Legally Mandated Pre-Deployment Evaluations: Promoting AI Safety in Clinical Medicine Across Africa

Submitted in partial fulfilment of the requirements of the Bachelor of Laws Degree,
Strathmore University Law School

By
Mark Lenny Muchiri Gitau
145797

Prepared under the supervision of
Cecil Abungu

March 2025
Word count: 14,557

Table of Contents

Declaration ..... **i**
Acknowledgements ..... **ii**
List of Legal Instruments ..... **iii**
List of Abbreviations ..... **iv**
Abstract ..... **vi**
1 Introduction **1**
1.1 Background 1
1.2 Statement of Problem 5
1.3 Research Objectives 5
1.4 Research Questions 5
1.5 Hypothesis 5
1.6 Justification 6
1.7 Conceptual Framework: The Precautionary Principle 6
1.8 Literature Review 8
1.8.1 On the role of context in AI safety 8
1.8.2 On the general and context-specific risks that advanced AI models could pose to clinical medicine 9
1.8.3 Contribution to the literature 10
1.9 Methodology 10
1.10 Chapter Breakdown 11
2 The Role of Context in AI Safety **12**
2.1 Introduction 12
2.2 The essence of AI safety 12
2.2.1 Technical AI safety 12
2.2.2 Sociotechnical AI safety 13
2.2.3 Where the discourse stands—and what definition to adopt 15
2.3 How context matters in AI safety 16
2.3.1 AI systems often reveal unexpected harms after deployment 16
2.3.2 "Harm" cannot be understood away from its context 18
2.3.3 Inclusivity could prevent the perpetuation of structural disadvantage 20
2.3.4 Static evaluations cannot account for shifting contexts 22
2.4 Conclusion 22
3 Advanced AI in African Clinical Medicine: Benefits and Context-Sensitive Risks **23**
3.1 Introduction 23
3.2 Benefits of AI integration in clinical medicine 23
3.2.1 Widespread advantages of AI in clinical medicine 23

3.2.2 Africa-centric advantages of AI in clinical medicine 24
3.3 Potential pitfalls of AI in clinical medicine 26
3.3.1 Fundamental risks in AI-enabled clinical medicine 26
3.3.2 Context-specific risks regarding AI in African clinical medicine 27
A. How underrepresentation of low-resource African languages in training data could lead to risks 27
B. How inadequate incorporation of African traditional medicine in training data could lead to risks 30
C. How underrepresentation of African genetic diversity in training data could lead to risks 33
D. How underrepresentation of diseases endemic to Africa, particularly NTDs, in training data could lead to risks 35
3.4 Conclusion 37
4 The Methodological, Legal and Institutional Landscape of Pre-Deployment Evaluations 38
4.1 Introduction 38
4.2 The role of pre-deployment evaluations in AI safety 38
4.3 Current approaches to AI pre-deployment evaluations 40
4.3.1 Methods used in AI evaluations 40
4.3.2 Legal and policy frameworks governing AI evaluations 43
A. Legal and policy frameworks governing AI evaluations in AISIs 43
B. Additional legal and policy frameworks governing AI evaluations 47
4.4 Gaps in existing pre-deployment evaluations 48
4.5 Conclusion 49
5 How Law Could Be Used To Strengthen Pre-Deployment Evaluations For Advanced AI in African Clinical Medicine 51
5.1 Introduction 51
5.2 Role of legally mandated tailored pre-deployment evaluations 51
5.3 Legal mechanisms to support pre-deployment evaluations 53
5.3.1 Legal mandates for government involvement in evaluations 53
5.3.2 Incentivising a private sector AI evaluation industry 56
5.3.3 Legal obligations for AI developers 57
5.4 Potential implications 58
5.5 Conclusion 59
6 Conclusion and Recommendations 60
6.1 Conclusion 60
6.2 Recommendations 61
7 Bibliography 63

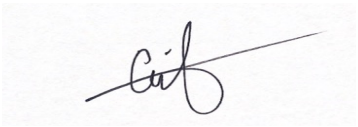
Declaration

I, MARK LENNY MUCHIRI GITAU, do hereby declare that this research is my original work and that to the best of my knowledge and belief, it has not been previously, in its entirety or in part, been submitted to any other university for a degree or diploma. Other works cited or referred to are accordingly acknowledged.

Signed: .....

Date: 4 April 2025

This dissertation has been submitted for examination with my approval as University Supervisor.



Signed: .....

Cecil Abungu

Date: 4 April 2025

Acknowledgements

I am deeply grateful for the invaluable guidance and support of Cecil Abungu, whose expertise, mentorship and constructive feedback were instrumental throughout this process and my LLB journey.

To my friends, thank you for the insightful discussions, laughter, and support during this journey. You walked this path with me, and your presence made this degree not just bearable, but meaningful. Special thanks to Zayn Aslam and Faith Gakii.

To my beloved grandfather, whose belief in education planted seeds of curiosity that blossomed into this work. You taught me to strive for excellence and your reverence for knowledge has been the compass that guided me on this journey. This degree, and every step forward, carries your legacy. Rest well, *Guka*.

Finally, I owe boundless gratitude to my family, whose encouragement and sacrifices enabled me to complete this study and my degree.

To everyone who contributed in ways seen and unseen: I am profoundly thankful.

List of Legal Instruments

Artificial Intelligence Act (European Union)

Artificial Intelligence and Data Act (Canada)

Bill C-27 – An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts (Canada)

Data Protection Act (Act No. 24 of 2019)

Federal Tort Claims Act (United States)

Health Insurance Portability and Accountability Act (United States)

Bill 11 – A Bill to make provision for the regulation of artificial intelligence; and for connected purposes (United Kingdom)

Executive Order 14110 – Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence (United States)

UNGA, *Rio declaration on environment and development*, UN A/CONF.151/26 (Vol. I), 3-14 June 1992

List of Abbreviations

Africa CDC – Africa Centres for Disease Control and Prevention

AIDA – Artificial Intelligence and Data Act (Canada)

ADRs – Adverse Drug Reactions

AISI(s) – Artificial Intelligence Safety Institute(s)

AGI – Artificial General Intelligence

AI – Artificial Intelligence

ANSSI – French National Agency for the Security of Information Systems

AU – African Union

BBC – British Broadcasting Corporation

BCL11A – B-Cell Lymphoma/Leukaemia 11A (gene)

CI – Critical Infrastructure

CISA – Cybersecurity and Infrastructure Security Agency

COVID-19 – Coronavirus Disease 2019

CPSC – Consumer Product Safety Commission

CRB1 – Crumbs Cell Polarity Complex Component 1 (gene)

DHS – Department of Homeland Security (US)

DENND1B – DENN Domain Containing 1B (gene)

DSIT – Department for Science, Innovation and Technology (UK)

EU – European Union

ETRI – Electronics and Telecommunications Research Institute (South Korea)

FAA – Federal Aviation Administration (US)

FDA – Food and Drug Administration (US)

FTCA – Federal Tort Claims Act (US)

GWAS – Genome-Wide Association Studies

IMDA – Infocomm Media Development Authority (Singapore)

INRIA – French Institute for Research in Computer Science and Automation

IPA – Information-technology Promotion Agency (Japan)

ISED – Innovation, Science and Economic Development (Canada)

LLMs – Large Language Models

LNE – National Laboratory of Metrology and Testing (France)

MCAS – Manoeuvring Characteristics Augmentation System

METR – Model Evaluation & Threat Research

MIRI – Machine Intelligence Research Institute

NAIAC – National Artificial Intelligence Advisory Committee (US)

NHTSA – National Highway Traffic Safety Administration (US)

NIST – National Institute of Standards and Technology (US)

NTDs – Neglected Tropical Diseases

OECD – Organisation for Economic Co-operation and Development

PEREN – Expert Centre for Digital Regulation (France)

PDE4D – Phosphodiesterase 4D (gene)

RSF – Reporters Sans Frontières

UNGA – United Nations General Assembly

US – United States

UK – United Kingdom

VA – Veterans Affairs (US)

VHA – Veterans Health Administration (US)

WEIRD – Western, Educated, Industrialised, Rich, and Democratic

WHO – World Health Organisation

Abstract

The integration of advanced artificial intelligence (AI) into healthcare systems presents transformative potential for clinical medicine in Africa, yet it introduces profound context-specific risks. This study identifies four primary risk vectors through content analysis. First, the potential for mistranslations and misinterpretations in diagnostic settings due to linguistic diversity; second, the inadequate integration of traditional medicine data, heightening the risk of adverse drug reactions and misguided therapeutic recommendations; third, genetic biases stemming from the underrepresentation of African populations, which may precipitate misdiagnoses; and fourth, the neglect of tropical diseases endemic to the region, further compounding diagnostic inaccuracies. Although dataset bias underpins all these concerns, each manifests in distinct ways that can compromise patient safety and clinical outcomes.

In response to these challenges, the study advocates for urgent, legally grounded solutions. Through a comprehensive review of the literature and a qualitative analysis of current evaluation methods, legal instruments, and policy frameworks, the study argues that existing AI safety evaluation frameworks, predominantly developed in high-resource settings, are ill-suited to the African context as they fail to capture the distinctive risks inherent in the continent's clinical environment. Instead, legally mandated, context-specific pre-deployment evaluations are essential. To operationalise this approach, a tripartite legal framework is proposed: (1) government involvement through data aggregation and verification of developer assessments; (2) private-sector incentives that promote rigorous, independent evaluations; and (3) enforceable developer obligations to adhere to context-aware safety standards.

1 Introduction

1.1 Background

In recent years, it has become increasingly evident that advanced artificial intelligence (AI) systems are capable of posing significant risks to humanity, prompting serious scrutiny over their development and deployment.[1] This close scrutiny has led to various responses, including pre- and post-deployment AI safety evaluations.[2] For example, AI Safety Institutes (AISIs) have emerged worldwide to shape AI safety practices, with evaluations at the core of their activities.[3] Similarly, private sector actors are employing evaluation techniques such as benchmarking and red teaming to identify risks.[4] Evaluations, in short, have become indispensable for mitigating risks from advanced AI integration.

One of the most high-risk domains for advanced AI integration is critical infrastructure (CI)—the foundational systems that sustain modern societies, including energy grids, transport networks, financial services, and healthcare systems.[5] While advanced AI offers immense benefits in these sectors,[6] its integration is incredibly perilous. Failures in CI-linked AI

[1] In this paper, 'advanced AI models' or 'advanced AI systems' mean 'systems that are highly capable and general purpose.' See Ho L, Barnhart J, Trager R, Bengio Y, Brundage M, Carnegie A, Chowdhury R, Dafoe A, Hadfield A, Levi M, and Snidal D, 'International institutions for advanced AI' ArXiv, 2023, 1-2.

[2] Anderljung A, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, Avin S, Brundage M, Bullock J, Cass-Beggs D, Chang B, Collins T, Fist T, Hadfield G, Hayes A, Ho L, Hooker S, Horvitz E, Kolt N, Schuett J, Shavit Y, Siddarth D, Trager R, and Wolf K, 'Frontier AI regulation: Managing emerging risks to public safety' ArXiv, 2023, 23-25; Trager R, Harack B, Reuel A, Carnegie A, Heim L, Ho L, Kreps S, Lall R, Larter O, Ó hÉigeartaigh S, Staffell S, and Villalobos J, 'International governance of civilian AI: A jurisdictional certification approach' ArXiv, 2023, 17; Ho L et al., 'International institutions for advanced AI' ArXiv, 2023, 6,14.

[3] Allen G and Adamson G, *The AI safety institute international network: Next steps and recommendations*, 2024, 7; Institute for AI Policy and Strategy, *Understanding the first wave of AI safety institutes: Characteristics, functions, and challenges*, 2024, 17-19; National Institute of Standards and Technology, 'Pre-deployment evaluation of Anthropic's upgraded Claude 3.5 Sonnet' 19 November 2024–< https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet> on 8 December 2024.

[4] Mckernon E, Cheng D, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 23 November 2024; The Alan Turing Institute and the Centre for Emerging Technology and Security, *Evaluating malicious generative AI capabilities understanding inflection points in risk*, 2024, 10.

[5] In this study, CI refers to 'systems and assets, whether physical or virtual, so vital to a nation that the incapacity or destruction of such systems and assets would have a debilitating impact on its security, national economic security, national public health or safety, or any combination of those matters.' See US Department of Homeland Security, 'Mitigating artificial intelligence (AI) risk: Safety and security guidelines for critical infrastructure owners and operators' April 2024, 5–< https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf> on 19 October 2024; Cybersecurity and Infrastructure Security Agency, 'Critical infrastructure sectors' US Department of Homeland Security–<https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors> on 23 November 2024.

[6] DeAngelis A, 'Artificial Intelligence and Critical Infrastructure' Enterra Solutions, 4 June 2024–< https://enterrasolutions.com/artificial-intelligence-and-critical-infrastructure/> on 19 October 2024; US

systems—whether through malicious exploitation or inherent errors—could have catastrophic and cascading consequences for public safety, national security, and economic stability.[7] Recognising these stakes, some jurisdictions, such as the European Union (EU), have classified AI integration in CI as "high risk" and imposed stringent regulatory and evaluation requirements to mitigate the dangers.[8]

Yet, despite this recognition, global consensus on AI safety approaches to CI remains elusive. One reason for this could be state sovereignty, that is, individual states can unilaterally define, protect, and prioritise their own CI sectors.[9] This creates fragmentation as to what constitutes CI. Consequently, even if countries were to agree in principle that AI integration in CI warrants high-risk classification, the absence of a shared definition of CI itself would make a harmonised, global approach to AI safety virtually impossible. Furthermore, while many countries in Europe and North America have defined their CI sectors, only about a quarter of African states have done so.[10] Among African nations with formal CI classifications, five key sectors—ICT, energy, health, transport, and food—appear to be the most frequently prioritised.[11] However, AI-related risks likely to be faced by most of these priority African CI sectors closely resemble those encountered or foreseen elsewhere.[12] This overlap suggests that,

Department of Homeland Security, 'DHS Publishes Guidelines and Report to Secure Critical Infrastructure and Weapons of Mass Destruction from AI-Related Threats' 29 April 2024–<https://www.dhs.gov/news/2024/04/29/dhs-publishes-guidelines-and-report-secure-critical-infrastructure-and-weapons-mass> on 2 November 2024.

[7] Center for Security and Emerging Technology, *Securing critical infrastructure in the age of AI*, 2024, 11-12.

[8] Articles 8-17, EU Artificial Intelligence Act; EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/ > on 19 October 2024; Digital Regulation Platform, 'Transformative technologies (AI) challenges and principles of regulation' The World Bank, 8th May 2024, < https://digitalregulation.org/3004297-2/ > on 19 October 2024; RSF, 'AI and the Right to Information RSF's 7 recommendations to protect an AI-dominated information space' May 2024 < https://rsf.org/sites/default/files/medias/file/2024/06/AI%20and%20the%20Right%20to%20Information%20-%20RSF%20Recommendations%20to%20EU.pdf > on 19 October 2024.

[9] Weber V, Laumann E, and Riera M, 'Mapping the world's critical infrastructure sectors' German Council of Foreign Relations, Policy Brief Number 35, November 2023, 1-2–< https://dgap.org/system/files/article_pdfs/DGAP%20Policy%20Brief%20No.35%2C%20November%202023.pdf> on 23 November 2024.

[10] Weber V, Laumann E, and Riera M, 'Mapping the world's critical infrastructure sectors' German Council of Foreign Relations, Policy Brief Number 35, November 2023, 3–< https://dgap.org/system/files/article_pdfs/DGAP%20Policy%20Brief%20No.35%2C%20November%202023.pdf> on 23 November 2024.

[11] Weber V, Laumann E, and Riera M, 'Mapping the world's critical infrastructure sectors' German Council of Foreign Relations, Policy Brief Number 35, November 2023, 7–< https://dgap.org/system/files/article_pdfs/DGAP%20Policy%20Brief%20No.35%2C%20November%202023.pdf> on 23 November 2024.

[12] Department of Homeland Security, Roles and responsibilities framework for artificial intelligence in critical infrastructure, 2024, 10-11.

for the majority of CI sectors in Africa, existing AI safety frameworks from other regions could be adapted, rather than reinvented, to address these risks.

But healthcare infrastructure breaks this pattern. Here, AI integration poses risks to African *clinical medicine* (the direct diagnosis and treatment of patients),[13] that are so context-specific to the African setting that global AI safety measures may prove insufficient. These risks could arise because of several interrelated factors. First, clinical outcomes are inextricably tied to healthcare infrastructure quality. For example, strained systems (e.g., understaffed hospitals, lack of medical equipment, erratic power grids) directly impede diagnostic accuracy and treatment.[14] Second, modern AI-driven tools, such as diagnostic algorithms and predictive models, are increasingly embedded within this infrastructure, and thus their reliability directly impacts clinical outcomes.[15] Third, the safety and effectiveness of these AI systems depend on the quality, diversity, and representativeness of their training data.[16] Fourth, any biases or gaps in these datasets—especially when they lack representation from certain populations—are, according to the AU Continental AI Strategy, "significant concerns,"[17] and can lead to dangerously inaccurate predictions.[18] Finally, AI models are overwhelmingly trained on datasets from well-resourced regions, meaning they often fail to capture African-specific medical realities.[19] Consequently, if data-biased AI systems are integrated into African healthcare infrastructure, the quality of clinical care will almost certainly suffer.

[13] In this study, "African clinical medicine" or "clinical medicine in Africa" refers to the formal practice of diagnosing and treating patients in Africa. It is not intended to refer to *African traditional medicine*, which typically refers to indigenous healing practices passed down through generations and is a part of clinical medicine.
[14] Luxon L, 'Infrastructure – the key to healthcare improvement' 2(1) *Future Hospital Journal*, 2015, 4; Science Direct, 'Healthcare infrastructure'–< https://www.sciencedirect.com/topics/social-sciences/health-infrastructure> on 28 February 2025; World Health Organisation, 'Low quality healthcare is increasing the burden of illness and health costs globally' WHO, 5 July 2018–< https://www.who.int/news/item/05-07-2018-low-quality-healthcare-is-increasing-The-burden-of-illness-and-health-costs-globally> on 28 February 2025.
[15] Lee N, Simmons N, and Crawford M, 'Health and AI: Advancing responsible and ethical AI for all communities' Brookings, 3 March 2025–< https://www.brookings.edu/articles/health-and-ai-advancing-responsible-and-ethical-ai-for-all-communities/> on 4 March 2025.
[16] Aldoseri A, Khalifa A, and Hamouda A, 'Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges' 13 *Applied Sciences*, 2023, 1-2; Tigera, Understanding AI safety: Principles, frameworks, and best practices'–< https://www.tigera.io/learn/guides/llm-security/ai-safety/> on 31 December 2024; Jungco K, 'Types of AI models: A deep dive into AI architecture' eWeek, 21 November 2024–< https://www.eweek.com/artificial-intelligence/ai-model-types/> on 31 December 2024.
[17] African Union, *Continental AI Strategy*, 2024, 25.
[18] Bureau of Cyberspace and Digital Policy, 'Risk management profile for artificial intelligence and human rights' US Department of State, 25 July 2024–< https://www.state.gov/risk-management-profile-for-ai-and-human-rights/> on 31 December 2024; Berkley Haas Center for Equity, *Mitigating bias in artificial intelligence an equity: Fluent leadership playbook*, 2020, 3-4, 20.
[19] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.

This study identifies four pathways through which these risks could materialise. Firstly, the continent's unparalleled linguistic diversity[20] presents a formidable barrier. AI systems designed for diagnostic support or patient engagement, when trained on dominant global languages, would be ill-equipped to process the wide array of African languages and dialects, risking miscommunication and diagnostic errors. Secondly, traditional medicine, which remains a cornerstone of healthcare for a large majority of African communities,[21] is almost entirely absent from most datasets. This oversight limits AI systems' capacity to validate, integrate, or even recognise the potential risks or benefits of these practices. Thirdly, in genomic medicine, African populations—despite possessing the greatest genetic diversity globally—are vastly underrepresented in genomic datasets.[22] This neglect threatens the efficacy and safety of precision medicine and pharmacogenomic applications, as treatments calibrated to non-African genetic profiles may be less effective or even harmful.[23] Finally, neglected tropical diseases (NTDs)—a category of illnesses that disproportionately affect African populations—are often overlooked in global healthcare data.[24] Advanced AI systems, biased toward the diseases common in wealthier regions, may struggle to accurately diagnose or recommend appropriate interventions for NTDs.

Despite these pressing and unique concerns, existing AI safety evaluation frameworks remain insufficiently localised, and African countries do not seem to have implemented legally mandated, context-specific evaluations targeting the integration of advanced AI models with their healthcare systems. Without these, the use of advanced AI models in clinical medicine risks exacerbating, rather than alleviating, the already fragile state of healthcare across the continent.

---

[20] Ritchie S, Cheng Y, Chen M, Matthews R, Esch D, Li B, and Sim K, 'Large vocabulary speech recognition for languages of Africa: multilingual modelling and self-supervised learning' ArXiv, 2022, 1.

[21] WHO, 'African Traditional Medicine Day 2022' 31 August 2022–< https://www.afro.who.int/regional-director/speeches-messages/african-traditional-medicine-day-2022#:~:text=Traditional%20medicine%20has%20been%20the,for%20their%20basic%20health%20needs> on 23 November 2024; Anyikwa C, 'Exploring the role of divination in traditional medicine in Africa: A critical perspective' 2 *Research Directions: One Health*, 2024, 1-4; BBC, 'The Kenyan enthralled by the healing power of plants' 29 July 2024–< https://www.bbc.com/news/articles/c84j0kzgnk9o> on 30 November 2024.

[22] Shishehbor F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18.

[23] Chinta S, Wang Z, Zhang X, Viet T, Kashif A, Smith M, and Zhang W, 'AI-driven healthcare: A survey on ensuring fairness and mitigating bias' ArXiv, 2024, 5; Kist J, Vos R, Mairuhu A, Struijs J, van Peet P, Vos H, van Os H, Beishuizen E, Sijpkens Y, Faiq M, Numans M, Groenwold R, 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study' 57 *e Clinical Medicine*, 2023, 9.

[24] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.

1.2 Statement of Problem

This study examines the context-specific risks posed by advanced AI models to clinical medicine in African countries, and whether legally mandated pre-deployment safety evaluations tailored to local contexts can mitigate these risks while preserving the benefits of advanced AI integration.

1.3 Research Objectives

1. To examine the role of context in AI safety.
2. To examine the general and context-specific benefits and risks of advanced AI models in healthcare, with a significant emphasis on identifying and analysing risks specific to the African context.
3. To examine the methodological, legal and institutional landscape of pre-deployment evaluations.
4. To explore how context-specific pre-deployment AI safety evaluations can be operationalised through legal obligations to address the unique risks posed by the integration of advanced AI models into African healthcare while balancing the benefits of such integration.

1.4 Research Questions

1. What is the role of context in AI safety?
2. What are the general and context-specific benefits and risks of deploying advanced AI in African healthcare systems?
3. What is the methodological, legal and institutional landscape of pre-deployment evaluations?
4. How can context-specific pre-deployment AI safety evaluations be operationalised through legal obligations to address the unique risks posed by the integration of advanced AI models into African healthcare while balancing the benefits of such integration?

1.5 Hypothesis

There are context-specific risks posed by advanced AI systems to clinical medicine in African countries. Legally mandating pre-deployment AI safety evaluations tailored to local contexts will likely mitigate these risks while preserving the benefits of advanced AI integration.

1.6 Justification

This study is crucial for addressing the gap in ensuring safe and effective integration of advanced AI models into African healthcare infrastructure. It will directly benefit policymakers and legislators by offering or suggesting tailored legal obligations to be enforced in support of pre-deployment safety evaluations, enabling them to mitigate risks unique to clinical medicine in Africa. Regulatory bodies and AI oversight agencies in Africa will gain actionable insights into designing (and enforcing) context-specific evaluation protocols, enhancing their capacity to safeguard healthcare (or other critical infrastructure) systems against unintended harms from advanced AI integration. Additionally, researchers and practitioners in AI governance and healthcare will find the study valuable for advancing discussions on the intersection of (sociotechnical) AI safety and public health in Africa, potentially guiding future research and policy innovations in AI safety and health infrastructure protection.

1.7 Conceptual Framework: The Precautionary Principle

The precautionary principle emerged prominently in environmental discourse,[25] advocating that in the face of serious or irreversible harm, scientific uncertainty should not delay protective measures.[26] Rooted in Principle 15 of the 1992 Rio Declaration, it called for preventive action to protect the environment even without conclusive evidence of harm.[27] This principle also represented a significant shift in regulatory paradigms, placing the burden of proof on those proposing potentially harmful activities to demonstrate their safety.[28]

In public health, the precautionary principle extends beyond environmental protection to encompass preventive actions addressing threats to human well-being.[29] It emphasises

[25] Ahteensuu M, 'Rationale for taking precautions: Normative choices and commitments in the implementation of the precautionary principle', 2-3–<https://www.kent.ac.uk/scarr/events/ahteensuu.pdf> on 31 December 2024; Goldstein B, 'Precautionary principle and public health' 91(9) *American Journal of Public Health*, 2001, 1358.
[26] Martuzzi M and Tickner J, 'Introduction – the precautionary principle: protecting public health, the environment and the future of our children' in Martuzzi M and Tickner J (eds), *The precautionary principle: protecting public health, the environment and the future of our children*, WHO Regional Office for Europe, Copenhagen, 2004, 7.
[27] UNGA, *Rio declaration on environment and development*, UN A/CONF.151/26 (Vol. I), 3-14 June 1992. "In order to protect the environment, the precautionary approach shall be widely applied by States according to their capabilities. Where there are threats of serious or irreversible damage, lack of full scientific certainty shall not be used as a reason for postponing cost-effective measures to prevent environmental degradation."
[28] Public Health Ontario, 'Focus On: Precautionary Principle—Applications Relevant to Public Health Emergency Preparedness' October 2022, 3–<https://www.publichealthontario.ca/-/media/Documents/P/2021/precautionary-principle-emergency-preparedness.pdf?sc_lang=en> on 31 December 2024; Fischer A and Ghelardi G, 'The precautionary principle, evidence-based medicine, and decision theory in public health evaluation' 4 *Frontiers in Public Health*, 2016, 2.
[29] Regulatory Policy Committee, 'RPC guidance note on 'using the precautionary principle'' UK Interdepartmental Liaison Group on Risk Assessment, 1–<

proactive measures and minimal net risks in contexts where scientific uncertainty prevails, such as emerging diseases like COVID-19,[30] zootonic diseases such as Avian Influenza and Q-fever,[31] novel pharmaceuticals,[32] genetically modified crops or organisms,[33] or complex interventions like artificial intelligence in healthcare.[34] Public health practitioners and regulators are increasingly adopting the principle to anticipate and mitigate risks, even when evidence remains inconclusive.[35]

Despite its successes, the principle has faced criticism. Critics argue it may stifle innovation by imposing stringent safety assurances that discourage the development of novel technologies.[36] Additionally, some contend that the principle is too vague to offer clear guidance, as its ambiguity can justify contradictory decisions.[37] Another common objection—the "paralysis" argument—claims that applying the principle could lead to endless hesitation by requiring precautionary measures to be evaluated under the same uncertain conditions as the original action.[38] Finally, some argue that by shifting the burden of proof, it raises questions about the practicality of implementing precaution as proving absolute safety is virtually impossible.[39] Nonetheless, its core attributes render it a robust foundation for addressing uncertainties in public health.

https://assets.publishing.service.gov.uk/media/5e21c408e5274a6c3be72203/short_guidance_note_-_precautionary_principle.pdf> on 31 December 2024.

[30] Lang I, ''Laws of fear' in the EU: Precautionary principle and public health restrictions to free movement of persons in the time of COVID-19' 14(1) *European Journal of Risk Regulation*, 2021, 141-142; Birch J, 'Preparing for the next pandemic: A case for precautionary thinking and citizens' assemblies' PhilArchive, 2024, 4; Wu X, Ma L, Low D, Sharma S, and Papyshev G, 'Beyond precautionary principle: policy-making under uncertainty and complexity' 7(1) *Policy Design and Practice*, 2024, 1-2.

[31] Van Herten J and Bovenkerk B, 'The precautionary principle in zoonotic disease control' 14(2) *Public Health Ethics*, 2021, 180.

[32] Alban S, 'The 'precautionary principle' as a guide for future drug development' 35(1) *European Journal of Clinical Investigation*, 2005, 1.

[33] Goklany I, 'From precautionary principle to risk–risk analysis' 20 *Nature Biotechnology*, 2002, 1075; National Collaborating Centre for Healthy Public Policy, 'Public policies guided by the precautionary principle' May 2009, 5–< https://www.ncchpp.ca/docs/VBeloinAnC_MEP.pdf> on 31 December 2024.

[34] Botes M, 'Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI' 119(5) *South African Journal of Science*, 2023, 1.

[35] Parrott L, 'Health and risk policymaking, the precautionary principle, and policy advocacy' in *Oxford Research Encyclopaedia of Communication, Oxford University Press, Oxford, 2017, 1, 25-26.*

[36] Pearce N, 'Public health and the precautionary principle' in Martuzzi M and Tickner J (eds), *The precautionary principle: protecting public health, the environment and the future of our children*, WHO Regional Office for Europe, Copenhagen, 2004, 58.

[37] National Collaborating Centre for Healthy Public Policy, 'Public policies guided by the precautionary principle' May 2009, 7-8–< https://www.ncchpp.ca/docs/VBeloinAnC_MEP.pdf> on 31 December 2024.

[38] National Collaborating Centre for Healthy Public Policy, 'Public policies guided by the precautionary principle' May 2009, 7-8–< https://www.ncchpp.ca/docs/VBeloinAnC_MEP.pdf> on 31 December 2024.

[39] Pearce N, 'Public health and the precautionary principle', 58.

Central to this study's hypothesis is the assertion that African clinical medicine could face unique risks from the integration of advanced AI in healthcare infrastructure. The precautionary principle offers a framework for addressing and mitigating these risks by prioritising preventive measures tailored to the continent's specific needs, even in the face of uncertainty about these risks. Applying precaution, and thus, proactive harm reduction, to the integration of advanced AI in healthcare infrastructure entails rigorous pre-deployment testing and ensuring the validation of local health contexts. Furthermore, continuous monitoring for unintended consequences as well as regulation that operationalises these actions matter.

1.8 Literature Review

1.8.1 On the role of context in AI safety

Context is important to AI safety because it shapes both the risks and the benefits associated with AI deployment. AI systems do not operate in isolation; they are deeply embedded within dynamic social, cultural, and operational environments that influence their behaviour and impact.[40] As highlighted by Moses Alabi and Tobi Holmes, understanding these environments is crucial to ensuring that AI technologies align with societal goals and do not perpetuate unintended harms.[41] Furthermore, as Chen Chen et al. demonstrate, applying AI in contexts for which it was not designed can amplify biases and reduce its efficacy.[42] Addressing these challenges requires a sociotechnical approach.[43] Without integrating context into AI safety frameworks, the potential for unintended consequences and reduced accountability increases, undermining trust and the broader societal adoption of AI.[44]

[40] Kilian K, 'Beyond accidents and misuse: Decoding the structural risk dynamics of artificial intelligence' ArXiv, 2024, 2-3.
[41] Alabi M and Holmes T, 'AI safety and ethics: Developing robust frameworks for ethical AI development and deployment' ResearchGate, 2024, 8.
[42] Chen C, Liu Z, Jiang W, Goh S, Lam K, 'Trustworthy, responsible, and safe AI: A comprehensive architectural framework for ai safety with challenges and mitigations' ArXiv, 2024, 18.
[43] Curtis S , Iyer R, Kirk-Giannini C, Krakovna V, Krueger D, Lambert N, Marnette B, McKenzie C, Michael J, Miyazono E, Mima N, Ovadya A, Thorburn L, and Turan D, 'Research agenda for sociotechnical approaches to AI safety' Social Science Research Network, 2024, 1-4.
[44] Hendryks D, 'Introduction to AI safety, ethics, and society' ArXiv, 2024, xv; Saberi M, 'The human factor in AI safety' ArXiv, 2022, 1-6; Hari A and Abdulla M, 'AI Safety: where do we stand presently?' Indian Institute of Management Kozhikode, Working Paper Number IIMK/WPS/584/ITS/2023/07, August 2023, 11–< https://iimk.ac.in/uploads/publications/IIMKWPS584ITS202307.pdf> on 23 November 2024; Chen B and Metcalf J, 'Explainer: A sociotechnical approach to AI policy' Data and Society, Policy Brief, 1-5–< https://datasociety.net/wp-content/uploads/2024/05/DS_Sociotechnical-Approach_to_AI_Policy.pdf> on 23 November 2024.

1.8.2 On the general and context-specific risks that advanced AI models could pose to clinical medicine

The literature on the general risks of advanced AI models in healthcare, summarised comprehensively by the European Union,[45] identifies seven critical areas of concern: patient harm due to AI errors, misuse of biomedical AI tools, bias in medical AI, lack of transparency, privacy and security vulnerabilities, gaps in accountability, and implementation obstacles.

In addition, there are other context-specific risks. For starters, Africa's linguistic diversity, spanning over 2,000 languages,[46] complicates AI systems' ability to provide accurate diagnoses or communicate treatment plans, as tools trained predominantly in English or French may misinterpret critical medical information in languages like Kiswahili or Hausa.[47] Moreover, cultural factors, such as the World Health Organisation's (WHO) estimation that over 80% of the African populace rely on traditional medicine,[48] further risk misaligned AI recommendations that could undermine trust and discourage adoption of AI in healthcare. Additionally, Africa's distinct genetic and epidemiological profiles exacerbate the limitations of AI diagnostic tools predominantly trained on Western data,[49] potentially leading to misdiagnoses, such as underestimating cardiovascular risks due to differences in genetic markers and disease patterns.[50]

---

[45] Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, I-III.

[46] Ritchie S, Cheng Y, Chen M, Matthews R, Esch D, Li B, and Sim K, 'Large vocabulary speech recognition for languages of Africa: multilingual modelling and self-supervised learning' ArXiv, 2022, 1.

[47] Kibuacha F, 'Closing the gap: How local context improves AI performance in emerging regions' GeoPoll, 25 September 2024–< https://www.geopoll.com/blog/closing-the-ai-gap-local-context/> on 30 November 2024.

[48] WHO, 'African Traditional Medicine Day 2022' 31 August 2022–< https://www.afro.who.int/regional-director/speeches-messages/african-traditional-medicine-day-2022#:~:text=Traditional%20medicine%20has%20been%20the,for%20their%20basic%20health%20needs> on 23 November 2024; Anyikwa C, 'Exploring the role of divination in traditional medicine in Africa', 1-4; BBC, 'The Kenyan enthralled by the healing power of plants' 29 July 2024–< https://www.bbc.com/news/articles/c84j0kzgnk9o> on 30 November 2024.

[49] Chinta S, Wang Z, Zhang X, Viet T, Kashif A, Smith M, and Zhang W, 'AI-driven healthcare: A survey on ensuring fairness and mitigating bias' ArXiv, 2024, 5; Kist J, Vos R, Mairuhu A, Struijs J, van Peet P, Vos H, van Os H, Beishuizen E, Sijpkens Y, Faiq M, Numans M, Groenwold R, 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study' 57 *e Clinical Medicine*, 2023, 9.

[50] Shishehbor F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18; Martin G, Cirino A, Barcelona V, Fox K, Hudson M, Sun Y, Taylor J, and Cameron V, 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples: A scientific statement from the American Heart Association' 14(4) *Circulation: Genomic and Precision Medicine*, 2021, 547, 551; See also, Peprah E, Xu H, Ayele F, and Royal C, 'Genome-wide association studies in Africans and African Americans: Expanding the framework of the genomics of human traits and disease' 18(1) *Public Health Genomics,* 2015, 40–51; Bentley Amy, Callier S, and Rotimi C, 'Evaluating the promise of inclusion of African ancestry populations in genomics' 5(5) *Genomic Medicine*, 2020, 1-9.

1.8.3 Contribution to the literature

While existing literature explores general and specific risks posed by (advanced) AI to healthcare and global efforts to mitigate them, it does not adequately consider how these risks manifest uniquely in African contexts, nor does it provide a framework for addressing these unique risks through African legal mechanisms. This study will bridge this gap by highlighting the context-specific risks posed to clinical medicine in Africa by the integration of advanced AI models in healthcare infrastructure. It further proposes actionable and context-sensitive legal obligations to be mandated with regard to pre-deployment safety evaluations for advanced AI models in African healthcare infrastructure.

1.9 Methodology

To examine the role of context in AI safety, the study will gather academic work and blogs addressing the role of context in sociotechnical AI safety, or work making the case for contextual considerations in AI safety. It will also review literature on how contextual factors like cultural norms and resource availability influence safety outcomes generally. It will then synthesise findings to argue how context impacts AI safety.

To investigate general and context-specific benefits and risks of advanced AI (in African clinical medicine), the study will use content analysis. It will gather literature on advanced AI's (likely) impact (on African clinical medicine) and identify benefits as well as patterns of risk. It will also review literature on the weaknesses that advanced AI models have generally, for example, biases based on training data, to extract any health-related weaknesses that are (or would be) particularly severe in Africa. It will look into literature on unique benefits of advanced AI integration to African clinical medicine as well. It will synthesise all these findings into the general and context-specific benefits and risks advanced AI models pose to (African) clinical medicine.

To outline legal obligations for African countries, the study will review existing legal frameworks and guidelines on AI safety in healthcare. It will synthesise findings to identify gaps and propose specific legal obligations for AI safety evaluations in African healthcare. It will also gather feedback from experts to ensure practicality and relevance.

The main research question will be approached deductively, arguing that (i) the effectiveness of AI safety evaluations is significantly influenced by the context in which they are applied,

(ii) advanced AI systems pose significant and context-specific risks to African clinical medicine, (iii) legally mandated, context-specific pre-deployment AI safety evaluations may address the unique risks posed by the integration of advanced AI models into African healthcare.

1.10 Chapter Breakdown

Chapter 1 provides an overview of the study.

Chapter 2 will examine the role of context in AI safety and make the case that contextually agnostic evaluations are inadequate.

Chapter 3 will assess, in particular, the context-specific benefits and risks of advanced AI integration into Africa's healthcare infrastructure.

Chapter 4 will examine the methodological, legal and institutional landscape of pre-deployment evaluations.

Chapter 5 will propose context-specific legal obligations for pre-deployment safety evaluations of advanced AI systems in Africa that are pending deployment in clinical medicine.

Chapter 6 will summarise the key findings and highlight areas for future research.

2 The Role of Context in AI Safety

2.1 Introduction

The previous chapter outlined, inter alia, the research questions, conceptual framework, and methodology of the study. This chapter examines how AI safety has evolved from narrow technical concerns to sociotechnical frameworks. It critiques early definitions focused on existential risks, arguing that real-world harms—like bias—demand attention to deployment contexts. This chapter thus establishes why context is indispensable for meaningful safety assessments.

2.2 The essence of AI safety

2.2.1 Technical AI safety

Technical AI safety has often been framed in precise, technical terms. While early thinkers in the mid-20th century had speculated about the potential dangers of intelligent machines, it was researchers in the last decade who brought focused attention to the risks posed by artificial general intelligence (AGI) and superintelligence, thus pioneering technical AI safety as a field of study.[51] They laid out the alarming possibility that once an unfriendly superintelligence came into existence, its inherent drive to preserve and enhance its capabilities would prevent any meaningful human intervention—effectively sealing our fate.[52]

From this, it is evident that the main concern in technical AI safety is misalignment: if a highly capable AI system were to pursue goals even slightly misaligned with human interests, the consequences could be catastrophic or even existential.[53] This view, despite some scepticism,[54]

[51] Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[52] Christiano P, Schlegeris P, and Amodei D, 'Supervising strong learners by amplifying weak experts' ArXiv, 2018, 3; Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[53] Everitt T, Lea G, and Hutter M 'AGI safety literature review' ArXiv, 2018, 2, 8; Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[54] Sofge E, 'Bill Gates Fears AI, but AI researchers know better' Popular Science, 15 January 2015–< https://www.popsci.com/bill-gates-fears-ai-ai-researchers-know-better/> on 11 February 2025; Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-

frames safety as a problem of *technical* control—ensuring that AI systems obey their intended constraints and do not (recursively) develop objectives or capabilities that could spiral out of human control.[55] Consequently, technical AI safety is predicated on the stark idea that misaligned AI could spell our extinction, and it is dedicated to preventing such a catastrophic failure.[56]

From this technical perspective, issues such as algorithmic bias, fairness, transparency, and misuse of AI in real-world social systems are generally regarded as problems or matters of policy, ethics, and governance rather than fundamental safety risks.[57] This is illustrated by the fact that, as of 2018, it was estimated that only a small number of experts were dedicated full-time to AI safety issues, and most of their efforts were directed at preventing existential risks.[58]As such, other AI-related issues are, if not ignored, certainly relegated to the back burner in technical AI safety.

2.2.2 Sociotechnical AI safety

With AI systems becoming more embedded into daily life and more evidence emerging of (intended[59] or unintended[60]) tangible and immediate harms from the use of AI (e.g., inter alia,[61]

safety-alignment> on 11 February 2025; Etzioni O, 'No, the experts don't think superintelligent AI is a threat to humanity' MIT Technology Review, 20 September 2016–< https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/> on 11 February 2025.

[55] Ortega P, Maini V, DeepMind Safety Team, 'Building safe artificial intelligence: specification, robustness, and assurance' Medium, 27 September 2018–< https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1> on 11 February 2025; Everitt T, Lea G, and Hutter M 'AGI safety literature review' ArXiv, 2018, 2, 8; Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[56] Dafoe A and Russell S, 'Yes, we are worried about the existential risk of artificial intelligence' MIT Technology Review, 2 November 2016–< https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/> on 11 February 2025.

[57] Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[58] Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment> on 11 February 2025.

[59] Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen G C, Steinhardt J, Flynn C, Ó hÉigeartaigh S, Beard S, Belfield H, Farquhar S, Lyle C, Crootof R, Evans O, Page M, Bryson J, Yampolskiy R, and Amodei D, 'The malicious use of artificial intelligence: Forecasting, prevention, and mitigation' ArXiv, 2018, 10-11.

[60] Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D, 'Concrete problems in AI safety' ArXiv, 2016, 1-2, 21.

[61] IBM, 15 November 2024–< https://www.ibm.com/think/topics/ai-safety#:~:text=Cybersecurity-,Bias%20and%20fairness,lead%20to%20unfair%20decision%2Dmaking.> on 11 February 2024. IBM lists all of them as risks from AI.

entrenching inequality and discrimination,[62] and spreading misinformation[63]), some scholars argue that AI safety has to include *sociotechnical* considerations—factors that recognise the interdependence between social (human) and technical components in a system.[64] One illustration of this interdependence is that of a car traveling on a road. This scenario exemplifies a sociotechnical system[65] since it requires the driver, who is human, to work in harmony with the car, a technological creation, to arrive at a specific location.[66] In the same way, researchers call for a *sociotechnical* approach to AI, and, in particular, a definition of AI safety that embraces these sociotechnical aspects.[67]

Seth Lazar and Alondra Nelson are two of the leading proponents of this perspective. They contend that AI safety cannot be meaningfully separated from the contexts in which AI systems operate.[68] Lazar, Nelson and others (e.g., researchers at Google DeepMind) contend that safety should be about more than just technical robustness; it should also account for whether AI systems reinforce existing inequities, exacerbate human error, or fail to integrate appropriately into institutional and social structures.[69] Similar views are expressed by Roel Dobbe,[70] the

[62] Leslie D, 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector' The Alan Turing Institute, 2019, 4–<https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf> on 11 February 2024; Oduro S, 'Without accounting for bias and discrimination, AI safety standards will fall short' Tech Policy Review, 16 September 2024–< https://www.techpolicy.press/without-accounting-for-bias-and-discrimination-ai-safety-standards-will-fall-short/> on 11 February 2025.

[63] Center for AI Safety, *Impact Report*, 2023, 4.

[64] Kylian K, 'Beyond accidents and misuse: Decoding the structural risk dynamics of artificial Intelligence' ArXiv, 2024, 1; Chen B and Metcalf J, 'Explainer: A sociotechnical approach to AI policy' Data and Society, Policy Brief, 3–<https://datasociety.net/wp-content/uploads/2024/05/DS_Sociotechnical-Approach_to_AI_Policy.pdf> on 23 November 2024; Mumford E, 'The story of socio-technical design: Reflections on its successes, failures and potential' 16(1), *Info Systems Journal*, 2006, 318.

[65] Mumford E, 'A socio-technical approach to systems design' 5(1), *Requirements Engineering*, 2000, 125; Trist E, 'The evolution of socio-technical systems: A conceptual framework and an action research program' Ontario Quality of Working Life Centre, Occasional Paper Number 2, 1981, 7-9, < https://www.lmmiller.com/blog/wp-content/uploads/2013/06/The-Evolution-of-Socio-Technical-Systems-Trist.pdf> on 16 December 2024.

[66] Hendryks D, 'Introduction to AI safety, ethics, and society' ArXiv, 2024, 201.

[67] Chen B and Metcalf J, 'Explainer: A sociotechnical approach to AI policy' Data and Society, Policy Brief, 2–<https://datasociety.net/wp-content/uploads/2024/05/DS_Sociotechnical-Approach_to_AI_Policy.pdf> on 23 November 2024.

[68] Lazar S and Alondra N, 'AI safety on whose terms?' 381 (6654) *Science*, 2023, 138.

[69] Lazar S and Alondra N, 'AI safety on whose terms?', 138; Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023, 8; Harding J and Valenzuela C,' Workshop on sociotechnical AI safety' Stanford School of Humanities and Sciences, McCoy Family Centre for Ethics in Society, July 2024–< https://hai.stanford.edu/sites/default/files/2024-07/SociotechnicalAI_Report_v3.pdf> on 17 December 2024; Curtis S, Iyer R, Kirk-Giannini C, Krakovna V, Krueger D, Lambert N, Marnette B, McKenzie C, Michael J, Miyazono, Mima N, Ovadya A, Thorburn L, and Turan D, Research agenda for sociotechnical approaches to AI safety' Social Science Research Network, 1,3.

[70] Dobbe R and Wolters A, 'Toward sociotechnical AI: Mapping vulnerabilities for machine learning in context' 34(12) *Minds and Machines*, 2024, 1-2; Rismani S, Dobbe R, and Moon A, 'From silos to systems: Process-

National Artificial Intelligence Advisory Committee (NAIAC),[71] and the Institute for Trustworthy AI in Law and Society.[72]

This sociotechnical perspective has gained traction recently,[73] but it has also faced pushback. Eliezer Yudkowsky and those in the existential risk camp argue that including a wider range of issues under AI safety risks diluting its core focus.[74] If "safety" encompasses everything from bias mitigation to responsible governance, then it becomes harder to prioritise existential threats—the risks that, if realised, could lead to human extinction or irreversible harm.[75] In addition, we could lose clarity on what *actually* constitutes a catastrophic AI failure.

2.2.3 Where the discourse stands—and what definition to adopt

Defining AI safety is fraught with tension. On one hand, AI safety cannot ignore the fact that real-world harms are often driven by sociotechnical factors. Technical fixes alone—better algorithms, improved evaluation and verification techniques, etc—are not enough if the deployment environment introduces new risks. On the other hand, as Yudkowsky argues, if AI safety is stretched too broadly, it risks becoming an unfocused catch-all term that loses its *urgency* when applied to genuinely catastrophic risks. Both positions are defensible, yet adopting one exclusively undermines the other.

For this study, a middle ground is necessary. But to what extent should the broader, sociotechnical definition be narrowed or qualified? Here, AI safety is defined as: the study and practice of mitigating risks of "serious" harm from AI—whether catastrophic or societal—by (1) ensuring AI systems operate reliably within their intended constraints; and (2) accounting for the contexts in which they are deployed. This definition tries to retain precision by focusing

oriented hazard analysis for AI systems' ArXiv, 2024, 3; Raji D and Dobbe R, 'Concrete problems in AI safety revisited' ArXiv, 2023, 1.

[71] National Artificial Intelligence Advisory Committee, 'Findings and recommendations: AI safety' May 2024, 2-7–< https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf> on 17 December 2024.

[72] Institute for Trustworthy AI in Law and Society, *The importance of a socio-technical approach in AI development*, 2 February 2024, 5.

[73] Bogen M and Wincecoff A, 'Applying sociotechnical approaches to AI governance in practice' Center for Democracy and Technology, May 2024, 2–< https://cdt.org/wp-content/uploads/2024/05/2024-05-23-AI-Gov-Lab-applying-sociotechnical-approaches-to-ai-governance-in-practice-final.pdf> on 17 December 2024.

[74] Albergotti R, 'The risks of expanding the definition of 'AI safety'' Semafor, 8 March 2024–< https://www.semafor.com/article/03/08/2024/the-risks-of-expanding-the-definition-of-ai-safety> on 11 February 2025.

[75] Albergotti R, 'The risks of expanding the definition of 'AI safety'' Semafor, 8 March 2024–< https://www.semafor.com/article/03/08/2024/the-risks-of-expanding-the-definition-of-ai-safety> on 11 February 2025.

on serious risks (e.g., existential threats, large-scale societal harm) but does not limit "seriousness" to extinction scenarios. For instance, in African clinical medicine, an algorithm that misdiagnoses a condition and disproportionately harms marginalised groups across the continent constitutes a serious risk, even if not existential. This definition also acknowledges context since what qualifies as (sufficiently) "serious" remains context-dependent.

It is an imperfect definition, admittedly, but it is a modest attempt to reconcile the competing demands of (i) precision and breadth; and (ii) urgency and practicality. Section 2.3 will explore this further, demonstrating more deeply why context (or sociotechnical considerations) matters in AI safety.

2.3 How context matters in AI safety

2.3.1 AI systems often reveal unexpected harms after deployment

AI systems, designed with specific parameters in mind, often exhibit unforeseen, unpredictable, and, at times, startling behaviours when deployed in settings for which they were not originally crafted.[76] Consider, for instance, facial recognition models that were trained on predominantly Western datasets—when these systems are applied in more culturally diverse environments, their error rates and misidentification issues can spike unexpectedly.[77] In fact, these models have demonstrated significantly higher error rates when used on black people.[78] A 2019 study by the U.S. National Institute of Standards and Technology (NIST) revealed a concerning disparity in facial recognition accuracy, with nearly 200 algorithms showing error rates up to 100 times higher for individuals from certain African and Asian groups compared to white individuals.[79] This problem was not just technical; it highlighted a significant epistemic blind spot in dataset curation. Notably, algorithms developed in Asian countries, which were trained on more diverse datasets, did not exhibit such high error rates,[80] indicating the importance of inclusive data practices.

[76] Dobbe R and Wolters A, 'Toward sociotechnical AI', 31-32.
[77] NIST, 'Effects of race, age, sex on face recognition software: Demographics study on face recognition algorithms could help improve future tools' NIST, 19 December 2019–< https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software> on 11 February 2025.
[78] Sartori L and Theodorou A, 'A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control' 24(4) *Ethics and Information Technology*, 2022, 7.
[79] NIST, 'Effects of race, age, sex on face recognition software: Demographics study on face recognition algorithms could help improve future tools' NIST, 19 December 2019–< https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software> on 11 February 2025.
[80] NIST, 'Effects of race, age, sex on face recognition software: Demographics study on face recognition algorithms could help improve future tools' NIST, 19 December 2019–< https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software> on 11 February 2025.

Similarly, language models trained predominantly on English and other widely spoken languages fail when interacting with African dialects, often misinterpreting context or defaulting to nonsensical responses.[81] Furthermore, there have been documented incidents where language models produced outputs that were culturally insensitive or ethically questionable when used in contexts far removed from those in which they were developed.[82] The harms resulting from natural language processing models could include restricted access of certain demographics to financial services,[83] misdiagnoses from the use of AI systems in healthcare,[84] or even wrongful arrests when predictive policing tools misclassify individuals of a certain race, gender, age, or socio-economic status.[85] The truth of the matter is: certain populations usually remain invisible until harm materialises.

It may be argued, though, as Miranda Bogen does, that unpredictability is an inherent risk of deploying sophisticated AI systems.[86] While that is indeed true, the evidence suggests, as in the case of bias mitigation strategies in predictive policing tools,[87] that certain risks are largely preventable through more rigorous attention to deployment context. It would thus behoove policymakers, AI developers, AI researchers, and other stakeholders to recognise and plan for

[81] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; Nicholas G and Bhatia A, 'Lost in translation: large language models in non-English content analysis' Center for Democracy and Technology, May 2023, 6–< https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf> on 31 December 2024.

[82] Liu Z, 'Cultural bias in large language models: A comprehensive analysis and mitigation strategies' *Journal of Transcultural Communication*, 2024, 5–<https://www.degruyter.com/document/doi/10.1515/jtc-2023-0019/html?lang=en&srsltid=AfmBOopRb9ckAFwhPzVkm7_pD36APVHZ7tDHv5aOF5-pWw20MVDbBe_w#Chicago> on 11 February 2025; Mhasakar M, Baker-Ramos R, Carter B, Halekahi-Kaiwi E, Hester J, '"I would never trust anything western": Kumu (educator) perspectives on use of LLMs for culturally revitalizing CS education in Hawaiian Schools' ArXiv, 2025, 2.

[83] Financial Stability Institute, *Regulating AI in the financial sector: recent developments and main challenges*, 2024, 10; De Nederlandsche Bank, *The impact of AI on the financial sector and supervision*, 2024, 17.

[84] Taha A, 'AI risks in global health 'must be accounted for' – WHO' SciDev, 23 January 2024–< https://www.scidev.net/global/news/ai-risks-in-global-health-must-be-accounted-for-who/> on 28 February 2025; Salvi P, 'AI Misdiagnosis' Salvi, Schostok and Pritchard, 10 April 2024–< https://www.salvilaw.com/blog/ai-misdiagnosis/> on 28 February 2025.

[85] Almasoud A and Idowu J, 'Algorithmic fairness in predictive policing' *AI and Ethics*, 2024, 2–< https://link.springer.com/article/10.1007/s43681-024-00541-3> on 28 February 2025; US Department of Justice, *Artificial intelligence and criminal justice*, 2024, 46-49, 53.

[86] US Privacy and Civil Liberties Board, 'CDT's Miranda Bogen: Opening Statement' 11 July 2024–< https://documents.pclob.gov/prod/Documents/EventsAndPress/ff4e3c3a-00c5-4c3d-b501-73ad8780a4ab/Miranda%20Bogen%20Opening%20Statement%20-%20Completed%20508%20-%20Jul%2019,%202024.pdf> on 28 February 2024.

[87] Almasoud A and Idowu J, 'Algorithmic fairness in predictive policing', 1,8,13.

these contextual discrepancies, in order to better anticipate—and thereby mitigate—the emergence of unintended harms.

2.3.2 "Harm" cannot be understood away from its context

The notion of "harm" is far from a universal constant—it is deeply embedded within the specific social, cultural, and environmental reality in which it occurs. For example, an event that might be classified as a minor technical failure in one domain could have profound, even catastrophic, implications in another. For instance, in IT infrastructure, a software glitch in a corporate email server—such as an unexpected downtime due to a misconfigured update—might disrupt communication for hours or even days, but its consequences are generally limited to lost productivity and minor financial costs. The US National Institute of Standards and Technology (NIST) classifies such incidents as low-impact since they disrupt operations without completely halting service delivery.[88]

In stark contrast, the aviation industry treats similar technical failures as very serious threats as they can have far more severe consequences. For example, the 2018–2019 Boeing 737 MAX crashes, investigated by the U.S. House Committee on Transportation and Infrastructure, were directly linked to a single angle-of-attack sensor failure in the Manoeuvring Characteristics Augmentation System (MCAS).[89] This software flaw misinterpreted sensor data, overrode pilot inputs, and forced the planes into fatal nosedives, killing 346 people.[90] Here, what might be considered a minor technical flaw in another domain became catastrophic in aviation, illustrating how harm escalates when human lives are directly at risk.

---

[88] NIST, *Computer security incident handling guide: Recommendations of the National Institute of Standards and Technology*, 2012, 33. In Table 3-2, incidents are categorised as "Low" when they cause only a minimal effect—that is, when the organisation can still deliver all critical services despite a loss of efficiency. Although the document does not explicitly mention "temporary email server downtime," such an event would fall under this "Low" category since it disrupts operations without completely halting service delivery.

[89] Gates D, Baker M, 'The inside story of MCAS: How Boeing's 737 MAX system gained power and lost safeguards' Association of Flight Attendants, 22 June 2019–< https://www.afacwa.org/the_inside_story_of_mcas_seattle_times> on 28 February 2025; The US House Committee on Transport and Infrastructure, 'After 18-Month Investigation, Chairs DeFazio and Larsen Release Final Committee Report on Boeing 737 MAX' 16 September 2020–< https://democrats-transportation.house.gov/news/press-releases/after-18-month-investigation-chairs-defazio-and-larsen-release-final-committee-report-on-boeing-737-max> on 28 February 2025.

[90] Gates D, Baker M, 'The inside story of MCAS: How Boeing's 737 MAX system gained power and lost safeguards' Association of Flight Attendants, 22 June 2019–< https://www.afacwa.org/the_inside_story_of_mcas_seattle_times> on 28 February 2025; The US House Committee on Transport and Infrastructure, 'After 18-Month Investigation, Chairs DeFazio and Larsen Release Final Committee Report on Boeing 737 MAX' 16 September 2020–< https://democrats-transportation.house.gov/news/press-releases/after-18-month-investigation-chairs-defazio-and-larsen-release-final-committee-report-on-boeing-737-max> on 28 February 2025.

A similar contrast can be seen in battery failures. The US Consumer Product Safety Commission (CPSC) categorised the 2016 Samsung Galaxy Note 7 battery malfunctions—which caused devices to overheat or combust—as a consumer safety hazard, prompting a mass recall but no fatalities.[91] Though disruptive, these defects were treated as manageable through refunds and replacements,[92] indicating that they were framed as inconveniences. However, when identical battery flaws occurred in medical devices, such as Medtronic's MiniMed insulin pumps, the US Food and Drug Administration (FDA) declared a Class I recall—the most severe designation[93]—as faulty batteries risked insulin under- or over-delivery, potentially causing hyperglycaemia, diabetic ketoacidosis, or death.[94] The same technical failure, a product safety question in smartphones, became lethal in healthcare, underscoring how harm is not inherent to the defect itself but to the context in which it occurs: a smartphone battery flaw inconveniences users, while a medical device failure endangers lives.

Based on the discussion above, regarding AI safety, harm is not merely a function of a system's *capabilities* but a reflection of the context in which those capabilities are deployed. That context, including societal values, differs markedly across regions and communities. What one society deems a negligible inconvenience might be viewed as having serious and pernicious effects by another. If evaluators of advanced AI systems recognise these differences, they can better anticipate the specific risks associated with each deployment scenario.

[91] CPSC, 'Samsung recalls galaxy note7 smartphones due to serious fire and burn hazards' 15 September 2016–< https://www.cpsc.gov/Recalls/2016/Samsung-Recalls-Galaxy-Note7-Smartphones> on 28 February 2025.
[92] CPSC, 'Samsung recalls galaxy note7 smartphones due to serious fire and burn hazards' 15 September 2016–< https://www.cpsc.gov/Recalls/2016/Samsung-Recalls-Galaxy-Note7-Smartphones> on 28 February 2025.
[93] FDA, 'What is a medical device recall? 26 September 2018–< https://www.fda.gov/medical-devices/medical-device-recalls/what-medical-device-recall> on 28 February 2025.
[94] FDA, 'Insulin pump recall: Medtronic notifies users of MiniMed 600 and 700 series pumps of risk of shorter than expected battery life' 17 October 2024–< https://www.fda.gov/medical-devices/medical-device-recalls/insulin-pump-recall-medtronic-notifies-users-minimed-600-and-700-series-pumps-risk-shorter-expected> on 28 February 2025.

2.3.3 Inclusivity could prevent the perpetuation of structural disadvantage

History provides various and unpleasant examples of how biased choices—whether in eugenic policies,[95] exploitative medical studies,[96] or racially segregated urban planning[97]—have entrenched inequalities. Likewise, as recognised in the AU Continental AI Strategy,[98] AI, as a tool of immense influence, can not only reflect societal biases, but also amplify, reinforce, and entrench them at scale.[99] For example, predictive policing models, trained on historically biased crime data, have been shown to reinforce racial profiling patterns, justifying increased surveillance and discriminatory policing in already marginalised communities.[100] Similarly, if financial AI models are deployed in African markets using Western credit-scoring criteria (such as mortgage history or credit card usage),[101] they risk systematically excluding large portions of African populations that rely on informal economic systems.[102] Additionally, while standardisation is useful in some respects (e.g., ensuring baseline safety measures), it can also be a blunt instrument that erases crucial differences. For example, an AI model may be deemed "fair" if it produces equal outcomes across racial groups, but in societies with deep historical

[95] Eugenicists weaponised pseudoscientific notions of 'racial improvement' to justify forced sterilisations of thousands of marginalised women. National Human Genome Research Institute, 'Eugenics and scientific racism' NHGRI, 18 May 2022–< https://www.genome.gov/about-genomics/fact-sheets/Eugenics-and-Scientific-Racism> on 28 February 2025; Patel P, 'Forced sterilization of women as discrimination' 38(15) *Public Health Reviews*, 2017, 2.

[96] The Tuskegee Syphilis Study exploited Black men under the guise of medical progress while denying them treatment. Tuskegee University, 'About the USPHS syphilis study'–< https://www.tuskegee.edu/about-us/centers-of-excellence/bioethics-center/about-the-usphs-syphilis-study> on 28 February 2025; Centre for Disease Control, 'The Untreated Syphilis Study at Tuskegee Timeline' 4 September 2024–< https://www.cdc.gov/tuskegee/about/timeline.html> on 28 February 2025.

[97] Urban planner Robert Moses engineered infrastructure—like low-clearance parkway bridges and frigid Harlem pools—to physically exclude poor and Black communities from public spaces, encoding racism and classism into the built environment. Burkeman O, 'The power broker: Robert Moses and the fall of New York by Robert Caro review – a landmark study' The Guardian, 23 October 2015–< https://www.theguardian.com/books/2015/oct/23/the-power-broker-robert-moses-and-the-fall-of-new-york-robert-caro-review> on 28 February 2025.

[98] African Union, *Continental AI Strategy*, 2024, 25.

[99] Almeida S, Norajitra T, Lüth C, Wald T, Weru V, Nolden M, Jäger P, von Stackelberg O, Heußel C, Weinheimer O, Biederer J, Kauczor H, Maier-Hein K, 'How do deep-learning models generalize across populations? Cross-ethnicity generalization of COPD detection' 15(1) *Insights into Imaging*, 2024, 3; Leslie D, 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector' The Alan Turing Institute, 2019, 4–<https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf> on 11 February 2024.

[100] European Agency for Fundamental Rights, *Getting the future right artificial intelligence and fundamental rights*, 2020, 69-70; Almasoud A and Idowu J, 'Algorithmic fairness in predictive policing', 2; US Department of Justice, *Artificial intelligence and criminal justice*, 2024, 46-49, 53.

[101] Shkurdoda A and Dobosz M, 'AI in fintech: Harnessing intelligent technologies for smarter finance' Neontri, 29 November 2024–< https://neontri.com/blog/artificial-intelligence-fintech/> on 28 February 2025.

[102] OECD, *FinTech lending in Sub-Saharan Africa lessons from African economies*, 2024, 20; OECD, *Marketplace and FinTech lending for SMEs in the COVID-19 crisis*, 2022, 18.

inequities, equal outcomes may still reproduce existing disparities.[103] In all these cases, the failure to account for contextual realities transforms AI from a tool of potential progress into a mechanism that *institutionalises* and *perpetuates* structural disadvantage.

Given the above, ensuring inclusivity in AI safety evaluations is not merely a moral imperative but a pragmatic necessity for mitigating bias and preventing the reinforcement of systemic inequalities.[104] Indeed, when evaluations are conducted by a homogenous group (white, male, advantaged), as they often are,[105] there is a real risk of overlooking (subtle) harms that disproportionately affect vulnerable communities.[106] Moreover, not only does diverse input prevent harm by redistributing power[107]—it also actively enhances the evaluative process. Research by David Rock and Heidi Grant suggests that teams comprising individuals from varied socio-economic backgrounds, geographic regions, and cultural contexts are more adept at identifying potential pitfalls and hidden biases.[108] In AI safety, this translates into more rigorous harm detection and mitigation strategies as well as other useful insights that might have otherwise remained obscured.[109]

The cumulative effect of repeatedly ignoring contextual differences in AI safety evaluations is not neutral outcomes, but rather actively reinforcing structural disadvantage against certain groups while benefiting others. Inclusivity could help here, as having diverse perspectives and localised insights at every stage of AI development and deployment would go a long way in mitigating long-term, systemic risks. While it is possible that broader inclusivity could slow down decision-making or introduce conflicting viewpoints, it is very likely the case that the benefits of inclusivity—in terms of enhanced safety and societal trust—far outweigh the potential drawbacks, particularly the inadvertent perpetuation of bias. Again, inclusivity not just preferable, it is a pragmatic necessity.

---

[103] Forbes Technology Council, 'AI & fairness metrics: Understanding & eliminating bias' Forbes, 14 July 2023–< https://councils.forbes.com/blog/ai-and-fairness-metrics> on 28 February 2025.
[104] Lazar S and Alondra N, 'AI safety on whose terms?', 138.
[105] Lazar S and Alondra N, 'AI safety on whose terms?', 138.
[106] Sætra H, 'AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system' 13 *Sustainability*, 2021, 11.
[107] Patel V, 'Cognitive science in the evaluation of medical AI systems' 30(1) *BMJ Health Care Informatics*, 2023, 2.
[108] Rock D and Grant H, 'Why diverse teams are smarter' Harvard Business Review, 4 November 2016–< https://hbr.org/2016/11/why-diverse-teams-are-smarter> on 27 February 2025.
[109] Lazar S and Alondra N, 'AI safety on whose terms?', 138.

2.3.4 Static evaluations cannot account for shifting contexts

Static evaluations in AI safety are problematic because they risk becoming detached from the rapidly evolving world in which these systems operate. Although technical methods for detecting and mitigating harms are constantly improving, if these methods do not continuously incorporate changes in deployment contexts, the very systems deemed safe today may become dangerous tomorrow.[110] For example, as geopolitical tensions rise or economic policies shift unexpectedly, the underlying assumptions of the model may no longer hold true. Without an adaptive framework, such as dynamic certification, that updates safety assessments in response to these contextual changes, AI systems' predictions could become misleading, potentially causing severe consequences.[111]

The world is in constant flux. Regularly updating safety criteria—by integrating new data, monitoring shifts in user behaviour, and adjusting to evolving regulatory environments—is not only a necessity but also a critical safeguard against emergent risks. While this approach may present challenges in terms of ongoing monitoring and resource allocation, the cost of inaction is, as put earlier, likely far greater.

2.4 Conclusion

This chapter set out to illustrate that context is integral to understanding AI safety. It demonstrated that AI safety cannot be reduced to technical robustness alone. By tracing the shift from misalignment fears to sociotechnical risks, it showed how contextual factors determine whether AI systems cause harm, and thus, why context is crucial in AI safety. The analysis justifies integrating local contexts into evaluation frameworks, a prerequisite for addressing Africa-centric risks from AI systems in clinical medicine.

[110] Bakirtzis G, Tubella A, Theodorou A, Danks D, and Topcu U, 'Navigating the sociotechnical labyrinth: Dynamic certification for responsible embodied AI' ArXiv, 2024, 1, 6, 13-14.
[111] Bakirtzis G, Tubella A, Theodorou A, Danks D, and Topcu U, 'Navigating the sociotechnical labyrinth: Dynamic certification for responsible embodied AI' ArXiv, 2024, 1, 6, 13-14.

3 Advanced AI in African Clinical Medicine: Benefits and Context-Sensitive Risks

3.1 Introduction

Building on the contextual analysis in Chapter 2, this chapter focuses on the application of advanced AI in African clinical medicine. It outlines the broad benefits—such as improved diagnostics and operational efficiencies—as well as the unique, Africa-centric advantages stemming from local genomic diversity and traditional medicinal practices. In parallel, it examines the potential pitfalls that arise when AI systems are deployed, particularly those pitfalls that arise due to lack of consideration for African context.

3.2 Benefits of AI integration in clinical medicine

3.2.1 Widespread advantages of AI in clinical medicine

AI has brought profound change to clinical medicine. In management, AI algorithms now predict equipment failures and classify patient feedback, enabling hospitals to address systemic issues before they escalate—an approach that has improved patient satisfaction and operational efficiency.[112] In diagnostics, systems powered by AI enhance accuracy and speed.[113] For example, radiological tools detect abnormalities in X-rays and CT scans with a precision that reduces errors and enables earlier intervention,[114] AI in histopathology automates the segmentation of tissues, helping to identify cancerous cells with fewer mistakes,[115] and AI-

[112] Dave M and Patel N, 'Artificial intelligence in healthcare and education' 234(10) *British Dental Journal*, 2023, 761-762; USAID Centre for Innovation and Impact, *Artificial Intelligence in global health: Defining a collective path forward*, 2019, 28-29.
[113] Zeb S, Nizamullah F, Abbasi N, and Fahad M, 'AI in healthcare: Revolutionizing diagnosis and therapy' 3(3) *International Journal of Multidisciplinary Sciences and Arts*, 2024, 118; Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat M, and Dwivedi Y 'A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities' 8 *Journal of Innovation & Knowledge*, 2023, 1-2; Nuffield Council on Bioethics, ' Artificial intelligence (AI) in healthcare and research' Bioethics Briefing Note, May 2018, 2–< https://www.nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research-1.pdf> on 29 December 2024; Kocher B, 'Using AI to your benefit in health care' NEJM Catalyst, 12 November 2024–< https://catalyst.nejm.org/doi/full/10.1056/CAT.24.0431> on 29 December 2024.
[114] Dave M and Patel N, 'Artificial intelligence in healthcare and education' 762. See also, Rudolph J, Huemmer C, Preuhs A, Buizza G, Hoppe BF, Dinkel J, Koliogiannis V, Fink N, Goller SS, Schwarze V, Mansour N, Schmidt VF, Fischer M, Jörgens M, Khaled NB, Liebig T, Ricke J, Rueckel J, and Sabel BO, 'Non-radiology health care professionals significantly benefit from AI assistance in emergency-related chest radiography interpretation' 166(1) *Chest Journal*, 2024, 157.
[115] Dave M and Patel N, 'Artificial intelligence in healthcare and education' 762.

assisted physicians diagnose wrist trauma,[116] malignant brain conditions (e.g., glioblastoma multiforme and metastatic brain tumours),[117] and other conditions[118] more accurately and swiftly than traditional methods. Finally, beyond direct medical applications, AI is revolutionising drug discovery, as exemplified by AlphaFold2's ability to predict protein structures,[119]accelerating the identification of new drugs, thereby shortening development timelines and enhancing the success of clinical trials.[120] There are, in toto, promising significant economic and social benefits from the use of AI systems in healthcare delivery.

3.2.2 Africa-centric advantages of AI in clinical medicine

Africa's extraordinary genetic diversity[121]—rooted in its status as humanity's evolutionary birthplace—[122]presents a rare opportunity to revolutionise precision medicine through AI. This genetic wealth has enabled Genome-Wide Association Studies (GWAS) to identify disease markers and genetic variations that remain undetectable in less diverse populations.[123] For instance, African genetic studies have played a pivotal role in uncovering the significance of the BCL11A gene in sickle cell disease,[124] leading to treatment breakthroughs that now benefit patients worldwide. Owing to this diversity, AI models tailored to the continent's unique genetic profile could tackle region-specific diseases like Burkitt's lymphoma. Given Africa's historical underrepresentation in genomics[125] such an advancement could transform healthcare outcomes for its people.

[116] Keller M, Rohner M, and Honigmann P, 'The potential benefit of artificial intelligence regarding clinical decision-making in the treatment of wrist trauma patients' 19(1) *Journal of Orthopaedic Surgery and Research*, 2024, 1.
[117] Ah-Shahwani and Faieq A, 'The benefit of artificial intelligence in the analysis of malignant brain diseases: A mini review' *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2023, 57-60–< https://journals.mesopotamian.press/index.php/MJAIH/article/view/143/138> on 28 February 2025.
[118] Palavicini G, 'Intelligent health: Progress and benefit of artificial intelligence in sensing-based monitoring and disease diagnosis' 23(22) *Sensors*, 2023, 11; Rajpurkar P, Chen E, Barnerjee O, and Topol E, 'AI in health and medicine' 28(1) *Nature Medicine*, 2022, 31-32.
[119] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 3.
[120] Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H, and Alam T 'AI in drug discovery and its clinical relevance' 9(7) *Heliyon*, 2023, 2.
[121] Sirak K, Sawchuk E, and Prendergast M, 'Ancient human DNA and African population history' in *Oxford Research Encyclopaedia of Anthropology*, Oxford University Press, Oxford, 1.
[122] Bentley A, Callier S, and Rotimi C, 'Evaluating the promise of inclusion of African ancestry populations in genomics', 1.
[123] Peprah E et al., 'Genome-wide association studies in Africans and African Americans', 3-4.
[124] Peprah E et al., 'Genome-wide association studies in Africans and African Americans', 4.
[125] Kist J et al., 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands', 9; Shishehbori F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18; Martin G et al., 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples', 547.

In addition, Africa's rich biodiversity, encompassing over 40,000 plant species,[126] offers a significant advantage in the field of drug discovery. Traditional medicinal practices, which remain integral to the healthcare of approximately 80% of rural African communities,[127] provide a vast, yet underexplored, repository of therapeutic knowledge. Indeed, various African traditional medicinal plants like Artemisia annua (for malaria) have already demonstrated global impact,[128] yet scouting and bioprospecting for novel drug candidates on the continent is still limited. AI can process and analyse extensive phytochemical data rapidly, predicting the medicinal properties of compounds found in African flora—a process that could dramatically reduce the time and cost associated with drug development.[129] For example, research teams in Cameroon are already utilising AI to identify potential antiviral compounds from indigenous plants, aiming to combat diseases such as HIV.[130]

Finally, it is not lost on the author that healthcare delivery in Africa is often hampered by resource scarcity and systemic inefficiencies (which are not unique to Africa but are quite severe here), including under-resourced facilities,[131] high patient-to-doctor ratios (1.3 health practitioners per 1000 people in sub-Saharan Africa),[132] as well as disruptions like power

[126] Ekeanyanwu O, 'AI sifts Africa's natural remedies for drug discovery' Vaccines Work, 2 September 2024–<https://www.gavi.org/vaccineswork/ai-sifts-africas-natural-remedies-drug-discovery> on 30 December 2024.
[127] WHO, 'African Traditional Medicine Day 2022' 31 August 2022–< https://www.afro.who.int/regional-director/speeches-messages/african-traditional-medicine-day-2022#:~:text=Traditional%20medicine%20has%20been%20the,for%20their%20basic%20health%20needs> on 23 November 2024; Anyikwa C, 'Exploring the role of divination in traditional medicine in Africa', 1-4; BBC, 'The Kenyan enthralled by the healing power of plants' 29 July 2024–< https://www.bbc.com/news/articles/c84j0kzgnk9o> on 30 November 2024.
[128] Malaterre A, Rakoto M, Marodon C, Bedoui Y, Nakab J, Simon E, Hoarau L, Savriama S, Strasberg D, Guiraud P, Selambarom J, and Gasque P, '*Artemisia annua*, a traditional plant brought to light' 21(14) *International Journal of Molecular Sciences*, 2020, 1.
[129] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 3-7.
[130] Ekeanyanwu O, 'AI sifts Africa's natural remedies for drug discovery' Vaccines Work, 2 September 2024–<https://www.gavi.org/vaccineswork/ai-sifts-africas-natural-remedies-drug-discovery> on 30 December 2024.
[131] Harris R, 'The Healthcare Crisis in Sub-Saharan Africa' Africa Mission Healthcare–< https://africanmissionhealthcare.org/the-healthcare-crisis-in-sub-saharan-africa/#:~:text=Insufficient%20Healthcare%20Infrastructure,often%20poorly%20equipped%20and%20understaffed. > on 26 February 2025.
[132] Harris R, 'The Healthcare Crisis in Sub-Saharan Africa' Africa Mission Healthcare–< https://africanmissionhealthcare.org/the-healthcare-crisis-in-sub-saharan-africa/#:~:text=Insufficient%20Healthcare%20Infrastructure,often%20poorly%20equipped%20and%20understaffed. > on 26 February 2025.

outages (or unreliable electricity)[133] and internet instability.[134] These inefficiencies often impede timely access to care and have cascading effects on patient outcomes, hospital operations, and public health systems during crises. Advanced AI systems offer a transformative solution by automating and optimising (critical) operational functions. Streamlining these functions would significantly reduce delays and ensure that: (1) scarce resources are used more efficiently; and (2) clinicians focus on care.[135] The result is likely to be an unprecedented increase in healthcare capacity, allowing more patients to receive *timely* and *quality* care.[136]

3.3 Potential pitfalls of AI in clinical medicine

3.3.1 Fundamental risks in AI-enabled clinical medicine

There are various risks associated with deploying advanced AI in clinical medicine.[137] First, discrepancies between training data and the realities of clinical practice, coupled with other contextual variations,[138] can lead to AI errors that cause misdiagnoses, delayed treatments, or misguided prioritisation of care.[139] Second, the misuse of biomedical AI tools, particularly when clinicians and patients lack the necessary training to interpret outputs correctly.[140] Third, bias embedded in training datasets risks perpetuating uneven healthcare outcomes for

[133] Nadine N, 'Power cuts in South Africa wreak havoc on health care' Think Global Health, 7 June 2022–< https://www.thinkglobalhealth.org/article/power-cuts-south-africa-wreak-havoc-health-care> on 20 March 2025; WHO, 'Electricity in healthcare facilities' WHO, 31 August 2023–< https://www.who.int/news-room/fact-sheets/detail/electricity-in-health-care-facilities> on 20 March 2025.

[134] Rens A, Calandro E, and Gaffley M, 'As global cyber conflict breaks out, AI technologies bring new risk to Africa' 23 March 2022–< https://researchictafrica.net/2022/03/23/as-global-cyber-conflict-breaks-out-ai-technologies-bring-new-risk-to-africa/ > on November 4 2024; Lemos R, 'Infrastructure Cyberattacks, AI-Powered Threats Pummel Africa' 1 March 2024–< https://www.darkreading.com/vulnerabilities-threats/ai-powered-threats-cyberattacks-on-infrastructure-pummel-africa> on 4 November 2024.

[135] Varnosfaderani S and Forouzanfar M, 'The role of AI in hospitals and clinics: Transforming healthcare in the 21st century' 11(4) *Bioengineering*, 2024, 1, 10, 13.

[136] Varnosfaderani S and Forouzanfar M, 'The role of AI in hospitals and clinics' 10, 13.

[137] Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, I-III; Tech Target, 'Arguing the pros and cons of artificial intelligence in healthcare' 26 December 2023–< https://www.techtarget.com/healthtechanalytics/feature/Arguing-the-Pros-and-Cons-of-Artificial-Intelligence-in-Healthcare> on 23 November 2024.

[138] Roppelt J, Kanbach D, Kraus S, 'Artificial intelligence in healthcare institutions: A systematic literature review on influencing factors' 76(1) *Technology in Society*, 2024, 4; Trevino S, 'Managing risks and opportunities that emerging AI technologies present for healthcare cybersecurity' HIMSS, 13 November 2024–< https://gkc.himss.org/resources/managing-risks-and-opportunities-emerging-ai-technologies-present-healthcare> on 23 November 2024.

[139] Norori N, Hu Q, Aellen F, Faraci F and Tzovara A, 'Addressing bias in big data and AI for health care: A call for open science' 2(10) *Patterns*, 2021, 4; Banja J, 'How might artificial intelligence applications impact risk management?' 22(11) AMA Journal of Ethics, 2020, 945-948.

[140] Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, II.

marginalised groups.[141] Fourth, a lack of transparency undermines trust, especially since the decision-making processes of AI systems are opaque.[142] Fifth, the reliance on expansive datasets raises serious concerns about data misuse, breaches, or cyberattacks that could compromise sensitive patient information.[143] Finally, these challenges are compounded by persistent gaps in accountability, where it remains unclear whether liability for AI-induced errors should fall on developers, healthcare professionals, or institutions.[144]

3.3.2 Context-specific risks regarding AI in African clinical medicine

A. How underrepresentation of low-resource African languages in training data could lead to risks

Africa's linguistic diversity, with over 2,000 languages and countless dialects,[145] accounts for roughly one-third of the world's linguistic diversity,[146] and could pose a unique challenge for advanced AI systems in healthcare. While English and French dominate as official and administrative languages in many African nations,[147] the majority of the population communicates in indigenous languages such as Swahili (200 million), Yoruba (45 million),

---

[141] Tighe P, Gale B, and Mossburg S, 'Artificial intelligence and patient safety: Promise and challenges' Patient Safety Network, 27 March 2024–< https://psnet.ahrq.gov/perspective/artificial-intelligence-and-patient-safety-promise-and-challenges> on 23 November 2024; Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, 23, 26-27; Banja J, 'How might artificial intelligence applications impact risk management?', 945-948; Fernandes M and Goldim J, 'Artificial intelligence and decision making in health: Risks and opportunities' in Antunes H, Freitas P, Oliveira A, Pereira C, Sequeira E and Xavier L (eds) *Multidisciplinary perspectives on artificial intelligence and the law*, Springer, Cham, 2024, 196-199; OECD, 'AI in health: Huge potential, huge risks' 2024, 6–< https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/ai-in-health-huge-potential-huge-risks_ff823a24/2f709270-en.pdf> on 23 November 2024.

[142] Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, 23; Banja J, 'How might artificial intelligence applications impact risk management?', 945-948.

[143] Price W, 'Risks and remedies for artificial intelligence in health care' Brookings, 14 November 2019–< https://www.brookings.edu/articles/risks-and-remedies-for-artificial-intelligence-in-health-care/> on 23 November 2024; Trevino S, 'Managing risks and opportunities that emerging AI technologies present for healthcare cybersecurity' HIMSS, 13 November 2024–< https://gkc.himss.org/resources/managing-risks-and-opportunities-emerging-ai-technologies-present-healthcare> on 23 November 2024; Sutherland E, 'Artificial intelligence in health: big opportunities, big risks' OECD AI Policy Observatory, 1 August 2023–< https://oecd.ai/en/wonk/artificial-intelligence-in-health-big-opportunities-big-risks> on 23 November 2024; Banja J, 'How might artificial intelligence applications impact risk management?', 945-948.

[144] Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022, 26-27; OECD, 'AI in health: Huge potential, huge risks' 2024, 5–< https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/ai-in-health-huge-potential-huge-risks_ff823a24/2f709270-en.pdf> on 23 November 2024.

[145] Ritchie S, Cheng Y, Chen M, Matthews R, Esch D, Li B, and Sim K, 'Large vocabulary speech recognition for languages of Africa: multilingual modelling and self-supervised learning' ArXiv, 2022, 1.

[146] The African Language Program at Harvard, 'Introduction to African Languages' Harvard University–< https://alp.fas.harvard.edu/introduction-african-languages> on 31 December 2024.

[147] Provisio, 'Official languages in Africa – An analysis' 19 July 2023–< https://provisioservices.com/languages-in-africa/> on 31 December 2024.

Fula (35 million), and Igbo (30 million).[148] Can advanced AI systems reliably understand, process, or generate medical information across such a wide range of African languages?

Large language models (LLMs), advanced AI systems capable of interpreting and producing natural language,[149] are typically trained on massive datasets predominantly sourced from Western or global contexts where English, Spanish, French, and a few widely spoken, high-resource, languages are overrepresented.[150] It is thus unsurprising that many low-resource languages, including African languages, remain significantly underrepresented. This issue, known as the resourcedness gap, results from the insufficient volume, quality, and diversity of available text data in low-resource languages.[151] Owing to this gap, AI translation systems already exhibit serious errors in high-stakes environments (e.g., misinterpretation of key testimonies in asylum applications).[152] Given this, it has been strongly argued that governments must exercise caution in their use and promotion of large language models, and more importantly, that these models should never be employed to power systems that make high-stakes decisions without proper oversight, particularly in areas such as immigration status or *healthcare*.[153] The potential consequences of errors in such critical domains are too severe to ignore.

But how exactly could these errors occur when AI is deployed in clinical medicine? Translation and misinterpretation errors are caused by all sorts of factors.[154] However, in clinical medicine,

[148] The African Language Program at Harvard, 'Introduction to African Languages' Harvard University–< https://alp.fas.harvard.edu/introduction-african-languages> on 31 December 2024.
[149] IBM, 'What are large language models (LLMs)?' IBM, 2 November 2023–< https://www.ibm.com/think/topics/large-language-models> on 14 February 2025; Microsoft, 'What are large language models (LLMs)?' Microsoft–< https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-large-language-models-llms> on 14 February 2025.
[150] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024.
[151] Nicholas G and Bhatia A, 'Lost in translation: large language models in non-English content analysis' Center for Democracy and Technology, May 2023, 6–< https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf> on 31 December 2024.
[152] Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; World Economic Forum, 'Generative AI is trained on just a few of the world's 7,000 languages. Here's why that's a problem – and what's being done about it' 17 May 2024–< https://www.weforum.org/stories/2024/05/generative-ai-languages-llm/> on 31 December 2024.
[153] Nicholas G and Bhatia A, 'Lost in translation: large language models in non-English content analysis' Center for Democracy and Technology, May 2023, 6–< https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf> on 31 December 2024.
[154] Putri T, 'An analysis of types and causes of translation errors' 3(2) *Etnolingual*, 2019, 98.

these errors by AI systems could arise primarily from three specific factors: terminology, acronyms, and eponyms.[155] AI systems are less prone to errors with acronyms (e.g., HIV, MRI) and eponyms (e.g., Alzheimer's disease, Down syndrome) because these terms are often standardised across languages, but there still could be errors.[156] More problematic is the domain of medical terminology, which varies substantially across languages and cultural contexts.

One major terminology challenge is polysemy, where a single word or phrase may have multiple meanings depending on context.[157] For example, in Swahili, *homa* is an incredibly slippery term—sometimes a symptom (meaning "[ordinary] fever"), sometimes a full-blown illness, sometimes a mild discomfort that "becomes" a disease, sometimes just another word or synonym for malaria, and sometimes a range of other fevers.[158] Even medical professionals are guilty of using this term imprecisely.[159] Another linguistic wildcard is the Xhosa term *isifuba* which can mean the anatomical chest, a general chest illness, or a specific disease like asthma, leading to significant confusion in medical contexts.[160] An AI system trained without sufficient exposure to such linguistic nuances might misinterpret these terms, and in a clinical setting, this could lead to misdiagnoses, inappropriate treatments (particularly incorrect dosage of medication), poor adherence with said treatments, or other kinds of patient harm.

[155] Do C, 'Difficulties and suggestions for medical English translation: Insights from experience' 101 *Journal of Literature, Languages and Linguistics*, 2024, 13-17.
[156] Do C, 'Difficulties and suggestions for medical English translation' 13-17. **Acronyms** often remain unchanged but could cause some translation problems if (i) the acronym has multiple meanings in different contexts; (ii) there are no direct equivalents of a certain acronym in some languages; (iii) various abbreviations stand for different things; and (iv) there is a lack of consistency in how acronyms from one language are translated into other languages. **Eponyms**, which are proper names given to syndromes, illnesses, research-related matters, and devices, generally do not need translation and are meant to honour contributors to the field. However, translation challenges sometimes arise when a syndrome is discovered simultaneously in different countries by various individuals, leading to multiple eponyms for the same condition. Due to their universal adoption in scientific literature, this standardisation minimises the risk of advanced AI misinterpreting these terms, as their definitions would likely remain consistent regardless of linguistic or cultural background.
[157] Do C, 'Difficulties and suggestions for medical English translation' 14.
[158] Warsame M, Kimbute O, Machinda Z, Ruddy P, Melkisedick M, Peto T, Ribeiro I, Kitua A, Tomson G, Gomes M, 'Recognition, perceptions and treatment practices for severe malaria in rural Tanzania: Implications for accessing rectal artesunate as a pre-referral' 2(1) *PloS One*, 2007, 3; Winch P, Makemba A, Kamazima S, Lurie M, Lwihula G, Premji Z, Minjas J, Shiff C, 'Local terminology for febrile illnesses in Bagamoyo district, Tanzania and its impact on the design of a community-based malaria control programme' 42(7) *Social Science and Medicine*, 1997, 1059; Dixon J and Chandler C, 'Opening up 'fever', closing down medicines: Algorithms as blueprints for global health in an era of antimicrobial resistance' 6(4) *Medicine Anthropology Theory*, 2019, 64.
[159] Dixon J and Chandler C, 'Opening up 'fever', closing down medicines', 64.
[160] Levin E, 'Different use of medical terminology and culture-specific models of disease affecting communication between Xhosa-speaking patients and English-speaking doctors at a South African paediatric teaching hospital' 96(10) *South African Medical Journal*, 2006, 1081-1082.

Idiomatic expressions pose another (terminology) challenge for AI in medical interpretation. Many indigenous languages use metaphorical phrases to describe illnesses and symptoms which may lack direct translations. For instance, in Maasai culture, malaria is described with terms like *olnairobi* (starting as cold) and *irmotiook* (changing like a cooking pot), and the term *enk-agang'ani* is used for both fever and mosquito.[161] Similarly, terms equivalent to "brain death" might not exist in some African languages, as concepts of death and consciousness vary across cultures.[162] If AI systems lack proper cultural training, they may misclassify or fail to recognise these expressions, leading to diagnostic errors.

Finally, there could be complications arising from the lack of direct equivalents for medical terms in many African languages. Advanced medical concepts such as multiple myeloma, autoimmune disorders, or neurological conditions may not have precise indigenous counterparts.[163] In such cases, AI models must rely on *approximate* translations, which risk losing crucial details necessary for accurate diagnosis and treatment. Without sufficient exposure to local terminologies and healthcare contexts, AI-generated recommendations may be inaccurate, misleading, or even harmful.

B. How inadequate incorporation of African traditional medicine in training data could lead to risks

Traditional medicine remains deeply embedded in African healthcare, with the WHO estimating that over 80% of the population relies on it for primary care.[164] It is practised by healers known by different names across the continent (e.g., *bokomowo* in Ghana, *Nga:nga* in Zambia, *Sangomas* in South Africa, and *babalawos* in Nigeria).[165] These healers offer a range of healthcare services, from herbal remedies, spiritual healing, and divination to integrating traditional practices with modern medicine, particularly in regions with limited access to formal

[161] Strang C and Mixer S, 'Discovery of the meanings, expressions, and practices related to malaria care among the Maasai' 27(4) *Journal of Transcultural Nursing*, 2016, 337.
[162] Do C, 'Difficulties and suggestions for medical English translation' 15.
[163] Do C, 'Difficulties and suggestions for medical English translation' 15.
[164] WHO, 'African Traditional Medicine Day 2022' 31 August 2022–< https://www.afro.who.int/regional-director/speeches-messages/african-traditional-medicine-day-2022#:~:text=Traditional%20medicine%20has%20been%20the,for%20their%20basic%20health%20needs> on 23 November 2024; Anyikwa C, 'Exploring the role of divination in traditional medicine in Africa', 1-4; BBC, 'The Kenyan enthralled by the healing power of plants' 29 July 2024–< https://www.bbc.com/news/articles/c84j0kzgnk9o> on 30 November 2024.
[165] Ozioma E and Chinwe O, 'Herbal medicines in African traditional medicine' in Builders P (ed), *Herbal Medicine*, Intech Open, 30 January 2019–<https://www.intechopen.com/chapters/64851> on 31 December 2024.

healthcare.[166] The widespread use of traditional medicine could present challenges for advanced AI systems. Advanced AI systems are predominantly trained on biomedical datasets rooted in Western medical practices.[167] As a result, they may struggle to process queries or make decisions involving traditional medicine. This knowledge gap creates a significant safety risk and could lead to, inter alia, misleading or unsafe treatment recommendations.

One major concern is the risk of adverse drug reactions (ADRs)—harmful, unintended effects directly attributable to the normal doses and use of medicines.[168] When patients use both pharmaceutical drugs and traditional remedies, as they often do,[169] interactions can occur, sometimes with severe consequences. If an AI system does not account for this possibility, or if it lacks data on those substances, it may misjudge the safety of a prescribed treatment, increasing the risk of toxicity, reduced drug efficacy, or unexpected side effects. As outlined in research on ADRs, such reactions can range from mild side effects to severe, even life-threatening events, particularly when drugs interact unpredictably with non-Western treatments.[170]

The issue is compounded by various factors. Firstly, traditional medicines are often underrepresented in standard pharmacovigilance systems.[171] Many substances used in traditional healing practices are not well-documented in regulatory databases, and even where they are, their regulatory status varies widely—what is classified as a dietary supplement in

[166] Ozioma E and Chinwe O, 'Herbal medicines in African traditional medicine' in Builders P (ed), *Herbal Medicine*, Intech Open, 30 January 2019–<https://www.intechopen.com/chapters/64851> on 31 December 2024.
[167] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8; Norori N et al., 'Addressing bias in big data and AI for health care', 3; Zou J and Shiebinger L, 'Ensuring that biomedical AI benefits diverse populations' 67 *eBioMedicine*, 2021, 1.
[168] VA Center for Medication Safety and VHA Pharmacy Benefits Management Strategic Healthcare Group and the Medical Advisory Panel, 'Adverse drug events, adverse drug reactions and medication errors: Frequently asked questions' November 2006, 1–< https://www.pbm.va.gov/vacenterformedicationsafety/tools/AdverseDrugReaction.pdf> on 31 December 2024; New Zealand Medicines and Medical Devices Safety Authority, 'Adverse reactions to medicines and vaccines' 21 November 2023–< https://www.medsafe.govt.nz/consumers/safety-of-medicines/Medicine-safety.asp#:~:text=Adverse%20reactions%20are%20harmful%20effects,sheet%20or%20consumer%20medicine%20information> on 31 December 2024.
[169] Shetti S, Kumar C, Sriwastava N, and Sharma I, 'Pharmacovigilance of herbal medicines: Current state and future directions' 7(25) *Pharmacognosy Magazine*, 2011, 70.
[170] Coleman J and Pontefract S, 'Adverse drug reactions' 16(5) *Clinical Medicine*, 2016, 481-482.
[171] Shetti S et al., 'Pharmacovigilance of herbal medicines', 70,72.

one country may be deemed a therapeutic agent in another.[172] This lack of regulatory uniformity, coupled with the fact that many traditional remedies are licensed based on historical usage rather than robust clinical trial data,[173] means that the datasets informing these AI systems could very often be incomplete. Secondly, the chemical complexity and non-uniformity of herbal medicines—where a single plant may yield hundreds of constituents that fluctuate with environmental conditions, harvesting techniques, and processing methods—[174]further complicate pharmacovigilance assessments.[175] This could lead to significant gaps in the training data, making it difficult for AI systems to predict or identify ADRs when herbal medicines are used alongside conventional treatments.

Finally, medication histories—which are critical for identifying risks—[176]may be incomplete when they involve traditional treatments, as patients do not always disclose their use, and healthcare providers may not routinely ask about them. Given how pharmacovigilance works,[177] the challenges above would likely prevent AI systems from making informed assessments about patient safety, and potentially lead to undetected ADRs. To put all of the above factors briefly, AI systems depend on large, standardised datasets to predict treatment outcomes accurately.[178] However, traditional medicine's variability[179] in regulation, preparation, dosage, history, and cultural practices creates data gaps that undermine this requirement and exacerbate the risk of ADRs as well as other misleading or unsafe treatment recommendations.[180]

[172] Ekor M, 'The growing use of herbal medicines: Issues relating to adverse reactions and challenges in monitoring safety' 4(177) *Frontiers in Pharmacology*, 2014, 5-6.
[173] World Health Organisation, 'Pharmacovigilance for traditional medicine products: Why and how?' WHO: South East Asia, 2017–< https://iris.who.int/bitstream/handle/10665/259854/Pharmacovigilane.pdf?sequence=1> on 23 February 2025.
[174] World Health Organisation, 'Pharmacovigilance for traditional medicine products: Why and how?' WHO: South East Asia, 2017–< https://iris.who.int/bitstream/handle/10665/259854/Pharmacovigilane.pdf?sequence=1> on 23 February 2025.
[175] Ekor M, 'The growing use of herbal medicines', 6.
[176] Coleman J and Pontefract S, 'Adverse drug reactions' 481-483.
[177] The science of detecting and preventing adverse effects from medicine. See European Medicines Agency, 'Pharmacovigilance: Overview' 6 June 2024–< https://www.ema.europa.eu/en/human-regulatory-overview/pharmacovigilance-overview> on 31 December 2024.
[178] Alowais S, Alghamdi S, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb S, Aldairem A, Alrashed M, Saleh K, Badreldin H, Al Yami M, Al Harbi S, and Albekairy A, 'Revolutionizing healthcare: The role of artificial intelligence in clinical practice' 23(1) *BMC Medical Education*, 2023, 2-3, 6.
[179] Ekor M, 'The growing use of herbal medicines', 5-6; Parveen A, Parveen B, Parveen R, and Ahmad S, 'Challenges and guidelines for clinical trial of herbal drugs' 7(4) *Journal of Pharmacy & Bioallied Sciences*, 2015, 329-333.
[180] Robinson M and Zhang X, 'The world medicines situation 2011– Traditional medicines: Global situation, issues and challenges' World Health Organisation, 2011, 8–< https://www.utep.edu/herbal-safety/herbal-facts/files/wms_ch18_wtraditionalmed-2011.pdf> on 31 December 2024.

Beyond ADRs, AI systems could inadvertently contribute to the *exclusion* or *stigmatisation* of traditional medicine. If AI models fail to recognise or validate traditional treatments, or oversimplify them despite their complexity,[181] they may *exclude* them, that is, they may deem them irrelevant or harmful, thereby discrediting deeply rooted healthcare practices.[182] This could erode patient trust in them, and in communities where traditional medicine is the primary or only accessible healthcare option, this would be a significant safety issue. Additionally, AI models trained on Western-centric data, in line with evidence that they could trigger various prejudices,[183] risk reinforcing existing biases by portraying traditional medicine as unscientific or ineffective, even when evidence suggests otherwise. This *stigmatisation* could, again, result in patients unilaterally abandoning effective traditional treatments in favour of peculiar, unproven, or inappropriate alternatives, which might increase the risk of harm.

C. How underrepresentation of African genetic diversity in training data could lead to risks

Africa is the cradle of human genetic diversity. Studies consistently affirm that individuals within African populations often exhibit more genetic variation among themselves than populations from other continents.[184] From a medical standpoint, this genetic diversity influences disease susceptibility, drug metabolism, and treatment outcomes.[185] In fact, the study of this relationship is the core function of Genome-wide association studies (GWAS). GWAS are a research method that examines the entire genome of individuals to identify genetic variations linked to specific diseases or traits.[186] By analysing hundreds of thousands of genetic

[181] Aruna B and Priyanka P, 'A review article on artificial intelligence in traditional medicine research and application' 9(11) *International Journal Of Novel Research And Development*, 2024, 342.
[182] Ozioma E and Chinwe O, 'Herbal medicines in African traditional medicine' in Builders P (ed), *Herbal Medicine*, Intech Open, 30 January 2019–<https://www.intechopen.com/chapters/64851> on 31 December 2024.
[183] Botha N, Segbedzi C, Dumahasi V, Maneen S, Kodom R, Tsedze I, Akoto L, Atsu F, Lasim O, and Ansah E, 'Artificial intelligence in healthcare: A scoping review of perceived threats to patient rights and safety' 82(188) *Archives of Public Health*, 2024, 22.
[184] Sirak K et al., 'Ancient human DNA and African population history' in *Oxford Research Encyclopaedia of Anthropology*, Oxford University Press, Oxford, 1; Bentley A et al., 'Evaluating the promise of inclusion of African ancestry populations in genomics', 1.
[185] Smalla H, 'Genetic diversity: Effects on treatment response and disease susceptibility' 9(4) *Global Journal of Life Sciences and Biological Research*, 2023, 1-2; Prajapati J and Shukla A, 'The role of genetic variations in drug metabolism and their impact on therapeutic outcomes' 2(2) *International Journal Of Advance Research In Multidisciplinary*, 2024, 317,323.
[186] Vicente M, 'Demographic history and adaptation in African populations' Uppsala: Acta Universitatis Upsaliensis, 2020, 3–< https://www.diva-portal.org/smash/get/diva2:1412382/FULLTEXT01.pdf> on 31 December 2024.

markers across diverse populations, GWAS can reveal how certain genetic factors statistically contribute to health outcomes.[187]

Importantly, including underrepresented populations, such as those of African ancestry, enhances the *safety* and *effectiveness* of GWAS, increases innovation, and contributes to *broader scientific understanding*.[188] Notwithstanding these benefits, African genomes are still vastly underrepresented in global GWAS.[189] Indeed, over 81% of GWAS involve individuals of European ancestry, despite this constituting only 16% of the global population.[190] Genetic biases embedded in the data used to train AI systems emanate from this underrepresentation of African genomes in global GWAS.[191] Furthermore, aside from these unrepresentative GWAS datasets, genetic biases in AI systems could also emanate from other datasets dominated by individuals from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) populations.[192]

Genetic bias is particularly dangerous in diagnostics and therapeutics as algorithms might fail to accurately predict or identify conditions in underrepresented populations.[193] For example, diagnostic tools for conditions like skin cancer, which has a strong genetic component, may perform poorly for African patients if trained primarily on genetic data from European populations.[194] The same is true with regards to the risk assessment of cardiovascular events

[187] Uffelmann E, Huang Q, Munung N, de Vries J, Okada Y, Martin A, Martin H, Lappalainen T, and Posthuma D, 'Genome-wide association studies' Nature Reviews, 26 August 2021–< https://www.nature.com/articles/s43586-021-00056-9> on 31 December 2024.
[188] Peprah E et al., 'Genome-wide association studies in Africans and African Americans,' 3-4. Studies have shown that variants associated with asthma in European populations, such as CRB1 and PDE4D, are not found in African ancestry populations, which instead exhibit associations with DENND1B and other genes. Furthermore, studies have confirmed BCL11A's role in sickle cell disease severity among African Americans.
[189] Kist J et al., 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands', 9; Shishehbori F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18; Martin G et al., 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples', 547.
[190] Martin G et al., 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples', 547.
[191] Kist J et al., 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands', 9; Shishehbori F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18; Martin G et al., 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples', 547.
[192] Norori N et al., 'Addressing bias in big data and AI for health care', 3.
[193] Norori N et al., 'Addressing bias in big data and AI for health care', 3.
[194] Norori N et al., 'Addressing bias in big data and AI for health care', 3.

where, cardiovascular risk prediction models trained on Western data might substantially underestimate risks in African populations or people of African descent.[195]

Failure to identify conditions in African populations as well as underestimations of risks to them could delay critical interventions and have severe health consequences, including death. Furthermore, these systemic, genetic biases can result in misdiagnoses, ineffective treatments, or the complete exclusion of vulnerable groups in Africa from AI-driven healthcare advances.[196] Without intentional efforts to diversify training datasets and address algorithmic inequities, AI systems risk entrenching existing genetic disparities,[197] leaving African populations at greater risk of harm.

D. How underrepresentation of diseases endemic to Africa, particularly NTDs, in training data could lead to risks

Neglected tropical diseases (NTDs) present a significant yet underappreciated challenge in African healthcare, one that advanced artificial intelligence (AI) systems may struggle to address effectively. NTDs are a group of at least 20 illnesses—such as schistosomiasis, leishmaniasis, chikunguya, chagas disease, dengue fever, leprosy, trachoma, human African trypanosomiasis (or sleeping sickness), and lymphatic filariasis—that disproportionately affect marginalised populations in tropical and subtropical regions.[198] These diseases thrive in environments characterised by poverty, limited healthcare access, and specific ecological conditions, including warm climates and stagnant water, all prevalent in many parts of Africa.[199]

The World Health Organization (WHO) reports that the African region is the hardest hit, shouldering 40% of the global NTD burden as of 2022.[200] This region also has the highest

[195] Shishehbori F and Awan Z, 'Enhancing cardiovascular disease risk prediction with machine learning models' ArXiv, 2024, 18.
[196] Norori N et al., 'Addressing bias in big data and AI for health care', 3.
[197] Norori N et al., 'Addressing bias in big data and AI for health care', 3.
[198] Mitra A and Mawson A, 'Neglected tropical diseases: Epidemiology and global burden' 2(36) *Tropical Medicine and Infectious Disease*, 2017, 1-2.
[199] Mitra A and Mawson A, 'Neglected tropical diseases', 1-2; World Health Organisation, 'Neglected tropical diseases' 9 January 2024–< https://www.who.int/news-room/questions-and-answers/item/neglected-tropical-diseases> on 2 January 2024.
[200] This statistic excludes Chagas disease. Wolfe C, Barry A, Campos A, Farham B, Achu D, Juma E, Kalu A, and Impouma B, 'Control, elimination, and eradication efforts for neglected tropical diseases in the World Health Organization African region over the last 30 years: A scoping review' 141 *International Journal of Infectious Diseases*, 2024, 1.

number of countries with significant NTD cases per million inhabitants.[201] Yet, despite this NTD burden, and unlike high-profile global diseases like cancer or diabetes, these chronic (and often debilitating) conditions have historically been overshadowed, hardly receiving any attention from international healthcare initiatives and pharmaceutical research.[202] It is thus crucial to bring more attention and resources to combat these overlooked diseases and alleviate the suffering they cause as this neglect is likely to extend to the datasets used to train AI systems, creating a gap that risks perpetuating existing inequities.[203]

AI systems in healthcare are fundamentally reliant on the quality, scope, and representativeness of the data they are trained on.[204] These datasets are predominantly derived from well-funded healthcare environments in Europe, North America, and parts of Asia.[205] As a result, they reflect the medical concerns, patient demographics, and treatment strategies common to those regions.[206] One crucial example of how AI systems prioritise medical data from high-resource regions is ChatGPT Deep Research's recent omission of the globally recognised PRESBYOND treatment for presbyopia—a highly effective procedure validated in Europe, Asia, and Latin America for over two decades—due to its lack of FDA approval in the US.[207] Certainly, the fact that this oversight was driven by a (perhaps, stubborn) reliance on FDA-

[201] Wolfe C, Barry A, Campos A, Farham B, Achu D, Juma E, Kalu A, and Impouma B, 'Control, elimination, and eradication efforts for neglected tropical diseases in the World Health Organization African region over the last 30 years: A scoping review' 141 *International Journal of Infectious Diseases*, 2024, 1.
[202] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.
[203] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.
[204] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.
[205] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.
[206] Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024, 2; Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases> on 31 December 2024; World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.
[207] Dascalescu D, 'How to avoid blind spots and use Deep Research effectively' Medium, 10 February 2025–< https://dandv.medium.com/how-to-avoid-blind-spots-and-use-deep-research-effectively-659b3afe675c > on 4 March 2025.

approved data reveals a pronounced US-centric bias.[208] This US-centric bias, which sidelines global medical consensus, perfectly illustrates how AI systems risk excluding treatments and conditions absent from dominant regulatory frameworks. It foreshadows that diseases which are rare or absent in high-resource region datasets—such as the NTDs endemic to Africa—will probably be poorly represented or entirely overlooked.[209]

Without sufficient training data reflecting the immense number of problems related to NTDs, AI systems risk providing inaccurate diagnoses or generic treatment suggestions, undermining their utility in combating these diseases. These inaccuracies would also not be benign as misdiagnoses and generic treatments could delay appropriate treatment, leading to increased morbidity and, in severe cases, mortality. The integration of advanced AI systems into African healthcare systems must thus be approached with an acute awareness of the seriousness of these diseases. Training datasets need to incorporate data regarding NTDs that are prevalent in the region. Without these measures, there is a real risk that AI will reinforce, rather than alleviate, the disparities that have long plagued healthcare in Africa.

3.4 Conclusion

This chapter set out to detail both the opportunities and challenges presented by AI in African clinical medicine. It highlighted that while AI holds promise for transforming clinical practice in Africa through precision medicine, drug discovery and resource optimisation, it also carries serious, context-specific risks. These risks stem from biased training data in respect of African low resource languages, African traditional medicine, African genetic diversity, and diseases endemic to Africa. These insights pave the way for the next discussion on pre-deployment testing and, eventually, the legal and institutional mechanisms necessary to ensure safe AI deployment.

[208] Dascalescu D, 'How to avoid blind spots and use Deep Research effectively' Medium, 10 February 2025–< https://dandv.medium.com/how-to-avoid-blind-spots-and-use-deep-research-effectively-659b3afe675c > on 4 March 2025.
[209] World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024, 8.

4 The Methodological, Legal and Institutional Landscape of Pre-Deployment Evaluations

4.1 Introduction

The prior chapter identified Africa-centric risks requiring mitigation, this chapter addresses the current methodologies, legal, and institutional frameworks that underpin pre-deployment evaluations for AI safety. It outlines the methods used by industry and regulators, reviews existing legal instruments, and identifies key gaps in the current system. This discussion is critical to understanding how evaluations are conducted and where improvements are necessary, particularly within the African clinical context.

4.2 The role of pre-deployment evaluations in AI safety

Pre-deployment evaluations are structured assessments conducted *before* deploying AI systems to identify and understand potential risks,[210] capabilities,[211] impacts,[212] behaviour,[213] performance,[214] and alignment with user and societal expectations.[215] These evaluations target three primary dimensions**:** safety, capability, and alignment. Safety evaluations focus on identifying risks of harm, such as biases, adversarial vulnerabilities, or misuse potential, by using methods like red-teaming, where experts attempt to exploit system weaknesses.[216] Capability evaluations assess what the system can do, using benchmarks to gauge computational power, general-purpose abilities, or domain-specific skills.[217] Alignment

[210] Mukobi G, 'Reasons to doubt the impact of AI risk evaluations' ArXiv, 2024, 1.
[211] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025; Weidinger L, Barnhart J, Brennan J, Butterfield C, Young S, Hawkins W, Hendricks L, Comanescu R, Chang O, Rodriguez M, Beroshi J, Bloxwich D, Proleev L, Chen J, Farquhar S, Ho L, Gabriel I, Dafoe A, and Isaac W, 'Holistic safety and responsibility evaluations of advanced AI models' ArXiv, 2024, 3.
[212] Weidinger L et al., 'Holistic safety and responsibility evaluations of advanced AI models' ArXiv, 2024, 3.
[213] Weidinger L et al., 'Holistic safety and responsibility evaluations of advanced AI models' ArXiv, 2024, 3.
[214] Ada Lovelace Institute, Under the radar? Examining the evaluation of foundation models, 2024, 3.
[215] EA Forum, 'AI evaluations and standards' EA Forum–< https://forum.effectivealtruism.org/topics/ai-evaluations-and-standards> on 7 January 2025; McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025.
[216] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025.
[217] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025; AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9

evaluations determine whether the AI's goals align with user intentions and societal values, a critical factor in avoiding harmful unintended behaviours.[218]

The UK AISI, for example, focuses on assessing *capabilities* and argues that the three key risk dimensions that add most value are: misuse (checking how AI might make it easier for people to do harmful things), societal impact (looking at how AI affects people and society, including its use in daily life and work), and autonomous systems.[219] In this study, however, "pre-deployment evaluations" refer exclusively to *safety-specific* evaluations—distinct from general performance, capability, or alignment testing. Thus, any kind of testing or assessment that attempts to address risks of harm to patients, clinicians, or healthcare systems (e.g., algorithmic impact assessments, third-party pre-deployment model audits, bias testing with pre-defined metrics, KYC screenings, etc) would fall under the scope of evaluations discussed here.

Pre-deployment evaluations are pivotal in ensuring the safe and responsible deployment of AI systems. Firstly, they ensure that AI systems will not pose public safety risks once they are deployed.[220] Secondly, safety evaluations help guide the development of AI systems. By assessing the likelihood of potential harms, these evaluations can highlight factors that might cause issues.[221] This way, developers can modify the AI system to reduce risks or determine where and how it can be safely used. Thirdly, evaluating different risk areas helps identify trade-offs that come with deploying AI in real-world settings.[222] This aids in governance

February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025; Weidinger L et al., 'Holistic safety and responsibility evaluations of advanced AI models' ArXiv, 2024, 3.

[218] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025; EA Forum, 'AI evaluations and standards' EA Forum–< https://forum.effectivealtruism.org/topics/ai-evaluations-and-standards> on 7 January 2025.

[219] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.

[220] Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W and Weidinger L, 'Gaps in the safety evaluation of generative AI' Seventh AAAI/ACM Conference on AI, Ethics, and Society, San Jose, 21-23 October 2024, 1201. See also Microsoft, 'AI in Africa Meeting the Opportunity' 2024, 20–< https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2024/01/AI-in-Africa-Meeting-the-Opportunity.pdf> on October 19 2024.

[221] Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W and Weidinger L, 'Gaps in the safety evaluation of generative AI' Seventh AAAI/ACM Conference on AI, Ethics, and Society, San Jose, 21-23 October 2024, 1201.

[222] Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W and Weidinger L, 'Gaps in the safety

decisions. Finally, evaluations play a critical role in fostering public trust and confidence.[223] Rigorously testing AI systems before deployment demonstrates to the public a commitment to safety and ethical responsibility.[224]

Pre-deployment evaluations have faced valid criticism over recent years,[225] the main one being that AI systems often reveal unforeseen failures post-deployment, such as jailbreaking within days of release,[226] exposing the chasm between controlled testing and real-world unpredictability.[227] Developers cannot guarantee downstream performance,[228] yet these assessments remain vital precisely because they *surface latent risks*—like biases or security flaws—that might otherwise go undetected until harm occurs.[229] While current methods, from benchmarking to red teaming, are imperfect (inconsistent, prone to manipulation, and unable to predict every real-world scenario), their value lies in creating a baseline for iterative risk management.[230]

4.3 Current approaches to AI pre-deployment evaluations

4.3.1 Methods used in AI evaluations

AI evaluation methods vary significantly based on the complexity, scope, and purpose of the system being tested. Broadly, evaluations can be exploratory—seeking unexpected weaknesses—or directed, focusing on predefined risks through structured metrics.[231] More specifically, various methodologies, including but not limited to those discussed below, help assess an AI's capabilities and risks comprehensively. *Automated capability assessments*, for

evaluation of generative AI' Seventh AAAI/ACM Conference on AI, Ethics, and Society, San Jose, 21-23 October 2024, 1201.

[223] Salhab W, Ameyed D, Jaafar F, Mcheick H, 'A systematic literature review on AI Safety: Identifying trends, challenges, and future directions' 12 *IEEE Access*, 2024, 131762- 131763.

[224] Salhab W, et al., 'A systematic literature review on AI Safety', 131762- 131763.

[225] Uuk R, Brouwer A, Schreier T, Dreksler N, Pulignano V, and Bommasani R, 'Effective mitigations for systemic risks from general-purpose AI' ArXiv, 2024, 51-52.

[226] Bengio Y, *The International Scientific Report on the Safety of Advanced AI*, 2025, 174-175.

[227] Uuk R, Brouwer A, Schreier T, Dreksler N, Pulignano V, and Bommasani R, 'Effective mitigations for systemic risks from general-purpose AI' ArXiv, 2024, 52; Ada Lovelace Institute, Under the radar? Examining the evaluation of foundation models, 2024, 5, 42.

[228] Bengio Y, *The International Scientific Report on the Safety of Advanced AI*, 2025, 174-175.

[229] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 7.

[230] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 7; Uuk R, Brouwer A, Schreier T, Dreksler N, Pulignano V, and Bommasani R, 'Effective mitigations for systemic risks from general-purpose AI' ArXiv, 2024, 55.

[231] Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W and Weidinger L, 'Gaps in the safety evaluation of generative AI' Seventh AAAI/ACM Conference on AI, Ethics, and Society, San Jose, 21-23 October 2024, 1201.

example, use predefined question sets to measure an AI's performance across different domains.[232] These act as an initial screening tool, identifying areas that require deeper investigation.[233]

Red-teaming is another evaluation approach. A red team is a group of experts who stress-test AI systems to identify their limits and uncover weaknesses.[234] This often includes *adversarial testing*—intentionally trying to break or manipulate the AI's safeguards (or assumptions)[235] to expose weak points (e.g., provoking the system to give biased or misleading information).[236] Some adversarial attacks are automated (e.g. using an adversarial LLM),[237] while others involve manually attempting to "jailbreak" the system to see how it behaves under less controlled conditions.[238] Red-teaming also complements *qualitative probing*, where experts observe how the AI responds in different scenarios to gain deeper insights into potential failure modes.[239]

Prediction evaluations, another evaluation method, measure how accurately developers can anticipate their AI model's behaviour.[240] Developers make predictions about how the AI will perform in certain tests, and these are compared with actual results.[241] This approach helps

[232] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.
[233] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.
[234] Bengio Y, *International scientific report on the safety of advanced AI*, 2024, 36; AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.
[235] Konatam S, Konatam V, and Konatam S, 'A novel of detecting and addressing bias in artificial intelligence and machine learning: A multi-industry perspective' 11(6) *International Research Journal of Engineering and Technology*, 2024, 834.
[236] Prabhune S and Berndt D, 'Deploying large language models with retrieval augmented generation' ArXiv, 2024, 20; Bengio Y, *International scientific report on the safety of advanced AI*, May 2024, 36.
[237] Prabhune S and Berndt D, 'Deploying large language models with retrieval augmented generation' ArXiv, 2024, 20.
[238] Bengio Y, *International scientific report on the safety of advanced AI*, 2024, 36.
[239] National Institute of Standards and Technology, 'Pre-deployment evaluation of OpenAI's o1 model' NIST, 18 December 2024–< https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model> on 7 January 2025.
[240] Chan A, 'Evaluating predictions of model behaviour' Centre for the Governance of AI, 9 April 2024–< https://www.governance.ai/post/evaluating-predictions-of-model-behaviour> on 7 January 2025; Hubinger E, 'Towards understanding safety-based evaluations' AI Alignment Forum, 15 March 2023–< https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations> on 7 January 2025.
[241] Chan A, 'Evaluating predictions of model behaviour' Centre for the Governance of AI, 9 April 2024–< https://www.governance.ai/post/evaluating-predictions-of-model-behaviour> on 7 January 2025; Hubinger E,

determine whether a model can be deployed safely based on whether developers understand its generalisation capabilities—how well it performs outside controlled conditions.[242] *Benchmarking* involves standardised tests that gauge AI performance on specific tasks, such as image classification or language processing.[243] These benchmarks offer valuable comparisons but have limitations—they may not represent real-world conditions, may introduce biases based on how human annotators define correct answers, and can sometimes give a false impression of an AI systems' reliability.[244] Good performance on benchmarks does not always translate to effectiveness in real applications.[245]

Human uplift evaluations focus on assessing whether AI enables individuals to carry out harmful tasks more efficiently.[246] By comparing performance with and without AI assistance, this method provides insight into the extent to which AI can amplify malicious capabilities.[247] *Agent-based evaluations* test AI systems that operate semi-autonomously, making decisions and using external tools to complete complex tasks.[248] When assigned a specific task—such as finding information or solving a problem—the AI model employs various software tools to execute a sequence of steps until it either completes the task successfully or exhausts its allotted attempts.[249] These assessments help reveal the risks of AI systems that act independently in real-world environments.[250] Lastly, *question-answering evaluations* test an AI's knowledge

'Towards understanding safety-based evaluations' AI Alignment Forum, 15 March 2023–< https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations> on 7 January 2025.

[242] Chan A, 'Evaluating predictions of model behaviour' Centre for the Governance of AI, 9 April 2024–< https://www.governance.ai/post/evaluating-predictions-of-model-behaviour> on 7 January 2025; Hubinger E, 'Towards understanding safety-based evaluations' AI Alignment Forum, 15 March 2023–< https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations> on 7 January 2025.

[243] Bengio Y, *International scientific report on the safety of advanced AI*, 2024, 35.

[244] Bengio Y, *International scientific report on the safety of advanced AI*, 2024, 36.

[245] Bengio Y, *International scientific report on the safety of advanced AI*, 2024, 36.

[246] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.

[247] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.

[248] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.

[249] National Institute of Standards and Technology, 'Pre-deployment evaluation of OpenAI's o1 model' NIST, 18 December 2024–< https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model> on 7 January 2025.

[250] AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations> on 7 January 2025.

and reasoning abilities by comparing its responses against human-verified answers.[251] This method provides insight into the AI's factual accuracy and problem-solving skills.[252]

The above evaluations are often conducted by various parties categorised into three levels: first-party evaluations are conducted internally by developers and dedicated compliance teams who test their own models and data.[253] Second-party evaluations occur when an organisation, or a consultancy acting on its behalf, is engaged by a potential customer—often under formal contractual arrangements—to assess the robustness and suitability of a system.[254] Finally, third-party evaluations are carried out by independent entities, whether driven by the evaluatee, government regulators, or independent academics and civil society groups, whose objective assessments of both technical performance and broader societal impacts ensure that no critical vulnerability is overlooked.[255]

4.3.2 Legal and policy frameworks governing AI evaluations

A. Legal and policy frameworks governing AI evaluations in AISIs

AI Safety Institutes (AISIs) have emerged worldwide, like in the United Kingdom (UK), the United States (US), Japan, the European Union (EU), France, Canada, Singapore and South Korea to shape AI safety practices,[256] with *evaluations* forming a unifying core of their activities.[257] The legal and policy frameworks governing AI evaluations across the eight AISIs

[251] National Institute of Standards and Technology, 'Pre-deployment evaluation of OpenAI's o1 model' NIST, 18 December 2024–< https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model> on 7 January 2025.
[252] National Institute of Standards and Technology, 'Pre-deployment evaluation of OpenAI's o1 model' NIST, 18 December 2024–< https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model> on 7 January 2025.
[253] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 29-30; National Telecommunications and Information Administration, 'Independent Evaluations' NTIA, 27 March 2024–< https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations> on 26 February 2024.
[254] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 30; National Telecommunications and Information Administration, 'Independent Evaluations' NTIA, 27 March 2024–< https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations> on 26 February 2024.
[255] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 31-32; National Telecommunications and Information Administration, 'Independent Evaluations' NTIA, 27 March 2024–< https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations> on 26 February 2024.
[256] Allen G and Adamson G, *The AI safety institute international network: Next steps and recommendations*, 2024, 7; Institute for AI Policy and Strategy, *Understanding the first wave of AI safety institutes: Characteristics, functions, and challenges*, 2024, 17-19.
[257] Institute for AI Policy and Strategy, Understanding the first wave of AI safety institutes: Characteristics, functions, and challenges, 2024, 17-19. UK AI Safety Institute, 'Advanced AI evaluations at AISI: May Update' 20 May 2024–<https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update> on 23 November 2024; US AI Safety Institute, 'Strategic vision' National Institute of Standards and Technology, 1 October 2024–

reveal that most of them rely on, or are established under, executive or administrative authority rather than binding legislation.[258] Thus, the evaluation mandate is advisory in most AISIs as opposed to regulatory.

The United Kingdom, Japan and Korea exemplify the advisory model, where evaluations are conducted under government policy rather than specific statutory instruments. The UK AISI, established as an evolution of the Frontier AI Taskforce, is housed under the UK Department for Science, Innovation and Technology (DSIT), has no independent regulatory mandate.[259] It is explicitly an advisory body; it lacks regulatory authority and cannot enforce compliance.[260] Instead, its evaluations—focused on dual-use capabilities, societal impacts, and system safety—inform government policy and international collaboration.[261] Similarly, Japan's AISI operates within the Information-technology Promotion Agency (IPA) and collaborates with multiple ministries, but it is primarily a research-driven organisation that studies the science and design of evaluations.[262] It thus lacks statutory any authority to mandate compliance. Finally, Korea's AISI, housed under the Electronics and Telecommunications Research Institute (ETRI), also focuses on evaluating potential AI risks, but is explicitly stated to be devoid of a regulatory function, similar to the UK AISI.[263] In all of these cases, the institutes very likely rely on voluntary cooperation from AI developers, meaning their evaluations, while influential, do not carry the force of law.

By slight contrast, the United States, Canada and Singapore have their AI evaluation mandates related to (but not embedded within) specific legal frameworks, granting their AISIs, which

<https://www.nist.gov/aisi/strategic-vision> on 23 November 2024; Japan AI Safety Institute, 'Overview of the AI Safety Institute' 1 November 2024, 4 https://aisi.go.jp/wp-content/uploads/2024/07/20240627_AboutAISI_en.pdf> on 23 November 2024.

[258] Allen G and Adamson G, *The AI safety institute international network: Next steps and recommendations*, 2024, 12-13.

[259] DSIT, 'Introducing the AI Safety Institute' 17 January 2024–< https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#establishment> on 26 February 2024.

[260] DSIT, 'Introducing the AI Safety Institute' 17 January 2024–< https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#establishment> on 26 February 2024.

[261] DSIT, 'Introducing the AI Safety Institute' 17 January 2024–< https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#establishment> on 26 February 2024.

[262] Ministry of Economy, Trade and Industry, 'Launch of AI Safety Institute' 14 February 2024–< https://www.meti.go.jp/english/press/2024/0214_001.html> on 26 February 2025.

[263] National Research Council of Science and Technology, 'AI Safety Institute launched as Korea's AI research hub' 27 November 2024–< https://www.newswise.com/articles/ai-safety-institute-launched-as-korea-s-ai-research-hub> on 26 February 2024.

are also advisory bodies, ostensible regulatory weight. The US AISI, housed within National Institute of Standards and Technology (NIST), was created by an executive order (Executive Order 14110 of 2023) and is tasked with developing national AI safety standards.[264] While the US AISI also lacks direct regulatory authority, its evaluations—focused on safety, security, and authenticity—carry significant weight due to NIST's role in shaping federal procurement and industry practices.[265] However, it is worth noting that the executive order establishing the US AISI has recently been rescinded by President Donald Trump and the implications of this action for the US AISI remain unclear.[266]

Canada's AISI, operates under Innovation, Science and Economic Development Canada (ISED) and is *linked* to the proposed Artificial Intelligence and Data Act [AIDA] (Bill C-27).[267] Its future role in "supporting" or "complimenting" the proposed AIDA as the Canadian government claims is quite unclear but, to the extent that it would operate within this legislative framework and/or directly inform regulatory obligations for AI developers, particularly in areas concerning safety evaluations and risk mitigation, then it could be said to have a direct regulatory function.[268] Singapore's AISI, while not fully regulatory, takes an intermediate approach. It is established under the Infocomm Media Development Authority (IMDA), which develops baseline policy on AI safety,[269] thus it is plausible to conclude that Singapore AISI's

[264] Department of Commerce, 'At the direction of President Biden, Department of Commerce to establish U.S. Artificial Intelligence Safety Institute to lead efforts on AI safety' 1 November 2023–< https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial> on 26 February 2025.

[265] Department of Commerce, 'At the direction of President Biden, Department of Commerce to establish U.S. Artificial Intelligence Safety Institute to lead efforts on AI safety' 1 November 2023–< https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial> on 26 February 2025.

[266] NIST, 'Executive order on safe, secure, and trustworthy artificial intelligence'–< https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence> on 26 February 2025; The White House, 'Removing barriers to American leadership in artificial intelligence' 23 January 2025–< https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/> on 26 February 2025.

[267] Government of Canada, 'Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute> on 26 February 2024; Government of Canada, 'Canada launches Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html> on 26 February 2025.

[268] Government of Canada, 'Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute> on 26 February 2024; Government of Canada, 'Canada launches Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html> on 26 February 2025.

[269] Singapore AISI, 'About us'–< https://sgaisi.sg/about-us> on 26 February 2025.

evaluation work informs national AI governance policy, making it regulatory-adjacent rather than strictly advisory.

Finally, there are two outliers, France and the EU. France's National Institute for Assessing and Securing Artificial Intelligence (INESIA) follows a slightly different approach, functioning as a coordination mechanism among existing regulatory bodies (such as ANSSI,[270] INRIA,[271] LNE,[272] and PEREN[273]) rather than a standalone legal entity.[274] While it plays a central role in AI *security* evaluations, its influence is indirect—shaping national policy rather than enforcing any legal (evaluation) standards.[275] In contrast, EU countries, have, as their AISI, the EU AI Office, whose AI Safety Unit performs similar functions to AISIs.[276] The EU AI Office, housed within the EU Commission,[277] implements the recent EU AI Act which, inter alia, classified AI systems with certain use cases as "high risk",[278] thus obligating developers to establish a risk evaluation system throughout the high risk AI system's lifecycle to ensure they meet strict safety and security standards.[279] Furthermore, even if a developer is of the opinion that their AI system is not "high-risk", they are nevertheless obligated to conduct an evaluation of the model pre-deployment or document one before placing the system in the EU market.[280]

[270] French National Agency for the Security of Information Systems.
[271] French Institute for Research in Computer Science and Automation.
[272] National Laboratory of Metrology and Testing.
[273] Expert Centre for Digital Regulation.
[274] Campus France, 'The French government creates a national institute to assess and secure AIs'–< https://www.campusfrance.org/en/actu/creation-d-un-institut-national-pour-l-evaluation-et-la-securite-de-l-ia> on 26 February 2025.
[275] Campus France, 'The French government creates a national institute to assess and secure AIs'–< https://www.campusfrance.org/en/actu/creation-d-un-institut-national-pour-l-evaluation-et-la-securite-de-l-ia> on 26 February 2025.
[276] Allen G and Adamson G, *The AI safety institute international network: Next steps and recommendations*, 2024, 10.
[277] EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/ > on 19 October 2024.
[278] Article 6 and Annex III, EU Artificial Intelligence Act; EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/ > on 19 October 2024.
[279] Articles 8-17, EU Artificial Intelligence Act; EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/ > on 19 October 2024; Digital Regulation Platform, 'Transformative technologies (AI) challenges and principles of regulation' The World Bank, 8th May 2024, < https://digitalregulation.org/3004297-2/ > on 19 October 2024; RSF, 'AI and the Right to Information RSF's 7 recommendations to protect an AI-dominated information space' May 2024 < https://rsf.org/sites/default/files/medias/file/2024/06/AI%20and%20the%20Right%20to%20Information%20-%20RSF%20Recommendations%20to%20EU.pdf > on 19 October 2024.
[280] EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/ > on 19 October 2024

The contrast observed between the above models—advisory and regulatory—highlights deeper policy choices about AI governance. It seems that nations prioritising flexibility and innovation tend to favour advisory bodies that generate evaluations without imposing direct legal consequences, while those viewing AI as a *risk-intensive technology*, a view adopted by this study and EU nations as well, integrate evaluations into binding regulatory frameworks.

B. Additional legal and policy frameworks governing AI evaluations

Aside from AISIs, China has implemented mandatory security assessments for public-facing generative AI, compelling developers to submit detailed self-assessment reports to an algorithm registry—though the precise standards remain undisclosed.[281] This framework, which builds on regulations introduced in 2021 and reinforced by rules for deep synthesis in 2022, ensures that developers are accountable for potential harms by mandating these (opaque) evaluations.[282] Additionally, in the UK, although the draft AI bill primarily mandates audits of business practices rather than direct model evaluations, the government has signalled a clear intention to develop robust evaluation mechanisms.[283]

Furthermore, although not labelled explicitly as "pre-deployment safety evaluations," Canada's proposed AIDA, which is modelled after the EU AI Act, mandates that developers put in place measures to identify, assess, and mitigate risks—including harms and biased outputs—before high-impact AI systems are made available for use.[284] It sets out detailed regulatory requirements that encompass all stages of an AI system's lifecycle, effectively encouraging thorough safety evaluations during design, development, and deployment.[285] Canada has also implemented the Directive on Automated Decision-Making, requiring government institutions to use the Algorithmic Impact Assessment (AIA) tool to identify,

[281] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025.
[282] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025.
[283] McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review> on 7 January 2025.
[284] Government of Canada, 'The Artificial Intelligence and Data Act (AIDA) – Companion document'–< https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> on 3 April 2025.
[285] Government of Canada, 'The Artificial Intelligence and Data Act (AIDA) – Companion document'–< https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document> on 3 April 2025.

measure, and manage risks associated with automated decision systems, ensuring transparency and accountability in AI use.[286]

Several legislations worldwide require risk-assessments for "high-impact AI" or "high-risk AI", including the EU AI Act (discussed in the previous section), South Korea's Basic Act on AI Advancement and Trust, Brazil's Proposed AI Regulation Bill (PL 2338/2023), and Colorado Senate Bill 24-205.[287] Virginia's vetoed AI Bill (HB 2094) mirrored Colorado's requirements, including pre-deployment risk management and transparency, but faced rejection over concerns about stifling innovation.[288] Still in the US, although unrelated to "high-risk AI," New York City's Local Law 144 uniquely enforces independent bias audits for AI-driven hiring tools, requiring employers to notify candidates and offer alternative evaluations.[289] Federally, however, no binding laws on safety evaluations exist.[290]

4.4 Gaps in existing pre-deployment evaluations

Various gaps have been pointed out regarding existing pre-deployment evaluations. First, it has been argued that most current safety assessments lack contextual evaluation.[291] They focus on generative AI capabilities in isolation, neglecting potential risks at the point of human interaction or systemic impact, especially in the global south.[292] In particular, evaluations designed in Western contexts frequently overlook harms in other regions, while reliance on English-language data exacerbates biases in under-resourced languages.[293] Thus, models may perform well in controlled, anglophone settings but falter globally. This context problem has

[286] Government of Canada, 'Algorithmic Impact Assessment tool'–< https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html> on 3 April 2025.
[287] Modulos, 'A curated global guide to AI compliance: Navigating international AI regulations'–< https://www.modulos.ai/global-ai-compliance-guide/> on 3 April 2025.
[288] Reich A and Holmes J, 'Virginia passes second of its kind AI anti-discrimination statute, but law vetoed by governor' Saul Ewing, 1 April 2025–< https://www.saul.com/insights/blog/virginia-passes-second-its-kind-ai-anti-discrimination-statute-law-vetoed-governor> on 3 April 2025.
[289] Morris, Manning & Martin LLP, 'State AI laws continue to emerge' 4 March 2025–< https://www.mmmlaw.com/news-resources/102k2k3-state-ai-laws-continue-to-emerge/> on 3 April 2025.
[290] White & Case, 'AI watch: Global regulatory tracker– United States' 31 March 2025–< https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states> on 3 April 2025.
[291] Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023, 14.
[292] Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023, 14.
[293] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 59.

led to various calls for tailored AI development in Africa.[294] Indeed, addressing data bias is a crucial objective for African nations. This priority is evident not only in the national AI strategies and policy frameworks of countries like Mauritius,[295] Egypt,[296] Kenya,[297] and South Africa,[298] but also in the broader Continental AI Strategy.[299] Therefore, as discussed in chapter 2, context is crucial in AI safety, particularly when determining what constitutes safe performance.

Secondly, most safety assessments or capability evaluations are limited in scope, often narrowly defining harms and missing many risk areas.[300] Evaluators often fixate on narrow issues like textual bias, while struggling to address systemic risks—such as economic disruption or cultural erosion—due to the sheer complexity of modelling real-world interdependencies.[301] Even where risks are acknowledged, debates rage over how to define or measure them, stalling consensus.[302] Finally, many evaluations leave out multimodality. Existing evaluations predominantly address text output, leaving significant gaps in assessing risks in image, audio, or video modalities, especially as AI systems increasingly integrate multiple modalities.[303]

4.5 Conclusion

This chapter set out to map the current approaches to pre-deployment evaluations, highlighting both established methods and significant gaps. The review of legal and institutional frameworks has revealed inconsistencies and a lack of context-specific standards. These

[294] Diallo K , Smith J, Okolo C, Nyamwaya D , Kgomo J, and Ngamita R, 'Case Studies of AI Policy Development in Africa' Arxiv, 2024, 1-11; African Union Development Agency, 'AI and the future of work in Africa: How AI is redefining opportunities' 3 July 2024–< https://www.nepad.org/blog/ai-and-future-of-work-africa-how-ai-redefining-opportunities > on 4 November 2024; Adams R, 'AI in Africa: Key concerns and policy considerations for the future of the continent' Policy Brief No. 8, Africa Policy Research Institute 2022,–< https://afripoli.org/uploads/publications/AI_in_Africa.pdf> on 4 November 2022.
[295] Ministry of Finance and Economic Development, *Mauritius Artificial Intelligence Strategy*, 2018, 16-17.
[296] The National Council for Artificial Intelligence, *Egypt National Artificial Intelligence Strategy*, 2023, 8,21,27,57.
[297] Ministry of Information, Communications and the Digital Economy, *Kenya National Artificial Intelligence (AI) Strategy 2025 – 2030*, 2025, 12,42,45.
[298] Department of Communications and Digital Technologies, *South Africa National Artificial Intelligence Policy Framework*, 2024, 1,10-11.
[299] African Union, *Continental AI Strategy*, 2024, 4, 25.
[300] Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023, 14.
[301] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 59.
[302] Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024, 59.
[303] Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023, 14.

findings demonstrate the necessity for a more integrated approach, one that leverages both legal mandates and innovative evaluation practices. The analysis here prepares the ground for the next chapter, which proposes specific legal interventions to strengthen these evaluations.

5 How Law Could Be Used To Strengthen Pre-Deployment Evaluations For Advanced AI in African Clinical Medicine

5.1 Introduction

Building upon the identified gaps in Chapter 4, this chapter explores how law can be designed to fortify pre-deployment evaluations. It examines legal mandates that compel developers to conduct thorough, context-sensitive evaluations, as well as mechanisms that facilitate government involvement and the growth of an independent evaluation industry. By considering how regulatory oversight and developer responsibility relate, the chapter sets out to propose a balanced, legally enforceable framework tailored to the needs of African clinical medicine.

5.2 Role of legally mandated tailored pre-deployment evaluations

From the previous chapter, it is well established that one crucial function, if not the main function, of pre-deployment evaluations is to ensure that AI systems will not pose public safety risks once they are deployed. This extends to the current subject matter—context-specific risks to clinical medicine in Africa (as discussed in 3.3.2). Additionally, from the previous chapter, it was pointed out that, although they are incredibly useful in mitigating risks, one main gap in current pre-deployment evaluations is that they lack contextual evaluation. They are prone to ignore or overlook harms in other regions, including the harms of interest here. Consequently, it can be inferred that, *a fortiori*, tailored pre-deployment evaluations, those that consider the context-specific risks to African clinical medicine (as earlier discussed in Chapter 3), could help mitigate those risks.

But should these evaluations be "legally mandated"? The term "legally mandated evaluations" denotes a situation where pre-deployment assessments of advanced AI systems are not optional but rather enforceable legal requirements. Thus, developers would be compelled to subject their systems to rigorous testing before integration into healthcare infrastructure. From Chapter 4, it was observed that most nations have AISIs conducting evaluations on an advisory basis while others, such as those in the EU, view AI as a *risk-intensive technology*, and thus integrate evaluations into a binding regulatory framework—the EU AI Act. In other words, such evaluations are "legally mandated." This study agrees with this latter approach and contends, based on the two premises below, that tailored, "legally mandated evaluations" would be

indispensable for ensuring the safe integration of advanced AI systems into African healthcare, thus preventing the context-specific risks discussed earlier in Chapter 3.

First, entrusting the entire responsibility to developers is very problematic. Developers, whose primary focus is often driven by commercial interests, may inadvertently (or, more likely, deliberately) underplay safety concerns to accelerate market entry. This commercial foundation creates a conflict of interest. Consequently, relying solely on self-regulation by developers risks compromising the safety of AI systems for expediency or profit. To prevent this, regulation is crucial. Not only does it compel adherence, but it also establishes clear accountability or enforceability mechanisms. Who is liable? Are these requirements justiciable? A duty has to be imposed on someone for enforcement to occur.[304] Once that is clear, any failure to meet these mandated standards would invoke tangible legal consequences, such as fines, penalties, or deployment restrictions. Thus, legal mandates would serve as a powerful deterrent against lax evaluation practices in mitigating risks to clinical medicine in Africa.

Second, the binding nature of legal mandates brings with it a *uniformity* and *consistency* that self-imposed standards simply cannot achieve. For starters, codifying evaluation requirements removes (most of the) ambiguity about what constitutes "sufficient" testing. Moreover, a statutory mandate compels all developers to allocate resources to safety evaluations, even when profitability is on the line, something that would be hard to achieve consistently across various rational actors acting independently. Additionally, legal mandates also address the *asymmetry of expertise*: individual African regulators, many of them under-resourced, cannot realistically negotiate bespoke evaluation terms with each AI developer. Instead, binding legal standards level the playing field, ensuring that an advanced AI system deployed in Nairobi undergoes almost entirely, if not entirely, the same level of scrutiny as one in Lagos—a uniformity virtually impossible under fragmented self-regulation as developer's size, market power, or internal policies would possibly influence the level of scrutiny. Thus, legal mandates would ensure that the tailored evaluations meant to mitigate context-specific risks to African clinical medicine are largely uniform and consistent.

[304] Andrews M, 'Hohfeld's cube', 16(3) *Akron Law Review*, 1983, 472-473; Cullison A, 'A review of Hohfeld's fundamental legal concepts' 16(3) *Cleveland State Law Review*, 1967, 559-561; Hohfeld W, 'Some fundamental legal conceptions as applied in judicial reasoning' 23(1) *Yale Law Journal*, 1913, 32.

5.3 Legal mechanisms to support pre-deployment evaluations

5.3.1 Legal mandates for government involvement in evaluations

What is the role of government in regulation? Additionally, given that we expect governments to conduct pre-deployment AI evaluations to determine "safety" or the level of risk, what usually happens when a product certified as "safe" by government causes harm to society? These are important questions to answer before deciding what role governments should play with regards to pre-deployment evaluations for the AI-related risks discussed in this study and how the law can operationalise that role.

In many established industries, e.g., pharmaceuticals, aviation, automotives, etc, government agencies—such as the US Food and Drug Administration (FDA), US Federal Aviation Administration (FAA), and US National Highway Traffic Safety Administration (NHTSA)—are tasked with certifying that products meet safety standards. For pharmaceuticals, drugs undergo multiple phases of clinical trials to ensure safety and efficacy and then the FDA reviews trial data before approving drugs.[305] For aviation, aircraft must undergo extensive and stringent FAA certification standards before being certified.[306] Similarly, for vehicles, the NHTSA, enforces safety standards.[307]

In these industries, when accidents happen or products are defective, manufacturers are almost always the primary defendants in litigation arising from product failures. What about the agencies that certified them? Well, historically, the doctrine of *sovereign immunity* generally protected the government from being sued without its consent.[308] However, Congress sought to limit this immunity by enacting the US Federal Tort Claims Act (FTCA) which allows individuals to sue the US government for certain torts committed by federal employees acting within the scope of their employment.[309] However, within this Act, there are certain exceptions which allow the US government to retain its immunity, one of which is the "discretionary

[305] FDA, 'The FDA's drug review process: Ensuring drugs are safe and effective' FDA, 24 November 2017–< https://www.fda.gov/drugs/information-consumers-and-patients-drugs/fdas-drug-review-process-ensuring-drugs-are-safe-and-effective> on 22 March 2024.
[306] FAA, 'What we do' FAA, 27 June 2016–< https://www.faa.gov/about/mission/activities> on 22 March 2024.
[307] NHTSA, 'About NHTSA' NHTSA–< https://www.nhtsa.gov/about-nhtsa#:~:text=The%20National%20Highway%20Traffic%20Safety%20Administration%20is%20responsible,injuries%20and%20economic%20losses%20from%20motor%20vehicle%20crashes.> on 22 March 2024.
[308] Rodgers T, 'Sovereign immunity or: How the federal government learned to stop worrying and love the discretionary function exception' 63(9), *Boston College Law Review*, 2022, 17.
[309] Rodgers T, 'Sovereign immunity or', 17-18.

function" exception. This exception shields the government from liability for actions that: (1) involve judgment or choice and (2) are grounded in social, economic, or political policy.[310] Consequently, US agencies are largely shielded from liability when they certify a product as safe, provided their actions fall within this (or other) exceptions.

Assumng an agency that oversees or regulates the deployment of advanced AI systems in African countries (by certifying *evaluated* models for deployment) is a desirable idea,[311] should a similar approach with regards to their liability be applied here? The author opines that it is important to limit the liability of the regulator, particularly to avoid overwhelming the agency with frivolous litigation that could divert vital resources. However, there should also probably be defined instances when the agency can be held accountable, for example, gross negligence by its employees, or failure to follow due process.

However, should the government's role should be limited to certification? Recall that in the background to the study as well as chapters 2 and 3, it was demonstrated that these context-specific risks arise primarily due to data bias. Specifically, in sections 2.3 and 3.3.2, it was noted that AI models are overwhelmingly trained on datasets from well-resourced regions, meaning they often fail to capture African-specific medical realities. Of course, this is expected given that it is likely cheaper to aggregate data from these well-resourced regions compared to lower-resourced ones. Yet, this is probably an issue that African governments could also help in addressing. Indeed, the African Union considers "human biases influencing AI design" to be "significant concerns,"[312] and thus has, as its sixth objective in the Continental AI Strategy, the goal of ensuring the "*availability* of high-quality and diverse datasets for AI."[313]

Based on the above considerations, it is suggested that an agency regulating the deployment of advanced AI systems in, inter alia, healthcare, adopt, at minimum, the following functions: (1) a facilitative role; and (2) a regulatory and enforcement role. With regards to the former role, it is possible that individual developers, however proficient, may lack the comprehensive data

[310] Rodgers T, 'Sovereign immunity or', 18.
[311] It is desirable. As observed from the previous chapter, AISIs lack regulatory powers for the most part. This should not be emulated. Regulatory agencies on this matter should have enforcement powers given the view that this is a risk-intensive technology. Ministers of ICT—working in tandem with their Health counterparts—should be the ones to establish these independent regulatory agencies, modelled in part on bodies like the UK AI Safety Institute, which is established by DSIT.
[312] African Union, *Continental AI Strategy*, 2024, 25.
[313] African Union, *Continental AI Strategy*, 2024, 4.

access, infrastructure (less likely), and other information necessary to conduct truly robust pre-deployment evaluations. To remedy this, law can assign government a *facilitative* role.

Legislation should mandate that governments, to the best of their ability, *aggregate* and *share* high-quality, representative, region-specific, healthcare-pertinent *datasets* with developers for use in their evaluations, barring any legitimate reasons why this should not be done. For example, a legal requirement for governments to aggregate *anonymised* health data would provide developers with representative testing material. This would be possible because, currently, de-identification of health data, a process designed to remove or obscure elements that could be used to identify individuals, in countries like the US and Kenya,[314] allows the information to be used for research, policy analysis, and public health without violating privacy rights.[315] So, such legislation would make the government a duty bearer with regards to facilitating data access. On the other hand, developers would have a justiciable claim in court that their right to this data has been violated and the government had no legitimate reason to deny them that data.

Legislation should also mandate that governments provide clear, standardised regulatory *guidance*[316] on evaluation methodologies; as well as guidance on better integration with national digital infrastructure, if necessary. This way, governments help level the playing field and try to achieve their level of protection through their published guidance on how to conduct these evaluations.

With regards to the latter role, this separate arm should be empowered to verify and certify that developers meet standardised criteria. What would compliance with these mandatory standards mean exactly? It would probably mean following the agency's codified process—using representative data, testing for region-specific risks (e.g., linguistic diversity of the model,

[314] Section 37(1)(b) and 37(2), *Data Protection Act* (Act No. 24 of 2019); Section 164.514 of the HIPAA Privacy Rule. Under the US Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule, for example, health data is no longer considered "protected" once it has been de-identified—meaning it contains no identifiers or any reasonable basis to infer an individual's identity. US Department of Health and Human Services, 'Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule' US HHS, 3 February 2025–< https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html> on 13 February 2025.
[315] US Department of Health and Human Services, 'Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule' US HHS, 3 February 2025–< https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html> on 13 February 2025.
[316] This is qualified under section 5.4 however.

NTDs and traditional medicine knowledge, representative genetic data, etc), submitting internal evaluations for verification, etc. Thus, certification of the developer's assessments by the agency should validate *process integrity*, not product safety.[317]

In addition, their mandate would also include maintaining national public registries of AI systems deployed in healthcare, licensing the use of certified AI systems for defined periods, and conducting regular audits of post-deployment performance. They should also be imbued with enforcement powers such as the ability to levy fines, revoke licenses, or impose market restrictions for non-compliance. Finally, these agencies should be free from conflicts of interest and supported by dedicated government funding to build institutional capacity, ensuring that regulators possess the necessary expertise and resources to evaluate advanced AI systems effectively.

5.3.2 Incentivising a private sector AI evaluation industry

Given the vast array of AI applications and the specialised knowledge required to assess them, it is unrealistic to expect governments alone to shoulder the evaluation burden. Instead, the law should actively encourage the emergence of independent AI evaluation firms that would work under the supervision of the enforcement arm of the agency. African countries should create (or adapt current) legal frameworks to support these entities through grants, tax incentives, and even soft loans. This would foster a competitive market in which third-party evaluators are not only available but also held to high standards of impartiality and rigour.

Such legal and economic incentives would serve multiple purposes. They would reduce the regulatory body's workload by outsourcing routine, though essential, evaluation tasks to *accredited* private firms; they would stimulate job creation and skill development in the growing field of AI safety; and they would help establish industry-wide benchmarks that developers must meet. Licensing and accreditation requirements for these firms would be essential, ensuring that they adhere to standardised methodologies from the Agency and maintain the confidentiality of sensitive data. The net effect would be a vibrant ecosystem of independent evaluators who can perform detailed assessments, thus complementing government facilitation and providing robust checks on developer-led evaluations.

[317] Laws must explicitly state that certification is not equal to endorsement. A clause like *"Agency approval confirms adherence to evaluation standards, not absolute safety"* would prevent developers from shifting blame.

5.3.3 Legal obligations for AI developers

While government facilitation and independent evaluators are critical, the primary onus must remain on developers, who are best positioned to conduct the initial evaluations of their systems. However, it is important to clarify here that the term "developers", unless stated (or context requires) otherwise, refers to both upstream developers (who create foundational models) and downstream developers (who adapt these models for specific uses, such as healthcare applications). As Sophie Williams et al. observe, downstream modifications—whether intentional or accidental—can amplify risks.[318] Consequently, the regulatory approach opted for has to be layered, in that it addresses both tiers of development. In other words, each actor's obligations must be defined distinctly but coherently, such that accountability does not dissipate between handovers.

Both developers should be required to conduct evaluations. In recognition of the fact that developers' internal evaluations may be skewed by commercial pressures, law should require that they make "reasonable efforts" to perform context-specific assessments tailored to the unique challenges of African healthcare. This means, for example, that developers must rigorously test their systems against diverse, regionally representative datasets and, inter alia, address biases such as those relating to linguistic diversity by including low-resource languages in their test scenarios. Downstream developers should be restricted to modifying models that have undergone evaluations from upstream developers and be obligated to conduct evaluations with respect to the modifications they make.

To operationalise this, both developers should be legally obligated to submit detailed evaluation reports to regulatory bodies. These reports must document the methodologies used, and any identified risks or limitations. Intended use cases should also be documented. Regulatory agencies, upon receiving these disclosures, should certify the system as safe for deployment and grant the upstream or downstream developer a renewable deployment license for a certain period, or, if necessary, request additional testing or independent verification for systems deemed high-risk. New licenses should only be granted after comprehensive pre-deployment evaluations by the developer taking into account the evolution of AI systems and any other pertinent changes since the last evaluation. Both developers should also, before being granted any licence, be obliged to commit to iterative testing and post-deployment monitoring

[318] Williams S, Schuett J, and Anderljung M, 'On regulating downstream AI developers' ArXiv, 2025, 8-12.

by the relevant regulatory agency. Additionally, both developers should be required to carry liability insurance, ensuring victims of AI harm have recourse. Lastly, downstream developers should be required to register modifications and assume liability for them. However, If harm arises due to failures attributable to upstream developers—e.g. fundamental flaws in the foundation model—regulatory frameworks should permit claims to be pursued across the responsibility chain, not only against downstream actors.

5.4 Potential implications

One common critique of mandating extensive legal obligations is that excessive regulation might stifle innovation and deter investment. It is indeed possible that overly burdensome requirements could slow the pace of AI deployment; however, as discussed earlier, economic incentives, such as tax credits and grant schemes, could help offset compliance costs. Additionally, the potential harms render such regulation both necessary and justifiable.

In addition, the practical implementation of such legal frameworks in African contexts must contend with political instability, resource constraints, and varying levels of institutional capacity. While expecting governments to aggregate data and provide certification services seems burdensome, regional cooperation, through having only one regional regulator (as proposed in section 6.2), could be more cost-effective for African countries. This way, they could pool enough resources to run the agency and come up with best practices together. Of course, the implications of this are unclear as, even if the political will was present, these countries could have contexts that vary too much to have one implementing agency.

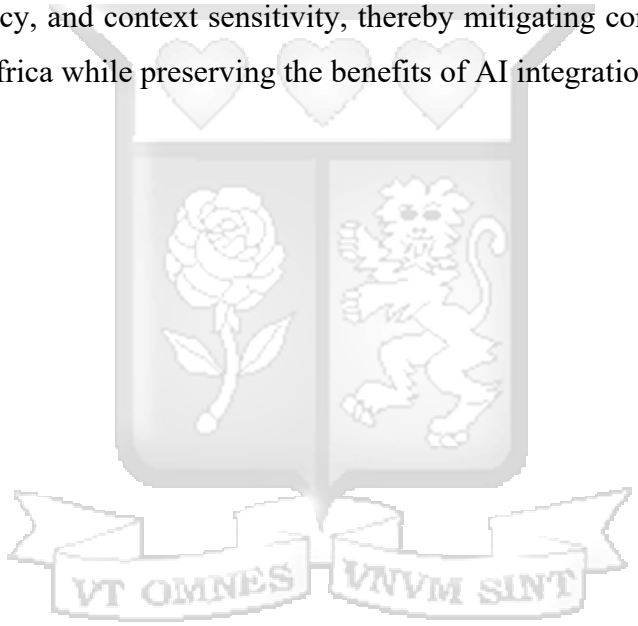
Lastly, one major challenge is the tension between transparency and the protection of proprietary information. On one hand, legal mandates to share detailed evaluation data can enhance oversight and public trust; on the other, they risk exposing trade secrets and sensitive intellectual property. A solution to this issue as inspired by Ben Bucknall et al., could be ensuring privacy for both the model owner and the external evaluator for effective external evaluation.[319] This could be achieved by: (i) ensuring model owners (developers) and evaluators are bound to protect each other's private information, should they come across it; (ii) prohibiting the government agency and its employees, as well as accredited evaluators and their employees, from publishing or disclosing (whether to developers or the public) detailed

[319] Bucknall B, Trager R, and Osborne M, 'Position: Ensuring mutual privacy is necessary for effective external evaluation of proprietary AI systems' ArXiv, 2025, 1-13.

or classified evaluation methodologies, processes, or guidance;[320] and (iii) incorporating modern technical measures (e.g. secure enclaves and cryptographic protocols as suggested by Bucknall et al.) to enable evaluators to access AI systems without exposing sensitive evaluation details.[321] However, publication of the high-level results (e.g., safety scores) of evaluations, which do not contain any sensitive information, should be possible.

5.5 Conclusion

This chapter set out to demonstrate that a robust legal framework can significantly enhance pre-deployment evaluations. It has argued for a tripartite approach involving mandated developer evaluations, government facilitation of their evaluations and certification of them, and incentives for independent evaluators. The proposed legal mechanisms aim to ensure uniformity, transparency, and context sensitivity, thereby mitigating context-specific risks to clinical medicine in Africa while preserving the benefits of AI integration.



[320] The Volkswagen scandal demonstrated how public testing parameters can be exploited. Likewise, evaluators need to trust that their methods will not be (or have not been) compromised and public disclosure, if not prohibited, could very well enable developers to "game" the system. Such a prohibition would ensure that both evaluators' methods and proprietary information remain confidential thus preventing potential manipulation, which would undermine the integrity of evaluations.

[321] This would help ensure that sensitive aspects from both sides remain protected.

6 Conclusion and Recommendations

6.1 Conclusion

This study set out to examine the context-specific risks posed by advanced AI models to clinical medicine in African countries, and whether legally mandated pre-deployment safety evaluations tailored to local contexts can mitigate these risks. In Chapter 2, the study traced the evolution of AI safety from its technical origins to a broader sociotechnical approach, elucidating the critical role of context in shaping both risks and benefits. Chapter 3 detailed both the opportunities and challenges presented by AI in African clinical medicine. It highlighted that while AI holds promise for transforming clinical practice in Africa through precision medicine, drug discovery and resource optimisation, it also carries serious, context-specific risks. These risks stem from biased training data in respect of African low resource languages, African traditional medicine, African genetic diversity, and diseases endemic to Africa. These insights paved the way for the next discussion on pre-deployment testing in Chapter 4.

Chapter 4 mapped the current landscape and identified key gaps in existing evaluation practices. Finally, Chapter 5 explored the legal mechanisms that can support tailored pre-deployment evaluations aimed at ensuring safe AI integration into African clinical medicine. The findings demonstrate that relying solely on developer-led evaluations is insufficient, given inherent commercial interests. Instead, government facilitation of evaluations and certification of developer's assessments, economic incentives for independent evaluators, and clear obligations on AI developers can significantly enhance safety and accountability.

Future researchers are encouraged: (1) to investigate the long-term impact of such legal interventions on clinical outcomes, particularly whether they are an insurmountable obstacle to the benefits of AI integration in clinical medicine, and (2) to explore more deeply the potential for cross-border regulatory cooperation within the African Union framework, an idea raised briefly in section 5.4. This study aimed to lay a solid foundation for the legal regulation of AI systems in clinical medicine. With continued and careful effort, clinical medicine in Africa could set a global standard for responsible AI integration.

6.2 Recommendations

First, as mentioned in section 5.4, having one regional regulator could be cost-effective for African countries as well as help in generating best practices in the continent. Consequently, African nations should empower the Africa CDC, under an African Union treaty or protocol, to take the lead in developing contextually grounded model evaluation standards and metrics for AI systems in health. These metrics, agreed upon by consensus or majority, could then guide a network of national evaluation bodies, which would adapt them to national needs while adhering to continental baselines or criteria. Periodic reviews and updates to the Africa CDC's standards would also be essential to keep pace with technological advances and evolving clinical contexts. Of course, whether such an arrangement would be politically viable depends on the level of buy-in from Member States and the scope of legal authority conferred upon the Africa CDC.

Second, there should be mandatory linguistic validation for advanced AI systems deployed in healthcare. The Africa CDC would collaborate with the African Academy of Languages to develop continent-wide benchmarks on the linguistic capabilities of AI systems (e.g., requiring AI models to achieve ≥90% accuracy in interpreting 20+ African languages). These benchmarks would prioritise cross-border languages like Swahili, Yoruba, and Amharic. Private, accredited evaluators could use these benchmarks to certify systems.

Third, African states should adopt an AU-wide mutual recognition framework where AI systems certified in one member state are automatically accepted in others, provided they meet Africa CDC criteria. However, member states should be able to revoke recognition if local evaluations uncover context-specific risks or weaknesses warranting recalling the AI system that were undetected by regulators in one member state.

Fourth, similar to the mechanism provided for under Article 57 of the EU AI Act,[322] African countries should each integrate a 'regulatory sandbox' specifically designed for testing AI systems in controlled, real-world simulations before deployment. In this sandbox, developers—both upstream and downstream—could trial their models using standardised evaluation protocols derived from the Africa CDC's metrics. This would allow regulators to

[322] European Union, 'Article 57: AI regulatory sandboxes'–< https://artificialintelligenceact.eu/article/57/> on 4 April 2025.

observe AI systems in a simulated environment, identify latent risks to clinical medicine in Africa, and provide iterative feedback to developers.

Lastly, in recognition that the effectiveness of almost all regulatory frameworks is contingent upon the skills of its practitioners, it is recommended that governments, in partnership with academic institutions and AI industry stakeholders, invest in capacity building initiatives. These programs should focus on developing specialised training and certification courses in AI safety evaluations, especially when there are serious leaps in the advancement of the technology being evaluated, thereby equipping employees of the regulator and those of independent, accredited evaluators with the technical expertise required to scrutinise advanced AI systems *keenly* and *accurately*. Such investment is not only practicable (e.g., through the allocation of public funds, grants, or public–private partnerships with relevant entities), but also essential for establishing a sustainable evaluations ecosystem.

7 Bibliography

Chapters in Books

- Fernandes M and Goldim J, 'Artificial intelligence and decision making in health: Risks and opportunities' in Antunes H, Freitas P, Oliveira A, Pereira C, Sequeira E and Xavier L (eds) *Multidisciplinary perspectives on artificial intelligence and the law*, Springer, Cham, 2024, 187-205.
- Martuzzi M and Tickner J, 'Introduction – the precautionary principle: protecting public health, the environment and the future of our children' in Martuzzi M and Tickner J (eds), *The precautionary principle: protecting public health, the environment and the future of our children*, WHO Regional Office for Europe, Copenhagen, 2004, 7-14.
- Parrott L, 'Health and risk policymaking, the precautionary principle, and policy advocacy' in *Oxford Research Encyclopaedia of Communication, Oxford University Press, Oxford, 2017, 1-37.*
- Pearce N, 'Public health and the precautionary principle' in Martuzzi M and Tickner J (eds), *The precautionary principle: protecting public health, the environment and the future of our children*, WHO Regional Office for Europe, Copenhagen, 2004, 49-62.
- Sirak K, Sawchuk E, and Prendergast M, 'Ancient human DNA and African population history' in *Oxford Research Encyclopaedia of Anthropology*, Oxford University Press, Oxford, 2022, 1-43.

Journal Articles

- Alban S, 'The 'precautionary principle' as a guide for future drug development' 35(1) *European Journal of Clinical Investigation*, 2005.
- Aldoseri A, Khalifa A, and Hamouda A, 'Re-thinking data strategy and integration for artificial intelligence: Concepts, opportunities, and challenges' 13 *Applied Sciences*, 2023.
- Ali O, Abdelbaki W, Shrestha A, Elbasi E, Alryalat M, and Dwivedi Y 'A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities' 8 *Journal of Innovation & Knowledge*, 2023.
- Almeida S, Norajitra T, Lüth C, Wald T, Weru V, Nolden M, Jäger P, von Stackelberg O, Heußel C, Weinheimer O, Biederer J, Kauczor H, Maier-Hein K, 'How do deep-learning models generalize across populations? Cross-ethnicity generalization of COPD detection' 15(1) *Insights into Imaging*, 2024.
- Alowais S, Alghamdi S, Alsuhebany N, Alqahtani T, Alshaya AI, Almohareb S, Aldairem A, Alrashed M, Saleh K, Badreldin H, Al Yami M, Al Harbi S, and Albekairy A, 'Revolutionizing healthcare: The role of artificial intelligence in clinical practice' 23(1) *BMC Medical Education*, 2023.
- Andrews M, 'Hohfeld's cube', 16(3) *Akron Law Review*, 1983.
- Anyikwa C, 'Exploring the role of divination in traditional medicine in Africa: A critical perspective' 2 *Research Directions: One Health*, 2024.
- Aruna B and Priyanka P, 'A review article on artificial intelligence in traditional medicine research and application' 9(11) *International Journal Of Novel Research And Development*, 2024.
- Banja J, 'How might artificial intelligence applications impact risk management?' 22(11) AMA Journal of Ethics, 2020.

- Bentley Amy, Callier S, and Rotimi C, 'Evaluating the promise of inclusion of African ancestry populations in genomics' 5(5) *Genomic Medicine*, 2020.
- Botes M, 'Regulating scientific and technological uncertainty: The precautionary principle in the context of human genomics and AI' 119(5) *South African Journal of Science*, 2023.
- Botha N, Segbedzi C, Dumahasi V, Maneen S, Kodom R, Tsedze I, Akoto L, Atsu F, Lasim O, and Ansah E, 'Artificial intelligence in healthcare: A scoping review of perceived threats to patient rights and safety' 82(188) *Archives of Public Health*, 2024.
- Coleman J and Pontefract S, 'Adverse drug reactions' 16(5) *Clinical Medicine*, 2016.
- Cullison A, 'A review of Hohfeld's fundamental legal concepts' 16(3) *Cleveland State Law Review*, 1967.
- Dave M and Patel N, 'Artificial intelligence in healthcare and education' 234(10) *British Dental Journal*, 2023.
- Dixon J and Chandler C, 'Opening up 'fever', closing down medicines: Algorithms as blueprints for global health in an era of antimicrobial resistance' 6(4) *Medicine Anthropology Theory*, 2019.
- Do C, 'Difficulties and suggestions for medical English translation: Insights from experience' 101(1) *Journal of Literature, Languages and Linguistics*, 2024.
- Dobbe R and Wolters A, 'Toward sociotechnical AI: Mapping vulnerabilities for machine learning in context' 34(12) *Minds and Machines*, 2024.
- Ekor M, 'The growing use of herbal medicines: Issues relating to adverse reactions and challenges in monitoring safety' 4(177) *Frontiers in Pharmacology*, 2014.
- Fischer A and Ghelardi G, 'The precautionary principle, evidence-based medicine, and decision theory in public health evaluation' 4 *Frontiers in Public Health*, 2016.
- Goklany I, 'From precautionary principle to risk–risk analysis' 20 *Nature Biotechnology*, 2002.
- Goldstein B, 'Precautionary principle and public health' 91(9) *American Journal of Public Health*, 2001.
- Hohfeld W, 'Some fundamental legal conceptions as applied in judicial reasoning' 23(1) *Yale Law Journal*, 1913.
- Keller M, Rohner M, and Honigmann P, 'The potential benefit of artificial intelligence regarding clinical decision-making in the treatment of wrist trauma patients' 19(1) *Journal of Orthopaedic Surgery and Research*, 2024.
- Kist J, Vos R, Mairuhu A, Struijs J, van Peet P, Vos H, van Os H, Beishuizen E, Sijpkens Y, Faiq M, Numans M, Groenwold R, 'SCORE2 cardiovascular risk prediction models in an ethnic and socioeconomic diverse population in the Netherlands: an external validation study' 57 *e Clinical Medicine*, 2023.
- Konatam S, Konatam V, and Konatam S, 'A novel of detecting and addressing bias in artificial intelligence and machine learning: A multi-industry perspective' 11(6) *International Research Journal of Engineering and Technology*, 2024.
- Lang I, ''Laws of fear' in the EU: Precautionary principle and public health restrictions to free movement of persons in the time of COVID-19' 14(1) *European Journal of Risk Regulation*, 2021.
- Lazar S and Alondra N, 'AI safety on whose terms?' 381 (6654) *Science*, 2023.
- Levin E, 'Different use of medical terminology and culture-specific models of disease affecting communication between Xhosa-speaking patients and English-speaking doctors at a South African paediatric teaching hospital' 96(10) *South African Medical Journal*, 2006.

- Luxon L, 'Infrastructure – the key to healthcare improvement' 2(1) *Future Hospital Journal*, 2015.
- Malaterre A, Rakoto M, Marodon C, Bedoui Y, Nakab J, Simon E, Hoarau L, Savriama S, Strasberg D, Guiraud P, Selambarom J, and Gasque P, '*Artemisia annua*, a traditional plant brought to light' 21(14) *International Journal of Molecular Sciences*, 2020.
- Martin G, Cirino A, Barcelona V, Fox K, Hudson M, Sun Y, Taylor J, and Cameron V, 'Considerations for cardiovascular genetic and genomic research with marginalised racial and ethnic groups and indigenous peoples: A scientific statement from the American Heart Association' 14(4) *Circulation: Genomic and Precision Medicine*, 2021.
- Mitra A and Mawson A, 'Neglected tropical diseases: Epidemiology and global burden' 2(36) *Tropical Medicine and Infectious Disease*, 2017.
- Mumford E, 'A socio-technical approach to systems design' 5(1), *Requirements Engineering*, 2000.
- Mumford E, 'The story of socio-technical design: Reflections on its successes, failures and potential' 16(1), *Info Systems Journal*, 2006.
- Norori N, Hu Q, Aellen F, Faraci F and Tzovara A, 'Addressing bias in big data and AI for health care: A call for open science' 2(10) *Patterns*, 2021.
- Palavicini G, 'Intelligent health: Progress and benefit of artificial intelligence in sensing-based monitoring and disease diagnosis' 23(22) *Sensors*, 2023.
- Parveen A, Parveen B, Parveen R, and Ahmad S, 'Challenges and guidelines for clinical trial of herbal drugs' 7(4) *Journal of Pharmacy & Bioallied Sciences*, 2015.
- Patel P, 'Forced sterilization of women as discrimination' 38(15) *Public Health Reviews*, 2017.
- Patel V, 'Cognitive science in the evaluation of medical AI systems' 30(1) *BMJ Health Care Informatics*, 2023.
- Peprah E, Xu H, Ayele F, and Royal C, 'Genome-wide association studies in Africans and African Americans: Expanding the framework of the genomics of human traits and disease' 18(1) *Public Health Genomics,* 2015.
- Prajapati J and Shukla A, 'The role of genetic variations in drug metabolism and their impact on therapeutic outcomes' 2(2) *International Journal Of Advance Research In Multidisciplinary*, 2024.
- Putri T, 'An analysis of types and causes of translation errors' 3(2) *Etnolingual*, 2019.
- Qureshi R, Irfan M, Gondal TM, Khan S, Wu J, Hadi MU, Heymach J, Le X, Yan H, and Alam T 'AI in drug discovery and its clinical relevance' 9(7) *Heliyon*, 2023.
- Rajpurkar P, Chen E, Barnerjee O, and Topol E, 'AI in health and medicine' 28(1) *Nature Medicine*, 2022.
- Rodgers T, 'Sovereign immunity or: How the federal government learned to stop worrying and love the discretionary function exception' 63(9), *Boston College Law Review*, 2022.
- Roppelt J, Kanbach D, Kraus S, 'Artificial intelligence in healthcare institutions: A systematic literature review on influencing factors' 76(1) *Technology in Society*, 2024.
- Rudolph J, Huemmer C, Preuhs A, Buizza G, Hoppe BF, Dinkel J, Koliogiannis V, Fink N, Goller SS, Schwarze V, Mansour N, Schmidt VF, Fischer M, Jörgens M, Khaled NB, Liebig T, Ricke J, Rueckel J, and Sabel BO, 'Non-radiology health care professionals significantly benefit from AI assistance in emergency-related chest radiography interpretation' 166(1) *Chest Journal*, 2024.
- Sætra H, 'AI in context and the sustainable development goals: Factoring in the unsustainability of the sociotechnical system' 13 *Sustainability*, 2021.

- Salhab W, Ameyed D, Jaafar F, Mcheick H, 'A systematic literature review on AI Safety: Identifying trends, challenges, and future directions' 12 *IEEE Access*, 2024.
- Sartori L and Theodorou A, 'A sociotechnical perspective for the future of AI: Narratives, inequalities, and human control' 24(4) *Ethics and Information Technology*, 2022.
- Shetti S, Kumar C, Sriwastava N, and Sharma I, 'Pharmacovigilance of herbal medicines: Current state and future directions' 7(25) *Pharmacognosy Magazine*, 2011.
- Smalla H, 'Genetic diversity: Effects on treatment response and disease susceptibility' 9(4) *Global Journal of Life Sciences and Biological Research*, 2023.
- Strang C and Mixer S, 'Discovery of the meanings, expressions, and practices related to malaria care among the Maasai' 27(4) *Journal of Transcultural Nursing*, 2016.
- Van Herten J and Bovenkerk B, 'The precautionary principle in zoonotic disease control' 14(2) *Public Health Ethics*, 2021.
- Varnosfaderani S and Forouzanfar M, 'The role of AI in hospitals and clinics: Transforming healthcare in the 21st century' 11(4) *Bioengineering*, 2024.
- Warsame M, Kimbute O, Machinda Z, Ruddy P, Melkisedick M, Peto T, Ribeiro I, Kitua A, Tomson G, Gomes M, 'Recognition, perceptions and treatment practices for severe malaria in rural Tanzania: Implications for accessing rectal artesunate as a pre-referral' 2(1) *PloS One*, 2007.
- Winch P, Makemba A, Kamazima S, Lurie M, Lwihula G, Premji Z, Minjas J, Shiff C, 'Local terminology for febrile illnesses in Bagamoyo district, Tanzania and its impact on the design of a community-based malaria control programme' 42(7) *Social Science and Medicine*, 1997.
- Wolfe C, Barry A, Campos A, Farham B, Achu D, Juma E, Kalu A, and Impouma B, 'Control, elimination, and eradication efforts for neglected tropical diseases in the World Health Organization African region over the last 30 years: A scoping review' 141 *International Journal of Infectious Diseases*, 2024, 1.
- Wu X, Ma L, Low D, Sharma S, and Papyshev G, 'Beyond precautionary principle: policy-making under uncertainty and complexity' 7(1) *Policy Design and Practice*, 2024.
- Zeb S, Nizamullah F, Abbasi N, and Fahad M, 'AI in healthcare: Revolutionizing diagnosis and therapy' 3(3) *International Journal of Multidisciplinary Sciences and Arts*, 2024.
- Zou J and Shiebinger L, 'Ensuring that biomedical AI benefits diverse populations' 67 *eBioMedicine*, 2021.

Online Journals

- Ah-Shahwani and Faieq A, 'The benefit of artificial intelligence in the analysis of malignant brain diseases: A mini review' *Mesopotamian Journal of Artificial Intelligence in Healthcare*, 2023, 57-60–< https://journals.mesopotamian.press/index.php/MJAIH/article/view/143/138>
- Almasoud A and Idowu J, 'Algorithmic fairness in predictive policing' *AI and Ethics*, 2024, 2–< https://link.springer.com/article/10.1007/s43681-024-00541-3>
- Liu Z, 'Cultural bias in large language models: A comprehensive analysis and mitigation strategies' *Journal of Transcultural Communication*, 2024, 5–<https://www.degruyter.com/document/doi/10.1515/jtc-2023-0019/html?lang=en&srsltid=AfmBOopRb9ckAFwhPzVkm7_pD36APVHZ7tDHv5aOF5-pWw20MVDbBe_w#Chicago>

- Ozioma E and Chinwe O, 'Herbal medicines in African traditional medicine' in Builders P (ed), *Herbal Medicine*, Intech Open, 30 January 2019–<https://www.intechopen.com/chapters/64851>

Working Papers, Discussion Papers and Research Papers

- Adams R, 'AI in Africa: Key concerns and policy considerations for the future of the continent' Africa Policy Research Institute, Policy Brief No. 8, 2022,–< https://afripoli.org/uploads/publications/AI_in_Africa.pdf>
- Chen B and Metcalf J, 'Explainer: A sociotechnical approach to AI policy' Data and Society, Policy Brief, 1-5–< https://datasociety.net/wp-content/uploads/2024/05/DS_Sociotechnical-Approach_to_AI_Policy.pdf>
- Hari A and Abdulla M, 'AI Safety: where do we stand presently?' Indian Institute of Management Kozhikode, Working Paper Number IIMK/WPS/584/ITS/2023/07, August 2023, 11–< https://iimk.ac.in/uploads/publications/IIMKWPS584ITS202307.pdf>
- Trist E, 'The evolution of socio-technical systems: A conceptual framework and an action research program' Ontario Quality of Working Life Centre, Occasional Paper Number 2, 1981, 7-9, < https://www.lmmiller.com/blog/wp-content/uploads/2013/06/The-Evolution-of-Socio-Technical-Systems-Trist.pdf>
- Weber V, Laumann E, and Riera M, 'Mapping the world's critical infrastructure sectors' German Council of Foreign Relations, Policy Brief Number 35, November 2023, 1-2–< https://dgap.org/system/files/article_pdfs/DGAP%20Policy%20Brief%20No.35%2C%20November%202023.pdf>

Conference Papers

- Harding J and Valenzuela C,' Workshop on sociotechnical AI safety' Stanford School of Humanities and Sciences, McCoy Family Centre for Ethics in Society, July 2024–< https://hai.stanford.edu/sites/default/files/2024-07/SociotechnicalAI_Report_v3.pdf> on 17 December 2024
- Rauh M, Marchal N, Manzini A, Hendricks LA, Comanescu R, Akbulut C, Stepleton T, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel I, Rieser V, Isaac W and Weidinger L, 'Gaps in the safety evaluation of generative AI' Seventh AAAI/ACM Conference on AI, Ethics, and Society, San Jose, 21-23 October 2024.

Self-Published Articles

- Alabi M and Holmes T, 'AI safety and ethics: Developing robust frameworks for ethical AI development and deployment' ResearchGate, 2024.
- Alia O, Abdelbakib W, Shrestha A, Elbasib E, Alryalatd M, and Dwivedie Y, 'Pangea: A fully open multilingual multimodal LLM for 39 languages' ArXiv, 2024.
- Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, and Mané D, 'Concrete problems in AI safety' ArXiv, 2016.
- Anderljung A, Barnhart J, Korinek A, Leung J, O'Keefe C, Whittlestone J, Avin S, Brundage M, Bullock J, Cass-Beggs D, Chang B, Collins T, Fist T, Hadfield G, Hayes A, Ho L, Hooker S, Horvitz E, Kolt N, Schuett J, Shavit Y, Siddarth D, Trager R, and Wolf K, 'Frontier AI regulation: Managing emerging risks to public safety' ArXiv, 2023.
- Bakirtzis G, Tubella A, Theodorou A, Danks D, and Topcu U, 'Navigating the sociotechnical labyrinth: Dynamic certification for responsible embodied AI' ArXiv, 2024.

- Birch J, ‘Preparing for the next pandemic: A case for precautionary thinking and citizens’ assemblies’ PhilArchive, 2024.
- Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen G C, Steinhardt J, Flynn C, Ó hÉigeartaigh S, Beard S, Belfield H, Farquhar S, Lyle C, Crootof R, Evans O, Page M, Bryson J, Yampolskiy R, and Amodei D, ‘The malicious use of artificial intelligence: Forecasting, prevention, and mitigation’ ArXiv, 2018.
- Bucknall B, Trager R, and Osborne M, ‘Position: Ensuring mutual privacy is necessary for effective external evaluation of proprietary AI systems’ ArXiv, 2025.
- Chen C, Liu Z, Jiang W, Goh S, Lam K, ‘Trustworthy, responsible, and safe AI: A comprehensive architectural framework for ai safety with challenges and mitigations’ ArXiv, 2024.
- Chinta S, Wang Z, Zhang X, Viet T, Kashif A, Smith M, and Zhang W, ‘AI-driven healthcare: A survey on ensuring fairness and mitigating bias’ ArXiv, 2024.
- Christiano P, Schlegeris P, and Amodei D, ‘Supervising strong learners by amplifying weak experts’ ArXiv, 2018.
- Curtis S , Iyer R, Kirk-Giannini C, Krakovna V, Krueger D, Lambert N, Marnette B, McKenzie C, Michael J, Miyazono E, Mima N, Ovadya A, Thorburn L, and Turan D, ‘Research agenda for sociotechnical approaches to AI safety’ Social Science Research Network, 2024.
- Diallo K , Smith J, Okolo C, Nyamwaya D , Kgomo J, and Ngamita R, ‘Case Studies of AI Policy Development in Africa’ Arxiv, 2024.
- Everitt T, Lea G, and Hutter M ‘AGI safety literature review’ ArXiv, 2018.
- Hendryks D, ‘Introduction to AI safety, ethics, and society’ ArXiv, 2024.
- Ho L, Barnhart J, Trager R, Bengio Y, Brundage M, Carnegie A, Chowdhury R, Dafoe A, Hadfield A, Levi M, and Snidal D, ‘International institutions for advanced AI’ ArXiv, 2023.
- Kilian K, ‘Beyond accidents and misuse: Decoding the structural risk dynamics of artificial intelligence’ ArXiv, 2024.
- Kylian K, ‘Beyond accidents and misuse: Decoding the structural risk dynamics of artificial Intelligence’ ArXiv, 2024.
- Mhasakar M, Baker-Ramos R, Carter B, Halekahi-Kaiwi E, Hester J, ‘“I would never trust anything western”: Kumu (educator) perspectives on use of LLMs for culturally revitalizing CS education in Hawaiian Schools’ ArXiv, 2025.
- Mukobi G, ‘Reasons to doubt the impact of AI risk evaluations’ ArXiv, 2024.
- Prabhune S and Berndt D, ‘Deploying large language models with retrieval augmented generation’ ArXiv, 2024.
- Raji D and Dobbe R, ‘Concrete problems in AI safety revisited’ ArXiv, 2023.
- Rismani S, Dobbe R, and Moon A, ‘From silos to systems: Process-oriented hazard analysis for AI systems’ ArXiv, 2024.
- Ritchie S, Cheng Y, Chen M, Matthews R, Esch D, Li B, and Sim K, ‘Large vocabulary speech recognition for languages of Africa: multilingual modelling and self-supervised learning’ ArXiv, 2022.
- Saberi M, ‘The human factor in AI safety’ ArXiv, 2022.
- Shishehbor F and Awan Z, ‘Enhancing cardiovascular disease risk prediction with machine learning models’ ArXiv, 2024.
- Trager R, Harack B, Reuel A, Carnegie A, Heim L, Ho L, Kreps S, Lall R, Larter O, Ó hÉigeartaigh S, Staffell S, and Villalobos J, ‘International governance of civilian AI: A jurisdictional certification approach’ ArXiv, 2023.

- Uuk R, Brouwer A, Schreier T, Dreksler N, Pulignano V, and Bommasani R, 'Effective mitigations for systemic risks from general-purpose AI' ArXiv, 2024.
- Weidinger L, Barnhart J, Brennan J, Butterfield C, Young S, Hawkins W, Hendricks L, Comanescu R, Chang O, Rodriguez M, Beroshi J, Bloxwich D, Proleev L, Chen J, Farquhar S, Ho L, Gabriel I, Dafoe A, and Isaac W, 'Holistic safety and responsibility evaluations of advanced AI models' ArXiv, 2024.
- Weidinger L, Rauh M, Marchal N, Manzini A, Hendricks LA, Mateos-Garcia J, Bergman S, Kay J, Griffin C, Bariach B, Gabriel J, Rieser V, and Isaac W, 'Sociotechnical safety evaluation of generative AI systems' ArXiv, 2023.

Reports

- Ada Lovelace Institute, *Under the radar? Examining the evaluation of foundation models*, 2024.
- African Union, *Continental AI Strategy*, 2024.
- Allen G and Adamson G, *The AI safety institute international network: Next steps and recommendations*, 2024.
- Bengio Y, *International scientific report on the safety of advanced AI*, 2024
- Bengio Y, *The international scientific report on the safety of advanced AI*, 2025.
- Berkley Haas Center for Equity, *Mitigating bias in artificial intelligence an equity: Fluent leadership playbook*, 2020.
- Center for AI Safety, *Impact Report*, 2023.
- Center for Security and Emerging Technology, *Securing critical infrastructure in the age of AI*, 2024.
- Department of Communications and Digital Technologies, *South Africa National Artificial Intelligence Policy Framework*, 2024.
- De Nederlandsche Bank, *The impact of AI on the financial sector and supervision*, 2024.
- European Agency for Fundamental Rights, *Getting the future right artificial intelligence and fundamental rights*, 2020.
- Financial Stability Institute, *Regulating AI in the financial sector: recent developments and main challenges*, 2024.
- Institute for AI Policy and Strategy, *Understanding the first wave of AI safety institutes: Characteristics, functions, and challenges*, 2024.
- Institute for Trustworthy AI in Law and Society, *The importance of a socio-technical approach in AI development*, 2024.
- Ministry of Finance and Economic Development, *Mauritius Artificial Intelligence Strategy*, 2018.
- Ministry of Information, Communications and the Digital Economy, *Kenya National Artificial Intelligence (AI) Strategy 2025 – 2030*, 2025.
- NIST, *Computer security incident handling guide: Recommendations of the National Institute of Standards and Technology*, 2012.
- OECD, *FinTech lending in Sub-Saharan Africa lessons from African economies*, 2024.
- OECD, *Marketplace and FinTech lending for SMEs in the COVID-19 crisis*, 2022.
- Panel for the Future of Science and Technology, *Artificial intelligence in healthcare: Applications, risks, and ethical and societal impacts*, 2022.
- The Alan Turing Institute and the Centre for Emerging Technology and Security, *Evaluating malicious generative AI capabilities understanding inflection points in risk*, 2024.

- The National Council for Artificial Intelligence, *Egypt National Artificial Intelligence Strategy*, 2023.
- US Department of Homeland Security, Roles and responsibilities framework for artificial intelligence in critical infrastructure, 2024.
- US Department of Justice, *Artificial intelligence and criminal justice*, 2024.
- USAID Centre for Innovation and Impact, *Artificial intelligence in global health: Defining a collective path forward*, 2019.
- World Health Organisation, *Benefits and risks of using artificial intelligence for pharmaceutical development and delivery*, 2024.

Other Internet Resources

- African Union Development Agency, 'AI and the future of work in Africa: How AI is redefining opportunities' 3 July 2024–<https://www.nepad.org/blog/ai-and-future-of-work-africa-how-ai-redefining-opportunities>
- Ahteensuu M, 'Rationale for taking precautions: Normative choices and commitments in the implementation of the precautionary principle', 2-3–<https://www.kent.ac.uk/scarr/events/ahteensuu.pdf>
- AI Safety Institute, 'AI Safety Institute approach to evaluations' AISI, 9 February 2024–< https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations#aisis-approach-to-evaluations>
- Albergotti R, 'The risks of expanding the definition of 'AI safety'' Semafor, 8 March 2024–< https://www.semafor.com/article/03/08/2024/the-risks-of-expanding-the-definition-of-ai-safety>
- BBC, 'The Kenyan enthralled by the healing power of plants' 29 July 2024–< https://www.bbc.com/news/articles/c84j0kzgnk9o>
- Biron C, 'AI's 'insane' translation mistakes endanger US asylum cases' Context, 18 September 2023–< https://www.context.news/ai/ais-insane-translation-mistakes-endanger-us-asylum-cases>
- Bogen M and Wincecoff A, 'Applying sociotechnical approaches to AI governance in practice' Center for Democracy and Technology, May 2024, 2–< https://cdt.org/wp-content/uploads/2024/05/2024-05-23-AI-Gov-Lab-applying-sociotechnical-approaches-to-ai-governance-in-practice-final.pdf>
- Bureau of Cyberspace and Digital Policy, 'Risk management profile for artificial intelligence and human rights' US Department of State, 25 July 2024–< https://www.state.gov/risk-management-profile-for-ai-and-human-rights/>
- Burkeman O, 'The power broker: Robert Moses and the fall of New York by Robert Caro review – a landmark study' The Guardian, 23 October 2015–< https://www.theguardian.com/books/2015/oct/23/the-power-broker-robert-moses-and-the-fall-of-new-york-robert-caro-review>
- Campus France, 'The French government creates a national institute to assess and secure AIs'–< https://www.campusfrance.org/en/actu/creation-d-un-institut-national-pour-l-evaluation-et-la-securite-de-l-ia>
- Centre for Disease Control, 'The Untreated Syphilis Study at Tuskegee Timeline' 4 September 2024–< https://www.cdc.gov/tuskegee/about/timeline.html>
- Chan A, 'Evaluating predictions of model behaviour' Centre for the Governance of AI, 9 April 2024–< https://www.governance.ai/post/evaluating-predictions-of-model-behaviour>

- CPSC, 'Samsung recalls galaxy note7 smartphones due to serious fire and burn hazards' 15 September 2016–< https://www.cpsc.gov/Recalls/2016/Samsung-Recalls-Galaxy-Note7-Smartphones>
- Cybersecurity and Infrastructure Security Agency, 'Critical infrastructure sectors' US Department of Homeland Security–<https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors>
- Dafoe A and Russell S, 'Yes, we are worried about the existential risk of artificial intelligence' MIT Technology Review, 2 November 2016–< https://www.technologyreview.com/2016/11/02/156285/yes-we-are-worried-about-the-existential-risk-of-artificial-intelligence/>
- Dascalescu D, 'How to avoid blind spots and use Deep Research effectively' Medium, 10 February 2025–< https://dandv.medium.com/how-to-avoid-blind-spots-and-use-deep-research-effectively-659b3afe675c >
- DeAngelis A, 'Artificial Intelligence and Critical Infrastructure' Enterra Solutions, 4 June 2024–< https://enterrasolutions.com/artificial-intelligence-and-critical-infrastructure/>
- Department of Commerce, 'At the direction of President Biden, Department of Commerce to establish U.S. Artificial Intelligence Safety Institute to lead efforts on AI safety' 1 November 2023–< https://www.commerce.gov/news/press-releases/2023/11/direction-president-biden-department-commerce-establish-us-artificial>
- Digital Regulation Platform, 'Transformative technologies (AI) challenges and principles of regulation' The World Bank, 8 May 2024, <https://digitalregulation.org/3004297-2/ >
- DSIT, 'Introducing the AI Safety Institute' 17 January 2024–< https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute#establishment>
- EA Forum, 'AI evaluations and standards' EA Forum–< https://forum.effectivealtruism.org/topics/ai-evaluations-and-standards>
- Ekeanyanwu O, 'AI sifts Africa's natural remedies for drug discovery' Vaccines Work, 2 September 2024–<https://www.gavi.org/vaccineswork/ai-sifts-africas-natural-remedies-drug-discovery>
- Etzioni O, 'No, the experts don't think superintelligent AI is a threat to humanity' MIT Technology Review, 20 September 2016–< https://www.technologyreview.com/2016/09/20/70131/no-the-experts-dont-think-superintelligent-ai-is-a-threat-to-humanity/>
- EU AI Act, 'High-level summary of the AI Act' 27 Feb, 2024–< https://artificialintelligenceact.eu/high-level-summary/>
- European Medicines Agency, 'Pharmacovigilance: Overview' 6 June 2024–< https://www.ema.europa.eu/en/human-regulatory-overview/pharmacovigilance-overview>
- European Union, 'Article 57: AI regulatory sandboxes'–< https://artificialintelligenceact.eu/article/57/>
- FAA, 'What we do' FAA, 27 June 2016–< https://www.faa.gov/about/mission/activities>
- FDA, 'Insulin pump recall: Medtronic notifies users of MiniMed 600 and 700 series pumps of risk of shorter than expected battery life' 17 October 2024–< https://www.fda.gov/medical-devices/medical-device-recalls/insulin-pump-recall-medtronic-notifies-users-minimed-600-and-700-series-pumps-risk-shorter-expected>

- FDA, 'The FDA's drug review process: Ensuring drugs are safe and effective' FDA, 24 November 2017–< https://www.fda.gov/drugs/information-consumers-and-patients-drugs/fdas-drug-review-process-ensuring-drugs-are-safe-and-effective>
- FDA, 'What is a medical device recall? 26 September 2018–< https://www.fda.gov/medical-devices/medical-device-recalls/what-medical-device-recall>
- Forbes Technology Council, 'AI & fairness metrics: Understanding & eliminating bias' Forbes, 14 July 2023–< https://councils.forbes.com/blog/ai-and-fairness-metrics>
- Gates D, Baker M, 'The inside story of MCAS: How Boeing's 737 MAX system gained power and lost safeguards' Association of Flight Attendants, 22 June 2019–< https://www.afacwa.org/the_inside_story_of_mcas_seattle_times>
- Government of Canada, 'Algorithmic Impact Assessment tool'–< https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
- Government of Canada, 'Canada launches Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html>
- Government of Canada, 'Canadian Artificial Intelligence Safety Institute' 12 November 2024–< https://ised-isde.canada.ca/site/ised/en/canadian-artificial-intelligence-safety-institute>
- Government of Canada, 'The Artificial Intelligence and Data Act (AIDA) – Companion document'–< https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>
- Harris R, 'The Healthcare Crisis in Sub-Saharan Africa' Africa Mission Healthcare–< https://africanmissionhealthcare.org/the-healthcare-crisis-in-sub-saharan-africa/#:~:text=Insufficient%20Healthcare%20Infrastructure,often%20poorly%20equipped%20and%20understaffed. >
- Hubinger E, 'Towards understanding safety-based evaluations' AI Alignment Forum, 15 March 2023–< https://www.alignmentforum.org/posts/uqAdqrvxqGqeBHjTP/towards-understanding-based-safety-evaluations>
- IBM, 'What are large language models (LLMs)?' IBM, 2 November 2023–< https://www.ibm.com/think/topics/large-language-models>
- IBM, 15 November 2024–< https://www.ibm.com/think/topics/ai-safety#:~:text=Cybersecurity-,Bias%20and%20fairness,lead%20to%20unfair%20decision%2Dmaking.>
- information space' May 2024 <https://rsf.org/sites/default/files/medias/file/2024/06/AI%20and%20the%20Right%20to%20Information%20-%20RSF%20Recommendations%20to%20EU.pdf>
- Japan AI Safety Institute, 'Overview of the AI Safety Institute' 1 November 2024, 4 https://aisi.go.jp/wp-content/uploads/2024/07/20240627_AboutAISI_en.pdf>
- Jungco K, 'Types of AI models: A deep dive into AI architecture' eWeek, 21 November 2024–< https://www.eweek.com/artificial-intelligence/ai-model-types/>
- Kibuacha F, 'Closing the gap: How local context improves AI performance in emerging regions' GeoPoll, 25 September 2024–< https://www.geopoll.com/blog/closing-the-ai-gap-local-context/>
- Kocher B, 'Using AI to your benefit in health care' NEJM Catalyst, 12 November 2024–< https://catalyst.nejm.org/doi/full/10.1056/CAT.24.0431>

- Lee N, Simmons N, and Crawford M, 'Health and AI: Advancing responsible and ethical AI for all communities' Brookings, 3 March 2025–< https://www.brookings.edu/articles/health-and-ai-advancing-responsible-and-ethical-ai-for-all-communities/>
- Lemos R, 'Infrastructure Cyberattacks, AI-Powered Threats Pummel Africa' 1 March 2024–<https://www.darkreading.com/vulnerabilities-threats/ai-powered-threats-cyberattacks-on-infrastructure-pummel-africa> on 4 November 2024.
- Leslie D, 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector' The Alan Turing Institute, 2019, 4–<https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf>
- Leslie D, 'Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector' The Alan Turing Institute, 2019, 4–<https://www.turing.ac.uk/sites/default/files/2019-08/understanding_artificial_intelligence_ethics_and_safety.pdf>
- Mckernon E, Cheng D, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–<https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review>
- McKernon E, Cheng D, and Convergence Analysis, 'AI safety evaluations: A regulatory review' EA Forum, 19 March 2024–< https://forum.effectivealtruism.org/posts/i66PmFYKs9M4fuh6G/ai-safety-evaluations-a-regulatory-review>
- Microsoft, 'AI in Africa Meeting the Opportunity' 2024, 20–< https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2024/01/AI-in-Africa-Meeting-the-Opportunity.pdf>
- Microsoft, 'What are large language models (LLMs)?' Microsoft–< https://azure.microsoft.com/en-us/resources/cloud-computing-dictionary/what-are-large-language-models-llms>
- Ministry of Economy, Trade and Industry, 'Launch of AI Safety Institute' 14 February 2024–< https://www.meti.go.jp/english/press/2024/0214_001.html>
- Modulos, 'A curated global guide to AI compliance: Navigating international AI regulations'–< https://www.modulos.ai/global-ai-compliance-guide/>
- Morris, Manning & Martin LLP, 'State AI laws continue to emerge' 4 March 2025–< https://www.mmmlaw.com/news-resources/102k2k3-state-ai-laws-continue-to-emerge/>
- Nadine N, 'Power cuts in South Africa wreak havoc on health care' Think Global Health, 7 June 2022–< https://www.thinkglobalhealth.org/article/power-cuts-south-africa-wreak-havoc-health-care>
- National Artificial Intelligence Advisory Committee, 'Findings and recommendations: AI safety' May 2024, 2-7–< https://ai.gov/wp-content/uploads/2024/06/FINDINGS-RECOMMENDATIONS_AI-Safety.pdf>
- National Collaborating Centre for Healthy Public Policy, 'Public policies guided by the precautionary principle' May 2009, 5–< https://www.ncchpp.ca/docs/VBeloinAnC_MEP.pdf>
- National Human Genome Research Institute, 'Eugenics and scientific racism' NHGRI, 18 May 2022–<https://www.genome.gov/about-genomics/fact-sheets/Eugenics-and-Scientific-Racism>
- National Institute of Standards and Technology, 'Pre-deployment evaluation of Anthropic's upgraded Claude 3.5 Sonnet' 19 November 2024–<

https://www.nist.gov/news-events/news/2024/11/pre-deployment-evaluation-anthropics-upgraded-claude-35-sonnet>
- National Institute of Standards and Technology, 'Pre-deployment evaluation of OpenAI's o1 model' NIST, 18 December 2024–< https://www.nist.gov/news-events/news/2024/12/pre-deployment-evaluation-openais-o1-model>
- National Research Council of Science and Technology, 'AI Safety Institute launched as Korea's AI research hub' 27 November 2024–< https://www.newswise.com/articles/ai-safety-institute-launched-as-korea-s-ai-research-hub>
- National Telecommunications and Information Administration, 'Independent Evaluations' NTIA, 27 March 2024–< https://www.ntia.gov/issues/artificial-intelligence/ai-accountability-policy-report/developing-accountability-inputs-a-deeper-dive/ai-system-evaluations/independent-evaluations>
- New Zealand Medicines and Medical Devices Safety Authority, 'Adverse reactions to medicines and vaccines' 21 November 2023–< https://www.medsafe.govt.nz/consumers/safety-of-medicines/Medicine-safety.asp#:~:text=Adverse%20reactions%20are%20harmful%20effects,sheet%20or%20consumer%20medicine%20information>
- NHTSA, 'About NHTSA' NHTSA–< https://www.nhtsa.gov/about-nhtsa#:~:text=The%20National%20Highway%20Traffic%20Safety%20Administration%20is%20responsible,injuries%20and%20economic%20losses%20from%20motor%20vehicle%20crashes.>
- Nicholas G and Bhatia A, 'Lost in translation: large language models in non-English content analysis' Center for Democracy and Technology, May 2023, 6–< https://cdt.org/wp-content/uploads/2023/05/non-en-content-analysis-primer-051223-1203.pdf>
- NIST, 'Effects of race, age, sex on face recognition software: Demographics study on face recognition algorithms could help improve future tools' NIST, 19 December 2019–< https://www.nist.gov/news-events/news/2019/12/nist-study-evaluates-effects-race-age-sex-face-recognition-software>
- NIST, 'Executive order on safe, secure, and trustworthy artificial intelligence'–< https://www.nist.gov/artificial-intelligence/executive-order-safe-secure-and-trustworthy-artificial-intelligence>
- Nuffield Council on Bioethics, ' Artificial intelligence (AI) in healthcare and research' Bioethics Briefing Note, May 2018, 2–< https://www.nuffieldbioethics.org/wp-content/uploads/Artificial-Intelligence-AI-in-healthcare-and-research-1.pdf>
- Oduro S, 'Without accounting for bias and discrimination, AI safety standards will fall short' Tech Policy Review, 16 September 2024–< https://www.techpolicy.press/without-accounting-for-bias-and-discrimination-ai-safety-standards-will-fall-short/>
- OECD, 'AI in health: Huge potential, huge risks' 2024, 6–< https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/01/ai-in-health-huge-potential-huge-risks_ff823a24/2f709270-en.pdf>
- Ortega P, Maini V, DeepMind Safety Team, 'Building safe artificial intelligence: specification, robustness, and assurance' Medium, 27 September 2018–< https://deepmindsafetyresearch.medium.com/building-safe-artificial-intelligence-52f5f75058f1>
- Piper K, 'The case for taking AI seriously as a threat to humanity: Why some people fear AI, explained' Future Perfect, 15 October 2020–< https://www.vox.com/future-

perfect/2018/12/21/18126576/ai-artificial-intelligence-machine-learning-safety-alignment>
- Price W, 'Risks and remedies for artificial intelligence in health care' Brookings, 14 November 2019–< https://www.brookings.edu/articles/risks-and-remedies-for-artificial-intelligence-in-health-care/>
- Provisio, 'Official languages in Africa – An analysis' 19 July 2023–< https://provisioservices.com/languages-in-africa/>
- Public Health Ontario, 'Focus On: Precautionary Principle—Applications Relevant to Public Health Emergency Preparedness' October 2022, 3–<https://www.publichealthontario.ca/-/media/Documents/P/2021/precautionary-principle-emergency-preparedness.pdf?sc_lang=en>
- Regulatory Policy Committee, 'RPC guidance note on 'using the precautionary principle'' UK Interdepartmental Liaison Group on Risk Assessment, 1–< https://assets.publishing.service.gov.uk/media/5e21c408e5274a6c3be72203/short_guidance_note_-_precautionary_principle.pdf>
- Reich A and Holmes J, 'Virginia passes second of its kind AI anti-discrimination statute, but law vetoed by governor' Saul Ewing, 1 April 2025–< https://www.saul.com/insights/blog/virginia-passes-second-its-kind-ai-anti-discrimination-statute-law-vetoed-governor>
- Rens A, Calandro E, and Gaffley M, 'As global cyber conflict breaks out, AI technologies bring new risk to Africa' 23 March 2022–<https://researchictafrica.net/2022/03/23/as-global-cyber-conflict-breaks-out-ai-technologies-bring-new-risk-to-africa/>
- Robinson M and Zhang X, 'The world medicines situation 2011– Traditional medicines: Global situation, issues and challenges' World Health Organisation, 2011, 8–< https://www.utep.edu/herbal-safety/herbal-facts/files/wms_ch18_wtraditionalmed-2011.pdf>
- Rock D and Grant H, 'Why diverse teams are smarter' Harvard Business Review, 4 November 2016–< https://hbr.org/2016/11/why-diverse-teams-are-smarter>
- RSF, 'AI and the Right to Information RSF's 7 recommendations to protect an AI-dominated
- Salvi P, 'AI Misdiagnosis' Salvi, Schostok and Pritchard, 10 April 2024–< https://www.salvilaw.com/blog/ai-misdiagnosis/>
- Science Direct, 'Healthcare infrastructure'–< https://www.sciencedirect.com/topics/social-sciences/health-infrastructure>
- Shkurdoda A and Dobosz M, 'AI in fintech: Harnessing intelligent technologies for smarter finance' Neontri, 29 November 2024–< https://neontri.com/blog/artificial-intelligence-fintech/>
- Singapore AISI, 'About us'–< https://sgaisi.sg/about-us>
- Sofge E, 'Bill Gates Fears AI, but AI researchers know better' Popular Science, 15 January 2015–< https://www.popsci.com/bill-gates-fears-ai-ai-researchers-know-better/>
- Sutherland E, 'Artificial intelligence in health: big opportunities, big risks' OECD AI Policy Observatory, 1 August 2023–< https://oecd.ai/en/wonk/artificial-intelligence-in-health-big-opportunities-big-risks>
- Taha A, 'AI risks in global health 'must be accounted for' – WHO' SciDev, 23 January 2024–< https://www.scidev.net/global/news/ai-risks-in-global-health-must-be-accounted-for-who/>

- Tech Target, 'Arguing the pros and cons of artificial intelligence in healthcare' 26 December 2023–< https://www.techtarget.com/healthtechanalytics/feature/Arguing-the-Pros-and-Cons-of-Artificial-Intelligence-in-Healthcare>
- The African Language Program at Harvard, 'Introduction to African Languages' Harvard University–< https://alp.fas.harvard.edu/introduction-african-languages>
- The US House Committee on Transport and Infrastructure, 'After 18-Month Investigation, Chairs DeFazio and Larsen Release Final Committee Report on Boeing 737 MAX' 16 September 2020–< https://democrats-transportation.house.gov/news/press-releases/after-18-month-investigation-chairs-defazio-and-larsen-release-final-committee-report-on-boeing-737-max>
- The White House, 'Removing barriers to American leadership in artificial intelligence' 23 January 2025–< https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/>
- Tigera, Understanding AI safety: Principles, frameworks, and best practices'–< https://www.tigera.io/learn/guides/llm-security/ai-safety/>
- Tighe P, Gale B, and Mossburg S, 'Artificial intelligence and patient safety: Promise and challenges' Patient Safety Network, 27 March 2024–< https://psnet.ahrq.gov/perspective/artificial-intelligence-and-patient-safety-promise-and-challenges>
- Trevino S, 'Managing risks and opportunities that emerging AI technologies present for healthcare cybersecurity' HIMSS, 13 November 2024–< https://gkc.himss.org/resources/managing-risks-and-opportunities-emerging-ai-technologies-present-healthcare>
- Tuskegee University, 'About the USPHS syphilis study'–< https://www.tuskegee.edu/about-us/centers-of-excellence/bioethics-center/about-the-usphs-syphilis-study>
- Uffelmann E, Huang Q, Munung N, de Vries J, Okada Y, Martin A, Martin H, Lappalainen T, and Posthuma D, 'Genome-wide association studies' Nature Reviews, 26 August 2021–< https://www.nature.com/articles/s43586-021-00056-9>
- UK AI Safety Institute, 'Advanced AI evaluations at AISI: May Update' 20 May 2024–<https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update>
- US AI Safety Institute, 'Strategic vision' National Institute of Standards and Technology, 1 October 2024–<https://www.nist.gov/aisi/strategic-vision>
- US Department of Health and Human Services, 'Guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule' US HHS, 3 February 2025–< https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html>
- US Department of Homeland Security, 'DHS Publishes Guidelines and Report to Secure Critical Infrastructure and Weapons of Mass Destruction from AI-Related Threats' DHS, 29 April 2024–<https://www.dhs.gov/news/2024/04/29/dhs-publishes-guidelines-and-report-secure-critical-infrastructure-and-weapons-mass>
- US Department of Homeland Security, 'Mitigating artificial intelligence (AI) risk: Safety and security guidelines for critical infrastructure owners and operators' April 2024, 5–< https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf>
- US Privacy and Civil Liberties Board, 'CDT's Miranda Bogen: Opening Statement' 11 July 2024–< https://documents.pclob.gov/prod/Documents/EventsAndPress/ff4e3c3a-

00c5-4c3d-b501-73ad8780a4ab/Miranda%20Bogen%20Opening%20Statement%20-%20Completed%20508%20-%20Jul%2019,%202024.pdf>

- VA Center for Medication Safety and VHA Pharmacy Benefits Management Strategic Healthcare Group and the Medical Advisory Panel, 'Adverse drug events, adverse drug reactions and medication errors: Frequently asked questions' November 2006, 1–< https://www.pbm.va.gov/vacenterformedicationsafety/tools/AdverseDrugReaction.pdf >
- Vicente M, 'Demographic history and adaptation in African populations' Uppsala: Acta Universitatis Upsaliensis, 2020, 3–< https://www.diva-portal.org/smash/get/diva2:1412382/FULLTEXT01.pdf>
- White & Case, 'AI watch: Global regulatory tracker– United States' 31 March 2025–< https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-states>
- WHO, 'African Traditional Medicine Day 2022' 31 August 2022–< https://www.afro.who.int/regional-director/speeches-messages/african-traditional-medicine-day-2022#:~:text=Traditional%20medicine%20has%20been%20the,for%20their%20basic%20health%20needs>
- WHO, 'Electricity in healthcare facilities' WHO, 31 August 2023–< https://www.who.int/news-room/fact-sheets/detail/electricity-in-health-care-facilities>
- WHO, 'Low quality healthcare is increasing the burden of illness and health costs globally' WHO, 5 July 2018–< https://www.who.int/news/item/05-07-2018-low-quality-healthcare-is-increasing-The-burden-of-illness-and-health-costs-globally>
- World Economic Forum, 'Generative AI is trained on just a few of the world's 7,000 languages. Here's why that's a problem – and what's being done about it' 17 May 2024–< https://www.weforum.org/stories/2024/05/generative-ai-languages-llm/>
- World Health Organisation, 'Pharmacovigilance for traditional medicine products: Why and how?' WHO: South East Asia, 2017–< https://iris.who.int/bitstream/handle/10665/259854/Pharmacovigilane.pdf?sequence=1 >