United Nations University

TECHNOLOGYBRIEF
No. 5, APRIL 2024

Framework for the Governance of Artificial Intelligence

Tshilidzi Marwala, United Nations University, Tokyo, Japan

Recommendations

- Define the values of AI by aligning them with the principles of human rights and the United Nations Charter.
- Utilize behavioural science to cultivate a culture that optimizes the potential of AI while limiting the associated risks.
- Implement strategies based on the discipline of mechanism design to foster a culture that maximizes the potential benefits of AI while limiting its associated risks.
- Implement frameworks and institutional governance structures to regulate AI.
- Establish policies and regulations to control AI.
- Establish AI standards.
- Create legislation on AI.
- Govern data, algorithms, computing systems, and AI applications.

Introduction

The swift advancement and implementation of artificial intelligence (AI) technologies have significant economic, societal, and ethical consequences.[1] Efficient governance is crucial to optimize the advantages of AI while minimizing its risks.[2,3] This technology brief offers policymakers a concise summary of essential governance matters, suggests strategic methodologies for the appropriate oversight of AI technologies, and presents a framework for AI governance.

Values of AI

The governance of AI must be based on sound values. In this section, I discuss some of these values.

Transparency: AI transparency fosters trust, accountability, and fairness.[4] Transparent AI systems allow stakeholders to understand and evaluate decision-making processes, identifying and correcting biases and assuring ethical and legal compliance. Transparent AI makes applications

safer and more trustworthy, especially in health care and transportation, where transparent AI behaviour is vital for public safety. Making AI systems more transparent helps developers and users collaborate, resulting in technological advances. This openness improves user engagement and regulatory compliance, and promotes informed and productive AI discussion.

Truth: For AI to be trustworthy, it must tell the truth.[5, 6] AI is a dependable decision-making tool in media, education, and science when it produces accurate and unbiased data. Truthful AI prevents misinformation and promotes critical thinking in public conversation. Integrating AI with accurate outputs reduces harm and manipulation by meeting ethical norms. This dedication to truth strengthens AI systems' legitimacy and builds the social trust needed for the widespread adoption and beneficial integration of AI technology into daily life.

Safety and Security: AI systems interact with people in different and profound ways, making their safety vital.[7] Safe AI protects people and communities, building trust and adoption. Prioritizing safety reduces accidents and errors in AI-driven industries, including health care, transportation, and finance, and safeguards lives and money. Safe AI techniques also encourage ethics and legal compliance, promoting social stability and preventing technology misuse. Safety in AI development and deployment ensures society benefits from this technology.

Ethics: Ethics are needed to design and deploy AI technologies that benefit society and minimize harm.[8] AI ethics include justice, transparency, accountability, privacy, and human rights. We may prevent discrimination and properly use AI systems by incorporating ethical values. Ethical AI improves user and stakeholder trust, regulatory compliance, and sustainable innovation. By accentuating ethical behaviours, AI developers and operators can help establish a just society where technology enhances human capabilities without compromising dignity or freedom.

Privacy: AI must respect privacy to secure personal data and retain trust in technology.[9] AI systems process massive volumes of data, including sensitive personal information, which can lead to privacy breaches and exploitation. Using strict privacy rules, AI protects personal data from unlawful access and exploitation, boosting AI technology trust.

Governance Model

The AI governance hierarchy for action proposed in this brief is in **Figure 1**. This diagram shows that AI governance should be based on AI values described in the previous section. On top of these values are human behaviour, mechanisms for incentives and disincentives, and institutional governance structures.[10] The Intergovernmental Panel on Climate Change (IPCC) and the International Atomic Energy Agency (IAEA) are examples of institutional structures with roles in governance.[11] After that are policies and regulations — some at the company level, some within professional organizations such as in the medical field, some at the government level and some at the international level. Standards are set at the national and international levels. Some standards require subject-specific expertise, while others require policy-level expertise. Next is the law. For example, AI governance with implications for human rights must be governed by laws, not regulations.[12] This AI governance hierarchy for action feeds into the AI governance model shown in **Figure 2**.

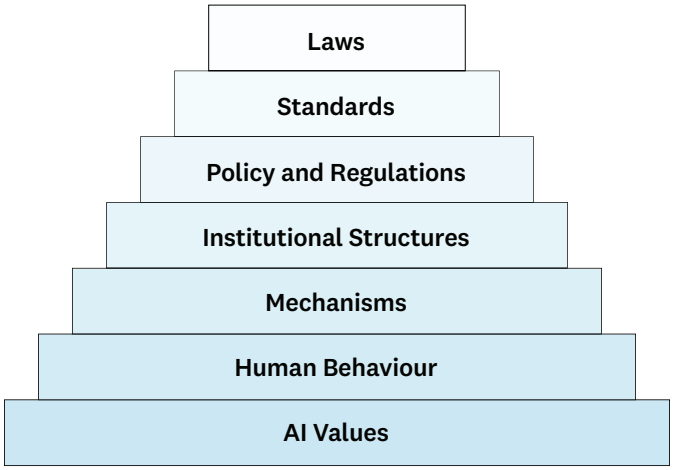

Figure 1 — AI governance hierarchy for action

Governance Areas

The AI governance model conforms to the AI governance hierarchy for action and consists of data feeding into algorithms, algorithms feeding into computing, and computing feeding into applications.

Data: AI data must be governed for ethical, accurate, and secure use. Robust governance systems allow us to oversee data collection, storage, analysis (e.g. synthetic data),[13, 14] and sharing (e.g. cross-border data flow),[15] while preserving sensitive data and preventing misuse. Good data governance

guarantees that AI systems use high-quality, unbiased data to make fair and productive judgments. It helps maintain public trust in AI technologies by meeting regulatory norms. Data governance promotes openness and accountability, allowing stakeholders to understand and dispute AI judgments. Well-governed data makes AI applications more trustworthy, accountable, and reliable.

AI Algorithmic Governance: AI algorithms must be governed to ensure ethical behaviour and impartial results.[16] AI judgments can significantly impact individuals and communities as they become more integrated into health care, banking, and law enforcement.[17, 18] Effective institutional governance structures may drive AI algorithm development, deployment, and monitoring to ensure ethical, transparent, and accountable behaviour. The selection, design, training, and testing of algorithms must be governed. This governance prevents algorithmic biases and discrimination, boosting user trust. It links AI techniques with cultural ideals and legal obligations, fostering technology equity and justice.

Computing Governance: To ensure safe, ethical, and valuable AI technology, computing must be governed.[19] As computer technologies permeate every part of life, from personal data management to essential infrastructure, clear governance frameworks avoid abuse, mitigate risks, and ensure ethical and legal compliance. Governance can help design and use computing technologies that respect privacy, improve security, and prevent harm. Well-defined governance mechanisms build public trust and promote more egalitarian technology access. Semiconductor chips, edge computing, cloud computing, ambient computing, quantum computing, computing energy, and computing water use should be governed.

Governance of AI Applications: To handle its enormous consequences and potential disruptions, AI must be governed across society, economy, and politics.[20] AI technologies affect labour markets, economic development, political decision-making, and public policy. Hence, robust governance mechanisms are needed. With this oversight, AI can improve social well-being, economic fairness, and democratic processes without damaging them. Effective governance prevents misuse, reduces unforeseen effects, and assures equitable benefit distribution. Governance also maintains public trust and promotes sustainable AI integration into daily life by aligning AI applications with ethical and regulatory requirements.

The AI governance areas are illustrated in **Figure 2**.

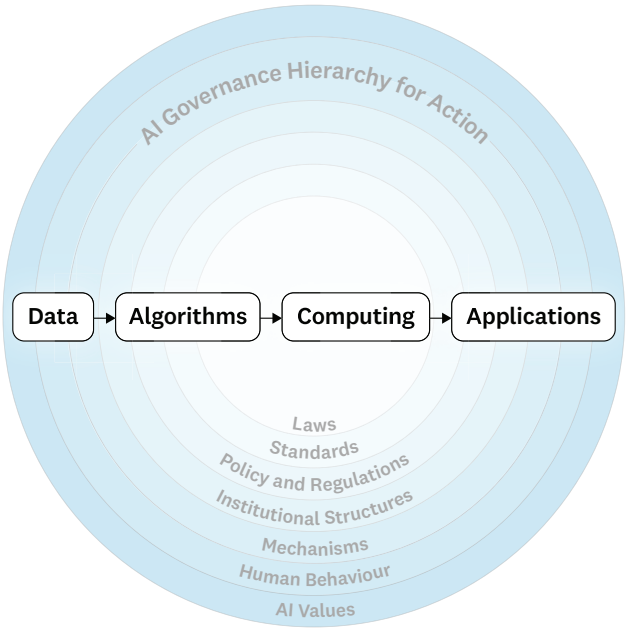

Figure 2 — AI governance model

Recommendations for Effective AI Governance

Define the values of AI by aligning them with the principles of human rights and the United Nations Charter: Defining AI values involves aligning them with human rights and the United Nations Charter to ensure AI development supports human dignity, equality, and freedom. AI can minimize discrimination and privacy risks by integrating fairness, transparency, and accountability. Moreover, incorporating the UN Charter's focus on peace, justice, and cooperation encourages AI to address global challenges, promoting ethical practices and international consensus on protective standards.

Utilize behavioural science to cultivate a culture that optimizes the potential of AI while limiting the associated risks: Utilizing behavioural science in AI involves using insights to shape a culture that maximizes AI's benefits while reducing risks. Organizations can encourage ethical AI usage and discourage harmful practices by implementing nudges and structured incentives. This approach promotes the adoption of privacy safeguards and bias mitigation, ensuring AI development aligns with safe and productive human interactions.

Implement strategies based on the discipline of mechanism design to foster a culture that maximizes the potential benefits of AI while limiting its associated risks: Implementing mechanism design strategies in AI can maximize benefits and minimize risks by aligning individual incentives

with societal goals. This economics and game theory field helps create systems where stakeholders are encouraged to uphold transparency, fairness, and accountability. For instance, algorithms could reward data accuracy and ethical behaviour to prevent biases. This approach ensures AI operates within a safeguarded framework, promoting trust and effectiveness in AI systems.

Implement frameworks and institutional governance structures to regulate AI: Establishing an institutional structure similar to the IPCC or the IAEA is recommended to regulate AI effectively. This entity would set global AI standards, monitor compliance, and foster international cooperation. It would address ethical guidelines, technological standards, and safety protocols, providing a platform for sharing best practices and promoting transparency. Such a structure would ensure that AI development and deployment are safe, ethical, and globally beneficial.

Establish policies and regulations to control AI: Establishing policies and rules to control AI involves crafting legal frameworks that ensure ethical AI use, focusing on transparency, accountability, and fairness while protecting privacy and preventing discrimination. Key measures could include mandatory AI audits and strict data protection requirements. Governments might also create dedicated bodies to oversee AI practices and enforce these regulations, ensuring AI benefits society and adheres to ethical standards.

Establish AI standards: Establishing AI standards involves creating uniform guidelines for AI design, development, and deployment, focusing on technical quality, ethical considerations, and compatibility. These standards should cover data privacy, algorithmic transparency, security, and bias prevention. Developed collaboratively by industry, academia, and regulators, these guidelines should be regularly updated to ensure global safety, effectiveness, and ethical compliance with AI technologies.

Create legislation on AI: Creating legislation on AI involves drafting laws to govern AI development, use, and impact, ensuring ethical operation, privacy protection, and discrimination prevention. Lawmakers should collaborate with experts and the public to craft adaptable, informed regulations. Such legislation might also set up oversight bodies to enforce these laws and manage disputes, safeguarding individual rights and societal welfare as AI technologies become more pervasive.

Govern data, algorithms, computing systems, and AI applications: Data, algorithms, computing systems, and AI applications demand a robust framework to ensure responsible and ethical use. This governance should include data protection measures, algorithmic transparency to avoid biases, strong security standards for computing systems, and specific regulations for AI applications in critical sectors like health care and finance. Such a comprehensive approach builds public trust and ensures that technological innovations contribute positively to society while minimizing risks.

Conclusion

AI governance frameworks must be dynamic and adaptable to technological advancements. Policymakers should collaborate with technologists, businesses, academia, and civil society to create comprehensive governance strategies that ensure AI serves the public good. Continued dialogue and iterative policy development will be vital in navigating the evolving landscape of AI technology and its impacts on society.

ENDNOTES

1 Marwala, T., 2022. *Closing the gap: The fourth industrial revolution in Africa*. Pan Macmillan South Africa.
2 Marwala, T., 2023. *Artificial Intelligence, Game Theory and Mechanism Design in Politics* (pp. 41-58). Singapore: Springer Nature Singapore.
3 Roberts, H., Hine, E., Taddeo, M. and Floridi, L., 2024. Global AI governance: barriers and pathways forward. *International Affairs*, p.iiae073.
4 Ali, A.E., Venkatraj, K.P., Morosoli, S., Naudts, L., Helberger, N. and Cesar, P., 2024. Transparent AI Disclosure Obligations: Who, What, When, Where, Why, How. *arXiv preprint arXiv:2403.06823*.
5 Hurwitz, E. and Marwala, T., 2007, October. Learning to bluff. In *2007 IEEE International Conference on Systems, Man and Cybernetics* (pp. 1188-1193).
6 Markowitz, D.M. and Hancock, J.T., 2024. Generative AI are more truth-biased than humans: A replication and extension of core truth-default theory principles. *Journal of Language and Social Psychology*, 43(2), pp.261-267.
7 Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A., Gupta, J., Hart, C. and Jirotka, M., 2021. Governing AI safety through independent audits. *Nature Machine Intelligence*, *3*(7), pp.566-571.
8 Coeckelbergh, M., 2020. *AI ethics*. MIT Press.
9 Elliott, D. and Soifer, E., 2022. AI technologies, privacy, and security. *Frontiers in Artificial Intelligence*, *5*, p.826737.
10 Marwala, T., 2024. *Mechanism Design, Behavioral Science, and Artificial Intelligence in International Relations*. Morgan Kaufmann.
11 Verbruggen, A. and Laes, E., 2015. Sustainability assessment of nuclear power: Discourse analysis of IAEA and IPCC frameworks. *Environmental Science & Policy, 51*, pp.170-180.
12 Marwala, T. and Mpedi, L.G., 2024. *Artificial intelligence and the law*. Palgrave Mcmillan.
13 Marwala, T., Fournier-Tombs, E. and Stinckwich, S., 2023. The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. *arXiv preprint arXiv:2309.00652*.
14 Sidogi, T., Mongwe, W.T., Mbuvha, R. and Marwala, T., 2022, December. Creating synthetic volatility surfaces using generative adversarial networks with static arbitrage loss conditions. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)* (pp. 1423-1429).
15 Tshilidzi Marwala, Eleonore Fournier-Tombs, Serge Stinckwich, "Regulating Cross-Border Data Flows: Harnessing Safe Data Sharing for Global and Inclusive Artificial Intelligence", UNU Technology Brief 3 (Tokyo: United Nations University, 2023).
16 Ebers, M. and Gamito, M.C., 2021. *Algorithmic Governance and Governance of Algorithms*. Springer.
17 Marwala, T., 2014. *Artificial intelligence techniques for rational decision making*. Springer.
18 Marwala, T. and Hurwitz, E., 2017. *Artificial intelligence and economic theory: skynet in the market (Vol. 1)*. Cham: Springer International Publishing.
19 Sasikala, P., 2012. Cloud computing and e-governance: Advances, opportunities and challenges. *International Journal of Cloud Applications and Computing (IJCAC), 2*(4), pp.32-52.
20 Marwala, T., 2023. *Intelligence, Game Theory and Mechanism Design in Politics* (pp. 135-155). Singapore: Springer Nature Singapore.

EDITORIAL INFORMATION

About the research
This technology brief is part of a UNU series highlighting specific areas of global technology governance related to the Global South and sustainable development.

Author biography
Professor Tshilidzi Marwala is the Rector of United Nations University, headquartered in Tokyo, and Under-Secretary-General of the United Nations. He was previously the Vice-Chancellor and Principal of the University of Johannesburg. Marwala has published over 300 research papers, over 250 articles in newspapers and magazines, 27 books on AI and related topics, and holds five patents. He is a member of the American Academy of Arts and Sciences, the Chinese Academy of Sciences, the World Academy of Sciences (TWAS) and the African Academy of Science.

Disclaimer
The views and opinions expressed in this technology brief do not necessarily reflect the official policies or positions of the United Nations University.

Citation
Tshilidzi Marwala, "*Framework for the Governance of Artificial Intelligence*", UNU Technology Brief 5 (Tokyo: United Nations University, 2024).

ISBN 978-92-808-9155-3