processes are transparent and accountable. To evaluate the effectiveness of these transparency measures, user studies and performance metrics can be employed to assess trust and safety in autonomous operations. By examining how well users understand and can verify AI decisions, as well as measuring the overall safety outcomes of these systems, stakeholders can gain valuable feedback on the transparency initiatives implemented, ultimately leading to more reliable and trustworthy autonomous shipping solutions[13].

The findings can be applied in healthcare settings to enhance patient-centred care by ensuring that AI systems provide transparent recommendations that patients and healthcare providers can easily understand.

To evaluate the effectiveness of these transparent AI systems, assessments can be conducted through patient feedback, which gauges their understanding and trust in the recommendations, as well as through clinical outcomes that measure the impact of AI transparency on patient decision-making and overall health improvements. This dual approach will help determine how well transparent AI supports informed patient choices and fosters trust in the healthcare process[11].

This review serves as a valuable resource for policymakers and organizations aiming to implement ethical AI practices in decision-making processes across various sectors. To evaluate the effectiveness of these practices, audits and assessments of AI systems can be conducted to measure their transparency and fairness. By systematically examining the adherence to ethical standards and the outcomes of AI implementations, stakeholders can gain insights into the impact of these practices, ensuring that AI systems operate responsibly and align with societal values[1].In the second phase, we apply the developed framework to a practical application within a clinical decision support system. We implement an explainable AI tool that utilizes machine learning algorithms to assist healthcare professionals in diagnosing patients. The system incorporates various XAI techniques, such as feature attribution and visualizations, to provide interpretable insights into the model's predictions[7].

To evaluate the effectiveness of the XAI tool, we conduct user studies involving healthcare practitioners. Participants engage with the system while providing feedback on the clarity and usefulness of the explanations generated.

We employ qualitative methods, such as think-aloud protocols and semi-structured interviews, to gather insights into user experiences and identify areas for improvement (Techniques for Transparent AI). This iterative evaluation process not only refines the XAI system but also contributes to a deeper understanding of how to design transparent AI solutions that effectively meet the needs of diverse stakeholders.

9. Ensuring Clarity in AI Systems:

Information asymmetry frequently exists between AI systems and human users, creating a gap in understanding that hinders the establishment of trust and effective collaboration. Additionally, varied user expectations regarding the desired level of transparency further complicate the design of AI systems, as different users may require different types of information to feel confident in the AI's decision-making processes. This dual challenge necessitates a thoughtful approach to AI design that addresses both the need for clear communication and the diverse expectations of users to foster trust and enhance collaboration[3].Transparency in AI necessitates collaboration across multiple disciplines, presenting challenges in effective coordination.

Furthermore, as AI technology continues to evolve, the standards for transparency also shift, complicating efforts to keep practices up to date and aligned with the latest advancements. This interplay between multidisciplinary collaboration and the dynamic nature of technological standards highlights the importance of ongoing dialogue and adaptability among stakeholders to ensure that transparency measures remain relevant and effective[12].

AI in public governance can exacerbate existing power imbalances, complicating efforts to ensure equitable outcomes. In this context, developing effective regulatory frameworks that promote transparency while also allowing for innovation presents a significant challenge. These frameworks must navigate the complexities of power asymmetries among various stakeholders, aiming to create a balanced approach that fosters accountability and fairness in the implementation of AI technologies within public governance[18].Achieving transparency in AI systems is fraught with challenges. One major hurdle is the inherent complexity of algorithms, which can obscure the rationale behind decisions, making it difficult for users to comprehend[7].

Additionally, the diversity of user backgrounds and varying levels of AI literacy complicate the design of universally understandable explanations[6]. Moreover, the potential for cognitive biases to distort user interpretation of AI outputs further undermines trust and acceptance[8]. Addressing these challenges is crucial for fostering effective human-AI collaboration.

10. Future Scope:

Establishing industry-wide standards for transparency in AI decision-making processes in autonomous shipping can be complemented by creating user-friendly interfaces that effectively communicate these decisions to stakeholders. By developing clear, accessible standards for how AI systems make decisions, stakeholders—including operators, regulators, and the public—can better understand and trust the technology. Enhanced user interfaces can serve as a key tool in this effort, translating complex AI algorithms and decision-making processes into visual or interactive formats that are easily interpretable. This dual approach ensures both technical consistency across the industry and practical clarity for end users, promoting accountability and fostering confidence in autonomous shipping systems[13].

Establishing common frameworks for studying human-AI decision-making, paired with empirical validation of existing theories, can significantly enhance our understanding of human-AI interactions. By developing standardized frameworks, researchers and practitioners can work from a shared foundation, ensuring consistency in studying how humans and AI systems collaborate in decision-making. These frameworks, when empirically validated through real-world studies, can confirm their effectiveness and highlight areas for improvement. This combination of structured guidance and evidence-based validation will not only advance research but also improve the practical application of human-AI collaboration in various fields[19].

The future of explainable AI (XAI) lies in enhancing user-centric design and integrating interdisciplinary approaches. Future research should focus on developing adaptive explanation methods that cater to diverse user needs and contexts, ensuring that explanations are both comprehensible and actionable[8]. Additionally, exploring the integration of XAI in high-stakes domains, such as healthcare and finance, will be crucial for building trust and facilitating informed decision-making[7].

Furthermore, ongoing evaluation of XAI systems through user feedback will help refine methodologies and improve transparency in AI applications[6].

11. Conclusion:

The paper concludes that effective AI governance is crucial for ensuring that automated decision-making systems are transparent, accountable, and fair. It advocates for frameworks that balance the benefits of AI with the need to protect citizens from potential harms[18].The study concludes that algorithmic explainability is crucial for building trust in automated decision-making processes. It emphasizes that transparency must be context-sensitive to effectively enhance trust[20].

The paper concludes that while explainable AI offers potential solutions to the challenges of algorithmic transparency, it must be approached as a socio-technical challenge that considers both technological and social factors[4].The paper concludes that black-box AI systems pose a threat to patient-centred care by undermining informed decision-making. It advocates for the development of transparent AI systems that align with the principles of patient-centred medicine [11].

The analysis underscores the need for AI-specific regulations in Ghana’s healthcare system to address gaps in trust and transparency. A multidisciplinary approach involving legal, ethical, and technical experts is crucial for creating guidelines ensuring ethical use and clear communication of AI capabilities. Trust and transparency are key for equitable access and preventing medical negligence. While AI presents challenges, it also offers opportunities for healthcare improvement. Recommendations include updating regulations, fostering multistakeholder engagement, building capacity, establishing a transparency framework, and continuous monitoring. Ultimately, successful AI integration requires a proactive, context-specific approach balancing technology, ethics, and societal readiness[21].

The exploration of explainable AI (XAI) reveals its critical importance in fostering trust and understanding in AI systems across various domains. As AI technologies become increasingly integrated into decision-making processes, the need for transparency and interpretability is paramount. This paper highlights the challenges associated with achieving transparency, including the complexity of machine learning models and the diverse needs of users, which can hinder effective communication of AI reasoning[6][7].

Future research should prioritize user-centric design, focusing on adaptive explanation methods that cater to different user backgrounds and contexts[8].

Additionally, interdisciplinary approaches that incorporate insights from psychology and cognitive science can enhance the effectiveness of XAI systems by addressing cognitive biases that affect user interpretation (Techniques for Transparent AI).

Ultimately, the ongoing development and evaluation of XAI will be essential in ensuring that AI systems are not only powerful but also comprehensible and trustworthy.

References:

1. Femi Osasona, Olukunle Oladipupo Amoo, Akoh Atadoga, Temitayo Oluwaseun Abrahams, Oluwatoyin Ajoke Farayola, and Benjamin Samson Ayinla, “REVIEWING THE ETHICAL IMPLICATIONS OF AI IN DECISION MAKING PROCESSES,” *International Journal of*

Management & Entrepreneurship Research, vol. 6, no. 2, pp. 322–335, Feb. 2024, doi: 10.51594/ijmer.v6i2.773.

2. A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, and A. Holzinger, "Fairness and Explanation in AI-Informed Decision Making," *Mach Learn Knowl Extr*, vol. 4, no. 2, pp. 556–579, Jun. 2022, doi: 10.3390/make4020026.
3. M. Vössing, N. Kühl, M. Lind, and G. Satzger, "Designing Transparency for Effective Human-AI Collaboration," *Information Systems Frontiers*, vol. 24, no. 3, pp. 877–895, Jun. 2022, doi: 10.1007/s10796-022-10284-3.
4. H. de Bruijn, M. Warnier, and M. Janssen, "The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making," *Gov Inf Q*, vol. 39, no. 2, Apr. 2022, doi: 10.1016/j.giq.2021.101666.
5. B. Kim, J. Park, and J. Suh, "Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information," *Decis Support Syst*, vol. 134, Jul. 2020, doi: 10.1016/j.dss.2020.113302.
6. K. de Fine Licht and J. de Fine Licht, "Artificial intelligence, transparency, and public decision-making: Why explanations are key when trying to produce perceived legitimacy," *AI Soc*, vol. 35, no. 4, pp. 917–926, Dec. 2020, doi: 10.1007/s00146-020-00960-w.
7. M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, "Building Explainable Artificial Intelligence Systems." [Online]. Available: www.aaai.org
8. D. Wang, Q. Yang, A. Abdul, and B. Y. Lim, "Designing theory-driven user-centric explainable AI," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2019. doi: 10.1145/3290605.3300831.
9. M. G. Core, H. C. Lane, M. Van Lent, D. Gomboc, S. Solomon, and M. Rosenberg, "Building Explainable Artificial Intelligence Systems." [Online]. Available: www.aaai.org
10. D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, "XAI-Explainable artificial intelligence," *Sci Robot*, vol. 4, no. 37, Dec. 2019, doi: 10.1126/scirobotics.aay7120.
11. J. Christian Bjerring and J. Busch, "Artificial intelligence and patient-centered decision making."
12. S. Larsson and F. Heintz, "Transparency in artificial intelligence," *Internet Policy Review*, vol. 9, no. 2, pp. 1–16, May 2020, doi: 10.14763/2020.2.1469.
13. A. N. Madsen and T. E. Kim,"A state-of-the-art review of AI decision transparency for autonomous shipping," 2024, *Informa UK Ltd*. doi: 10.1080/25725084.2024.2336751
14. A. . Todupunuri, "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*, vol. 13, no. 09, pp. 1–14, Sep. 2024, doi: 10.15680/IJIRSET.2024.1309002.
15. R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and N. Shadbolt, "'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, Apr. 2018. doi: 10.1145/3173574.3173951.
16. K. Martin and B. Parmar, "AI and the Creation of Knowledge Gaps: The ethics of AI transparency."
17. R. Bommasani, D. Zhang, T. Lee, and P. Liang, "Key Takeaways Improving Transparency in AI Language Models: A Holistic Evaluation," 2023.

18. M. Kuziemski and G. Misuraca, “AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings,” *Telecomm Policy*, vol. 44, no. 6, Jul. 2020, doi: 10.1016/j.telpol.2020.101976.
19. V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, “Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies,” Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.11471
20. S. Grimmelikhuijsen, “Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making,” *Public Adm Rev*, vol. 83, no. 2, pp. 241–262, Mar. 2023, doi: 10.1111/puar.13483.
21. G. B. Mensah, “Trust and Transparency in AI Health Tools”, doi: 10.13140/RG.2.2.21323.40489.
22. W. Samek and K.-R. Müller, "Towards Explainable Artificial Intelligence," in Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Lecture Notes in Computer Science, vol. 11700, Springer, Cham, 2019, pp. 5–22. doi: 10.1007/978-3-030-28954-6_1
23. P. Schmidt, F. Biessmann, and T. Teubner, "Transparency and trust in artificial intelligence systems," Journal of Decision Systems, Sep. 2020. doi: 10.1080/12460125.2020.1819094
24. M. Langer et al., "What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research," Artificial Intelligence, 2021. doi: 10.1016/j.artint.2021.103473
25. F. Emmert-Streib, O. Yli-Harja, and M. Dehmer, "Explainable Artificial Intelligence and Machine Learning: A reality-rooted perspective," Jan. 2020
26. H. W. Loh et al., "Application of Explainable Artificial Intelligence for Healthcare: A Systematic Review of the Last Decade (2011–2022)," Journal of Biomedical Informatics, 2022. doi: 10.1016/j.jbi.2020.103523
27. P. J. Phillips et al.,Four Principles of Explainable Artificial Intelligence, NISTIR 8312, National Institute of Standards and Technology, U.S. Dept. of Commerce, Sep. 2021. Available: https://doi.org/10.6028/NIST.IR.8312
28. V. Pillai, "Enhancing Transparency and Understanding in AI Decision-Making Processes,"IRE Journals, vol. 8, no. 1, Jul. 2024. Available: https://doi.org/10.6028/NIST.IR.8312
29. D. K. Chettri, "Explainable Artificial Intelligence for Decision-Making Systems,"International Journal of Modern Developments in Engineering and Science, vol. 2, no. 2, Feb. 2023. Available: https://www.ijmdes.com
30. J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation," IEEE Conference on Computational Intelligence and Games (CIG), 2021
31. D. D. W. Praveenraj et al., "Exploring Explainable Artificial Intelligence for Transparent Decision Making," E3S Web of Conferences, vol. 399, 2023. doi: 10.1051/e3sconf/202339904030
32. R. Confalonieri, L. Coba, B. Wagner, and T. R. Besold, "A Historical Perspective of Explainable Artificial Intelligence," WIREs Data Mining and Knowledge Discovery, vol. 11, 2021. doi: 10.1002/widm.1391
33. S. Laato, M. Tiainen, A. K. M. N. Islam, and M. Mäntymäki, "How to Explain AI Systems to End Users: A Systematic Literature Review and Research Agenda," Internet Research, vol. 32, no. 7, 2022. doi: 10.1108/INTR-08-2021-0600

34. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, 2018, pp. 52138-52160. doi: 10.1109/ACCESS.2018.2870052
35. J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the Need for Explainable Artificial Intelligence (XAI)," in Proceedings of the 54th Hawaii International Conference on System Sciences, 2021. Available: https://hdl.handle.net/10125/70768
36. J. Zerilli, A. Knott, J. Maclaurin, and C. Gavaghan, "Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard?" Philosophy & Technology, vol.
37. H. Surden, "Artificial Intelligence and Law: An Overview," *Georgia State University Law Review*, vol. 35, no. 4, pp. 1305–1337, 2019. Available at: https://readingroom.law.gsu.edu/gsulr/vol35/iss4/8
38. D. M. West and J. R. Allen, *How Artificial Intelligence is Transforming the World*, Brookings Institution, Apr. 24, 2018. Available at: https://www.brookings.edu/research/how-artificial-intelligence-is-transforming-the-world/
39. F. Zuiderveen Borgesius, *Discrimination, Artificial Intelligence, and Algorithmic Decision-Making*, Council of Europe, Directorate General of Democracy, 2018. Available at: https://rm.coe.int/discrimination-artificial-intelligence-and-algorithmic-decision-making/1680925d73
40. P. Boddington, *Towards a Code of Ethics for Artificial Intelligence*, Springer, 2017, doi: 10.1007/978-3-319-60648-4