Envisaging Ethical Artificial Intelligence Governance Frameworks for Public Sector Applications: Addressing Accountability, Transparency and Fairness

Mwaaza Tembo
Digital Transformation and Strategic Information Research Centre
National Institute for Scientific and Industrial Research
Lusaka, Zambia
mwaazabelindatembo@gmail.com

Jameson Mbale
School of Information Communication & Technology, Department of Computer Science
The Copperbelt University
Kitwe, Zambia
jameson.mbale@gmail.com

Abstract - The integration of Artificial Intelligence (AI) into the public sector presents significant ethical and governance challenges that necessitate a comprehensive framework to ensure responsible implementation. This paper aims to develop a robust ethical AI governance framework tailored for government and public sector institutions. The research addresses critical issues such as accountability, transparency, and fairness in AI systems, focusing on maintaining public trust and mitigating potential biases. We propose ethical principles for AI development, evaluate existing regulatory models, and offer recommendations for effective oversight and public engagement. By analysing current policy models and integrating ethical considerations into the AI lifecycle, this study seeks to balance innovation with ethical imperatives. The findings provide actionable insights for public sector leaders and policymakers to establish governance frameworks that promote ethical AI usage, manage associated risks, and enhance societal benefits, ensuring equitable outcomes in public sector applications.

Keywords – Artificial Intelligence, Transparency, Accountability, Fairness, Ethics, Governance Frameworks

I. INTRODUCTION

Artificial Intelligence (AI) is an interdisciplinary research field that has recently gained special importance in society, economics and the public sector [1]. AI can be seen as "a capital-labor hybrid and can replicate labour activities at much greater scale and speed, and to even perform some tasks beyond human capabilities [2]. The applications of AI have encompassed several functions of the public sector , which include, but are not limited to, public health, transport, security , communications [3], mining and even the armed forces (Ayoub & Payne, 2016). These technological advancements in AI and the associated value potential are gaining importance in the context of governments. For example, the government of China has invested $147.8 billion to become a global innovator in the field of AI by 2030. This investment is aimed to promote the country's technology, economy, social welfare, maintain national security, and contribute to the world [4]. The United States in 2016 spent approximately $1.2 billion on research and development of AI related technologies and AI education programs [5]. The increasing investment by governments in AI is evidence of the confidence that the public sector has in AI. Consequently, AI has the ability to bring change and benefits to the public sector [6], however, it brings along ethical challenges particularly regarding transparency, accountability and fairness. These challenges potentially threaten the successful AI use and respective creation of value for the public sector and society as a whole. This paper explores the need for robust governance frameworks that ensure the ethical use of AI in the public sector, focusing on three key dimensions: accountability, transparency, and fairness.

A. Problem Statement

It is easy to see that AI will become pervasive in the public sector. This will certainly bring many benefits in terms of scientific progress, human wellbeing, economic value, and the possibility of exploring solutions to major social and environmental problems [7]. However, such a powerful technology also raises ethical implications such as lack of accountability, transparency and bias in decision making due to unfairness. Current governance frameworks do not sufficiently address these challenges. This paper aims to develop a robust ethical AI governance framework tailored for government and public sector institutions, addressing issues of accountability, fairness and transparency.

B. Objectives

1) To investigate the ethical challenges such as accountability, transparency and fairness posed by AI in public sector applications
2) To examine existing AI governance frameworks and review their effectiveness in addressing these ethical concerns.

 ISBN: 978-9982-95-500-3

3) To propose a comprehensive governance framework that integrates accountability, transparency, and fairness for public sector AI services.

C. Research Questions

The goal of this study is to address the following research questions:

1) How do ethical challenges related to accountability, transparency, and fairness manifest in the use of AI in public sector applications?
2) What existing AI governance frameworks are currently being used in public sector applications?
3) What are the key components of an ethical AI governance framework that can ensure accountability, transparency, and fairness in public sector AI applications?

II. LITERATURE REVIEW

A. Ethical Challenges of Artificial Intelligence

1) Accountability in AI

Accountability has many definitions but, at its core, it is defined as an obligation to inform about, and justify one's conduct to an authority [8]. It refers to the idea that one is responsible for their action—and as a corollary their consequences— and must be able to explain their aims, motivations, and reasons. Understood as a relation, accountability often counterbalances another relation that logically precedes it. Generally speaking, accountability in AI relates to the expectation that designers, developers, and deployers will comply with standards and legislation to ensure the proper functioning of AIs during their lifecycle [8]. However, AIs are neither mere artifacts nor traditional social systems: technological properties often make the outcome of AIs opaque and unpredictable, hindering the detection of causes and reasons for unintended outcomes [9]. Various factors lead to uses of AIs perpetrating wrongdoings, consider for example the case of AI perpetrating undue discrimination, this can result from biased training data, system bugs, programmer errors, misuses, or the replication of social discrimination; and sometimes a combination of these factors. The nature of AIs makes it problematic to assess accountability for such outcomes. This is because opaque and unpredictable outcomes of AIs have similar consequences to the 'many hands' problem [10]. Accountability issues are a major concern because some decisions made by AI systems may have real world implications, especially in sensitive areas such as healthcare, law enforcement and other sectors. That is why there must be an emphasis on developing comprehensive frameworks to ensure responsible and accountable AI implementation.

2) Transparency in AI

Transparency in AI plays a very important role in the overall strive to develop more trustworthy AI as applied to the society and public sector [11]. Generally, transparency is the quality of being easily seen through, while transparency in a business or governance context refers to being open and honest. As part of corporate governance best practices, this requires disclosure of all relevant information so that others can make informed decisions. Recent research highlights the importance of transparency and interpretability in AI systems, particularly as they become more integrated into society and critical domains like healthcare. As AI systems become increasingly integrated into various aspects of society, there is an urgent need to transform these 'black box' models into more transparent and understandable 'glass-box' systems, addressing ethical concerns and promoting trust [12]. Many of the artificial intelligence systems used nowadays have a very high level of accuracy but fail to explain their decisions. This is critical, especially in sensitive areas such as medicine and the health area at large but also for applications of the law, finance etc., where explanations for certain decisions are needed and are often same useful and valuable as the decision itself [13]. Various explanatory methods have been developed to interpret deep learning models, from computing input sensitivities to meaningfully decomposing decisions in terms of input variables [14].The medical field, which requires high accountability, has become a focus for XAI research, with efforts to categorize interpretability approaches and encourage data-driven, mathematically grounded medical education [15]. Ongoing research seeks to balance transparency with other objectives like accuracy and privacy. That is why there must be an emphasis on developing comprehensive frameworks to ensure transparency in AI systems.

3) Fairness in AI

Fairness can be defined as impartial and just treatment or behaviour without favouritism or discrimination. A plethora of work has shown that AI systems can systematically and unfairly be biased against certain populations in multiple scenarios [16]. For example, in the medical sector, it has been detected that AI models fed with chest radiography for pathology classification have a higher rate of underdiagnosis for under-served sub-populations, including Black patients, this could increase the probability of those patients being sent home without receiving the care they need [16]. Various tools and practices have been developed to support practitioners in identifying, assessing, and mitigating fairness-related harms caused by AI systems [17], however, a comprehensive governance framework is imperative to ensure AI does not negatively impact the public sector.

B. Ethical Ai Governance Frameworks

AI governance is a system of rules, processes, frameworks, and tools within an organization to ensure the ethical and responsible development of AI. It is a system of rules, processes, frameworks, and technological tools that are employed in an organization to ensure that the use of AI aligns with the organizational principles, legal requirements, as well as social and ethical standards [18].

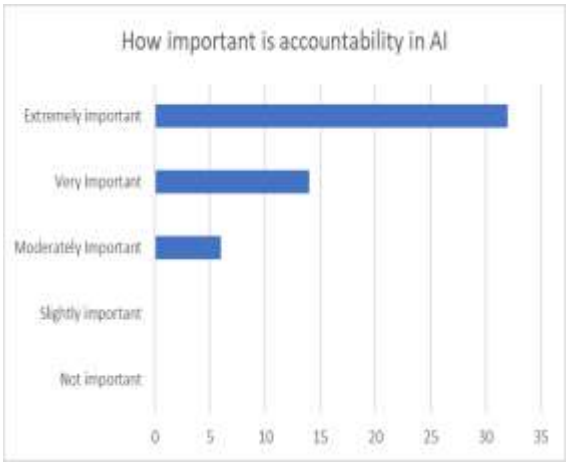
 ISBN: 978-9982-95-500-3

1) The European Union's Ethics Guidelines For Trustworthy AI

In 2019, a High-level Expert Group (HLEG) developed guidelines on trustworthy AI that acted as a basis for the policy recommendations in preparation for the EU AI Act. They define trustworthy AI to be lawful, ethical, and robust. It is based on seven key requirements which include transparency, accountability and diversity – non-discrimination and fairness [19]. It additionally introduces the concept of human oversight and enlarges the well-being requirement from societal to environmental.

2) The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

The framework is a consortium of different standards, including specific documents, for example, regarding system design, certification, and bias. This framework generally consists of eight principles: transparency, accountability, awareness of limitations, safety and well-being, reliability and dependability, equity, inclusivity, and privacy protection.



In addition to the eight principles, it also includes a set of metrics to assess the extent to which AI systems adhere to these principles [18].

3) The Montreal Declaration of Responsible AI

The Declaration's first objective consists of identifying the ethical principles and values that promote the fundamental interests of people and groups. The code for algorithms, whether public or private, must always be accessible to the relevant public authorities and stakeholders for verification and control purposes. In accordance with the transparency requirement for public decisions, the code for decision-making algorithms used by public authorities must be accessible to all. AIS must be designed and trained so as not to create, reinforce, or reproduce discrimination based on — among other things — social, sexual, ethnic, cultural, or religious differences

*Figure SEQ Figure * ARABIC 3: Importance of Accountability in AI*

III. METHODS

A. Data Collection

An online survey was also administered to collect quantitative data. Online surveys offer quick data collection, efficient service, easy data processing and wide coverage [20]. The survey included Likert scale questions related to the key dimensions – accountability, transparency and fairness.

The participants of this survey comprised of individuals from various fields and backgrounds, ensuring diverse perspectives and views, making the study more inclusive.

This allowed for comprehensive insights into the ethical concerns particularly transparency, accountability and fairness, surrounding AI implementation.

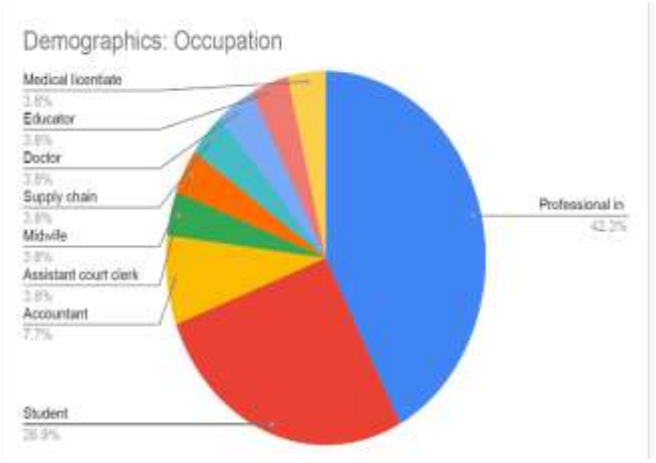
B. Data Analysis, Results and Discussion

1) Demography

A total of 52 respondents participated in this survey, this group comprised diverse backgrounds, including; students, IT professional, educators, supply chain professionals and many others. Fig 1 illustrates the occupational distribution of the respondents. The majority of respondents were moderately familiar with AI technologies, with about 10% indicating at least basic knowledge of AI applications in the public sector.

2) Accountability in AI

According to responses gathered in Fig 2, 60% emphasized the critical importance of accountability in AI decision making processes. This underscores a prevalent concern regarding the need for accountable and transparent mechanisms to ensure that AI systems operate within established ethical guidelines. As indicated in Fig 3 only, 15% indicated that they were "very confident" in AI accountability mechanisms, see Fig 2. These results suggest a perceived lack of clarity regarding who is responsible for AI decisions in the public sector. This finding suggests that accountability is a major ethical concern in the public sector



 ISBN: 978-9982-95-500-3

and to enhance public trust, this aspect must be integrated in a suitable AI governance framework.

Figure 3: Confidence levels in AI Accountability

3) Transparency

According to Fig 4 and 5, the findings indicated a large portion of respondents view AI systems as moderately transparent and would feel more comfortable if AI decision making processes would be accessible to the public. This points to a significant gap in public understanding of how AI operates in decision-making processes.

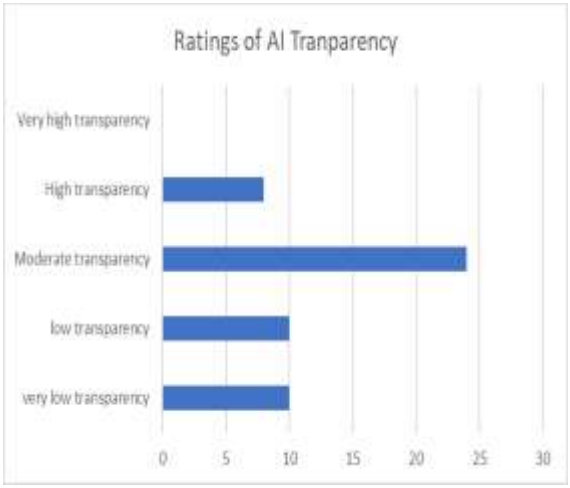


Figure 4: Ratings of AI Transparency

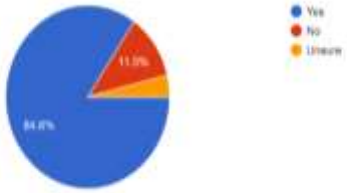


Figure 5: AI decision making processes accessible to public

4) Fairness

According to the responses gathered, over 90% expressed deep concerns about fairness in AI decision making. Several key issues emerged, such as, AI can amplify political opinions based on the data they have been trained on. This raises concerns about the fairness of AI in applications like content moderation or decision-making tools in public policy, where unbiased objectivity is crucial. Additionally, another issue that emerged was gender bias, many virtual assistants are designed with female voices, potentially reinforcing the stereotype of women being in subordinate or service- oriented roles.

IV. FINDINGS AND RECOMMENDATION

A. Summary of Findings

Overall, the survey revealed the need for stronger accountability, transparency and fairness in AI systems. Furthermore, it revealed the need for a comprehensive governance framework to address these key dimensions.

B. Recommendations For An Ethical AI Governance Framework

Based on the findings of this study, it is clear that issues of transparency, accountability and fairness raise major concerns in the use of AI in the public sector. Therefore, it is imperative to a comprehensive governance framework that will address these concerns and foster public trust.

1) Ensuring Accountability in Artificial Intelligence

A fundamental aspect of AI governance is ensuring clear accountability for decisions made by AI systems. To achieve this, the proposed framework includes the following key elements:

Role-Based Accountability: Public sector organizations must establish clear guidelines for human oversight at each stage of AI deployment. Specific individuals or teams should be designated as responsible for the outcomes generated by AI systems, ensuring that any errors or biases can be traced to accountable parties.

 ISBN: 978-9982-95-500-3

Audit Trails and Decision Documentation: AI systems must generate detailed records of their decision-making processes. These records should be subject to regular audits by independent bodies to ensure compliance with ethical standards.

Liability Framework: In cases where AI systems lead to adverse outcomes, there should be a well-defined process to assign liability. This framework must outline both individual and organizational responsibility, ensuring that those affected by AI decisions have access to recourse mechanisms.

2) Ensuring Transparency in AI Systems

The trade-off between the performance of AI algorithms and their explainability, commonly called explainable Artificial Intelligence (XAI), helps to improve trust in AI applications [21]. Explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand [22]. Public sector organizations should prioritize the use of explainable AI (XAI) technologies that can provide stakeholders with insights into how specific outcomes were reached. Additionally, AI algorithms and datasets used in public sector AI systems must be accessible to the public and regulatory bodies to allow them to understand how decisions are made.

3) Ensuring Fairness and Mitigating Bias

To avoid the risk of bias having a negative impact on the public sector, the proposed framework prioritizes fairness at all stages of AI development and deployment. All AI systems must be coerced to implement bias mitigation strategies at all main stages: before, during and after training. *Before training*, one must seek to rebalance datasets by collecting more representative data [16] that is appropriate with the demographics of the target population. *During training*, several alternatives exist to mitigate model biases, such as the use of data augmentation and adversarial training [16]. Finally, *after training*, model outcomes can be post-processed so as to calibrate the predictions across the different sub-groups. In order to develop AI systems that are trustworthy, it is advisable to consult stakeholders who may directly or indirectly be affected by the system throughout its life cycle. It is beneficial to solicit regular feedback even after deployment and set up longer term mechanisms for stakeholder participation, for example by ensuring workers information, consultation and participation throughout the whole process of implementing AI systems at organisations. Additionally, fairness audits must be caonducted regularly to assess the fairness of AI systems for public sector applications.

1) Regulatory and Legal Integration

Compliance is about binding AIs to align with ethical, legal, or technical norms. It defines the design, development, and deployment standards to be met throughout the entire lifecycle of an Artificial Intelligence systems (Claudio Novelli et al., 2023).

Governance frameworks must ensure that Artificial Intelligence systems adhere to relevant laws and regulations such as privacy, diversity, human rights and data protection. To ensure compliance, independent ethical committees must be established within public sector organizations. These committees would see to it that Artificial Intelligence systems are accountable, transparent and fair in their operations.

V. CONCLUSION

As AI continues to play a larger role in public sector applications, ethical governance frameworks that emphasize accountability, transparency, and fairness are essential. By implementing the principles outlined in this paper, public sector agencies can ensure that AI systems are used responsibly, maintaining trust and protecting the rights of citizens.

REFERENCES

[1] Boyd M and Wilson N, "Rapid Development in Artificial Intelligence: How mght the New Zealand Government Respond?," *Policy Quartely*, vol. 13, no. 4, pp. 36–44, 2017.

[2] Purdy M. and Daugherty P., "Why Artificial Intelligence is the Future of Growth," 2016.

[3] Weslei Gomes de Sousa, Elis Regina Pereira de Melo, Paulo Henrique De Souza Bermejo, Rafael Araújo Sousa Farias, and Adalmir Oliveira Gomes, "How and where is artificial intelligence in the public sector going? A literature review and research agenda," *Gov. Inf. Quartely*, vol. 36, no. 4, Oct. 2019.

[4] "The State Council The People's Republic of China," China issues guideline on artificial intelligence development. [Online]. Available: english.gov.cn

[5] "The Guardian," AI programs exhibit racial and gender biases, research reveals. [Online]. Available: https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals

[6] Bernd W. Wirtz, Jan C. Weyerer, and Carolin Geyer, "Artificial Intelligence and the Public Sector - Applications and Challenges," *Int. J. Public Adm.*, vol. 42, no. 7, pp. 596–615, 2019, doi: https://doi.org/10.1080/01900692.2018.1498103.

[7] Francesca Rossi, "BUILDING TRUST IN ARTIFICIAL INTELLIGENCE," *J. Int. Aff.*, vol. 72, no. 1, pp. 127–134, 2018, doi: https://www.jstor.org/stable/26588348.

[8] Lindberg SI, "Mapping accountability: core concept and subtypes," *Int Rev Adm Sci*, vol. 79, no. 2, pp. 202–226, 2013, doi: https:// doi. org/ 10. 1177/ 0020852313477761.

[9] Tsamados A *et al*., "The ethics of algorithms: key problems and solutions.," *AI Soc*, vol. 37, no. 1, pp. 215–230, 2022.

[10] Claudio Novelli, Mariarosaria Taddeo, and Luciano Floridi, "Accountability in artificial intelligence: what it is and how it works," *OPEN FORUM*, vol. 39, pp. 1871–1882, 2023.

[11] Stefan Larsson and Fredrik Heintz, "Transparency in Artificial Intelligence," May 2020, doi: DOI: 10.14763/2020.2.1469.

[12] Valentina Franzoni, "From Black Box to Glass Box: Advancing Transparency in Artificial Intelligence Systems for Ethical and Trustworthy AI," 2023, doi: https://doi.org/10.1007/978-3-031-37114-1_9.

 ISBN: 978-9982-95-500-3

[13] Lars Quakulinsk, A. Koumpis, and O. Beyan, "Establishing Transparency in Artificial Intelligence Systems," *Fourth Int. Conf. Transdiscipl. AI TransAI*, 2022.

[14] Leilani H, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael A. Specter, and Lalana K, "Explaining Explanations: An Overview of Interpretability of Machine Learning," *Int. Conf. Data Sci. Adv. Anal.*, 2018, doi: https://doi.org/10.1109/DSAA.2018.00018.

[15] Erico Tjoa and Cuntai Guan, "A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI," *IEEE Trans. Neural Netw. Learn. Syst.*, 2019, doi: https://doi.org/10.1109/TNNLS.2020.3027314.

[16] María Agustina Ricci Lara, Rodrigo Echeveste, and Enzo Ferrante, "Addressing fairness in artificial intelligence for medical imaging," *Comment*, vol. 13, no. 451, 2022, doi: https://doi.org/10.1038/s41467-022-32186-3.

[17] Michael Madaio, Lisa Egede, Hariharan Subramonyam, Jennifer Wortman Vaughan, and Hanna Wallach, "Assessing the Fairness of AI Systems: AI Practitioners' Processes, Challenges, and Needs for Support," *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSW1, pp. 1–26, Apr. 2022, doi: https://doi.org/10.1145/3512899.

[18] "Implementing AI Governance: from Framework to Practice," trail. [Online]. Available: https://www.trail-ml.com/blog/implementing-ai-governance

[19] "European AI Alliance," European Commission. [Online]. Available: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal

[20] Tejpal Singh, Avinash Kumar, Kumari Bhawna Vihar, and Aparna Singh, "ONLINE SURVEY MANAGEMENT SYSTEM," 2016.

[21] Masialeti Masialeti, Amir Talaei-Khoei, and Alan T. Yang, "Revealing the role of explainable AI: How does updating AI applications generate agility-driven performance," *Int. J. Inf. Manag.*, vol. 77, 2024, doi: https://doi.org/10.1016/j.ijinfomgt.2024.102779.

[22] Arrieta, A. B, Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., and Tabik, S., "Explainable