Recalibrating Responsible AI: Non-Western Ethics, Sustainable Machine Learning, and Human Oversight

Lawal Wasiu *

Department of Information and Communication Technology, The Federal University of Technology, Akure, Nigeria

* E-mail of the corresponding author: wlawal@futa.edu.ng

Abstract

The proliferation of smart systems in all spheres of society has called attention to the ethical issues of responsibility, fairness, and transparency associated with the development and deployment of AI at a global level. The more that machine learning becomes enmeshed in processes of decision-making, the more new ethical dilemmas that reflect the possibility for algorithmic bias, information vulnerability, socio-cultural exclusion, and environmental considerations begin to surface. This review provides an organised, humane, systematic examination of responsible AI, connecting ethical considerations with pragmatic concerns as derived from existing models, policies or problems in execution. Apart from standard controversies, the contribution of the paper is to single out little explored concerns that are related to non-Western ethical perspectives, sustainable AI practices, and the ethical reasoning of superintelligent systems. It is a critique of the existing limitations of ethical codes, which put forward ideas and directions to connect innovation and social good. By bringing in alternative perspectives and interdisciplinary thinking, it presents a rich base for the ethical recalibration of intelligent systems to work for human betterment, in a more inclusive, equitable and forward-looking manner.

Keywords: Responsible artificial intelligence; Algorithmic fairness and non-discrimination; Transparency and accountability; Data privacy and governance; Global AI governance

DOI: 10.7176/JIEA/15-2-09

Publication date: September 28th 2025

1. Introduction

Recently, the increased awareness of ethical issues on Responsible AI (RAI) and Machine Learning (ML) has become relevant [1]. This is typically referred to as responsible AI research and development (RAI) [2], or in the installation of systems that are fair; transparent; and secure, and accountable to human values. ML ethics expands this purview to question the impacts of AI from a social standpoint by calling for the development of AI "for the good of all humanity" [3]. Machine Learning has spurred an "industrial revolution" in many of the areas that relate to people's work, life, and communication [4]. "ML is in the air" – in terms of voice recognition as well as photos, it is undeniable. But, new technologies also pose new and greater ethical dilemmas as well [6]. Recent failures such as, but not limited to, bias in algorithmic decision making and mass privacy violations have demonstrated the need for more responsible handling of this technology [7]. Growing public awareness has also prompted developers and policymakers to be more cautious in maintaining moral standards [8]. This includes highly publicised situations, such as Facebook-Cambridge Analytica, and the use of AI in criminal justice to produce biased results, which has also spurred a market demand for responsible AI [9]. Events such as these have elicited demands for increased accountability and transparency, as well as the need for AI to represent cultural norms and values [10]. Also, the speed at which AI technologies have matured has not always being matched by a corresponding growth in ethical guidelines and proposes the need to have responsible AI accompanied by impactful transformative technology [11]

As AI and ML increasingly predominate various industry sectors, the ethical considerations are becoming more and more pressing. In this context the discussion below endeavors to sketch the contours of what responsible AI entails and some of the ethical issues that are implicated, highlighting the concepts of humane design, value alignment, and continuous supervision [12] . By conducting a literature review of the recent case studies regarding responsible AI and the identified challenges and accomplishments, the goal of this paper is to explore the avenues in which AI is currently governed and where it might be heading in order to integrate AI as something capable of improving individual lives while maintaining our core values [13].

2. Ethical Considerations in Machine Learning (ML)

Machine learning is not neutral like any other technology. Its design and how it is deployed affect the populations and the societies it targets. If not properly supervised, ML may exacerbate the existing prejudices, discrimination and social injustices [14]. This is why we need to embed ethical considerations into at the very least the development of systems rather than trying to retrofit ethics into them; in order to keep ML systems in line with human values and to keep them committed to serving the public interest [15].

A. Bias and Discrimination in Data and Algorithms

These outputs are representative of these inputs; if the data used to train an ML system is biased, the system's assessments will be as well. Facial recognition, for example, has been reported as several times less accurate for dark skinned individuals, raising several alarms in high sensitive areas like Law enforcement and borders control [16]. Likewise, models for processing language that are learned from raw text online can also inadvertently replicate gender and racial biases, thus encoding destructive associations in automated decision making [17] These issues are indicative of the ways in which latent biases within the data can be experienced by hundreds of thousands of users at once. In order to address these problems, diverse, representative datasets should be used, coupled with fairness-aware algorithms, and procedural methods to mitigate bias in the process of designing and evaluating model performance [18].

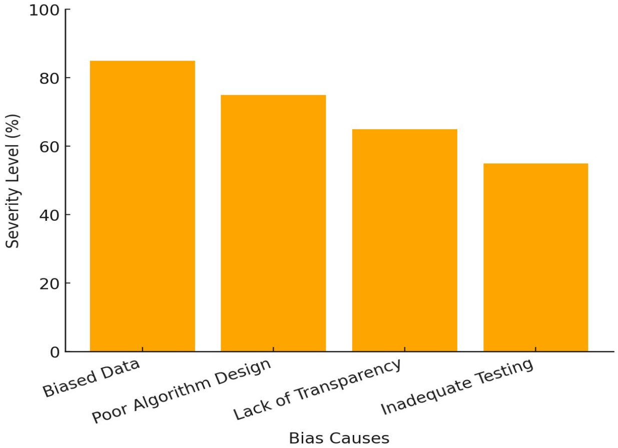

Figure 1. Key Sources of Bias in Machine Learning Systems.

Figure 1, revealed major factors of bias in machine learning systems. The bar graph includes factors along the y-axis in terms of the incidence of bias, with biased data the most prevalent followed by flawed algorithm design, limited transparency, and insufficient testing capabilities. It also illustrates that low quality data and model development/lack of oversight can engender systemic fairness and accuracy issues, which underscores the importance of meticulous oversight at each stage in the AI development process.

B. Privacy Concerns and Data Security

Also, these systems require huge amounts of personal and sensitive data, posing severe privacy concerns [19]. Safeguarding this knowledge, and not enabling it to be misused or accessed or exploited in any unauthorized capacity, has ramifications that can be harmful to individuals and organizations. They are also major risks for a loss of public trust, which is another good ethical reason to protect privacy. Strong privacy protective measures, like good data protection frameworks, strong encryption and effective anonymisation are needed to help in addressing these issues [20]. These mechanisms do not only prevent privacy breaches but also strengthen trust in the ethical implementation of ML technologies [21].

C. Explainability and Transparency in Decision-Making Processes

Complexity of ML models, and understanding them, can lead to lack of transparency and trust [22]. And this unexplainability can further allow bias and discrimination to continue. This is an issue that is being attempted to deal with, alongside the development of more 'transparent' AI methods and to make the process of decision making visible [23].

D. Human Oversight and Accountability

Given these potential dangers, machine learning needs to be designed for meaningful human control and mechanisms for accountability [24]. It consists in the use of validation and monitoring processes, which are considered a necessary mechanism for feedback, in that allow human beings to detect hidden errors or biases [25]. Transparency, through a clear audit trail or decision log, allows for lines of accountability when things go wrong [26]. The integration of human judgment at several stages of the lifecycle of ML will enhance ethical control, increase trust and prevent over-reliance on automatically generated output [27].

E. Environmental Impact and Sustainability

This is a central and often neglected issue when addressing ML. Energy used during the training and use of AI as well as the disposal of AI hardware are environmentally degrading processes [28]. This gives additional urgency to the need for sustainable AI practices, including the move to renewable energy and elimination of electronic waste [29].

F. Cultural Sensitivity and Inclusivity

This fact highlights the caution in designing ML systems to be culturally sensitive, as they may perpetuate stereotypes or marginalize one group or another [30]. The ignorance in this particular field can incept deleterious perceptions and mimic social disparities in the community [31]. This means incorporating a variety of cultural understandings into the data pools for the design processes, which ultimately will result in more inclusive AI technologies and AI better suited to address global challenges [32]. This is not only a way to avoid bias but also to create a beneficial system that does the least harm [33].

G. Equity and non- discrimination

There is a strong case to be made for the mandate of fairness in ML systems. The challenge of course, is that developers must also be aware of a variety of biases in data and algorithms and work towards not reinforcing social inequalities [34]. The means through which this technology gets created and used to ensure fairness and inclusion is important [35].

H. Human Rights and Digital Ethics.

AI systems must reflect respect for human rights and ethical standards in digital technologies, and protect human rights as ensuring user autonomy, privacy and freedom of expression [36]. “AI systems should not be designed or used to propagate harmful behaviors such as hate speech or discrimination” [37].

I. Accountability and Responsibility

ML practitioners and consumers need to take responsibility for the effects of AI related systems. This includes clear mechanisms of accountability and extensive testing of AI systems prior to deployment [38].

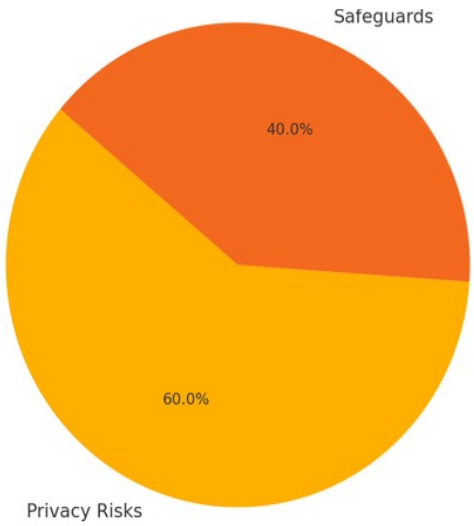

Figure 2. The privacy risks and safeguards trade-off in AI systems

This chart, Figure 2, shows that concerns for user privacy are still much greater than measures that have been adopted to try to protect privacy, and thus highlights that the problem of the protection of user data in the realm of machine learning still remains a significant one. With developments in encryptions, anonymizations and oversight procedures, these areas of weakness have increasingly been minimized, but nevertheless pose serious problems, and there is clearly a need for more accountability and sustainable data governance mechanisms.

3. Responsible AI Practices

AI ethics are a critical concern because it will be necessary that AI can help and not harm us. The section above contains important best practices, which may be utilized by developers and organizations in order to ensure their AI development meets ethical criteria.

A. Fairness and Inclusivity in AI Development and Deployment

The close relationship between rights, fairness and non-discrimination should be fundamental to AI development, to prevent its use from simply reinforcing existing biases and discriminations [39]. In order to create an inclusive experience, it is important to be intentional about incorporating a wide range of voices and positions to create the design [40]. To enable this, representative and balanced datasets, along with tests designed to preserve fairness prior to deployment are essential [41]. Further, continual oversight even once implemented is also critical as it helps to prevent systems from being further entrenched, as bias and exclusion can ossify over time [42].

B. Human-Centered Design and Value Alignment

Human-centered design specifically calls for AI systems to incorporate human values as well as be transparent, accountable, and explainable [43]. A design process that keeps stakeholders "in-the loop" is able to identify user needs that can be directly injected into the design of the system [44].

C. Environmental and Social Impact Assessments

AI impacts can be both beneficial and harmful. Energy consumption of AI systems is another monitored aspect, in order to reduce the harm and to ensure sustainability [45] . The environmental costs creates a deeper look into long term risks as well [46]. Considering social impacts to guarantee that the application will contribute positively to society and minimize unintentional negative impacts on it [47].

D. Continuous Monitoring and Evaluation

Continuous monitoring and evaluation would make sure that AI systems are fair, inclusive, and reliable [48]. Putting in place mechanisms for continuous validation helps establish accountability of the system [64]. The addition of feedback increases responsiveness and trust in actual real-word scenarios. [50]

E. Transparency and Accountability in AI Monitoring and Evaluation

This is something that should be made explicit and transparent in how AI monitoring and evaluation are done by developers [51]. One way to maintain trust in deployed systems is to establish and enforce accountability measures, such as audits and reporting [52]. Ongoing independent oversight and open lines of communication help to assure governance remains viable and successful over the life of the system [53]

F. Stakeholder Engagement and Participation

The inclusion of affected stakeholders in the development process is necessary for the inclusivity and co-responsibility of AI systems [54]. This enables regular interaction with policy makers, civil society, industry, and end- users, which leads to a range of voices that can bring a more balanced and nuanced understanding of issues and make research more responsive and immune to bias [55]. At an earlier stage they could take in this type of feedback to realize possible pitfalls and develop socially fair and socially acceptable systems [56].

G. Maintenance and support of such assurance.

"Sustainability", or the viability and ability to maintain and continue something into the future, should be an important determinant in the initial design of AI systems, if they are to maintain long term reliability and be able to provide relevance and adapt to changing contexts . This means developing systems and structures that can easily be updated, incorporate new data, and adjust to changing laws or social mores. Frequent testing and incremental validation also allow early identification of defects and avoid that a technical malfunction escalate into failure at systemic levels [58]. This includes transparency, documentation, modularity, and version control, which all contribute to the performance and auditability of the AI solution over time [59].

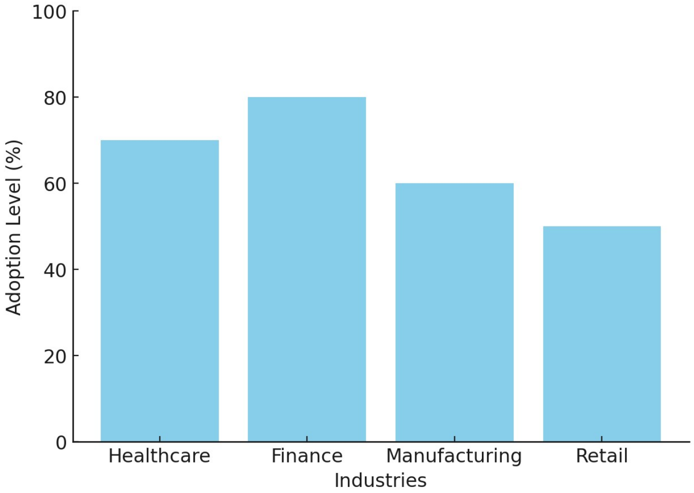

Figure 3. Industry adoption of responsible AI practices.

The chart shows differences in adoption levels across key sectors, with finance and healthcare leading in implementation, while manufacturing and retail lag behind. This variation reflects how regulatory demands, risk exposure, and investment priorities shape the adoption of responsible AI in different industries.

4. Case Studies and Examples

A. Successful Implementations of Responsible AI

Healthcare: AI-powered diagnosis tools improve outcomes in healthcare by facilitating the early detection of diseases like cancer [60]. Artificial intelligence applications may pose potential to reach populations otherwise unreached in mental health care [61].

Finance: AI systems are now beginning to be applied through all of the financial systems in order to help prevent fraudulent transactions in real-time as they happen which is an added protection for people and companies [62] In addition to fraud detection, these systems also help in the assessment of credit risk by being able to analyze big financial datasets allowing to actually look deeper into borrowers. These types of applications create a more fair and inclusive environment, especially in the case where they are designed specifically to reduce bias in lending decisions [63].

Education: Adaptive learning systems aim to make education more about each individual student's needs [64]. They focus on the performance pattern, and recommend resources to help build on weaknesses rather than try to fix deficits. The eventual beneficiaries are the teachers who receive feedback which helps to adapt classroom strategies as well as target the help/support where it is most needed [65]. This engagement between technology and teachers leads to an overall increase in learning outcomes and student interaction.

Manufacturing: In manufacturing, predictive analytics can help to reduce machine failures by predicting when machines need maintenance [66] This all reduces overhead, limits the down time and therefore lowers cost making the center more productive. At the same time, automation is problematic due to the threat of employment-related displacement – machines gain the ability to address what was previously human labor. In the ethical discussion on the implications for the labor force the role of the machine in transforming productivity versus human benefits is stressed [67] . The only way to address this balance is by investing in retraining the workforce and policies that protect these workers. Thus, for manufacturing, responsible AI practices should look into both technical success and social sustainability on the long run [68].

B. Lessons Learned from Failures or Controversies

Biased AI Systems: Examples such as facial recognition flops show the dangers of not training AI on diverse datasets [69]. In Africa and other Sub-Saharan regions, there have been similar worries about a lack of representation in the training data, which has magnified biases in very important systems of identification and decision making [70].

Data Security: Concerns regarding privacy because of the ways in which AI might be used show the need for transparency, user consent, and strong data security measures [71].

C. International Examples and Comparisons

European Union's GDPR: The GDPR sets unprecedented standards for privacy, responsibility, and data protection at a global level [72].

China's Social Credit System: Examines the ethical concerns of surveillance, autonomy, and privacy rights in this system [73] .

India's Aadhaar Program: Similarly, while Aadhaar enhances access to services, it raises concerns on the protection of biometric data and privacy issues [74].

United States' Sectoral AI Regulation: The U.S. follows a sectoral approach rather than implementing one, overarching regulation [75].

Africa and Sub-Saharan Region: The AI governance frameworks that are emerging are evolving, but there readiness for the technology is still lacking of infrastructure to support it, proper regulation to regulate it, or widespread access and concerns with questions of fairness in how it is applied to such a diverse population [76].

5. Regulatory and Governance Frameworks

A. Governmental Regulations

The European GDPR and U.S. sectoral methods to regulation focus on accountability and the protection of individual rights alongside innovation [77].

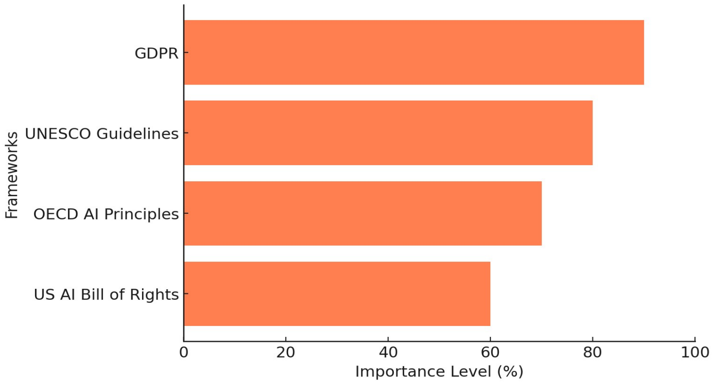

Figure 4. Relative focus of global AI governance frameworks.

The Figure 4, illustrates how different international approaches, including GDPR, UNESCO guidelines, OECD principles, and the U.S. AI Bill of Rights, prioritize accountability, fairness, and human rights. The variation in importance levels reflects regional priorities, showing how governance models adapt to local ethical, legal, and cultural contexts.

B. Industry Standards and Best Practices

Automotive: The NHTSA AI guidelines have notably incorporated measures like the AV STEP program which offer a prescriptive method for maintaining accountability and minimum safety standards for autonomous vehicles [78].

Technology: The IEEE's Ethically Aligned Design framework promotes the development of human-centric AI that is transparent, fair, and innovated responsibly [79].

C. International Cooperation and Agreements

OECD's AI Principles: Global AI norms focused on fairness, transparency, and accountability across jurisdictions [80].

UNESCO AI Ethics Recommendations: The UNESCO recommendations focus on inclusivity and human dignity as a way of promoting common ethics for the governance of AI [81].

6. Social Implications and Job Displacement

A. Impacts on Employment

AI reduces the role of routine labor in many fields and jobs by automating repetitive work, while also generating jobs in new areas like data science and oversight of AI [82]. The future of work may not be in traditional industries such as manufacturing and retail as automation increasingly changes these production and service delivery modes [83]. Simultaneously, the deployment of AI supports increased specialization in technology and analytics, which creates new pathways to high-skill employment as well [84].

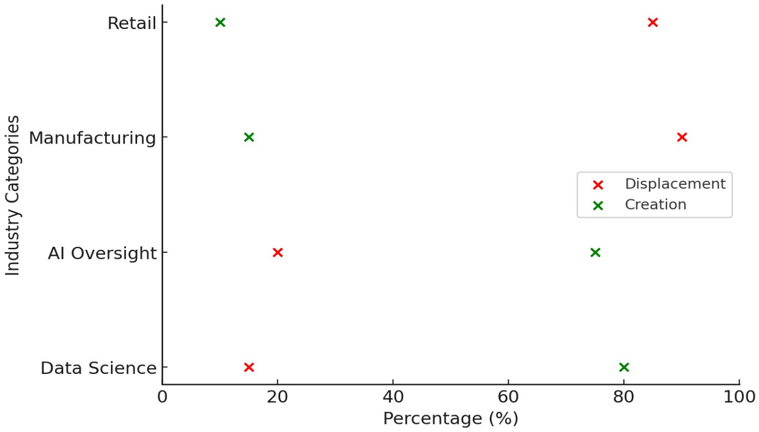

Figure 5. Job displacement and Job Creation in AI - driven sectors.

While there may be some job creation in related industries; the Figure 5, illustrates the expansion of traditional industries like manufacturing and retail but also to some extent job growth in data science or oversight of AI and the contraction in others caused by AI. This distinction reveals the dark side and the bright side of layoff experiences in knowledge intensive work.

B. Strategies for Mitigating Displacement

Upskilling Programs: we need to develop programs that train workers for the jobs of the future in AI technology and oversight [85].

Entrepreneurial Opportunities: AI-induced innovation generates start-ups that address new needs like personalized health care and smart technologies [86]. In addition to developing new products and services that drive economic growth, these businesses also create necessary employment for these new markets. The entrepreneurial activity in this space has the potential to encourage competitive economic development, as well as absorb workers displaced by the erosion of traditional industrial goods production.

C. Inclusive Solutions

Inclusive AI systems development is crucial to guarantee that those who are under and misrepresented are fairly benefiting from and not unduly suffering from technological advancements (Anderson, 2016) .This includes creating datasets and models that are representative of society, as well as policies to ensure access to and participation across various communities.

7. Future Directions and Research Opportunities

A. Emerging Trends and Ethical Frameworks

Explainable AI (XAI): In critical domains such as health and emergency response, XAI provides important transparency and trust by explaining the decision-making of an AI system [88].

Human-AI Teaming: The proposed paradigm of human-autonomy teaming highlights the need for dynamic partnerships whereby AI interfaces allow for incorporation of human cues and co-constructed decisions while preserving the authority of the human [89].

AI for Social Good: AI is applied to tackle important global issues such as climate prediction and fair access to healthcare, in line with the UN's sustainable development goals [90].

B. Open Challenges

Overcoming Bias: There is a need to address more sensitive and flexible methods for identifying and correcting biases [91].

Transparency: Making complex models auditable and interpretable remain crucial for trust in AI deployment [92].

8. Conclusion

The promise artificial intelligence holds for improving human well-being can only be met if it is governed responsibly. If designed to be fair, inclusive, and transparent, AI can be used to make and reinforce societal values rather than challenge them. This means that developers, policy makers and end-users must be working consciously to ensure that future systems are capacitating at the micro level and discussing global concerns at the macro level.

References

[1] R. K. Jaiswal, S. S. Sharma, and R. Kaushik (2023), "ETHICS IN AI AND MACHINE LEARNING," *Journal of Nonlinear Analysis and Optimization*, vol. 14, no. 1,. [Online]. Available: https://doi.org/10.36893/jnao.2023.v14i1.0008-0012

[2] C. Rigotti and E. Fosch-Villaronga, (2024), "Fairness, AI & recruitment," *Computer Law & Security Review*, vol. 53, p. 105966, [Online]. Available: https://doi.org/10.1016/j.clsr.2024.105966

[3] H. R. Saeidnia, S. G. H. Fotami, B. D. Lund, and N. Ghiasi (2024), "Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact," *Social Sciences*, vol. 13, no. 7 [Online]. Available: https://doi.org/10.3390/socsci13070381

[4] S. Upmanyu, A. Upmanyu, A. Jamwal, and R. Agrawal (2021), "Machine Learning in CAD/CAM: What We Think We Know So Far and What We Don't," *Lecture Notes in Mechanical Engineering*, [Online]. Available: https://doi.org/10.1007/978-981-16-5281-3_48

[5] F. Firouzi, B. Farahani, F. Ye, and M. Barzegari, "Machine Learning for IoT," *Machine Learning and Cognitive Computing for Mobile Communications and Wireless Networks*, [Online]. Available: https://doi.org/10.1007/978-3-030-30367-9_5

[6] V. Dhenge and K. Dorshetwar (2020), "Overview of Ethics in Artificial Intelligence: Using Case Studies Approach," in *2024 IEEE Int. Conf. on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)*, pp. 1–6, 2024. [Online]. Available: https://doi.org/10.1109/IATMSI60426.2024.10502940

[7] K. Kakarala and R. Rongali (2025), "Existing challenges in ethical AI: Addressing algorithmic bias, transparency, accountability and regulatory compliance," *World Journal of Advanced Research and Reviews*, vol. 25, no. 3, [Online]. Available: https://doi.org/10.30574/wjarr.2025.25.3.0554

[8] R. Awashreh and M. Khademizadeh (2025), "Navigating the Ethical Frontier: Privacy, Bias, and Regulation in AI Development," *Arab Journal of Administrative Sciences*, vol. 31, no. 3, [Online]. Available: https://doi.org/10.34120/ajas.v31i3.1273

[9] J. O. Arowosegbe (2023), "Data bias, intelligent systems and criminal justice outcomes," *Int. J. Law Inf. Technol.*, vol. 31, pp. 22–45, [Online]. Available: https://doi.org/10.1093/ijlit/eaad017

[10] G. Manias *et al.* (2023), "AI4Gov: Trusted AI for Transparent Public Governance Fostering Democratic Values," in *Proc. 19th Int. Conf. Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*, 2023, pp. 548–555. [Online]. Available: https://doi.org/10.1109/DCOSS-IoT58021.2023.00090

[11]R. Kottur (2024), "Responsible AI Development: A Comprehensive Framework for Ethical Implementation in Contemporary Technological Systems," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. [Online]. Available: https://doi.org/10.32628/cseit241061197

[12] C. Rigotti and E. Fosch-Villaronga (2024), "Fairness, AI & recruitment," *Computer Law & Security Review*, vol. 53, p. 105966, [Online]. Available: https://doi.org/10.1016/j.clsr.2024.105966

[13] J. Wang, Y. Huo, J. Mahe, Z. Ge, Z. Liu, W. Wang, and L. Zhang (2024), "Developing an Ethical Regulatory Framework for Artificial Intelligence: Integrating Systematic Review, Thematic Analysis, and Multidisciplinary Theories," *IEEE Access*, vol. 12, pp. 179383–179395, [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3501332

[14] C. Chen, R. Napolitano, Y. Hu, B. Kar, and B. Yao (2024), "Addressing machine learning bias to foster energy justice," *Energy Research & Social Science*, vol. 110, p. 103653, [Online]. Available: https://doi.org/10.1016/j.erss.2024.103653

[15] M. Pflanzer, Z. Traylor, J. B. Lyons, V. Dubljević, and C. S. Nam (2022), "Ethics in human–AI teaming: principles and perspectives," *AI and Ethics*, vol. 2, pp. 1–19, [Online]. Available: https://doi.org/10.1007/s43681-022-00214-z

[16] M. Gentzel (2021), "Biased Face Recognition Technology Used by Government: A Problem for Liberal Democracy," *Philosophy & Technology*, vol. 34, pp. 1639–1663, [Online]. Available: https://doi.org/10.1007/s13347-021-00478-z

[17] M. A. Palacios Barea, D. Boeren, and J. F. Ferreira Goncalves (2023), "At the intersection of humanity and technology: a technofeminist intersectional critical discourse analysis of gender and race biases in the natural language processing model GPT-3," *AI & Society*, pp. 1–19,

[18] N. Shahbazi, Y. Lin, A. Asudeh, and H. Jagadish (2022), "Representation Bias in Data: A Survey on Identification and Resolution Techniques," *ACM Computing Surveys*, vol. 55, pp. 1–39, [Online]. Available: https://doi.org/10.1145/3588433 ACM JOURNAL NOTE

[19] S. Z. El Mestari, G. Lenzini, and H. Demirci (2023), "Preserving data privacy in machine learning systems," *Computers & Security*, vol. 137, p. 103605, [Online]. Available: https://doi.org/10.1016/j.cose.2023.103605

[20] S. Myneni, G. Agrawal, Y. Deng, A. Chowdhary, N. Vadnere, and D. Huang (2022), "SCVS: On AI and Edge Clouds Enabled Privacy-preserved Smart-city Video Surveillance Services," *ACM Transactions on Internet of Things*, vol. 3, no. 3, pp. 1–26, [Online]. Available: https://doi.org/10.1145/3542953

[21] A. K. Nair, J. Sahoo, and E. D. Raj (2023), "Privacy preserving Federated Learning framework for IoMT based big data analysis using edge computing," *Computer Standards & Interfaces*, vol. 86, p. 103720, [Online]. Available: https://doi.org/10.1016/j.csi.2023.103720

[22] S. Shobeiri (2024), "Enhancing Transparency in Healthcare Machine Learning Models Using Shap and Deeplift: a Methodological Approach," *Iraqi Journal of Information and Communication Technology*, vol. 7, no. 2, pp. 1–12, [Online]. Available: https://doi.org/10.31987/ijict.7.2.285 IJICT(Iraqi Journal of Information and Communication Technology)

[23] R. Deokar, P. Nanjundan, and S. Mohanty (2024), "Transparency in Translation: A Deep Dive into Explainable AI Techniques for Bias Mitigation," in *Asia Pacific Conf. on Innovation in Technology (APCIT)*, pp. 1–6, [Online]. Available: https://doi.org/10.1109/APCIT62007.2024.10673712

[24] A. Holzinger, K. Zatloukal, and H. Müller (2024),, "Is Human Oversight to AI Systems still possible?," *New Biotechnology*, [Online]. Available: https://doi.org/10.1016/j.nbt.2024.12.003

[25] J. L. Cross, M. A. Choma, and J. Onofrey (2024), "Bias in medical AI: Implications for clinical decision-making," *PLOS Digital Health*, vol. 3, p. e0000651, [Online]. Available: https://doi.org/10.1371/journal.pdig.0000651

[26] Y. Li and S. Goel (2024), "Making It Possible for the Auditing of AI: A Systematic Review of AI Audits and AI Auditability," *Information Systems Frontiers*, [Online]. Available: https://doi.org/10.1007/s10796-024-10508-8

[27] J. Nuamah and Y. Seong (2019),, "A Machine Learning Approach to Predict Human Judgments in Compensatory and Noncompensatory Judgment Tasks," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 326–336, 2019. [Online]. Available: https://doi.org/10.1109/THMS.2019.2892436

[28] H. Järvenpää, P. Lago, J. Bogner, G. A. Lewis, H. Muccini, and I. Ozkaya (2023),, "A Synthesis of Green Architectural Tactics for ML-Enabled Systems," in *IEEE/ACM 46th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pp. 130–141, [Online]. Available: https://doi.org/10.1145/3639475.3640111

[29] D. Ueda, S. Walston, S. Fujita, Y. Fushimi, T. Tsuboyama, K. Kamagata, A. Yamada, M. Yanagawa, R. Ito, N. Fujima, M. Kawamura, T. Nakaura, Y. Matsui, F. Tatsugami, T. Fujioka, T. Nozaki, K. Hirata, and S. Naganawa (2024),, "Climate change and artificial intelligence in healthcare: Review and recommendations towards a sustainable future," *Diagnostic and Interventional Imaging*,[Online]. Available: https://doi.org/10.1016/j.diii.2024.06.002

[30] G. Franklin, R. Stephens, M. Piracha, S. Tiosano, F. LeHouillier, R. Koppel, and P. L. Elkin (2024),, "The Sociodemographic Biases in Machine Learning Algorithms: A Biomedical Informatics Perspective," *Life*, vol. 14, no. 6, p. 652, [Online]. Available: https://doi.org/10.3390/life14060652

[31] B. Fraile-Rojas, C. De-Pablos-Heredero, and M. Méndez-Suárez (2025),, "Female perspectives on algorithmic bias: implications for AI researchers and practitioners," *Management Decision*, [Online].

Available: https://doi.org/10.1108/md-04-2024-0884 write paper review on female gender EMERALD JOURNAL

[32] X. Ge, C. Xu, D. Misaki, H. R. Markus, and J. L. Tsai (2024),, "How Culture Shapes What People Want From AI," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, [Online]. Available: https://doi.org/10.1145/3613904.3642660

[33] M. Russo, Y. Chudasama, D. Purohit, S. Sawischa, and M.-E. Vidal (2024),, "Employing Hybrid AI Systems to Trace and Document Bias in ML Pipelines," *IEEE Access*, vol. 12, pp. 96821–96847, [Online]. Available: https://doi.org/10.1109/ACCESS.2024.3427388

[34] Z. Chen and S. Zhu (2025),, "Fine-Tuning a Biased Model for Improving Fairness," *IEEE Transactions on Big Data*, vol. 11, pp. 1397–1410, [Online]. Available: https://doi.org/10.1109/TBDATA.2024.3460537

[35] S. Uddin, H. Lu, A. Rahman, and J. Gao (2024),, "A novel approach for assessing fairness in deployed machine learning algorithms," *Scientific Reports*, vol. 14, no. 68651, pp. 1–15, [Online]. Available: https://doi.org/10.1038/s41598-024-68651-w

[36] S. Teo, (2024), "How to think about freedom of thought (and opinion) in the age of AI," *Computer Law & Security Review*, vol. 53, p. 105969, [Online]. Available: https://doi.org/10.1016/j.clsr.2024.105969

[37] N. N. M. Saufi, S. Kamaruddin, A. M. Mohammad, N. A. Abd Jabar, W. R. W. Rosli, and Z. M. Talib (2024), "Disruptive AI Technology and Hate Speech: A Legal Redress in Malaysia," in *Proc. 2023 Int. Conf. on Disruptive Technologies (ICDT)*, pp. 759–763, [Online]. Available: https://doi.org/10.1109/ICDT57929.2023.10150942

[38] I. Heider, J. Baumgärtner, A. Bott, R. Ströbel, A. Puchta, and J. Fleischer (2024), "Towards a Testing Framework for Machine Learning Model Deployment in Manufacturing Systems," *Procedia CIRP*, vol. 128, pp. 373–379, [Online]. Available: https://doi.org/10.1016/j.procir.2024.07.022

[39] J. González-Sendino and A. Serrano (2024),, "A Review of Bias and Fairness in Artificial Intelligence," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 9, no. 5, pp. 1–14,[Online]. Available: https://doi.org/10.9781/ijimai.2023.11.001

[40] R. Shams, D. Zowghi, and M. Bano (2023),, "AI and the quest for diversity and inclusion: a systematic literature review," *AI and Ethics*, vol. 5, pp. 411–438, [Online]. Available: https://doi.org/10.1007/s43681-023-00362-w

[41] R. Görge, M. Mock, and M. Akila (2024),, "Inspecting and Measuring Fairness of Unlabeled Image Datasets," in *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*, pp. 191–200, [Online]. Available: https://doi.org/10.1109/ICDEW61823.2024.00031

[42] F. Liao, S. Adelaine, M. Afshar, and B. Patterson (2022), "Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes," *Frontiers in Digital Health*, vol. 4, 2022. [Online]. Available: https://doi.org/10.3389/fdgth.2022.931439

[43] U. A. Usmani, A. Happonen, and J. Watada (2023), "Human-Centered Artificial Intelligence: Designing for User Empowerment and Ethical Considerations," in *2023 5th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–5, [Online]. Available: https://doi.org/10.1109/HORA58378.2023.10156761

[44] H. V. Subramanian, C. Canfield, and D. B. Shank (2024), "Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review," *Artificial Intelligence in Medicine*, vol. 149, p. 102780, [Online]. Available: https://doi.org/10.1016/j.artmed.2024.102780

[45] A. K. Burki, M. N. A. Mafaz, Z. Ahmad, A. Zulfaka, and M. Y. B. Isa (2024), "Artificial Intelligence and Environmental Sustainability: Insights from PLS-SEM on Resource Efficiency and Carbon Emission Reduction," *OPSearch: American Journal of Open Research*, [Online]. Available: https://doi.org/10.58811/opsearch.v3i10.141

[46] A. Halsband (2024), "Sustainable AI and Intergenerational Justice," *Sustainability*, vol. 14, no. 7, p. 3922, [Online]. Available: https://doi.org/10.3390/su14073922

[47] V. Maphosa (2024), "The Rise of Artificial Intelligence and Emerging Ethical and Social Concerns," *AI, Computer Science and Robotics Technology*, [Online]. Available: https://doi.org/10.5772/acrt.20240020

[48] E. S. Andersen, J. B. Birk-Korch, R. Hansen, L. H. Fly, R. Röttger, D. Arcani, C. Brasen, I. Brandslund, and J. S. Madsen (2024), "Monitoring performance of clinical artificial intelligence in health care: a scoping review," *JBI Evidence Synthesis*, [Online]. Available: https://doi.org/10.11124/jbies-24-00042

[49] C. Jacob, N. Brasier, E. Laurenzi, S. Heuss, S.-G. Mougiakakou, A. Çöltekin, and M. K. Peter (2024), "AI for IMPACTS Framework for Evaluating the Long-Term Real-World Impacts of AI-Powered Clinician Tools: Systematic Review and Narrative Synthesis," *Journal of Medical Internet Research*, vol. 27, p. e67485, [Online]. Available: https://doi.org/10.2196/67485

[50] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. C. Fong, and A. Monroy-Hernández, "Humans (2023), AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application," in *Proc. 2023 ACM Conf. on Fairness, Accountability, and Transparency (FAccT)*, pp. 1–13, [Online]. Available: https://doi.org/10.1145/3593013.3593978

[51] H. R. Saeidnia, S. G. H. Fotami, B. D. Lund, and N. Ghiasi (2024), "Ethical Considerations in Artificial Intelligence Interventions for Mental Health and Well-Being: Ensuring Responsible Implementation and Impact," *Social Sciences*, vol. 13, no. 7, p. 381, [Online]. Available: https://doi.org/10.3390/socsci13070381

[52] I. Naja, M. Markovic, P. Edwards, W. Pang, C. Cottrill, and R. Williams (2022), "Using Knowledge Graphs to Unlock Practical Collection, Integration, and Audit of AI Accountability Information," *IEEE Access*, vol. 10, pp. 74383–74411, [Online]. Available: https://doi.org/10.1109/ACCESS.2022.3188967

[53] G. V. Musyoka, R. M. Mutua, and J. Tocho (2024), "AI-Based Framework for Government Oversight of Personal Data Consent Compliance: A Case Study of Nairobi County," *International Journal of Professional Practice*, vol. 12, no. 6, [Online]. Available: https://doi.org/10.71274/ijpp.v12i6.498

[54] K. Maheshwari, C. Jedan, I. Christiaans, M. V. van Gijn, E. Maeckelberghe, and M. Plantinga (2024), "AI-Inclusivity in Healthcare: Motivating an Institutional Epistemic Trust Perspective," *Cambridge Quarterly of Healthcare Ethics*, pp. 1–15, [Online]. Available: https://doi.org/10.1017/S0963180124000215

[55] V. Matta, G. Bansal, F. Akakpo, S. Christian, S. Jain, D. Poggemann, J. Rousseau, and E. Ward (2022), "Diverse perspectives on bias in AI," *Journal of Information Technology Case and Application Research*, vol. 24, no. 3, pp. 135–143, [Online]. Available: https://doi.org/10.1080/15228053.2022.2095776

[56] O. E. Ademola (2024), "Detailing the Stakeholder Theory of Management in the AI World: A Position Paper on Ethical Decision-Making," *Advances in Multidisciplinary and Scientific Research Journal Publication*, vol. 10, no. 1, pp. 1–8, [Online]. Available: https://doi.org/10.22624/AIMS/SIJ/V10N1P1

[57] D. Nelson (2024), "Use Cases for AI in Reliability and Maintainability," in *2025 Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–6, [Online]. Available: https://doi.org/10.1109/RAMS48127.2025.10935281

[58] A. Sajid Mohammed, V. R. Saddi, S. K. Gopal, S. Dhanasekaran, and M. S. Naruka (2024), "AI-Driven Continuous Integration and Continuous Deployment in Software Engineering," in *2024 2nd Int. Conf. on Disruptive Technologies (ICDT)*, pp. 531–536, [Online]. Available: https://doi.org/10.1109/ICDT61202.2024.10489475

[59] B. Lund, Z. Orhan, N. R. Mannuru, R. V. K. Bevara, B. Porter, M. K. Vinaih, and P. Bhaskara (2025), "Standards, frameworks, and legislation for artificial intelligence (AI) transparency," *AI and Ethics*, [Online]. Available: https://doi.org/10.1007/s43681-025-00661-4

[60] H. R. Saeidnia, F. Firuzpour, M. Kozak, and H. S. Majd (2025), "Advancing cancer diagnosis and treatment: integrating image analysis and AI algorithms for enhanced clinical practice," *Artificial Intelligence Review*, vol. 58, p. 105, [Online]. Available: https://doi.org/10.1007/s10462-025-11117-w

[61] M. Casu, S. Triscari, S. Battiato, L. Guarnera, and P. Caponnetto (2024), "AI Chatbots for Mental Health: A Scoping Review of Effectiveness, Feasibility, and Applications," *Applied Sciences*, vol. 14, no. 13, p. 5889, [Online]. Available: https://doi.org/10.3390/app14135889

[62] M. K. L. Ismaeil (2024), "Harnessing AI for Next-Generation Financial Fraud Detection: A Data-Driven Revolution," *Journal of Ecohumanism*, vol. 3, no. 7, [Online]. Available: https://doi.org/10.62754/joe.v3i7.4248

[63] Á. Vieira-Barboza (2025), "Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review," *Journal of Risk and Financial Management*, vol. 18, no. 5, p. 228, [Online]. Available: https://doi.org/10.3390/jrfm18050228

[64] M. González (2025), "Personalization of Learning through Artificial Intelligence: An Analysis of Adaptive Models in Digital Education," *Journal of Information Systems Engineering and Management*, vol. 10, no. 30s, [Online]. Available: https://doi.org/10.52783/jisem.v10i30s.4922

[65] N. S. Sumanth, S. Vishnu Priya, M. Sankari, and K. S. Kamatchi (2024), "AI-Enhanced Learning Assistant Platform," in *2024 Int. Conf. on Inventive Computation Technologies (ICICT)*, pp. 846–852, [Online]. Available: https://doi.org/10.1109/ICICT60155.2024.10545011

[66] T. Lu, W. Li, and Y. Zhang (2024), "Predictive Maintenance in Smart Manufacturing Using Deep Learning and IoT: Review and Challenges," *Journal of Manufacturing Systems*, vol. 72, pp. 1–18, [Online]. Available: https://doi.org/10.1016/j.jmsy.2024.03.008

[67] M. Taleghani and M. Jabreilzadeh Sola (2024), "Examining the Social Consequences of Automation and Artificial Intelligence in Industrial Management," *Research in Economics and Management*, vol. 9, no. 3, pp. 35–46, [Online]. Available: https://doi.org/10.22158/rem.v9n3p35

[68] L. Babashahi, C. E. Barbosa, Y. Lima, A. Lyra, H. Salazar, M. Argôlo, M. Almeida, and J. M. de Souza (2024), "AI in the Workplace: A Systematic Review of Skill Transformation in the Industry," *Administrative Sciences*, vol. 14, no. 6, p. 127, [Online]. Available: https://doi.org/10.3390/admsci14060127

[69] I. P. Basheer (2024), "Bias in the Algorithm: Issues Raised Due to Use of Facial Recognition in India," *Journal of Development Policy and Practice*, [Online]. Available: https://doi.org/10.1177/24551333241283992

[70] A. Oluka (2024), "Mitigating Biases in Training Data: Technical and Legal Challenges for Sub-Saharan Africa," *International Journal of Applied Research in Business and Management*, vol. 5, no. 1, pp. 1–14, [Online]. Available: https://doi.org/10.51137/ijarbm.2024.5.1.10

[71] G. V. Musyoka, R. M. Mutua, and J. Tocho (2024), "AI-Based Framework for Government Oversight of Personal Data Consent Compliance: A Case Study of Nairobi County," *International Journal of Professional Practice*, vol. 12, no. 6, [Online]. Available: https://doi.org/10.71274/ijpp.v12i6.498

[72] A. Corning (2024), "The diffusion of data privacy laws in Southeast Asia: learning and the extraterritorial reach of the EU's GDPR," *Contemporary Politics*, vol. 30, no. 3, pp. 656–677, [Online]. Available: https://doi.org/10.1080/13569775.2024.2310220

[73] V. Agrawal (2022), "Demystifying the Chinese Social Credit System: A Case Study on AI-Powered Control Systems in China," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 36, no. 11, pp. 13124–13125, [Online]. Available: https://doi.org/10.1609/aaai.v36i11.21698

[74] D. Sadhya and T. Sahu (2024), "A critical survey of the security and privacy aspects of the Aadhaar framework," *Computers & Security*, vol. 140, p. 103782, [Online]. Available: https://doi.org/10.1016/j.cose.2024.103782

[75] A. Mishra, M. Gowrav, V. Balamuralidhara, and K. S. Reddy (2021), "Health in Digital World: A Regulatory Overview in United States," *Journal of Pharmaceutical Research International*, vol. 33, no. 43B, pp. 120–130, [Online]. Available: https://doi.org/10.9734/jpri/2021/v33i43b32573

[76] G. Ayana, K. Dese, H. Daba, B. Mellado, K. Badu, E. I. Yamba, S. L. Faye, M. Ondua, D. Nsagha, D. Nkweteyim, and J. D. Kong (2024), "Decolonizing global AI governance: Assessment of the state of decolonized AI governance in Sub-Saharan Africa," *Royal Society Open Science*, vol. 11, no. 2, p. 231994, [Online]. Available: https://doi.org/10.1098/rsos.231994

[77] J. Laksito, B. Pratiwi, and W. Ariani (2025), "Harmonizing Data Privacy Frameworks in Artificial Intelligence: Comparative Insights from Asia and Europe," *Perkara: Jurnal Ilmu Hukum dan Politik*, vol. 2, no. 4, [Online]. Available: https://doi.org/10.51903/perkara.v2i4.2229

[78] H. Uhlemann (2022), "Legislation Supports Autonomous Vehicles But Not Connected Ones [Connected and Automated Vehicles]," *IEEE Vehicular Technology Magazine*, vol. 17, no. 2, pp. 112–115, [Online]. Available: https://doi.org/10.1109/MVT.2022.3159987

[79] C. Bura, S. Kamatala, and P. K. Myakala (2025), "Ethical Challenges in Data Science: Navigating the Complex Landscape of Responsibility and Fairness," *International Journal of Current Science Research and Review*, vol. 8, no. 3, pp. 221–230, [Online]. Available: https://doi.org/10.47191/ijcsrr/v8-i3-09

[80] Z. Li (2024), "AI Ethics and Transparency in Operations Management: How Governance Mechanisms Can Reduce Data Bias and Privacy Risks," *Journal of Applied Economics and Policy Studies*, vol. 13, [Online]. Available: https://doi.org/10.54254/2977-5701/13/2024130

[81] N. Allahrakha (2024), "UNESCO's AI Ethics Principles: Challenges and Opportunities," *International Journal of Law and Policy*, [Online]. Available: https://doi.org/10.59022/ijlp.225

[82] B. Danso William and E. Hanson (2023), "Artificial intelligence disruption and its impacts on future employment in Africa - A case of the banking and financial sector in Ghana," *i-manager's Journal on Software Engineering*, vol. 18, no. 1, pp. 1–12, [Online]. Available: https://doi.org/10.26634/jse.18.1.20082

[83] T. Yaşar (2024), "Artificial Intelligence in Business Operations: Exploring How AI Technologies Are Reshaping Processes, Enhancing Decision-Making, and Driving Efficiency Across Various Industries," *Human Computer Interaction*, [Online]. Available: https://doi.org/10.62802/r78f9a37

[84] L. Babashahi, C. E. Barbosa, Y. Lima, A. Lyra, H. Salazar, M. Argôlo, M. Almeida, and J. M. de Souza (2024), "AI in the Workplace: A Systematic Review of Skill Transformation in the Industry," *Administrative Sciences*, vol. 14, no. 6, p. 127, [Online]. Available: https://doi.org/10.3390/admsci14060127

[85] S. Joshi (2025), "Retraining US Workforce in the Age of Agentic Gen AI: Role of Prompt Engineering and Up-Skilling Initiatives," *International Journal of Advanced Research in Science, Communication and Technology*, [Online]. Available: https://doi.org/10.48175/ijarsct-23272

[86] W. Ali (2024), "The Role of Artificial Intelligence Technology and Innovation in Disrupting Traditional Business Models and Startup Ecosystem," *International Journal of Scientific Research in Engineering and Management*, [Online]. Available: https://doi.org/10.55041/ijsrem37268

[87] S. Kondra, S. Medapati, M. Koripalli, S. R. S. C. Nandula, and J. Zink (2025), "AI and Diversity, Equity, and Inclusion (DEI): Examining the Potential for AI to Mitigate Bias and Promote Inclusive Communication," *Journal of Artificial Intelligence and Machine Learning*, vol. 3, no. 1, [Online]. Available: https://doi.org/10.55124/jaim.v3i1.249

[88] R. K. Yekollu, T. B. Ghuge, S. S. Biradar, S. V. Haldikar, and O. F. M. A. Kader (2024), "Explainable AI in Healthcare: Enhancing Transparency and Trust in Predictive Models," in *2024 5th Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1660–1664, [Online]. Available: https://doi.org/10.1109/ICESC60852.2024.10690121

[89] M. Ouled Sghaier, M. Hadzagic, and E. Shahbazian (2024), "Verifiable Human Autonomy Teaming for NORAD C2 Operations," in *2024 IEEE Conf. on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pp. 137–144, [Online]. Available: https://doi.org/10.1109/CogSIMA61085.2024.10553806

[90] R. Kumar, R. Goel, N. Sidana, A. P. Sharma, S. Ghai, T. Singh, R. Singh, N. Priyadarshi, B. Twala, and V. Ahmad (2024), "Enhancing climate forecasting with AI: Current state and future prospect," *F1000Research*, [Online]. Available: https://doi.org/10.12688/f1000research.154498.1

[91] V. Vieira Barboza (2025), "Towards Fair AI: Mitigating Bias in Credit Decisions—A Systematic Literature Review," *Journal of Risk and Financial Management*, vol. 18, no. 5, p. 228, [Online]. Available: https://doi.org/10.3390/jrfm18050228

[92] V. Orobinskaya, T. N. Mishina, A. P. Mazurenko, and V. V. Mishin (2024), "Problems of Interpretability and Transparency of Decisions Made by AI," in *Proc. 2024 6th Int. Conf. on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, pp. 667–671, [Online]. Available: https://doi.org/10.1109/SUMMA64428.2024.10803745