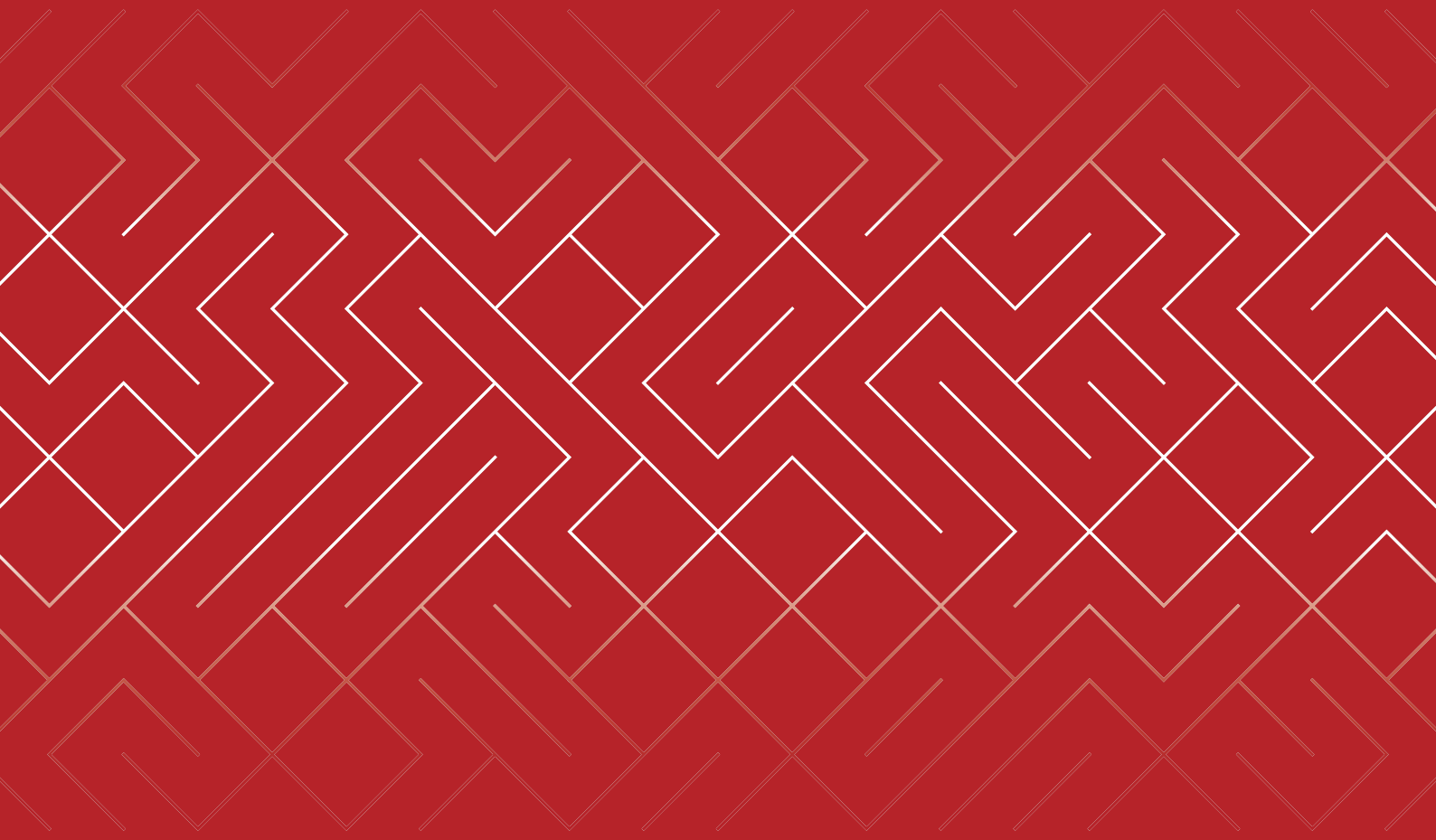
Ethical AI Development in Kenya: The Role of Ethical Data Sourcing and Governance

Strathmore University
Centre for Intellectual Property and Information Technology Law



ETHICAL AI DEVELOPMENT IN KENYA: THE ROLE OF ETHICAL DATA SOURCING AND GOVERNANCE

Prepared by;

The Centre for Intellectual Property and Information Technology Law (CIPIT)

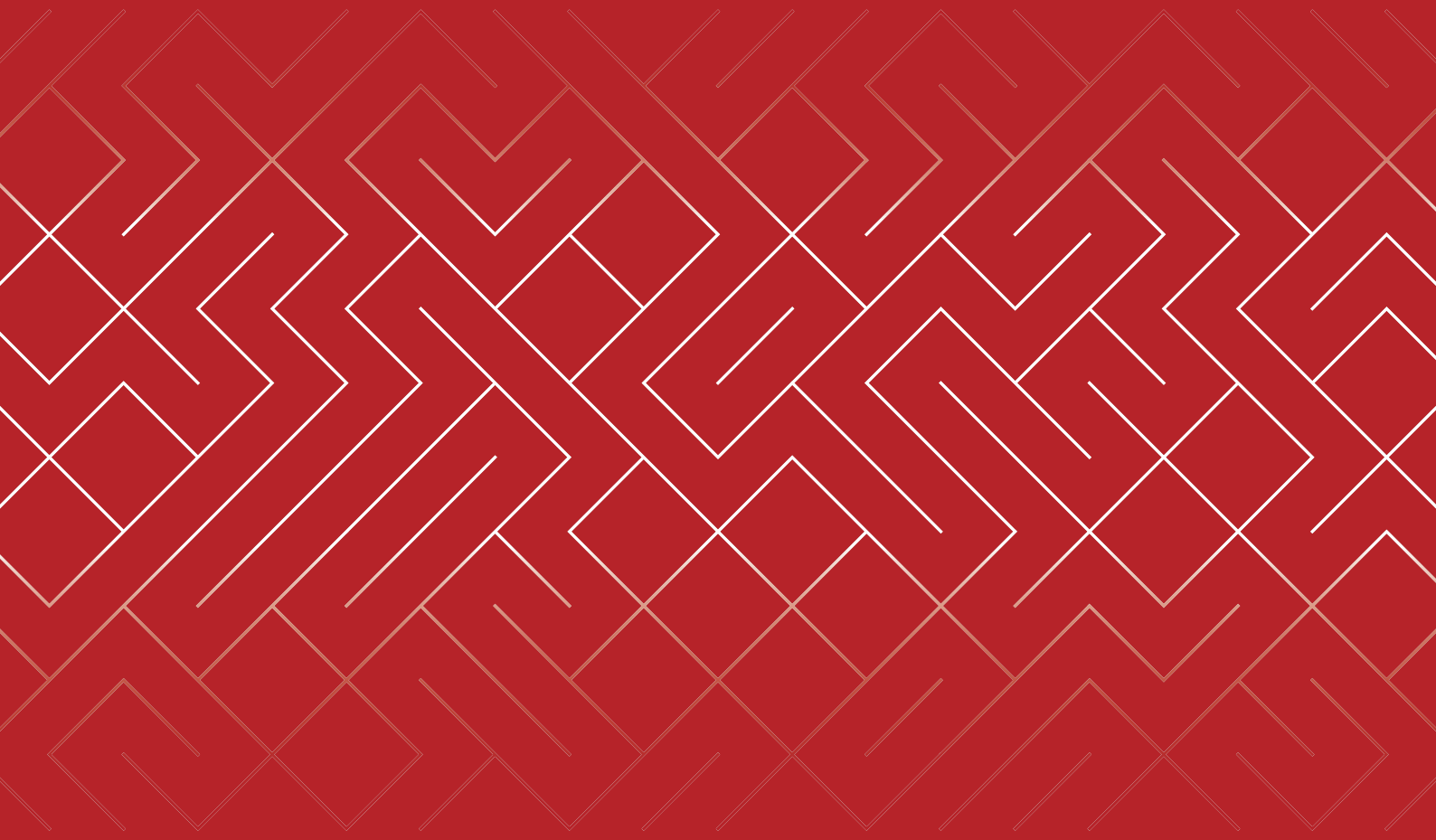
Table of Contents

Acknowledgment

CIPIT sincerely thanks the dedicated team whose teamwork and commitment made this report possible. We are especially grateful to **Ms. Josephine Kaaniru** for her valuable expertise and thoughtful contributions that shaped the content of this report. We also extend our heartfelt thanks to **Ms. Florence Ogonjo** for her careful editorial work, which greatly improved the clarity and quality of the report. We would also like to thank our **Director, Dr. Melissa Omino**, for her leadership and clear vision, which guided the direction of this report and helped maintain its focus and impact. Our appreciation goes to **Jacala Solutions Ltd** for their creative design that gave the report its attractive and professional look. Lastly, we are thankful to the **administrative team** for their steady support and guidance throughout the research and writing process. Their help was essential in bringing this project to completion.

Image Source: vecteezy.com

Ethical AI Development in Kenya: The Role of Ethical Data Sourcing and Governance



I. Background

Achieving ethical Artificial Intelligence (AI) begins long before an algorithm is trained. It starts with how data is gathered and annotated,[1] demanding fairness, transparency, and respect for rights from the outset.[2] Ethical AI, understood as the development and application of AI systems in a manner consistent with moral standards to prevent harm,[3] depends on integrity in these early stages of the AI lifecycle. As Kenya positions itself as a hub for AI innovation in Africa, ensuring ethical practices in data sourcing presents both a challenge and a major opportunity. AI is already making an impact in Kenya's education, agriculture, health, and business sectors,[4] and AI-powered Assistive Technologies (ATs) hold transformative potential for disability inclusion.[5] However, most of the AI tools deployed in these sectors have historically originated from large foreign technology companies with the financial capacity to develop them.[6] Due to the lack of high-quality datasets that fully reflect local populations and concerns, Kenyan developers have often relied on global foundational tools and datasets to build local solutions.[7] This reliance can perpetuate bias, as training data sourced from Western contexts often fails to account for diverse demographics. A striking example is Joy Buolamwini's 2021 documentary *Coded Bias*, which revealed how AI systems struggled to detect darker faces and inaccurately categorised women.[8]

While African languages have long been underrepresented in AI applications, recent initiatives such as Masakhane and AI4D Africa, are working to address this gap by building open-source NLP tools, leveraging transfer learning, and leading community-driven data collection to improve inclusivity.[9] More recently, discussions in Kenya and across Africa have highlighted the practical consequences of bias, including the potential for AI-driven tools in finance or healthcare to disadvantage individuals if algorithms are not audited and adapted for local contexts.[10] Addressing these risks through local, representative datasets offers a critical pathway for African developers to design AI solutions tailored to the continent's realities.

While stakeholders have embarked on the well-intentioned mission of collecting and preparing data for training AI systems, this process has faced challenges that have long plagued relations between Africa and the de-

1 Shahmar Mirishli, 'Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy' (2024) 6 International Scientific Journal 40 <https://arxiv.org/pdf/2503.14539>.

2 Neha Panchal, 'Ethical Considerations in AI Data Annotation' (Damco Solutions15 May 2023) <https://www.damcogroup.com/blogs/understanding-ethical-considerations-in-ai-data-annotation> accessed 12 August 2025.

3 IBM, 'AI Ethics' (Ibm.com17 September 2024) <https://www.ibm.com/think/topics/ai-ethics>.

4 CIPIT, 'The State of AI in Africa Report 2023' (2023) <https://cipit.strathmore.edu/wp-content/uploads/2023/05/The-State-of-AI-in-Africa-Report-2023-min.pdf>.

5 Josephine Kaaniru, 'AI Assistive Technologies (ATS) for Persons with Disabilities (PWDS) in Africa - Centre for Intellectual Property and Information Technology law 31 October 2023) <https://cipit.strathmore.edu/ai-assistive-technologies-ats-for-persons-with-disabilities-pwds-in-africa/>.

6 ibid.

7 United Nations Development Programme (UNDP) and Italian G7 Presidency, 'AI Hub for Sustainable Development Strengthening Local AI Ecosystems through Collective Action' (2024) <https://www.undp.org/sites/g/files/zskgke326/files/2024-07/ai_hub_report_digital.pdf>.

8 PBS, 'Coded Bias | Films | PBS' (Independent Lens2021) <https://www.pbs.org/independentlens/documentaries/coded-bias/>.

9 Victoria Reed, 'African Languages in AI: Breaking Barriers with NLP' (AICompetence.org23 December 2024) <https://aicompetence.org/african-languages-in-ai-barriers-with-nlp/> accessed 30 July 2025.

10 Notice Pasipamire and Abton Muroyiwa, 'Navigating Algorithm Bias in AI: Ensuring Fairness and Trust in Africa' (2024) 9 Frontiers in Research Metrics and Analytics <https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2024.1486600/full>.

veloped world.[11] A key issue is the tendency of foreign companies, and even local contractors, to extract data from local populations at little or no cost, only for the resulting technology systems to be sold back to these communities at a price.[12] This "digital colonisation" presents itself in Africa in various forms, such as cheap digital labour, data extraction, and Africans being relegated to the roles of beta testers for global platforms.[13] The challenge of social media data is particularly relevant. Since the rise of platforms like Facebook, Twitter, Instagram, and TikTok in Kenya, Kenyans have built a thriving digital ecosystem encompassing businesses, human progress, problem-solving, innovation, cultural expression, and dynamic social interactions, making their data invaluable for AI training. Recently, a Meta official confirmed that the company has scraped all public data posted on its platforms since 2007 for training AI, and only users in the European Union are granted an opportunity to opt out of this violation.[14] Beyond social media data, AI systems have also been trained on vast datasets from books, music, and global news archives, often without the creators' consent.[15] This widespread use of data raises serious ethical concerns around intellectual property rights, exploitative data extraction, and the reinforcement of biases in AI models. Ethical AI development depends on inclusive, well-regulated data governance mechanisms that ensure representativeness and accountability and foster responsible innovation.[16]

In response to these ethical challenges, various frameworks have emerged globally to guide AI development and deployment. Key among these is the European Union's AI Act, which emphasises a human-centric approach, categorising AI systems based on risk levels and imposing stringent requirements for high-risk applications.[17] The UNESCO Recommendation on the Ethics of AI has been instrumental in informing global governance frameworks for AI, by espousing principles such as fairness and non-discrimination, safety, protection of privacy, transparency and explainability, and responsibility and accountability.[18] At the continental level, the African Union's Continental AI Strategy (Adopted in July 2024) outlines a vision for responsible, ethical, and inclusive AI development in Africa.[19] This strategy is complemented by a key document concerning data governance, the AU Data Policy Framework. It provides a blueprint for data governance, emphasising data sovereignty, protection, and the need for a balanced approach to data access and use.[20] The Framework champions ethical AI by mandating a people-centric approach to data, emphasising ethical

11 Benedikt Erforth, 'Data Extraction, Data Governance and Africa- Europe Cooperation: A Research Agenda' (2024)<https://www.megatrends-afrika.de/assets/afrika/publications/MTA_working_paper/MTA_WP14_Erforth_Digital_Cooperation.pdf>.

12 Nima Elmi, 'Is Big Tech Setting Africa Back?' (Foreign Policy11 November 2020) <https://foreignpolicy.com/2020/11/11/is-big-tech-setting-africa-back/>.

13 Benedikt Erforth, 'Data Extraction, Data Governance and Africa- Europe Cooperation: A Research Agenda' (2024)

14 ibid.

15 Adil S Al-Busaidi and others, 'Redefining Boundaries in Innovation and Knowledge Domains: Investigating the Impact of Generative Artificial Intelligence on Copyright and Intellectual Property Rights' (2024) 9 Journal of Innovation & Knowledge 100630 <https://www.sciencedirect.com/science/article/pii/S2444569X24001690#:~:text=As%20previously%20noted%2C%20GenAI%20technologies,explicit%20permission%20of%20copyright%20holders.>.

16 Oakley Parker, 'Data Governance and Ethical AI: Developing Legal Frameworks to Address Algorithmic Bias and Discrimination' <https://www.researchgate.net/publication/384966994_Data_Governance_and_Ethical_AI_Developing_Legal_Frameworks_to_Address_Algorithmic_Bias_and_Discrimination>

17 European Parliament, 'European Parliament P9_TA(2024)0138 Artificial Intelligence Act European Parliament Legislative Resolution of 13 March 2024 on the Proposal for a Regulation of the European Parliament and of the Council on Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts (COM(2021)0206 -C9-0146/2021 -2021/0106(COD)) (Ordinary Legislative Procedure: First Reading)' (2024) <https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf>.

18 UNESCO, 'Recommendation on the Ethics of Artificial Intelligence | UNESCO' (www.unesco.org16 May 2023) <https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence>.

19 African Union, 'Continental Artificial Intelligence Strategy Harnessing AI for Africa's Development and Prosperity' (2024) <https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf>.

20 'AU Data Policy Framework_An Integrated, Prosperous and Peaceful Africa' (2022) <https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf>.

sourcing through consent, purpose limitation, and data justice to prevent bias and discrimination.[21] For ethical sharing, it promotes trustworthy, secure, and accountable data environments with robust safeguards, advocating for openness, interoperability, and data quality, while suggesting mechanisms like data trusts and common categorisation frameworks.[22] These principles collectively aim to ensure that data, the foundation of AI, is managed to uphold human rights, foster equitable benefit sharing, and support the responsible development of AI across Africa. Building on these developments, approximately 16 African nations have by the time of publishing this report, established dedicated AI policies, strategies, or frameworks.[23] This signifies a growing recognition across the continent of the importance of establishing robust AI strategic paths and governance structures.

The recently endorsed (June 2025) Hamburg Declaration on Responsible AI for Sustainable Development Goals (SDGs) is also a key governance milestone. The Declaration emphasises the significant role of AI in advancing the SDGs, despite potential risks such as widening inequalities and human rights violations, and calls for a human-centric, human-rights-based, inclusive, open, sustainable, and responsible AI future.[24] It also emphasises the need to engage meaningfully with all stakeholders, especially those from emerging markets and developing economies, such as African countries. The declaration is built upon five pillars aligned with the 2030 Agenda for Sustainable Development:

- *People*, hence a commitment to designing, developing, and deploying AI systems that respect human rights, gender equality, and inclusion, particularly benefiting emerging markets and developing economies;[25]
- *Planet*, aiming for AI usage aligned with global climate and environmental goals, emphasising energy efficiency, reduced carbon footprints, and the use of sustainable energy in AI infrastructure;[26]
- *Prosperity*, prioritising AI systems that support inclusive economic and social development by fostering local innovation and minimizing economic divides;[27]
- *Peace*, committing to ensuring AI systems are not misused to disrupt peace or undermine governance, particularly through information manipulation, and;[28]
- *Partnerships*, strengthening multi-stakeholder international partnerships to create a globally inclusive AI ecosystem that upholds equity and shared responsibility.[29]

This basis of achieving the Sustainable Development Goals by leveraging on Responsible AI is uniquely positioned to help African countries align AI adoption with local development needs tailored to African populations.

Kenya, recognising both the potential and inherent risks of AI, is proactively developing its AI ecosystem with a strong ethical emphasis. The Kenya AI Strategy 2025-2030 outlines a commitment to "ethical and responsible AI" and emphasises the importance of "data sovereignty and ethical AI practices to build a technological future that is safe, accounta-

21 ibid.

22 ibid.

23 'Africa Technology Policy Tracker' (Carnegie Endowment for International Peace2024) <https://carnegieendowment.org/features/africa-digital-regulations?lang=en> accessed 28 July 2025.

24 Federal Ministry for Economic Cooperation and Development, Germany (BMZ), 'Hamburg Declaration on Responsible Artificial Intelligence (AI) for the Sustainable Development Goals (SDGs)' (2025) <https://www.bmz-digital.global/wp-content/uploads/2025/06/250603_Hamburg_Declaration.pdf> accessed 7 July 2025.

25 ibid.

26 ibid.

27 ibid.

28 ibid.

29 ibid.

ble, and beneficial for all Kenyans."[30] This commitment is further emphasised in Kenya's AI Playbook for Diplomats, which guides the nation's representatives in navigating the ethical dimensions of AI on the global stage.[31] It is therefore clear that by embedding ethical considerations into its national AI strategy and other frameworks, Kenya aims to cultivate an AI ecosystem that is innovative, trustworthy, and aligned with its societal values.

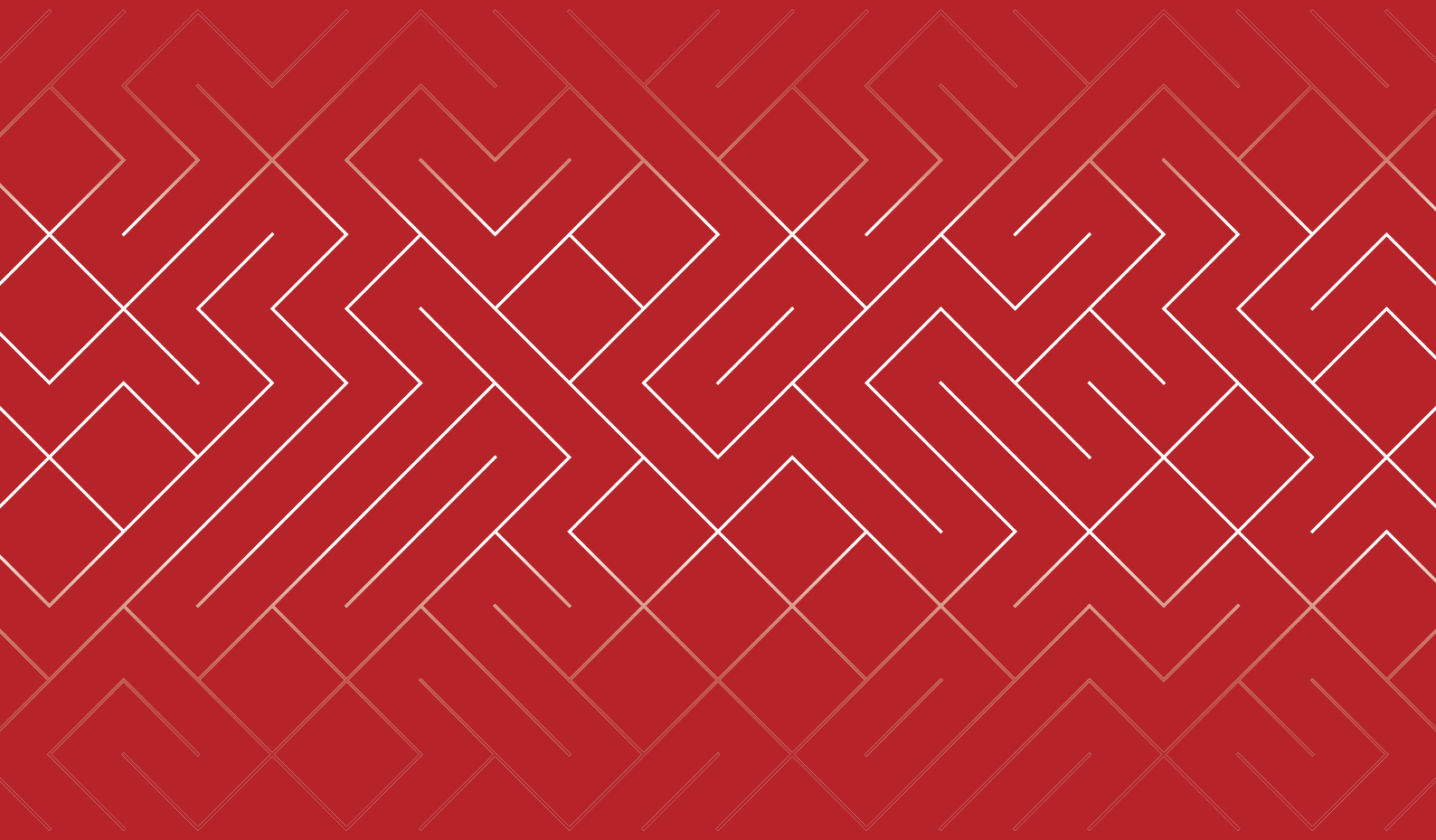
This research project addresses the pressing concern of how unethical data sourcing directly contributes to unethical AI. The report begins by comprehensively examining how unethical data sourcing has manifested in various global and local contexts. Following this analysis of existing problems, it then establishes what constitutes ethical data sourcing, providing a benchmark for ideal practices and the place of AI governance in addressing ethical data sourcing. This foundational understanding is subsequently illuminated through an in-depth analysis of three case studies exploring real-world examples of problematic and ethical data sourcing approaches. Through this multi-faceted examination, the study seeks to articulate actionable strategies that promote ethical AI in Africa by championing ethical data sourcing and effective governance.


Image Source: vecteezy.com

30 'Kenya AI Strategy 2025-2030' (The Ministry of Information, Communications and The Digital Economy (MICDE) 2025) <https://ict.go.ke/sites/default/files/2025-03/Kenya%20AI%20Strategy%202025%20-%202030.pdf>.

31 'Diplomat's Playbook on Artificial Intelligence' (Ministry of Foreign Affairs 2025) <https://mfa.go.ke/sites/default/files/2025-01/DIPLOMATS%20AI%20PLAYBOOK%20FINAL.pdf>.

II. How Unethical Data Sourcing Manifests



OpenAI, arguably the most recognisable AI company in the world, has faced its latest battle over how it sourced data to train the GPT models. The company is now entangled in a copyright lawsuit filed by Indian book publishers, joining a growing list of legal challenges from authors, news agencies, and musicians worldwide.[32] These cases highlight ongoing concerns over AI models training on proprietary content without authorisation, threatening intellectual property rights and undermining traditional revenue streams. The lawsuit in India, led by the Federation of Indian Publishers, demands that OpenAI either secure proper licensing agreements or delete datasets containing copyrighted material, reflecting broader global anxieties about the ethical sourcing of AI training data.[33]

Other global companies that have faced this challenge include Stability AI, sued by Getty Images for scraping millions of copyrighted photos; Anthropic, targeted by music publishers over song lyrics; and Microsoft, dragged into OpenAI-related suits due to its partnership, all underscoring a mounting tension between AI innovation and the rights of content creators across industries and borders.[34] Stability AI's legal woes began in February 2023 when Getty Images accused it of using over 12 million images to train its Stable Diffusion model, claiming this unauthorised use fueled a generative AI that competes with Getty's stock photo business.[35] Anthropic, founded by ex-OpenAI researchers, faced a lawsuit in October 2023 from Universal Music Group and others, who alleged that its Claude model was trained on copyrighted lyrics, and that it often produces almost verbatim reproductions of copyrighted materials, without proper licensing.[36] Meanwhile, companies like Nvidia[37] were hit with a 2024 authors' lawsuit over pirated books. Perplexity AI, sued by Dow Jones and the New York Post[38] for repurposing news articles, further illustrate how the AI industry has relied on vast datasets sourced from what they term "publicly available information", which is constantly clashing with established intellectual property frameworks. These developments have raised significant questions about fair use, the sourcing of training data for AI development, the future of the creative industry, and intellectual property considerations.

Like much of Africa, Kenya is experiencing extensive data extraction, often in ways that raise ethical concerns. Research by CIPIT on the Intellectual Property dilemma in sourcing AI training data, researchers found that AI developers require vast amounts of data to train models, and much of this data is sourced through methods like web crawling and scraping.[39] The authors note that while some datasets, such as the Demographic and Health Survey (DHS) and the Malaria Indicator Survey (MIS), are publicly available, others, like detailed reports from the Senegal National Malaria Control Program, may be copyrighted or sensitive but still end up being used without proper authorisation.[40] Many African coun-

32 Aditya Kalra, Arpan Chaturvedi and Munsif Vengattil, 'OpenAI Faces New Copyright Case, from Global Book Publishers in India' Reuters (24 January 2025) <https://www.reuters.com/technology/artificial-intelligence/openai-faces-new-copyright-case-global-publishers-india-2025-01-24/>.

33 ibid.

34 Bruce Barcott, 'AI Lawsuits Worth Watching: A Curated Guide | TechPolicy.Press' (Tech Policy Press1 July 2024) <https://www.techpolicy.press/ai-lawsuits-worth-watching-a-curated-guide/>.

35 Bruce Barcott, 'AI Lawsuits Worth Watching: A Curated Guide | TechPolicy.Press' (Tech Policy Press1 July 2024)

36 ibid.

37 Mark Hill and Courtney Benard, 'Nvidia Faces Class-Action Lawsuit for Training AI Model on "Shadow Library"' (Lexology 30 April 2024) <https://www.lexology.com/library/detail.aspx?g=3a665ce3-3db6-40a3-899e-10c2cf606a71> accessed 19 March 2025.

38 Dawn Chmielewski and Katie Paul, 'Murdoch's Dow Jones, New York Post Sue Perplexity AI for "Illegal" Copying of Content' Reuters (21 October 2024) <https://www.reuters.com/legal/murdoch-firms-dow-jones-new-york-post-sue-perplexity-ai-2024-10-21/>.

39 Natasha Karanja and Chebet Koros, 'Artificial Intelligence (AI) Training Data and the Copyright Dilemma: Insights for African Developers - Centre for Intellectual Property and Information Technology law12 February 2025) <https://cipit.strathmore.edu/artificial-intelligence-ai-training-data-and-the-copyright-dilemma-insights-for-african-developers/>.

40 Natasha Karanja and Chebet Koros, 'Artificial Intelligence (AI) Training Data and the Copyright Dilemma: Insights for African Developers - Centre for Intellectual Property and Information Technology law12 February 2025) <https://cipit.strathmore.edu/artificial-intelligence-ai-training-data-and-the-copyright-dilemma-insights-for-african-developers/>.

tries, including Kenya, have copyright laws that do not explicitly address AI training, leaving room for exploitation. Kenya's Copyright Act allows for fair dealing in scientific research but does not clarify whether AI model training falls under this exception.[41] In contrast, South Africa's Copyright Amendment Bill attempts to define fair use,[42] though its impact on AI development remains uncertain.

This lack of clear legal and ethical frameworks means AI developers often extract data without consent or transparency, which undermines privacy and erodes trust in AI systems. This unchecked data sourcing leads to unethical AI systems, trained on data acquired through questionable means, violating individual rights and reinforcing existing societal inequities.[43] Without stronger data governance, the continued extraction of African data to train AI will likely continue to be based on a foundation of exploitation rather than fairness and accountability.

Further exploitation in data sourcing has positioned Africa as a low-wage data annotation hub, where global tech companies outsource AI training tasks. In Kenya, young, unemployed workers, often desperate for economic opportunities, are hired by companies like Meta and OpenAI through third-party firms such as Sama to label vast datasets - images, text, and videos - for minimal pay, sometimes as low as $2 per hour. This work frequently exposes them to disturbing and harmful content, including violence and explicit material, to refine AI algorithms used worldwide.[44] The unethical nature of this sourcing is evident in the economic vulnerability it exploits, offering little job security or mental health support despite the psychological toll.[45] Global companies bypass stricter labour laws in wealthier nations, exploiting Kenya's weaker regulations.[46] Workers have brought legal action against Meta and Sama, citing long hours, precarious contracts, and a stark power imbalance that limits their rights and avenues for redress.[47] Allegations of retaliation against workers attempting to unionise further highlight the exploitative conditions.[48] This case exemplifies the digitisation of global inequality, where countries like Kenya serve as low-wage labour centres for AI advancements,[49] yet see little of the economic benefits despite their critical role in fueling tools like OpenAI's ChatGPT [50]

Kenya is taking steps to curb the longstanding exploitation of its citizens' data for AI training by foreign entities. The Ministry of Information, Communications, and the Digital Economy (MICDE) has acknowledged Kenya's vulnerability to data exploitation, where international companies harvest local data with little to no benefits returning to the population.[51] To counter this, the government has emphasised national data sovereignty through the Kenya AI Strategy, the recently released Diplomat's Playbook on Artificial Intelligence, and pro-

41 ibid.

42 Copyright Amendment Bill (South Africa) <https://www.parliament.gov.za/storage/app/media/uploaded-files/Copyright%20Amendment%20Bill%20Draft.pdf >

43 Natasha Karanja and Chebet Koros, 'Artificial Intelligence (AI) Training Data and the Copyright Dilemma: Insights for African Developers - Centre for Intellectual Property and Information Technology law12 February 2025)

44 Billy Perrigo, 'Exclusive: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic' (Time18 January 2023) <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

45 ibid.

46 Raksha Vasudevan, 'A Lawsuit against Meta Shows the Emptiness of Social Enterprises' (Wired 20 July 2022) <https://www.wired.com/story/social-enterprise-technology-africa/>

47 Business and Human Rights Resource Centre, 'Meta & Sama Lawsuit (Re Poor Working Conditions & Human Trafficking, Kenya) - Business & Human Rights Resource Centre' (Business & Human Rights Resource Centre 2022) <https://www.business-humanrights.org/fr/latest-news/meta-sama-lawsuit-re-poor-working-conditions-human-trafficking-kenya/>

48 Business and Human Rights Resource Centre, 'Meta & Sama Lawsuit (Re Poor Working Conditions & Human Trafficking, Kenya) - Business & Human Rights Resource Centre' (Business & Human Rights Resource Centre 2022)

49 'WeeTracker' (WeeTracker25 November 2024) <https://weetracker.com/2024/11/25/openai-sama-kenyan-workers-controversy/> accessed 9 April 2025.

50 Billy Perrigo, 'Exclusive: OpenAI Used Kenyan Workers on Less than $2 per Hour to Make ChatGPT Less Toxic' (Time18 January 2023) <https://time.com/6247678/openai-chatgpt-kenya-workers/>.

51 'Kenya to Restrict Use of Locals' Data for Foreign AI Training' (The East African21 January 2025) <https://www.theeastafrican.co.ke/tea/sustainability/innovation/kenya-to-restrict-use-of-locals-data-for-foreign-ai-training-4896508> .

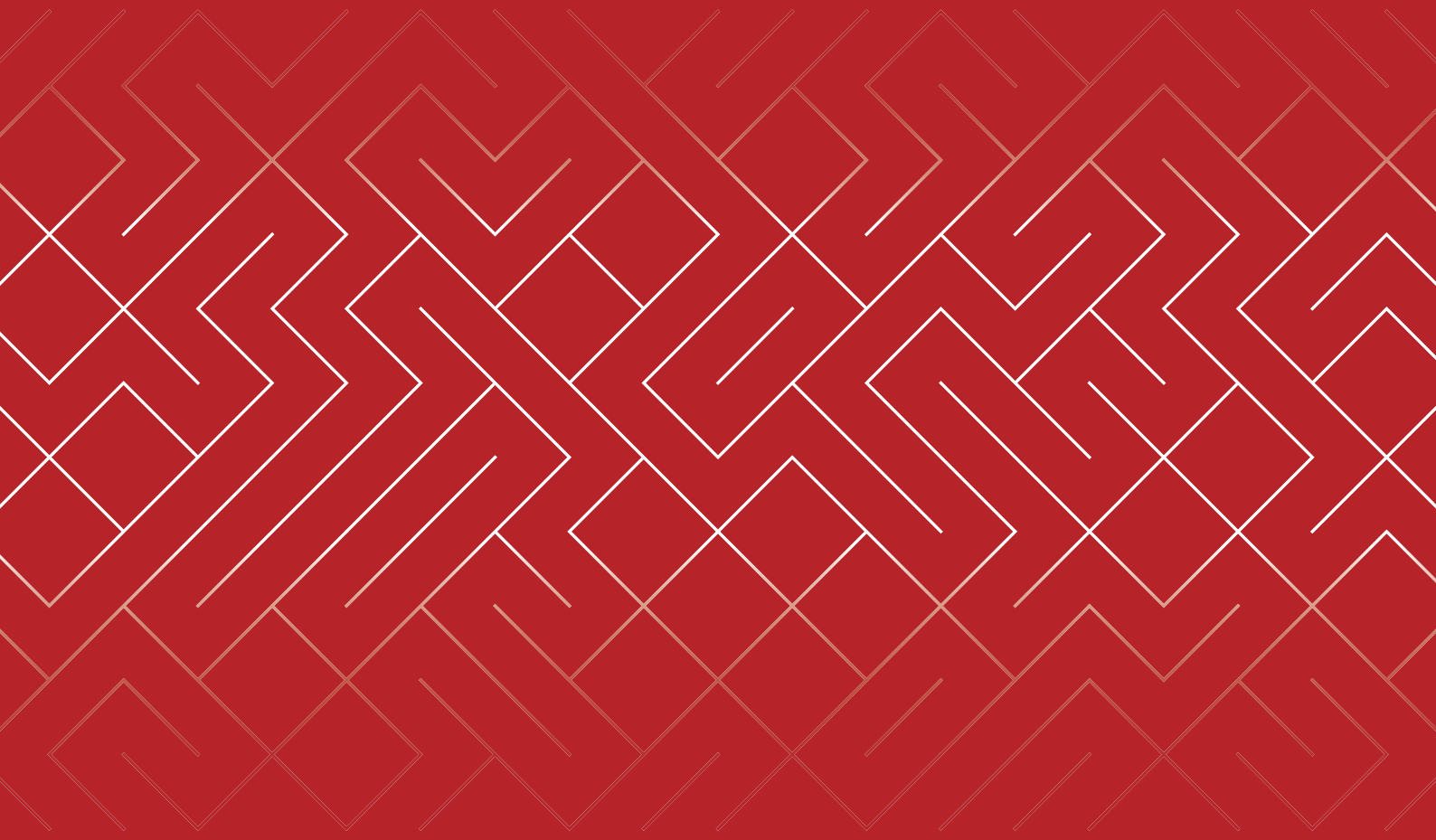
posed[52] data governance legislation. These initiatives aim to halt exploitative data practices and ensure that AI development aligns with ethical standards while empowering Kenyans rather than serving solely foreign interests.

Beyond national policy efforts, equitable data practices must also extend to direct benefit-sharing with affected communities. The implementation of Benefit Sharing Agreements can provide a structured way for African communities to receive fair compensation for their data contributions, as exemplified by the Masakhane *Pelargonium* case. Here, the Masakhane community in South Africa disputed their inclusion under a traditional authority's Benefit Sharing Agreements (BSA) with Schwabe Pharmaceuticals, demanding a separate BSA to secure benefits from the harvesting of the *Pelargonium sidoides* and *Pelargonium reniforme* plants on their land. This effort was rooted in their assertion of self-representation through a Communal Property Association.[53] This case exemplifies how unethical data sourcing manifests, where the Masakhane community's data was initially included under a traditional authority's Benefit Sharing Agreement (BSA) with a pharmaceutical company without their direct consent, undermining their agency. This case underscores how superficial consultations can erode community self-representation and limit their ability to secure fair benefits, highlighting the broader impact of exploitative data practices on community autonomy and rights.

52 'Report of the Information, Communications and the Digital Economy Sectoral Working Group Republic of Kenya Ministry of Information, Communications and the Digital Economy' (2024) <https://ict.go.ke/sites/default/files/2024-09/MICDE%20Sector%20Working%20Group%20Report%20-%20June%202024.pdf> .

53 Zuziwe Msomi and Sally Matthews, 'Protecting Indigenous Knowledge Using Intellectual Property Rights Law: The Masakhane Pelargonium Case' (2016) 45 Africanus: Journal of Development Studies 62 <https://unisapressjournals.co.za/index.php/Africanus/article/download/645/432/4917>.

III. Ethical Artificial Intelligence (AI) through Ethical Data Sourcing



Ethical AI governance frameworks in Africa, although still in development, have demonstrated a clear intention to prioritise Afro-centric values and a human-rights-based approach.[54] As such, the ethical development and deployment of AI systems must consider the local contexts of African populations that these systems intend to serve, while respecting people's rights of privacy, non-discrimination, human dignity and equitable benefit-sharing.[55] The early stages of the AI development lifecycle, as discussed, heavily rely on data collection, selection, annotation and validation, which must adhere to these priorities of achieving ethical AI development and deployment. This places ethical data sourcing as a cornerstone of responsible innovation, by ensuring that development in Kenya is anchored in robust AI governance frameworks that prioritise African-centeredness and a human rights-based approach. By centering African perspectives, as championed by frameworks like the African Union's Continental AI Strategy and the Malabo Convention, AI governance integrates principles of data sovereignty, informed consent, and fairness to address historical inequities and prevent exploitative data practices. The protection of the right to privacy is also reflected in *Article 31 of the Kenyan Constitution*, which is operationalised by the Data Protection Act, 2019. Adhering to the principles espoused by these existing governance frameworks sets a strong foundation for the responsible governance of data across the AI lifecycle.

Ethical data sourcing is fundamental to developing trustworthy AI systems, integrating data privacy, informed consent, inclusiveness, and digital rights within robust data governance frameworks.[56] The performance and legitimacy of AI systems rely on the quality and ethical integrity of their data sources, a principle increasingly codified in global legal and regulatory frameworks.[57] Central to ethical data sourcing is informed consent, which the Malabo Convention defines as data processing based on the "principle of consent and legitimacy," ensuring individuals explicitly agree to data use with exceptions only when legally permitted, grounding AI in African values of community agency and dignity.[58] Kenya's Data Protection Act reinforces this by requiring personal data processing to occur only with the data subject's express, unequivocal, free, specific, and informed consent, prioritising local empowerment and data sovereignty.[59] The GDPR provides a broader context, defining informed consent as a freely given, specific, and unambiguous acceptance of data processing, signalled through clear statements or actions, ensuring individuals fully understand data usage.[60] These overlapping regulations highlight a universal commitment to empowering data subjects, though implementation varies across jurisdictions. Core ethical considerations such as transparency in AI development, robust data protection, and fairness and non-discrimination, reinforce the integrity of AI systems, fostering trust and accountability.

Informed consent is critical to ethical data sourcing and ethical AI development, where expansive datasets shape model behaviour and societal outcomes.[61] In the context of personal data, the Kenyan law defines consent

54 'AI and the Future of Work in Africa: White Paper' (2024) <https://aspyee.org/sites/default/files/2024-06/AIandTheFutureofWorkinAfricaWhitePaper-June2024-666aa97e6eb8d.pdf>.

55 Damian Okaibedi Eke and others, 'African Perspectives of Trustworthy AI: An Introduction' [2025] Trustworthy AI 1 <https://link.springer.com/chapter/10.1007/978-3-031-75674-0_1>.

56 Swetha Sistla, 'AI with Integrity: The Necessity of Responsible AI Governance' (2024) Journal of Artificial Intelligence & Cloud Computing SRC/JAICC-E180 https://doi.org/10.47363/JAICC/2024(3)E180 accessed 24 January 2025.

57 Morgan Sullivan, 'Key Principles for Ethical AI Development' (20 October 2023) Transcend Blog https://transcend.io/blog/ai-ethics accessed 24 January 2025.

58 African Convention on Cyber Security and Personal Data Protection, Article 13

59 Data Protection Act, s 30

60 Article 4(11) General Data Protection Regulation <https://gdpr-info.eu/> accessed 24 January 2025.

61 Petar Radanliev, 'AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development' (2025) 39 Applied Artificial Intelligence <https://www.tandfonline.com/doi/full/10.1080/08839514.2025.2463722>.

to mean "express, unequivocal, free, specific and informed" sign that the owner of the data agrees for their data to be processed.[62] Informed consent thus requires transparent disclosure of the purpose of data collection,[63] such as training algorithms, alongside clear explanations of storage, usage, and risks like unintended profiling or data leaks. When assessing the validity of consent, it is crucial to consider the socioeconomic context in which Africans may feel compelled to provide personal data, such as biometric information, in exchange for modest economic incentives, as this can compromise the voluntariness required for ethical data collection.[64] This complexity in determining consent was explored in *Republic v Tools for Humanity Corporation (US) & 8 others,* where the Kenyan High Court ruled that Worldcoin's biometric data collection violated the Data Protection Act. The court found that offering cryptocurrency tokens worth about seven thousand Kenyan Shillings in exchange for iris scan data constituted inducement, invalidating consent.[65] While this case concerns personal data, the broader trend of offering minimal or no compensation for any data undermines genuine consent by exploiting economic vulnerabilities, particularly in marginalised communities. Thus, taking consent to mean an "express, unequivocal, free, specific, and informed" indication that the data owner agrees to their data being processed in the specified ways,[66] would render the consent granted invalid. Therefore, ethical data sourcing for AI must prioritise strong, transparent consent mechanisms to ensure voluntariness, fostering trust and safeguarding societal outcomes. Furthermore, compensation models should avoid taking advantage of communities' economic vulnerability, and prioritize fairness and community empowerment. This form of AI development process has been witnessed on the continent, such as with Kenya's AI Early Warning System developed by the Local Development Research Institute (LDRI). This project collects hyperlocal geo-referenced data from farmers to monitor agricultural activities, with the algorithm and initial dataset made open source for public access and use, while farmers receive access to climate-smart agricultural advice and digital tools, improving yields.[67] Such a practice exemplifies how equitable benefit-sharing ensures ethical data sourcing by fostering trust and genuine consent from data sources.

Achieving ethical AI also critically involves bias mitigation, as bias in AI systems is a significant concern. Bias originates from unrepresentative datasets, flawed algorithm design, or skewed user interactions, which can lead to discriminatory outcomes in machine learning models.[68] Ethical data sourcing is crucial to mitigate these biases, prioritising the collection and use of diverse, high-quality, and representative datasets that accurately reflect varied populations and demographic groups.[69] Furthermore, establishing strong user feedback mechanisms allows for ongoing monitoring and correction of biases that may emerge post-deployment.[70] Open source data science approaches can also contrib-

62 The Data Protection Act, section 2

63 Adam J Andreotta, Nin Kirkham and Marco Rizzi, 'AI, Big Data, and the Future of Consent' (2021) 37 AI & SOCIETY <https://link.springer.com/article/10.1007/s00146-021-01262-5>.

64 Carlos Mureithi, 'Five Things to Learn from Kenya's Inquiry into Worldcoin's Activities in the Country' (Tech Policy Press24 April 2024) <https://www.techpolicy.press/five-things-to-learn-from-kenyas-inquiry-into-worldcoins-activities-in-the-country/>

65 Republic v Tools for Humanity Corporation (US) & 8 others; Katiba Institute & 4 others (Exparte Applicants); Data Privacy & Governance Society of Kenya (Interested Party) (Judicial Review Application E119 of 2023) [2025] KEHC 5629 (KLR) (Judicial Review) (5 May 2025) (Judgment)

66 Data Protection Act, s 2

67 GIZ, 'Harnessing the Power of AI to Improve Harvests' (www.giz.de 28 May 2025) <https://www.giz.de/en/mediacenter/better-harvest-with-AI.html>.

68 Athena Vakali and Nicoleta Tantalaki, 'Rolling in the Deep of Cognitive and AI Biases' [2019] Arxiv.org <https://arxiv.org/html/2407.21202v1>.

69 George Benneh Mensah, 'Artificial Intelligence and Ethics: A Comprehensive Review of Bias Mitigation, Transparency, and Accountability in AI Systems' [2023] ResearchGate <https://www.researchgate.net/publication/375744287_Artificial_Intelligence_and_Ethics_A_Comprehensive_Review_of_Bias_Mitigation_Transparency_and_Accountability_in_AI_Systems>.

70 Matthew Hanna and others, 'Ethical and Bias Considerations in Artificial Intelligence/Machine Learning' (2024) 38 Modern Pathology 1 <https://www.sciencedirect.com/science/article/pii/S0893395224002667>.

ute by fostering collaboration, allowing users to access datasets, identifying data bugs like missing data for underrepresented minorities, and collectively working to fix datasets through crowdsourcing or submitting appropriate data points.[71] AI development and deployment can significantly mitigate bias and create more accountable, trustworthy, and equitable AI systems by systematically embedding these legal and ethical principles into the data sourcing process.

71 Abby Seneor and Matteo Mezzanotte, 'Open-Source Data Science: How to Reduce Bias in AI' (World Economic Forum14 October 2022) <https://www.weforum.org/stories/2022/10/open-source-data-science-bias-more-ethical-ai-technology/>.

Image Source: vecteezy.com

IV. Case Studies

Having contextualised how unethical data sourcing manifests and described the prerequisites for ethical data practices, the report presents case studies of Techworker Community Africa (TCA), Masakhane, and FAIR Forward's Common Voices Project. These projects, operating primarily within African contexts, exemplify concerted efforts to counteract the previously identified challenges by embedding ethical considerations into the AI data lifecycle. By exploring their goals, how they work, and who is involved, these cases offer real examples of how principles such as informed consent, bias mitigation, equitable compensation, and community empowerment are implemented. This would highlight actionable strategies and potential models for fostering ethical AI development grounded in responsible data handling across Africa.

a. Techworker Community Africa (TCA): Advancing Ethical AI by Embedding Ethics in Data Annotation, Processing, and Model Training

Techworker Community Africa (TCA), a Kenya-based organisation with over 4,000 members, is at the forefront of fostering an ethical tech ecosystem in Africa. Operating as both a non-governmental organisation and a data agency, TCA empowers tech workers through advocacy, training, and community-building, particularly in AI and content moderation.[72] By ensuring data is governed responsibly from collection through processing to use in training AI models, TCA addresses biases, protects worker rights, and promotes equitable practices, offering a model for ethical AI development in Africa and beyond. The lawsuit against Meta and its main subcontractor, Sama, in Kenya brought to light allegations of severe worker exploitation, including poor working conditions, inadequate mental health support, and union busting affecting content moderators.[73] Such events, alongside a broader awareness of systemic issues in the AI labour supply chain, underscore the paramount need for organisations like Techworker Community Africa to champion ethical practices, consequently upholding tech workers' human rights and psychological well-being while safeguarding AI users from harms like disinformation, bias, and exploitation.[74]

With a stakeholder composition of tech workers, content moderators, and community members across Africa, working to amplify tech worker voices, and partnerships with global organisations, the stakeholder pool collaborates to promote ethical data practices through worker training on rights advocacy and community engagement.[75] This is a key demonstration of how purpose-driven, human-centred approaches can inform ethical data sourcing for AI, by addressing why ethical data governance is essential, how it can be achieved through training and advocacy, and who drives these efforts, aiming for data to be responsibly managed from collection to model training.

b. Masakhane: Addressing Linguistic Exclusion and Biases in AI through Community-Driven, Ethical Data Creation

Masakhane, a grassroots African NLP community, is continually revolutionising natural language processing (NLP) by creating inclusive, locally relevant datasets for African languages. With over 2,000 contributors across 35+ Afri-

72 Techworker Community Africa, 'Techworker Community Africa' (Techworker Community Africa2025) <https://www.techworkercommunityafrica.com/> accessed 11 May 2025.

73 Business and Human Rights Resource Centre, 'Meta & Sama Lawsuit (Re Poor Working Conditions & Human Trafficking, Kenya) - Business & Human Rights Resource Centre' (Business & Human Rights Resource Centre 2022)

74 Techworker Community Africa, 'Techworker Community Africa' (Techworker Community Africa2025) <https://www.techworkercommunityafrica.com/> accessed 11 May 2025.

75 Techworker Community Africa, 'Techworker Community Africa' (Techworker Community Africa2025) <https://www.techworkercommunityafrica.com/> accessed 11 May 2025.

can countries, Masakhane fosters participatory research to address the underrepresentation of African languages in AI.[76] The collective prioritises community-driven data collection and open-source contributions to ensure culturally sensitive, representative datasets, promoting equitable AI development for African contexts.[77] As such, Masakhane's mission is to advance NLP for African languages and tackle the systemic exclusion of African users in AI, where major AI platforms fail to recognise a majority of the over 2,000 African languages.[78]

By building datasets like MasakhaNER, a named entity recognition dataset for 10 African languages, Masakhane ensures AI serves local needs while preserving linguistic heritage. The MasakhaNER dataset was developed through collaborative workshops, where contributors translated and annotated text in languages like Swahili, Yoruba, and Hausa.[79] Masakhane develops their datasets through community-driven data collection methods involving collaborative workshops, where contributors, including native speakers, translate and annotate text in languages such as Swahili, Yoruba, and Hausa, ensuring cultural and linguistic accuracy.[80] This method adheres to ethical data collection principles, including informed consent, where participants voluntarily contribute with clear understanding of the dataset's purpose for AI development, and inclusivity, ensuring diverse representation of African languages to mitigate biases. Data collection parameters prioritise high-quality, real-world language data, validated by community members to reflect authentic usage and minimise errors, aligning with transparency and fairness.[81] Masakhane also promotes transparency by making datasets publicly available on open-source platforms, fostering trust and enabling global researchers to build on its work.[82] In that regard, Masakhane demonstrates how community-driven, ethical data creation can transform NLP for underrepresented languages. Consequently, they address why inclusive datasets matter, how they are built through participatory methods, and who can continually drive these efforts to pave the way for a more inclusive digital future, where African languages are fully represented in AI systems.

c. *The Common Voices Project: Contributing to Ethical AI through Participatory and Open-Source Data Creation of African Language Datasets*

Mozilla Foundation's Common Voice initiative is a flagship global project aimed at democratising voice technology and mitigating bias in AI.[83] It addresses the critical issue that most existing voice datasets are proprietary, which stifles innovation, and significantly under-represents the vast majority of the world's languages and demographic communities.[84] By mobilising people everywhere to contribute their voices, Common Voice is building the world's most diverse, open-source voice data-

76 MasakhaNER Know, 'Masakhane - MasakhaNER: Know Our Names' (Masakhane.io2020) <https://www.masakhane.io/ongoing-projects/masakhaner-know-our-names>.

77 David Ifeoluwa Adelani and others, 'MasakhaNER: Named Entity Recognition for African Languages' (arXiv.org2021) <https://arxiv.org/abs/2103.11811> accessed 14 May 2025.

78 Andrew Paul, 'AI Programs Often Exclude African Languages. These Researchers Have a Plan to Fix That.' (Popular Science11 August 2023) <https://www.popsci.com/technology/african-language-ai-bias/>.

79 masakhane-io, 'GitHub - Masakhane-Io/Masakhane-Ner' (GitHub2020) <https://github.com/masakhane-io/masakhane-ner> accessed 14 May 2025.

80 Masakhane - Organisation Strategy, 'Masakhane - Organisation Strategy' (Masakhane.io2020) <https://www.masakhane.io/organisation-strategy> accessed 10 July 2025.

81 Doyin Akindotuni, 'Resource Asymmetry in Multilingual NLP: A Comprehensive Review and Critique' (2025) 13 Journal of Computer and Communications 14 <https://www.scirp.org/journal/paperinformation?paperid=143854>.

82 Rachael Wambua, 'How Open Source Is Powering the Future of African NLP - Lanfrica Blog' (Lanfrica Blog8 May 2025) <https://lanfrica.com/blog/how-open-source-is-powering-the-future-of-african-nlp/> accessed 10 July 2025.

83 'Common Voice' (Mozilla Foundation2025) <https://www.mozillafoundation.org/en/common-voice/> accessed 14 May 2025.

84 'Platform and Dataset' (Mozilla Foundation2025) <https://www.mozillafoundation.org/en/common-voice/platform-and-dataset/> accessed 14 May 2025.

set, making valuable data available to developers, researchers, and communities to foster more inclusive speech technology.[85] This open approach is crucial for ensuring that the future of voice technology serves every region, not just speakers of dominant languages.

In Africa, the Common Voice extends through the FAIR Forward initiative, which is advancing ethical AI by creating community-driven open voice data and technology for East African languages, including Kinyarwanda, Kiswahili, and Luganda.[86] This initiative exemplifies ethical data sourcing by prioritising informed consent, inclusivity, transparency, data sovereignty, benefit sharing, and bias mitigation. Native speakers, local startups, and researchers are engaged through participatory workshops to record, annotate, and validate speech data, ensuring cultural relevance and minimising bias.[87] For example, the collection of over 2,000 hours of Kinyarwanda speech data was implemented with partners like Rwanda's Umuganda.[88] Informed consent is ensured by engaging native speakers, who voluntarily contribute speech data. This is, at the same time, a demonstration of data sovereignty of Africans, as communities retain control over their linguistic data. The resulting datasets are released on open-source platforms, fostering transparency and global reuse. Furthermore, these efforts include training in data annotation and AI development to build local capacity and ensure sustainable data governance. This model, prioritising community trust, data sovereignty, and alignment with ethical frameworks, demonstrates a scalable approach to making AI more inclusive and equitable for underrepresented languages in Africa and beyond.

85 'Common Voice' (Mozilla Foundation2025) <https://www.mozillafoundation.org/en/common-voice/> accessed 14 May 2025.

86 Jan Krewer, 'Creating Community-Driven Datasets: Insights from Mozilla Common Voice Activities in East Africa' (2023) <https://www.bmz-digital.global/wp-content/uploads/2023/03/Creating-Community-Driven-Datasets-Report-032023-GIZ-Mozilla.pdf>.

87 common-voice-kiswahili-awards, 'Common-Voice-Kiswahili-Awards' (Mozilla Foundation2019) <https://www.mozillafoundation.org/en/what-we-fund/programs/common-voice-kiswahili-awards/> accessed 16 May 2025.

88 Kathleen Siminyu, 'Lessons from Building for Kinyarwanda on Common Voice' (Mozilla Foundation5 April 2022) <https://www.mozillafoundation.org/nl/blog/lessons-from-building-for-kinyarwanda-on-common-voice/> accessed 15 May 2025.


Image Source: vecteezy.com

V. The Open Data Dilemma

A majority of the AI and data initiatives aimed at promoting ethical data sourcing release their resulting datasets on open-source platforms, fostering transparency and enabling global reuse.[89] In practice, this means that any developer, researcher, or community organisation can access, inspect, and build upon the datasets as guided by less restrictive licensing.[90] Such openness enables independent verification of data quality, encourages collaboration across borders, and reduces duplication of effort. In Africa, these ambitions align with open data programmes that aim to democratise information access, spur innovation, and empower local developers, hence reducing reliance on proprietary or unethically sourced datasets.[91] Kenya's Open Data Initiative (KODI), launched in 2011 to make public data like census and budget figures openly available, was initially hailed as a pioneering effort.[92] However, the reality of maintaining such government-led open data platforms has proven challenging, with KODI, for example, experiencing stalled updates. The 2022 Global Data Barometer highlighted significant deficiencies, pointing to outdated, poorly managed, and incomplete datasets which undermine their utility for AI innovation and erode public trust.[93]

Beyond issues of data maintenance and quality, open data efforts face critical challenges related to privacy risks and pronounced governance gaps.[94] The drive to make data publicly available often outpaces the development and enforcement of adequate data protection frameworks and privacy-enhancing technologies. In addition, limited institutional capacity to oversee open data initiatives and ensure compliance with data protection laws would also lead to privacy breaches.

Finally, a complex ethical challenge lies in unresolved data ownership and the inequitable sharing of benefits, particularly concerning data derived from communal resources. Open data initiatives frequently lack clear legal and policy frameworks[95] to determine who legitimately controls this data once digitised and how any economic or social benefits from its use will be shared with communities. Addressing this requires moving beyond individual consent models towards developing participatory governance structures that recognise collective data rights and ensure equitable benefit distribution, a crucial step for fostering truly inclusive and ethical AI development on the continent.

89 'Access Open-Source Speech Datasets for African Languages' (Way With Words 27 February 2024) <https://waywithwords.net/resource/speech-datasets-for-african-languages/> accessed 12 August 2025.

90 ibid.

91 ICT Works, '14 Barriers to Using Open Data for Better Development Decisions - ICTworks' (ICTworks 3 April 2024) <https://www.ictworks.org/open-data-development-decisions/> accessed 19 March 2025.

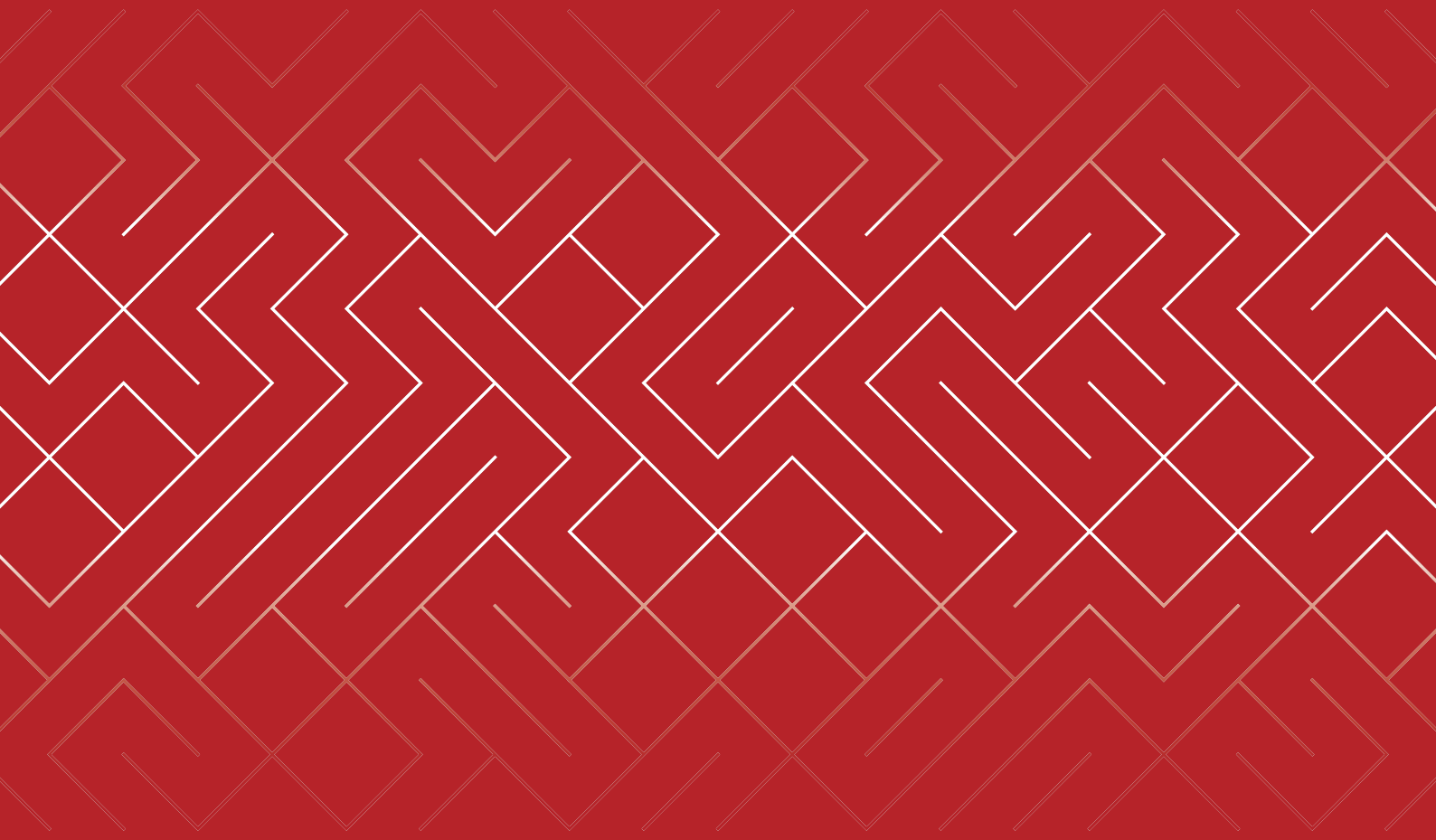
92 OGP, 'Open Data for Development (KE0034)' (Open Government Partnership2022) <https://www.opengovpartnership.org/members/kenya/commitments/KE0034/> accessed 28 February 2025.

93 'The Kenya Open Data Initiative - Centre for Public Impact' (Centre for Public Impact26 September 2024) <https://centreforpublicimpact.org/public-impact-fundamentals/the-kenya-open-data-initiative/> accessed 10 April 2025.

94 Mehdi Barati, 'Open Government Data Programs and Information Privacy Concerns: A Literature Review' (2023) 15 JeDEM - eJournal of eDemocracy and Open Government 73 <https://jedem.org/index.php/jedem/article/view/759/556>.

95 ICT Works, '14 Barriers to Using Open Data for Better Development Decisions - ICTworks'

VI. Recommendations

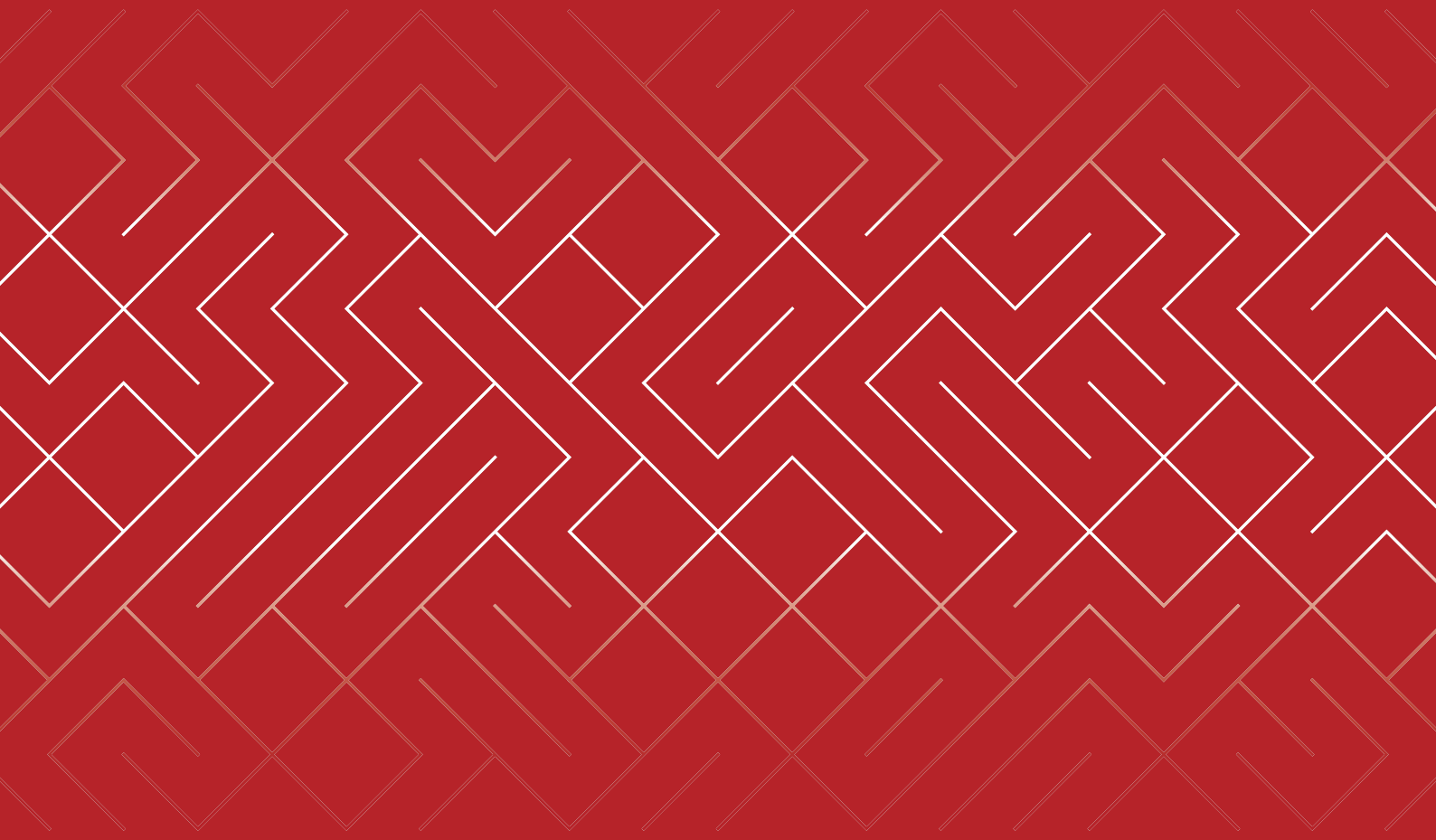


As Africa advances toward ethical AI development, robust data governance within AI regulation offers a critical framework for balancing innovation, inclusivity, and privacy. Addressing exploitative data extraction, enforcing consent-driven collection practices, and ensuring that AI development benefits local communities are critical steps in this process. While a catalyst for local AI solutions, open data requires robust governance safeguards to prevent misuse and inequity. By embedding equitable data-sharing agreements, strengthening regional collaboration, and prioritising transparency, Africa can chart a path toward ethical AI development that empowers its people and ensures that technological advancements align with ethical principles and digital sovereignty. Such an ethical AI policy, grounded in strong data governance, not only drives responsible AI but also empowers Africa's people, setting a global standard for fair benefit sharing and accountability.

To advance ethical data sourcing efforts for ethical AI, African countries must consider:

- ***Amending copyright laws to clarify AI training data use, defining fair use principles and robust consent requirements for data subjects*** - this is crucial to address the current legal ambiguities that allow AI models to be trained on vast amounts of copyrighted material and personal data without explicit authorisation or compensation, as seen in numerous global lawsuits. Clearer laws would protect creators' intellectual property and individuals' data rights while providing legal certainty for AI developers committed to ethical data sourcing.
- ***Establishing and operationalising independent national AI ethics bodies with mandates to oversee data governance and enforce equitable practices, including benefit-sharing agreements for African communities*** - such bodies would ensure compliance with national and international ethical principles, such as those from UNESCO and the African Union, providing a mechanism for redress against exploitative data extraction and ensuring that local communities gain tangible economic and social benefits from the use of their data in AI development.
- ***Strengthening local government data sources, like Kenya's Open Data Initiative (KODI), by ensuring datasets are consistently updated, accurately curated, and protected with robust privacy safeguards and clear usage guidelines -*** this is vital to support ethical AI innovation by providing reliable, safe, and locally relevant data, which can reduce dependency on potentially biased or unethically sourced international datasets and help rebuild public trust in open data platforms.
- ***Promoting robust regional collaboration to develop and share ethical data practices, harmonise regulatory standards, and collectively enhance data sovereignty across the continent -*** by working together, African nations can establish common frameworks for data governance, facilitate the creation of larger and more diverse regional datasets, share best practices for mitigating bias and ensuring fairness, and strengthen their collective negotiating position against potential external data exploitation, ensuring that data generated in Africa primarily benefits Africans.

VII. Conclusion



Kenya's emergence as a potential hub for AI innovation in Africa depends on the critical importance of ethical data sourcing and governance. As this report has shown, the pervasive challenges of unethical data practices, including exploitative extraction, lack of consent, and insufficient representation of local populations, have perpetuated biases and inequities in AI systems, eroding trust. Yet the case studies of Techworker Community Africa, Masakhane, and the Common Voices Project demonstrate that community-driven, transparent, and inclusive approaches to data collection and governance address these issues, fostering AI systems that are fair, culturally relevant, and aligned with Africa's unique needs.

However, the open data dilemma, exemplified by Kenya's Open Data Initiative, persistent gaps in privacy safeguards, data ownership clarity, and equitable benefit sharing, making reforms urgent. Kenya and other African nations should prioritise ethical data sourcing by embedding principles of consent, transparency, and fairness into national and regional frameworks. Strengthening copyright protections, embedding AI ethics oversight, investing in resilient local data ecosystems, and fostering cross-border collaboration are not optional; they are essential to a fair and sustainable AI future. By committing to consent, transparency, and fairness at every stage of the AI lifecycle, Kenya can not only protect its citizens and safeguard its sovereignty but also set a benchmark for developing an AI ecosystem rooted in human rights and shared benefits. Done right, this will ensure that AI's benefits are widely shared, its risks are responsibly managed, and its development reflects the values and aspirations of the Kenyan people — shaping a more inclusive and equitable digital future.

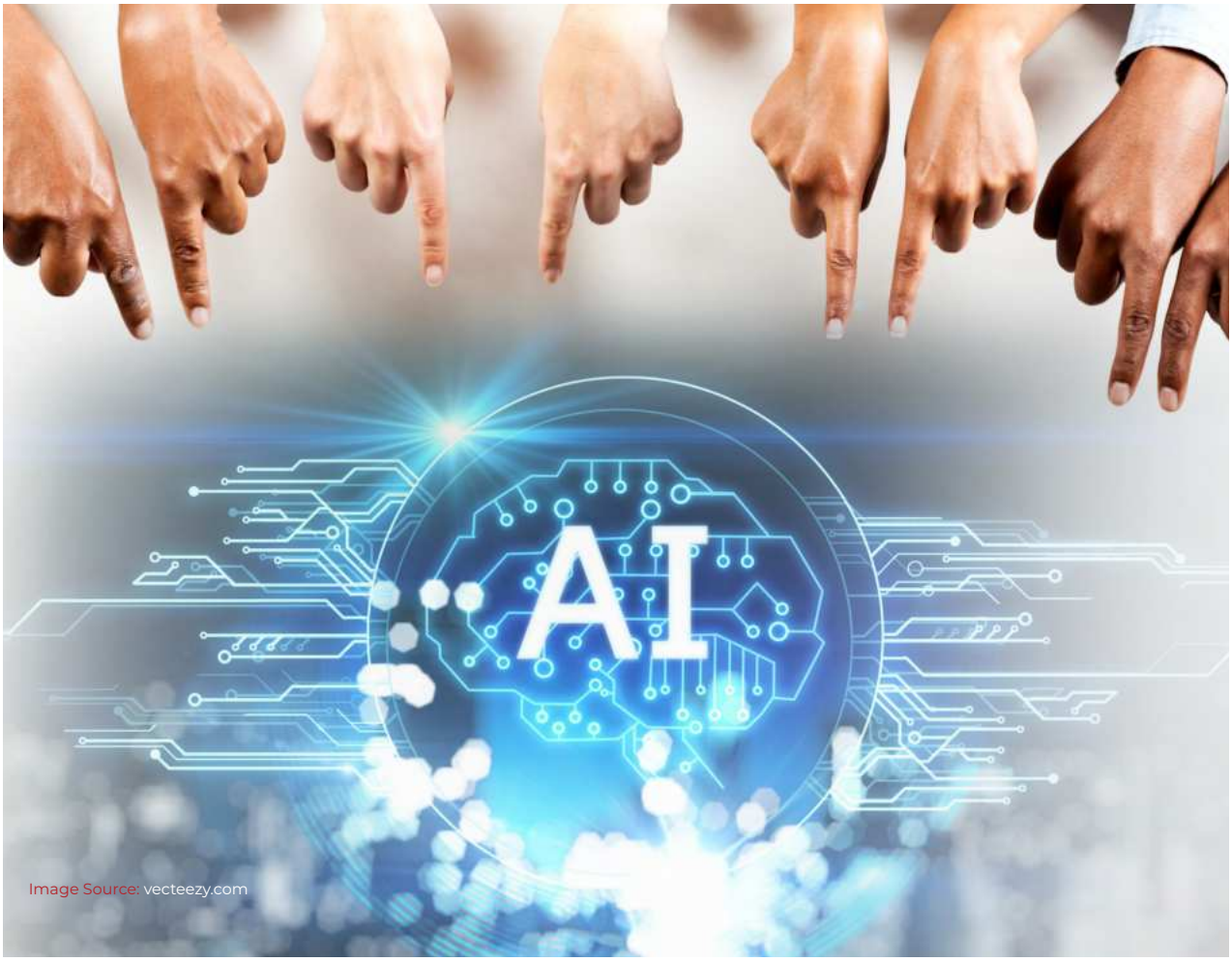


Image Source: vecteezy.com

LIST OF REFERENCES

Republic v Tools for Humanity Corporation (US) & 8 others; Katiba Institute & 4 others (Exparte Applicants); Data Privacy & Governance Society of Kenya (Interested Party) \ [2025] KEHC 5629 (KLR) (Judicial Review, 5 May 2025).

'AI and the Future of Work in Africa: White Paper' (2024) https://aspyee.org/sites/default/files/2024-06/AIandTheFutureofWorkinAfricaWhitePaper-June2024-666aa97e6eb8d.pdf.

Abby Seneor and Matteo Mezzanotte, 'Open-Source Data Science: How to Reduce Bias in AI' (World Economic Forum, 14 October 2022) https://www.weforum.org/stories/2022/10/open-source-data-science-bias-more-ethical-ai-technology/.

Adam J Andreotta, Nin Kirkham and Marco Rizzi, 'AI, Big Data, and the Future of Consent' (2021) 37 *AI & Society* https://link.springer.com/article/10.1007/s00146-021-01262-5.

Adil S Al-Busaidi and others, 'Redefining Boundaries in Innovation and Knowledge Domains: Investigating the Impact of Generative Artificial Intelligence on Copyright and Intellectual Property Rights' (2024) 9 *Journal of Innovation & Knowledge* 100630 https://www.sciencedirect.com/science/article/pii/S2444569X24001690.

Aditya Kalra, Arpan Chaturvedi and Munsif Vengattil, 'OpenAI Faces New Copyright Case, from Global Book Publishers in India' *Reuters* (24 January 2025) https://www.reuters.com/technology/artificial-intelligence/openai-faces-new-copyright-case-global-publishers-india-2025-01-24/.

African Convention on Cyber Security and Personal Data Protection, art 13.

African Union, *Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity* (2024) https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf.

African Union, *Data Policy Framework: An Integrated, Prosperous and Peaceful Africa* (2022) https://au.int/sites/default/files/documents/42078-doc-AU-DATA-POLICY-FRAMEWORK-ENG1.pdf.

Andrew Paul, 'AI Programs Often Exclude African Languages. These Researchers Have a Plan to Fix That' *Popular Science* (11 August 2023) https://www.popsci.com/technology/african-language-ai-bias/.

Athena Vakali and Nicoleta Tantalaki, 'Rolling in the Deep of Cognitive and AI Biases' \[2019] *arXiv.org* https://arxiv.org/html/2407.21202v1.

Benedikt Erforth, 'Data Extraction, Data Governance and Africa-Europe Cooperation: A Research Agenda' (2024) https://www.megatrends-afrika.de/assets/afrika/publications/MTA_working_paper/MTA_WP14_Erforth_Digital_Cooperation.pdf.

Billy Perrigo, 'Exclusive: OpenAI Used Kenyan Workers on Less than \$2 per Hour to Make ChatGPT Less Toxic' *Time* (18 Janu-

ary 2023) https://time.com/6247678/openai-chatgpt-kenya-workers/.

Bruce Barcott, 'AI Lawsuits Worth Watching: A Curated Guide' (Tech Policy Press, 1 July 2024) https://www.techpolicy.press/ai-lawsuits-worth-watching-a-curated-guide/.

Business and Human Rights Resource Centre, 'Meta & Sama Lawsuit (Re Poor Working Conditions & Human Trafficking, Kenya)' (2022) https://www.business-humanrights.org/fr/latest-news/meta-sama-lawsuit-re-poor-working-conditions-human-trafficking-kenya/.

Carlos Mureithi, 'Five Things to Learn from Kenya's Inquiry into Worldcoin's Activities in the Country' (Tech Policy Press, 24 April 2024) https://www.techpolicy.press/five-things-to-learn-from-kenyas-inquiry-into-worldcoins-activities-in-the-country/.

Carnegie Endowment for International Peace, 'Africa Technology Policy Tracker' (2024) https://carnegieendowment.org/features/africa-digital-regulations?lang=en accessed 28 July 2025.

Centre for Intellectual Property and Information Technology Law (CIPIT), *The State of AI in Africa Report 2023* (2023) https://cipit.strathmore.edu/wp-content/uploads/2023/05/The-State-of-AI-in-Africa-Report-2023-min.pdf.

Centre for Public Impact, 'The Kenya Open Data Initiative' (26 September 2024) https://centreforpublicimpact.org/public-impact-fundamentals/the-kenya-open-data-initiative/ accessed 10 April 2025.

Damian Okaibedi Eke and others, 'African Perspectives of Trustworthy AI: An Introduction' \[2025] *Trustworthy AI* 1 https://link.springer.com/chapter/10.1007/978-3-031-75674-0_1.

Data Protection Act (Kenya)

David Ifeoluwa Adelani and others, 'MasakhaNER: Named Entity Recognition for African Languages' (arXiv.org, 2021) https://arxiv.org/abs/2103.11811 accessed 14 May 2025.

Dawn Chmielewski and Katie Paul, 'Murdoch's Dow Jones, New York Post Sue Perplexity AI for "Illegal" Copying of Content' *Reuters* (21 October 2024) https://www.reuters.com/legal/murdoch-firms-dow-jones-new-york-post-sue-perplexity-ai-2024-10-21/.

Doyin Akindotuni, 'Resource Asymmetry in Multilingual NLP: A Comprehensive Review and Critique' (2025) 13 *Journal of Computer and Communications* 14 https://www.scirp.org/journal/paperinformation?paperid=143854.

European Parliament, 'P9_TA(2024)0138 Artificial Intelligence Act: European Parliament Legislative Resolution of 13 March 2024' (2024) https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

Federal Ministry for Economic Cooperation and Development (Germany), 'Hamburg Declaration on Responsible Artificial Intelligence (AI) for the Sustainable Development Goals (SDGs)' (2025) [https://www.bmz-digital.

global/wp-content/uploads/2025/06/250603_Hamburg_Declaration.pdf](https://www.bmz-digital.global/wp-content/uploads/2025/06/250603_Hamburg_Declaration.pdf) accessed 7 July 2025.

General Data Protection Regulation, art 4(11) https://gdpr-info.eu/ accessed 24 January 2025.

George Benneh Mensah, 'Artificial Intelligence and Ethics: A Comprehensive Review of Bias Mitigation, Transparency, and Accountability in AI Systems' \[2023] *ResearchGate* https://www.researchgate.net/publication/375744287_Artificial_Intelligence_and_Ethics_A_Comprehensive_Review_of_Bias_Mitigation_Transparency_and_Accountability_in_AI_Systems.

GIZ, 'Harnessing the Power of AI to Improve Harvests' (GIZ, 28 May 2025) https://www.giz.de/en/mediacenter/better-harvest-with-AI.html.

IBM, 'AI Ethics' (IBM.com, 17 September 2024) https://www.ibm.com/think/topics/ai-ethics.

ICTworks, '14 Barriers to Using Open Data for Better Development Decisions' (ICTworks, 3 April 2024) https://www.ictworks.org/open-data-development-decisions/ accessed 19 March 2025.

Jan Krewer, 'Creating Community-Driven Datasets: Insights from Mozilla Common Voice Activities in East Africa' (2023) https://www.bmz-digital.global/wp-content/uploads/2023/03/Creating-Community-Driven-Datasets-Report-032023-GIZ-Mozilla.pdf.

Josephine Kaaniru, 'AI Assistive Technologies (ATS) for Persons with Disabilities (PWDS) in Africa' (Centre for Intellectual Property and Information Technology Law, 31 October 2023) https://cipit.strathmore.edu/ai-assistive-technologies-ats-for-persons-with-disabilities-pwds-in-africa/.

Kathleen Siminyu, 'Lessons from Building for Kinyarwanda on Common Voice' (Mozilla Foundation, 5 April 2022) https://www.mozillafoundation.org/nl/blog/lessons-from-building-for-kinyarwanda-on-common-voice/ accessed 15 May 2025.

Mark Hill and Courtney Benard, 'Nvidia Faces Class-Action Lawsuit for Training AI Model on "Shadow Library"' (Lexology, 30 April 2024) https://www.lexology.com/library/detail.aspx?g=3a665ce3-3db6-40a3-899e-10c2cf606a71 accessed 19 March 2025.

Masakhane, 'Organisation Strategy' (Masakhane.io, 2020) https://www.masakhane.io/organisation-strategy accessed 10 July 2025.

Masakhane.io, 'Masakhane-NER' (GitHub, 2020) https://github.com/masakhane-io/masakhane-ner accessed 14 May 2025.

MasakhaNER Know, 'Masakhane - MasakhaNER: Know Our Names' (Masakhane.io, 2020) https://www.masakhane.io/ongoing-projects/masakhaner-know-our-names.

Matthew Hanna and others, 'Ethical and Bias Considerations in Artificial Intelligence/Machine Learning' (2024) 38 *Modern Pathology* 1 [https://www.sciencedirect.com/sci-

ence/article/pii/S0893395224002667](https://www.sciencedirect.com/science/article/pii/S0893395224002667).

Mehdi Barati, 'Open Government Data Programs and Information Privacy Concerns: A Literature Review' (2023) 15 *JeDEM – eJournal of eDemocracy and Open Government* 73 https://jedem.org/index.php/jedem/article/view/759/556.

Ministry of Foreign Affairs (Kenya), *Diplomat's Playbook on Artificial Intelligence* (2025) https://mfa.go.ke/sites/default/files/2025-01/Diplomats%20AI%20Playbook%20Final.pdf.

Ministry of Information, Communications and the Digital Economy (Kenya), *Kenya AI Strategy 2025–2030* (2025) https://ict.go.ke/sites/default/files/2025-03/Kenya%20AI%20Strategy%202025%20-%202030.pdf.

Ministry of Information, Communications and the Digital Economy (Kenya), *Report of the Information, Communications and the Digital Economy Sectoral Working Group* (2024) https://ict.go.ke/sites/default/files/2024-09/MICDE%20Sector%20Working%20Group%20Report%20-%20June%202024.pdf.

Morgan Sullivan, 'Key Principles for Ethical AI Development' (Transcend Blog, 20 October 2023) https://transcend.io/blog/ai-ethics accessed 24 January 2025.

Mozilla Foundation, 'Common Voice' (2025) https://www.mozillafoundation.org/en/common-voice/ accessed 14 May 2025.

Mozilla Foundation, 'Common-Voice-Kiswahili-Awards' (2019) https://www.mozillafoundation.org/en/what-we-fund/programs/common-voice-kiswahili-awards/ accessed 16 May 2025.

Mozilla Foundation, 'Platform and Dataset' (2025) https://www.mozillafoundation.org/en/common-voice/platform-and-dataset/ accessed 14 May 2025.

Natasha Karanja and Chebet Koros, 'Artificial Intelligence (AI) Training Data and the Copyright Dilemma: Insights for African Developers' (Centre for Intellectual Property and Information Technology Law, 12 February 2025) https://cipit.strathmore.edu/artificial-intelligence-ai-training-data-and-the-copyright-dilemma-insights-for-african-developers/.

Neha Panchal, 'Ethical Considerations in AI Data Annotation' (Damco Solutions, 15 May 2023) https://www.damcogroup.com/blogs/understanding-ethical-considerations-in-ai-data-annotation accessed 12 August 2025.

Notice Pasipamire and Abton Muroyiwa, 'Navigating Algorithm Bias in AI: Ensuring Fairness and Trust in Africa' (2024) 9 *Frontiers in Research Metrics and Analytics* https://www.frontiersin.org/journals/research-metrics-and-analytics/articles/10.3389/frma.2024.1486600/full.

Oakley Parker, 'Data Governance and Ethical AI: Developing Legal Frameworks to Address Algorithmic Bias and Discrimination' [https://www.researchgate.net/publication/384966994_Data_Governance_and_Ethical_AI_Developing_Legal_Frameworks_to_Address_

Algorithmic_Bias_and_Discrimination](https://www.researchgate.net/publication/384966994_Data_Governance_and_Ethical_AI_Developing_Legal_Frameworks_to_Address_Algorithmic_Bias_and_Discrimination).

Open Government Partnership, 'Open Data for Development (KE0034)' (2022) https://www.opengovpartnership.org/members/kenya/commitments/KE0034/ accessed 28 February 2025.

PBS, 'Coded Bias' (Independent Lens, 2021) https://www.pbs.org/independentlens/documentaries/coded-bias/.

Petar Radanliev, 'AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development' (2025) 39 *Applied Artificial Intelligence* https://www.tandfonline.com/doi/full/10.1080/08839514.2025.2463722.

Rachael Wambua, 'How Open Source Is Powering the Future of African NLP' (Lanfrica Blog, 8 May 2025) https://lanfrica.com/blog/how-open-source-is-powering-the-future-of-african-nlp/ accessed 10 July 2025.

Raksha Vasudevan, 'A Lawsuit against Meta Shows the Emptiness of Social Enterprises' *Wired* (20 July 2022) https://www.wired.com/story/social-enterprise-technology-africa/.

Republic of South Africa, *Copyright Amendment Bill* https://www.parliament.gov.za/storage/app/media/uploaded-files/Copyright%20Amendment%20Bill%20Draft.pdf.

Shahmar Mirishli, 'Ethical Implications of AI in Data Collection: Balancing Innovation with Privacy' (2024) 6 *International Scientific Journal* 40 https://arxiv.org/pdf/2503.14539.

Swetha Sistla, 'AI with Integrity: The Necessity of Responsible AI Governance' (2024) *Journal of Artificial Intelligence & Cloud Computing* SRC/JAICC-E180 [https://doi.org/10.47363/JAICC/2024(3)E180](https://doi.org/10.47363/JAICC/2024%283%29E180) accessed 24 January 2025.

Techworker Community Africa, 'Techworker Community Africa' (2025) https://www.techworkercommunityafrica.com/ accessed 11 May 2025.

The East African, 'Kenya to Restrict Use of Locals' Data for Foreign AI Training' (21 January 2025) https://www.theeastafrican.co.ke/tea/sustainability/innovation/kenya-to-restrict-use-of-locals-data-for-foreign-ai-training-4896508.

UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (UNESCO, 16 May 2023) https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence.

United Nations Development Programme and Italian G7 Presidency, *AI Hub for Sustainable Development: Strengthening Local AI Ecosystems through Collective Action* (2024) https://www.undp.org/sites/g/files/zskgke326/files/2024-07/ai_hub_report_digital.pdf.

Victoria Reed, 'African Languages in AI: Breaking Barriers with NLP' (AICompetence.org, 23 December 2024) [https://aicompetence.org/african-languages-in-ai-barriers-with-nlp/](https://aicompetence.org/african-languag-

es-in-ai-barriers-with-nlp/) accessed 30 July 2025.

Way With Words, 'Access Open-Source Speech Datasets for African Languages' (27 February 2024) https://waywithwords.net/resource/speech-datasets-for-african-languages/ accessed 12 August 2025.

WeeTracker, 'OpenAI–Sama Kenyan Workers Controversy' (25 November 2024) https://weetracker.com/2024/11/25/openai-sama-kenyan-workers-controversy/ accessed 9 April 2025.

Zuziwe Msomi and Sally Matthews, 'Protecting Indigenous Knowledge Using Intellectual Property Rights Law: The Masakhane Pelargonium Case' (2016) 45 *Africanus: Journal of Development Studies* 62 https://unisapress-journals.co.za/index.php/Africanus/article/download/645/432/4917.

This study was made possible by a grant provided by the International Development Research Center (IDRC). We thank the organization for their continued support.

Canada

Supported by

Ole Sangale Rd, Madaraka Estate.
P.O Box 59857-00200, Nairobi, Kenya.
Tel: +254 (0)703 034612
Email: cipit@strathmore.edu
Website: www.cipit.strathmore.edu