Operationalising AI safety: A lifecycle approach

February 2026

nlu DELHI

CENTRE FOR COMMUNICATION GOVERNANCE

KONRAD ADENAUER STIFTUNG

Published by: Centre for Communication Governance at National Law University Delhi, Sector 14, Dwarka, New Delhi – 110078

Patrons: Professor (Dr.) G.S. Bajpai, Vice Chancellor, NLUD, and Professor (Dr.) Ruhi Paul, Registrar, NLUD

Faculty Director, CCG: Dr. Daniel Mathew

Executive Director, CCG: Jhalak M. Kakkar

Supported by: Konrad-Adenauer-Stiftung

Authors: Angelina Dash, Rishiti Choudaha, Vidya Subramanian, Tavishi

Editing and Supervision: Tavishi**,** Joanne D'Cunha

Review: Joanne D'Cunha, Jhalak M. Kakkar

Conceptualisation: Tavishi, Joanne D'Cunha

Design by: Shivani Mago

Acknowledgements: This report was made possible by support provided by the Konrad-Adenauer-Stiftung. We are continuously grateful for the generous support we received from the National Law University Delhi (NLUD). The Centre for Communication Governance (CCG) would therefore like to thank our patrons, the Vice Chancellor, Professor (Dr.) G.S. Bajpai, and the Registrar, Professor (Dr.) Ruhi Paul of NLUD, for their constant guidance. CCG would also like to thank our Faculty Director, Dr. Daniel Mathew, for his continuous direction and mentorship. We would also like to thank Jhalak M. Kakkar, Executive Director at CCG, for her unwavering support, encouragement, and mentorship, which have been instrumental in publishing this research. We would also like to thank all members of CCG for the many ways in which they supported the report. Special thanks to the ever-present and ever-patient CCG Administration Team for their unwavering support in all the work we do at CCG.

Suggested Citation: Angelina Dash, Rishiti Choudaha, Vidya Subramanian, Tavishi 'Operationalising AI Safety: A Lifecycle Approach' (Centre for Communication Governance, National Law University Delhi 2026)

Operationalising AI Safety: A Lifecycle Approach

NLU DELHI

CENTRE FOR COMMUNICATION GOVERNANCE

॥ न्यायस्तत्र प्रमाणं स्यात् ॥

NLU

DELHI

NATIONAL LAW UNIVERSITY DELHI

Prof. (Dr.) G.S. Bajpai
Vice Chancellor

Foreword

The unprecedented proliferation of AI systems has led to their integration across several high-stakes functional applications. While the merits of AI are extensive, ranging from scientific breakthroughs to enhanced productivity, they are often accompanied by unique challenges that require rigorous oversight and preparedness. It therefore becomes essential to strike a fine balance between acknowledging the extraordinary potential of artificial intelligence and the responsibility of managing the impending risks. From a Global South perspective, where the impact of AI is deeply intertwined with infrastructural challenges and unique cultural contexts, establishing a robust framework for responsible AI deployment is an imperative for identifying risks that lie beyond the scope of technical metrics.

AI safety encompasses robust standards and evaluative methodologies to align AI development with the collective interests of society, thereby minimizing the probability of unanticipated detrimental outcomes. The field of AI safety can play a major role in shaping the trajectory of responsible AI adoption in the Global South, which is characterized by linguistic diversity, resource constraints, and an increasing need to focus on social empowerment.

AI safety is no longer a field solely falling within the realm of theoretical computer science, having far-reaching implications facilitated by the large-scale adoption of AI across different domains. With AI safety having established itself as a prominent field, this new report - *'Operationalising AI Safety: A lifecycle Approach'* from the Centre for Communication Governance at the National Law University Delhi with support from Konrad-Adenauer-Stiftung presents a meticulous and systematic evaluation of the evolving landscape of this field. By attempting to offer insights to a diverse range of stakeholders, this report can aid in creating awareness and contributing meaningfully in shaping the global dialogue within international governance forums. Part I of the report explores the various conceptions of AI safety followed by delving into the evolution of the concept. By examining the contemporary debates and challenges with interpretations within the field of AI safety, this report advocates embracing a socio-technical approach to it associated with the contextual nuances of the Global South.

Part II of this report explores how AI safety is operationalised through a lifecycle approach. By mapping the AI lifecycle, this report identifies specific risks and harms inherent to each developmental stage. The report also examines key corresponding governance mechanisms for each stage, while providing insights on how these mechanisms are reflected in legal frameworks. The report concludes with elaborating on key considerations for the Global South with respect to the implementation of the governance mechanisms by highlighting the contextualised challenges, local realities, and experiences from the Global South.

This timely report is intended to serve as a reference for a wide spectrum of actors across the AI ecosystem including academicians, technologists, civil society groups and policymakers. As Vice Chancellor of the National Law University Delhi, I commend the Centre for Communication Governance's vision in synthesizing scholarship that engages closely with emerging global issues. This report showcases the Centre's commitment to policy-relevant research that integrates scholarly research, governance and public advocacy.

॥ न्यायस्तत्र प्रमाणं स्यात् ॥

NLU

DELHI

NATIONAL LAW UNIVERSITY DELHI

Prof. (Dr.) G.S. Bajpai
Vice Chancellor

With India set to preside over the AI Impact Summit this year, we have a pivotal window to prioritize the operationalization of AI safety while ensuring that Global South equities are at the heart of the discourse. The Summit also offers a critical opportunity for India to help shape the trajectory for responsible development and deployment of AI for the Global South. It is imperative for the Global South priorities and interests to play a major part in informing and tuning the global AI safety policy discourse.

(G S Bajpai)

Prof. (Dr.) G. S Bajpai

Vice Chancellor, National Law University Delhi

Note to Reader

This report is part of the Konrad-Adenauer-Stiftung (KAS) Rule of Law Programme Asia's ongoing work on the governance, safety, and social impact of artificial intelligence.

It is released at a particularly timely moment, as it is launched alongside the India AI Impact Summit, when global debates on AI are marked by heightened momentum, visibility, and urgency. The AI Impact Summit has brought together policymakers, industry leaders, civil society, researchers, and technical experts from across regions, generating strong traction and dynamic exchanges on how AI systems should be developed, governed, and deployed. Launching this report in parallel with the Summit allows its analysis to resonate with ongoing discussions, benefit from heightened interest and attention, and contribute meaningfully to shaping emerging narratives and policy priorities around AI safety at both regional and global levels.

AI safety is closely linked to AI security and AI ethics. While AI safety refers to the practice of ensuring that AI systems operate reliably, predictably, and without causing unintended harm, AI security deals with protecting systems from misuse, attacks, and malicious actors. AI ethics, in turn, addresses broader questions of fairness, accountability, transparency, and respect for human rights. In practice, these areas often overlap and cannot be treated separately.

The goal of this report is to move beyond purely technical understandings of AI safety. It adopts a holistic, lifecycle-based and socio-technical approach, analysing how risks and harms can arise at every stage of the AI lifecycle, from early planning and data collection to deployment and post-deployment monitoring. For instance, during the data collection stage, AI developers are encouraged to keep detailed records of where data comes from, how it was collected, and whether consent was given. They should also audit the data for fairness and accuracy. This ensures that models are trained on trustworthy and representative data, reducing the risk of bias or unsafe outcomes later. This example illustrates how AI safety is a step-by-step process that combines technical measures with good governance.

The report shows how these risks are shaped not only by how models are built, but also by laws, governance structures, institutional choices, and social and political contexts. By focusing on real-world, longer-term and systemic risks, and by paying particular attention to Global South contexts, the report aims to support more practical and inclusive approaches to AI safety.

For the KAS Rule of Law Programme Asia (RLPA), working on AI safety is both timely and central to its mission. AI systems are increasingly used in areas that are fundamental to the rule of law, including public administration, law enforcement, education, healthcare, elections, or access to justice. When AI systems are poorly governed, opaque, or unsafe, they risk reinforcing existing inequalities, enabling discrimination, weakening procedural fairness, and undermining trust in public institutions. These risks are particularly serious in contexts where legal safeguards and oversight mechanisms are still developing.

At the same time, when AI is being developed and governed responsibly, it can strengthen institutions and improve the delivery of public services. Engaging with AI safety allows the RLPA to connect technological developments with core rule of law principles such as legality, accountability, transparency, proportionality, and the protection of fundamental rights.

The outcomes of the AI Impact Summit, whether new governance frameworks, institutional initiatives, or shifts in international policy debates, will strongly influence how AI safety is understood and implemented in the years ahead, including the role of Global South leadership in AI governance. Insights and developments from the Summit will also inform the next phase of research under the KAS Rule of Law Programme Asia, helping to identify new research priorities, and deepen engagement with regional perspectives.

In this sense, this report is not a final statement, but a contribution to an ongoing and evolving body of work. It is intended to support continued learning, dialogue, and collaboration as AI systems become more deeply embedded in societies across Asia and beyond.

Stefan Samse
Director, KAS Rule of Law Programme Asia

Olivia Schlouch
Programme Manager, KAS Rule of Law Programme Asia

About NLUD

The National Law University Delhi is one of the leading law universities in the capital city of India. Established in 2008 by an Act of the Delhi legislature (Act. No. 1 of 2009), the University is ranked second in the National Institutional Ranking Framework for the last five years. Dynamic in vision and robust in commitment, the University has shown terrific promise to become a world-class institution in a very short span of time. It follows a mandate to transform and redefine the process of legal education. The primary mission of the University is to create lawyers who will be professionally competent, technically sound and socially relevant, and will not only enter the Bar and the Bench but also be equipped to address the imperatives of the new millennium and uphold the constitutional values. The University aims to evolve and impart comprehensive and interdisciplinary legal education which will promote legal and ethical values, while fostering the rule of law.

The University offers a five-year integrated B.A., LL.B. (Hons.), a one-year postgraduate masters in law (LL.M), and a Ph.D. program, along with professional programs, diploma and certificate courses for both lawyers and non-lawyers. The University has made tremendous contributions to public discourse on law through pedagogy and research. Over the last decade, the University has established many specialised research centres, and this includes the Centre for Communication Governance (CCG), Centre for Innovation, Intellectual Property and Competition, Centre for Corporate Law and Governance, and Centre for Criminology and Victimology. The University has made submissions, recommendations, and worked in advisory/consultant capacities with government entities, universities in India and abroad, think tanks, private sector organisations, and international organisations. The University works in collaboration with other international universities on various projects and has established MoU's with several other academic institutions.

About CCG

The Centre for Communication Governance at the National Law University Delhi (CCG) was established in 2013 to ensure that Indian legal education establishments engage more meaningfully with information technology law and policy and contribute to improved governance and policy making. CCG is a leading academic research centre dedicated to undertaking rigorous academic research in India on information technology law and policy in India. Through its academic and policy research, CCG engages meaningfully with policy making in India by participating in public consultations, contributing to parliamentary committees and other consultation groups, and holding seminars, courses and workshops for capacity building of different stakeholders in the technology law and policy domain.

The Centre has had multiple publications over the years including reports on Exploring AISIs for the Global South, The Road to WSIS+20: Key Country Perspectives in the Twenty-Year Review of the World Summit on the Information Society (India Chapter), Platform Transparency under the EU's Digital Services Act: Opportunities and Challenges for the Global South, Social Media Regulation and the Rule of Law: Key Trends in Sri Lanka, India and Bangladesh, Intermediary Liability in India, a report Mapping the Blockchain Ecosystem in India and

Australia, a UNDP Guide on Drafting Data Protection Legislation, a book on Privacy and the Indian Supreme Court. The Centre has launched freely accessible online databases - Privacy Law Library (PLL) and High Court Tracker (HCT) to track privacy jurisprudence across the country and more than sixteen jurisdictions across the globe. CCG also has an online 'Teaching and Learning Resource' database for sharing research-oriented reading references on information technology law and policy. In recent times, the Centre has also offered courses on AI Law and Policy, Technology and Policy, and first principles of cybersecurity.

Table of Contents

Section I: Conceptualisation and Elements of AI Safety 1

A. What does AI Safety mean? 1

B. Methodology 2

C. How did the field of AI safety evolve? 3

- Addressing existential/catastrophic risks 4
- The alignment problem 5
- The need for a Socio-technical approach 7

D. How is AI Safety defined? 12

- Delineation and convergence within the fields of AI ethics, AI safety and AI security 13

E. Status quo and relevance for the Global South 15

Section II: AI Risks & Governance - A Lifecycle Approach 16

A. Inception and Planning 25

i. Identified Risks 25

ii. Governance Measures 26

Purpose Specification and Context Definition Mechanisms 27

28

Structured Stakeholder and Community Engagement 29

Early Risk Identification 30

Governance & Accountability Allocation 32

iii. Considerations for the Global South

B. Data Collection 33

i. Identified Risks 34

ii. Governance Measures 37

Data Provenance 38

Privacy Protections 41

Data governance and management best practices towards Fairness and Bias Mitigation 42

iii. Considerations for the Global South 45

C. Model Design and Training 47

i. Identified Risks 47

ii. Governance Measures 50

Risk Management 51

Transparency and Explainability mechanisms 59

iii. Considerations for the Global South 64

D. Verification and Validation 66

i. Identified Risks 67

ii. Governance Measures 70

Sandboxing 72

Benchmarking 74

Red teaming 76

Audits 77

Trials/ Testing in Real Environment 78

Regulatory Indicators 78

iii. Considerations for the Global South 80

E. Deployment and Post Deployment 81

i. Identified Risks 82

ii. Governance Measures 86

Monitoring and Maintenance 87

Incident Reporting 89

Feedback Loops 91

Human Oversight 91

Post-Deployment Audits 92

Model Retraining 93

iii. Considerations for the Global South 94

Section I

Conceptions and Elements of AI Safety

A. What does AI safety mean?

AI has achieved ubiquity across various domains, establishing itself as an effective tool for enhancing human capabilities and propelling its adoption across several high-stakes scenarios such as healthcare, financial markets, and public welfare. However, these rapid advancements have also paved the way for unchecked harms, inconsistent outcomes, and safety concerns raising apprehension among users, governance bodies, civil society groups, academia, AI researchers and policymakers.[1] The accelerated developments in the field of AI and the widespread adoption of Large Language Models (LLMs) have also highlighted its propensity for perpetuating harms including bias, privacy breaches and rampant spread of misinformation.[2] Notably, the perceived existential dangers from advanced AI have also contributed towards prioritising research on AI alignment[3] within major tech companies. The imperative to deliberate on AI safety has escalated, necessitating a comprehensive examination of the multifaceted risks inherent in AI design and deployment, as well as the development of efficacious mitigation strategies to address attendant or associated harms.[4]

In two sections, this report explores conceptions and discourse related to the notion of AI safety and subsequently, examines how key elements of AI safety are operationalised within the lifecycle of AI development and deployment. Section I of the report explores the various conceptions of AI safety followed by examining the evolution of the concept of AI safety. This section synthesises current debates and articulates a shift towards a socio-technical approach, hence ensuring that safety mechanisms are contextualised to the unique complexities of the Global South. Section II of the report delineates the methodologies for operationalising AI safety through a lifecycle approach. By dissecting the different stages of the AI lifecycle, the report offers insights on the critical risks relevant to each stage, the corresponding technical and governance interventions that can be designed for risk mitigation, and key considerations for the Global South.

B. Methodology

For the purpose of Section I in this report, we have examined existing literature to summarily capture the evolution of conceptions and risks associated with the field of AI safety. In Section II, we have adopted a skeletal lifecycle approach to identify and situate risks and governance measures in the development and deployment of an AI system. This discussion has involved examining existing academic and technical literature, policy reports, and legal analyses of relevant regulatory instruments. We have merged certain stages of the lifecycle since the risks and governance measures across these stages shared key qualities and had significant similarities. The stages classified have been isolated based on distinct identifiable harms and corresponding mitigation measures. These stages can be broken down further, but for the purposes of our research, the five stages identified are Inception and Planning, Data Collection and Preparation, Model Design and Training, Verification and Validation, and Deployment and Post-Deployment. In each stage, we identify pertinent risks, discuss key mitigation measures from a technical lens and where applicable, global regulatory efforts. Critically, we highlight the socio-technical harms that can present with AI systems, and what these harms and mitigations efforts look like, in particular, for the Global South. With a sub-section on 'Considerations for the Global South' under each stage, we seek to contextualise and highlight the unique challenges and lived realities to account for when deploying AI systems in this part of the world.

It is pertinent to note that the report was designed with certain limitations in consideration. Given the nascent stage of regulatory frameworks for AI, and the limited publicly available information on initiatives beyond the EU, UK, and US, this report draws heavily on regulatory provisions and examples from these three regions. Moreover, while this report introduces and acknowledges the importance of key actors in the AI lifecycle, it does not discuss risk and mitigation measures through this lens.

As part of our research for this report, we engaged with expert stakeholders and conducted interviews. Insights from our stakeholders have been anonymously attributed. Our stakeholders include Tarunima Prabhakar, Founder, Tattle Civic Technologies; Rafael Zanatta, Director, Data Privacy Brasil; Catherine Setiawan, Country Coordinator, Global Index on Responsible AI; Diane Chang, Senior Fellow at Tech Global Institute; and Kalika Bali, Senior Principal Researcher at Microsoft. This research has also involved stakeholder engagement through an event on 'Operationalising AI Safety' held at New Delhi on 6 November 2025. The panel and roundtables saw participation from a diverse range of stakeholders from the Global South, across sectors including industry, government, academia, and civil society. This event marked the report launch of our report, titled 'Exploring AISIs for the Global South.' The event provided an introduction to CCG's exploration and research findings from the report, including considerations for upcoming AISIs in the Global South. Building on this, the discussions at the panel and roundtables shaped insights for this report and provided further clarity on how AI safety is operationalised across the life cycle, and key considerations for the Global South.

C. How did the field of AI safety evolve?

The rise of AI automation sparked theories regarding machines' capability to generate unforeseeable decisions and surpass human control as early as the 1960s.[5] [6] Research highlights how some artificially intelligent agents, in their quest to accomplish and optimise for certain concrete end-in-itself goals,[7] were touted to value self-preservation and work strategically towards ensuring their own survival.[8] Commencing in the early 2000s, the conversations around mitigating the potential protracted effects of superintelligence and futuristic propositions related to Artificial General Intelligence (AGI)[9] marked the beginning of AI safety as a formal discipline.[10]

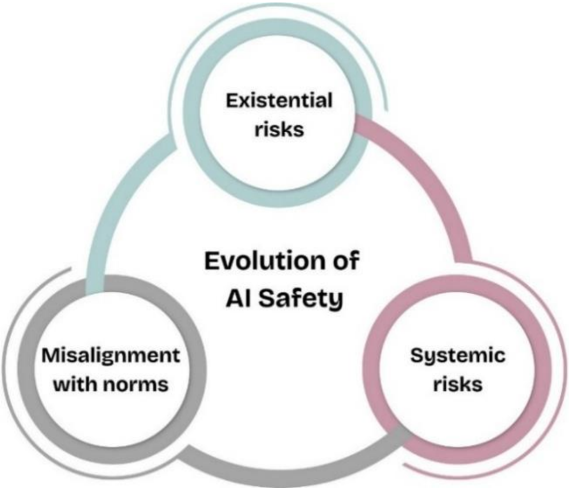


Figure 1: Evolution of AI Safety: Risks

i. *Addressing existential/ catastrophic risks*

By the mid-2000s, the prominent objective of the AI safety domain was to ensure that the use of AI machines did not culminate in harmful outcomes while contributing positively to the progress of the human race.[11] This period was followed by the beginnings of a growing consensus among a section of AI researchers that advanced AI would likely become a major source of existential risk.[12] The capability of AI tools to deliver critical information for designing weapons of mass destruction, launching cyberattacks, and affecting political stability through manipulation and misinformation are some of the examples cited by recent literature to reinforce focus on existential risks.[13] The narrative around AI potentially culminating in existential and catastrophic risks to humankind remains a heavily debated topic in research circles and continues to be instrumental in shaping the public discourse on AI safety.[14] [15]

The inflated projections around the inevitability of AGI-led apocalypse and the potential gains stemming from AI's transformative potential have also garnered interest and support from a movement, popularly termed as effective altruism.[16] The effective altruism community prioritises the promotion of a research agenda on technical AI safety, thereby steering the focus away from the consequences of an increasing corporate power asymmetry.[17] Extensively funded by tech billionaire donors and primarily driven by Silicon Valley AI companies, this ideology has been instrumental in racing to create precarious AI systems under the garb of AI safety. For instance, under the pretext of developing AI applications beneficial to humankind, tech companies have been contributing to unchecked risks by prioritising rapid deployment over rigorous safety testing, and obfuscating ambitions of securing supremacy in the global race for AGI.[18] [19] This heightened and insular focus on mitigation of AI-fuelled techno-utopianism[20] (a claim largely unfounded) associated with emerging AI also distracts and takes the attention away from genuine immediate harms warranting attention. [21] One of the most prominent focus areas for the effective altruism community also includes addressing technical alignment issues, identified as one of the many factors potentially leading to catastrophic risks to humanity.[22]

ii. The alignment problem

AI systems persistently encounter challenges related to alignment - a concept dating back to the 1960s, and which has emerged as one of the prominent sub-fields in the field of computer science.[23] The alignment problem translates as the inability of AI systems to internalise, optimise and represent the values and objectives defined by the human developer.[24] Ensuring that an AI system captures human values and norms, while understanding and behaving as expected, also forms a crucial objective in the field of

AI safety. Mis-specification of subgoals and misaligned objectives during the training and development of an AI system can prove risky, especially in scenarios where a system begins to pursue objectives outside the scope of its human-defined mission.[25] Reward hacking highlights one such instance of misalignment in AI systems, where in its quest to achieve a higher reward (separate from the intended reward outcomes), an AI agent is observed to resort to shortcuts by exploiting flaws in the design of the reward function.[26] Additionally, the complexity and subjectivity of human values also presents a significant hurdle for aligning AI in the absence of a universal agreement on the set of values that should be encoded into AI systems.[27]

iii. The need for a Socio-technical approach

Technical literature on AI safety takes a narrow approach while enumerating the objectives under the discipline of AI safety. It broadly categorises challenges to accuracy, robustness, reliability and security as the crucial components under this field.[28] This framework further lends credence to the assertion that AI safety efforts are exclusively technical. However, contemporary studies on AI safety highlight how the discipline also forms an intrinsic part of the larger structure of trust and responsibility,[29] and comprises ethical considerations including fairness, accountability, data privacy, bias mitigation and transparency, which are some of the key tenets of Safe AI development.[30] Algorithmic fairness is defined as a social construct, involving designing solutions to address the systemic harm that algorithms can inflict disproportionately on specific groups.[31] While robustness is characterised as a model's resilience against adverse circumstances,[32] the reliability of a model is measured in terms of its consistency of outputs and reduction in system failures.[33] Model transparency is the property that enables understanding the inner workings of a model, a crucial facet due to an AI system's inherent "black-box" nature. Explainability, on the other hand, aids in enhancing clarity on the rationale behind the decision-making process employed by an AI system.[34]

Further, the interdependencies between technical design decisions, system deployment contexts, and existing social hierarchies have also contributed to how technical systems shape and in return are shaped by societal power structures.[35] This perspective necessitates reframing traditional approaches to facilitate better management of risks under the domain of safety by adopting an alternative paradigm rooted in socio-technical discourse.

The adoption of AI has culminated in broader systemic risks[36] across the AI lifecycle due to intermingling with real-world social, political, and cultural dynamics. This has in turn contributed to structural harms being embedded right from the training stage of an AI system. The utilisation of underrepresentative datasets and the consequent effects of algorithmic bias have also resulted in skewed outcomes, discrimination, privacy breaches[37] and copyright violations.[38] The deployment of algorithmic decision-making for public welfare has also led to concerns of the amplification of centralised forms of state and private control.[39] Additionally, the widespread integration of AI across different applications has led to labour market disruption,[40] and human rights concerns owing to the harsh, inequitable and extractive working conditions under which the data annotators are made to operate.[41] This large-scale deployment of AI also carries substantial environmental implications, chiefly driven by the energy demands of AI infrastructure such as data centres and powerful transformer models.[42] All these contribute to an expanding carbon footprint,[43] [44] negatively impacting the living conditions of vulnerable communities inhabiting the surrounding areas.[45] [46]

Given that the array of risks falling within the ambit of AI clearly encompasses systemic risks and unforeseen downstream harms to society, safety cannot be conceptualised to hinge on purely technical approaches and emphasis on combating existential risks. Interdisciplinary scholarship also highlights how the domain of AI safety cannot solely depend on optimising technical design and focusing on the model property of an AI system[47] Emphasising that safety of an AI system is also largely contingent upon its deployment context, concerned stakeholders and institutional environment, research points out the significance of embracing a socio-technical framing to AI safety.[48] A sociotechnical approach understands that the efficacy and safety associated with any technology is always the result of not just technical features but also larger societal forces, and takes into consideration factors such as institutional governance, human labour and social conditions.[49] Taking a socio-technical approach can therefore enable the creation of solutions embedded with societal insights by ensuring stakeholder engagement from diverse domains. Opting for participatory models involving impacted communities can also support AI systems in functioning effectively by minimising reliance on technological workarounds alone.[50] [51]

The field of AI safety needs to be broadly conceptualised to incorporate a holistic framework encompassing a wide spectrum of harms and a diverse range of values and approaches to address these harms effectively.[52] Research has expanded the scope of this academic discipline by necessitating consideration of not just technical aspects such as model alignment, or research on existential risks, but also recommendations for regulating the governance of AI for mitigating social harms.[53] In other words, AI safety research must accord equal significance to the full spectrum of AI-generated risks, encompassing both technical issues (such as unsafe exploration [54] and distributional shift[55]) and social harms (including bias and discrimination) to examine how it engages with people, systems and processes.[56] The notion of AI safety must be contestable for it to be conceptualized and reinterpreted by different communities in accordance with their goals and aspirations.[57]

D. How is AI Safety defined?

AI safety is a nascent, emerging and conceptually disputed field open to various interpretations by different stakeholders. It has been defined as a research discipline entrusted with the prevention or minimisation of risks emanating from the development and implementation of AI systems.[58] In the technical sense, it has also been described as an area investigating causes of unanticipated outcomes in AI systems[59] and conducting research focused on optimising the benefits of AI integration.[60] This field also encompasses mitigation of accidental risks resulting from unforeseen divergence of AI systems from their intended behaviour.[61]

Delineation and convergence within the fields of AI ethics, AI safety and AI security

Literature on AI attempts to delineate AI safety with two closely related but distinct fields - AI ethics and AI security - depending on their core functions and the category of harms these fields seek to address. The field of AI ethics is well-established and has been defined as a collection of principles, values, and methods that applies accepted standards of morality to the development and application of AI.[62] Principles including fairness, transparency, explainability and privacy have been categorised under the ambit of AI ethics. This discipline typically concerns itself with mitigation of present-day harms[63] (bias, discrimination and privacy invasion, for instance) momentarily plaguing algorithmic systems.[64] AI safety, on the other hand, has been ostensibly linked with the design of interventions aimed at resolving prospective challenges (strengthening robustness of a system, for instance) inherent in evolving algorithmic systems.[65]

Research also points to a lack of consensus in the category of risks falling under the ambit of AI safety and AI security. AI security is a characteristic of a system for exhibiting resilience against malicious attacks on the system components and operations, and ensuring that the integrity of the system is preserved in the face of adversarial actions.[66]

According to scholarly analyses, AI safety must typically concern itself with unintended or accidental harms arising out of AI systems, such as misalignment, design flaws and embedded biases, while AI security is purportedly entrusted with the aim to build resilience in AI systems against malicious actors and intentional adversarial acts (jailbreaking, for instance).[67]

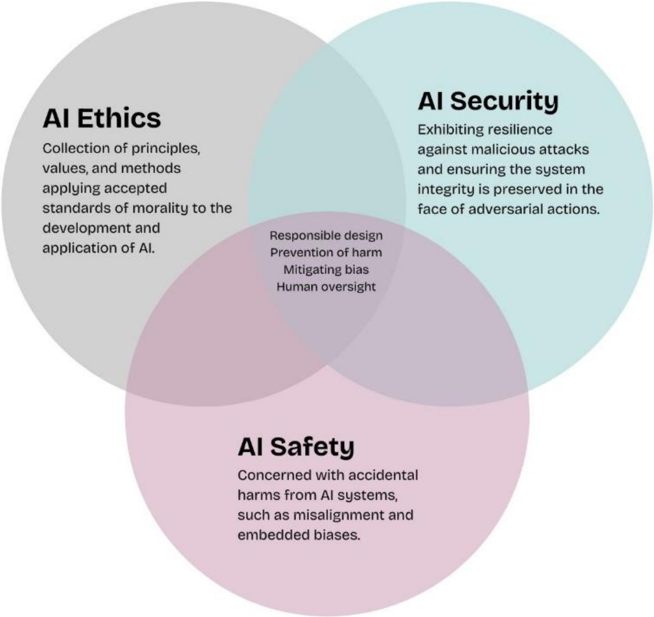

AI Ethics
Collection of principles, values, and methods applying accepted standards of morality to the development and application of AI.
AI Security
Exhibiting resilience against malicious attacks and ensuring the system integrity is preserved in the face of adversarial actions.
Responsible design
Prevention of harm
Mitigating bias
Human oversight
AI Safety
Concerned with accidental harms from AI systems, such as misalignment and embedded biases.

Figure 2: Delineation and convergence within the fields of AI ethics, AI safety and AI security

However, recent discourse on AI emphasises the necessity of integrating these disciplines given their overlapping boundaries and challenges in delineating both risks and solutions, to more effectively mitigate the contextual harms posed by AI systems.[68] [69]

E. Status quo and relevance for the Global South

The meteoric popularity and large-scale deployment of LLM applications has also acted as a catalyst for prompting urgent discussions amongst jurisdictions and policymakers worldwide around developing the field of AI safety.[70] The first AI Summit hosted by the United Kingdom at Bletchley Park in November 2023 was instrumental in ushering the issue of AI safety and the risks posed by frontier AI to the forefront.[71] This Summit also led to the establishment of the first AI Safety Institute[72] (AISI) in the UK, which was followed by multiple jurisdictions launching their own AISIs.[73] The International AI Safety Report released by the UK AISI also attempted to define and widen the scope of AI safety by including the mitigation of intentional risks, malicious uses and harms resulting from deepfake exploitation, dissemination of misinformation and adversarial attacks.[74] However, despite the initial two AI Summits at Bletchley Park and Seoul heavily focusing on the safety aspect, there was a noticeable shift and expansion of focus at the AI Summit held at Paris earlier in 2025.[75] [76] The emphasis in some jurisdictions has also gradually shifted from safety to addressing potential threats to national security,[77] including cybersecurity risks and implications from development of biological warfare.[78] Another noticeable shift in the AI discourse is the growing emphasis on driving AI innovation, with safety measures considered a roadblock by some governments in achieving global AI leadership.[79] With the nature and gravity of harms resulting from AI use compounding at exponential rates, it is crucial that AI safety receives the critical level of importance it warrants.

The Global South populations remain exposed to a distinct set of AI-related risks, ranging from fragmentation of local labour markets to the perpetuation of existing social biases.[80] [81] These vulnerabilities are further amplified by a dearth of indigenous datasets, algorithmic bias and inadequate digital infrastructure, exposing the populations to greater systemic risks often neglected by global frameworks on AI safety.[82] It becomes critical, therefore, to develop effective practices and mechanisms for risk mitigation contextualised to address the unique realities of the Global South.

Section II

AI Risk and Governance: A Lifecycle Approach

Section II of the report explores how AI safety can be operationalised through targeted interventions across the AI lifecycle, encompassing inception and planning, data collection and preparation, model design and training, verification and validation, and deployment and post-deployment. The stages of a lifecycle, and the associated risks and mitigation measures depend on the size, scale and scope of the AI system. Key safety risks are examined at each stage, such as adversarial vulnerability in the model design and training stage, and model degradation in the post deployment stage. The section also maps governance measures and notable technical interventions at each stage, including robust auditing procedures, red-teaming, and oversight mechanisms to highlight mitigation efforts and exercises. Additionally, this section highlights existing regulatory practices as part of mitigation efforts where relevant. Highlighting major considerations for the Global South underlines the research and is discussed at the end of every stage of the lifecycle. These insights for the Global South discuss emphasising equitable access, the necessity for context-specific adaptations, and inclusive and expansive training data, to prevent exacerbating regional inequalities.

Adopting a lifecycle approach allows stakeholders to study and manage the sequential progression of decisions and tasks leading up to and including the deployment of AI solutions.[83] Lifecycle management includes evaluations of the response time, quality, fairness and explainability, among others, of AI systems in context.[84] Key factors contributing to risks, namely privacy, cybersecurity, trust, interpretability, explainability, robustness, usability and wider social implications can be accounted for, and managed, given the comprehensive nature of a lifecycle approach.[85]

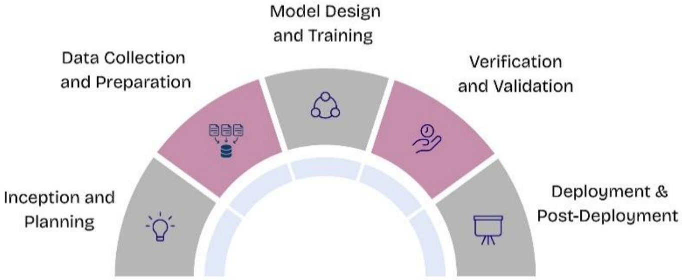

Inception and Planning
Data Collection and Preparation
Model Design and Training
Verification and Validation
Deployment & Post-Deployment

Figure 3: Stages of the AI Lifecycle

Another valuable addition of observing AI systems through this lens is the ease of identifying key actors involved across various stages, like data scientists, engineers, IT managers and compliance teams. Recognising these human parties in the lifecycle is vital to ensure appropriate allocation of responsibilities and adequate oversight of the stages, thereby supporting transparency and traceability efforts during the development and deployment.[86] Given the high number of actors involved, assigning specific roles and responsibilities is necessary for accountability, especially to address concerns during failures or breaches. By breaking down the process into a lifecycle and also identifying responsible actors, we can visualise a holistic image of the AI system in order to assess how and by whom safety is implemented. This report acknowledges the importance of key actors, but does not granularly approach risk and mitigation measures through this sole lens.

Model developers, also known as the architects, lay the foundational capabilities and conceptualise the AI system and its behaviours.[87] They build the models, crucially contributing to the potential output and its impact. System deployers are the party bridging the gap between the abilities of an AI system and its real-world application.[88] They actively customise and operationalise the system, depending on its abilities and uses within a targeted industry. Finally, the users interact with the system and uniquely introduce new information, influencing the real-world performance of the AI tool. They are most impacted by the AI system, and hence, clear explanations and directions on the performance and impact of the tool should be made available to them.

Beyond this operationalisation, two other human parties are also involved with the AI lifecycle.

Theorists are accountable for developing and advancing AI theory through a technical lens, and typically belong to academic or industrial research settings.[89] Ethicists are concerned with the principles of safe AI systems and include policymakers, commentators and critics from multiple disciplines like academia, law, journalism, economics and politics.[90] They are concerned with research beyond the technical qualities, and evaluate AI systems in light of external auditability, societal impact and wider legal compliance.[91]

Visualising the development and deployment of an AI system as an iterative lifecycle with built in testing, verification, and feedback methodologies allows for a process of continuous improvement. Distinct risks emerge at different stages of the lifecycle, necessitating tailored interventions to address them proactively. Biases and privacy violations often manifest during data collection and processing, while deception arises during model training.[92] In post-deployment, issues like unintended societal harms or jailbreaking vulnerabilities are revealed.[93] Corresponding safety measures include technical safeguards like careful data curation, adversarial robustness testing in the early stages, and governance tools such as auditing and regulatory sandboxes in the later stages. Breaking down risks and mitigation measures via stages proves useful by enabling targeted, scalable mechanisms that provide oversight, facilitate accountability and prevent compounded issues that may only manifest once the AI system is released to its context-specific users.[94] However, there are limitations to this approach. AI systems vary in their model design (foundational models as opposed to deployed applications, low-risk vs high-risk systems, etc.), which complicate generalised categorisation of stages, risks and mitigation measures.[95] Rapid technological advancements can modify these classifications further, especially given the adaptive nature of AI models. The stages overlap and are interconnected, defying neat delineations in how risks can present. This report aims to provide a broad mapping that can be referential in developing detailed and comprehensive studies as the research in this field, especially in the Global South, continues to grow.

These following two sub-sections provide a brief overview of the governance frameworks and technical safety measures related to AI safety, which we will explore in greater detail subsequently.

Governance frameworks for AI safety

Globally, various governance frameworks are being established to address and contain harms arising from the proliferation of AI. Voluntary standards, in the form of ISO/IEC 42001:2023 and the Risk Management Framework (RMF) conceived by the National Institute of Science and Technology (NIST) have proven useful in proposing mechanisms to organisations for internal governance and undertaking effective risk oversight procedures.[96] The NIST Risk Management Framework (RMF) provides a flexible mechanism for organisations involved in the development and use of AI models to identify, evaluate and effectively incorporate risk mitigation measures for harms arising through the entire lifecycle.[97] The Organisation for Economic Co-operation and Development (OECD) principles on AI also necessitate ensuring robustness, security and safety in the functioning of AI systems. These principles also recommend the creation of adequate safeguards across all the stages for prevention of unforeseen risks and unintended behaviour on the part of AI.[98]

The EU AI Act proposes a risk-based approach to AI regulation and imposes stringent requirements on developers of high-risk AI systems, which include conducting mandatory periodic risk assessments.[99] California's recently enacted AI safety legislation - the Transparency in Frontier Artificial Intelligence Act (SB 53), also implements a comprehensive governance framework by mandating transparency and disclosure requirements from developers of large frontier AI models regarding the safety framework followed by them.[100] The legislation also provides for whistleblower protections for workers of companies involved in model development and necessitates immediate reporting of a 'critical safety incident'.[101]

Technical safety measures under the realm of AI Safety

AI safety is also interpreted as a specialised domain within the ambit of system safety engineering, which enables investigating and offering solutions to AI-generated social and ethical harms through the application of foundational safety engineering principles.[102] Safety engineering prescribes taking a 'safety-by-design' approach at the outset, enabling better risk management and avoiding retrofitting safeguards until after the risk becomes apparent post the model deployment.[103] Designing technical safety measures, therefore, forms an integral part of ensuring the safety of AI systems and withstanding adversarial attacks in the form of data poisoning, jailbreaks and prompt injections.[104] Enhancing system safety also necessitates considering the behaviour of the numerous components of the AI system working in unison. This can facilitate crafting interventions throughout the system lifecycle in the broader context of its functioning and the environment under which it operates.[105] Dissecting the different stages of the AI lifecycle can also prove valuable in designing effective mitigation strategies to contain the downstream risks associated with AI implementation. A multitude of safety techniques including employing anomaly detection and adversarial robustness (such as red teaming) to ascertain malicious use[106] have been developed to minimise the risks arising out of the various stages of the AI lifecycle, which shall be dealt with subsequently.

A. Inception and Planning

The inception and planning stage is the earliest stage of the AI lifecycle. This stage involves context setting, identifying objectives and purposes that involves defining use and constraints or perceiving risks such as identifying affected communities. These tasks are critical for assessing feasibility and foreseeable misuses, and setting the accountability, oversight, and risk-management structures that will guide the entire lifecycle.[107] This stage is closely tied to NIST's AI RMF "Govern" and "Map" functions. The "Govern" function relates to outlining processes for the AI lifecycle to identify and account for foreseeable risks that may arise. This function involves steps such as documenting legal requirements and establishing accountability structures.[108] The "Govern" function is supported by the "Map" function which provides further context on the outlining of the AI lifecycle through steps such as mapping impacts for individuals, groups, communities and society.[109] The inception stage plays a critical role in determining and accounting for an AI system's long-term safety and social impacts because it defines the trajectory of the system.

i. Identified Risks

During the inception and planning stage, AI safety risks may emerge when problem definitions, objectives, potential risks and contextual parameters are insufficiently examined and articulated.[110] The absence of meaningful engagement with relevant stakeholders and affected communities further compounds these risks by obscuring local concerns, domain constraints, and lived-experience insights.[111] Early risks also stem from incomplete anticipation of failure modes,[112] security threats, and downstream impacts on individuals, groups, and institutions.[113]

ii. Governance Measures

These challenges can be addressed through rigorous purpose-and context-definition practices, structured stakeholder and community consultation, systematic early-stage risk identification, threat modelling and impact assessment processes, and clear allocation of governance roles and accountability mechanisms. It is relevant to note that while frameworks often recommend some of these mechanisms to be carried out across the AI lifecycle, it is critical that they are employed at the outset to prevent irreversible safety harms at later stages.[114]

Some of the key governance mechanisms under the inception and planning stage include purpose specification and context definition mechanisms, structured stakeholder and community engagement, early risk identification, and governance and accountability allocation.

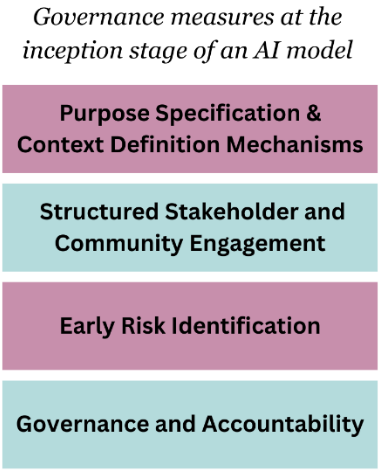

Governance measures at the inception stage of an AI model
Purpose Specification & Context Definition Mechanisms
Structured Stakeholder and Community Engagement
Early Risk Identification
Governance and Accountability

Figure 4: Inception Stage of AI Model: Governance measures

Purpose Specification and Context Definition Mechanisms

It is imperative to outline the purpose and set the context at the beginning of the AI lifecycle. This includes delineating and defining the AI system's purpose, intended use, non-intended uses, operational context, and constraints before development begins.[115] Doing so is relevant because several challenges, particularly at the design and training stage and at deployment, arise from misalignment between the objectives of the AI systems and the priorities of the designer, discussed in subsequent sections.

The 'Map' Function within the NIST AI RMF recognises the need to outline and define the purpose and context of the AI system, by specifying the environmental, organisational, and functional conditions that shape how risks arising from the AI system should be understood. These conditions relate to practices like documenting the organisation's mission and relevant goals,[116] intended purposes, anticipated beneficial use cases. It can also relate to contextual parameters relevant while deploying an AI system, like norms, and legal frameworks.[117]

Structured Stakeholder and Community Engagement

Through this mechanism, AI developers and deployers must establish formal processes to engage with diverse stakeholders in consultative processes. These stakeholders may include users, workers, citizens, domain experts, academics, civil society, grassroots organisations and marginalised communities who will also be impacted by the use of AI or whose data has been used to train models. Early engagement reduces downstream fairness risks, misalignment with societal values, and contextual misalignment.[118]

Identification and engagement with stakeholders is an integral aspect of the 'Map' function under the NIST AI RMF.[119] This framework articulates how engagement with relevant stakeholders and communication towards shared understanding can support AI developers and deployers in understanding the likelihood and magnitude of potential impact.[120]

Early Risk Identification

Organisations must identify risks and impacts that may potentially arise from the AI system through threat modelling and impact assessments. While it is critical to evaluate the system for potential risks throughout the lifecycle, it is particularly imperative for these assessments to be conducted at a foundational stage.

For instance, the 'Map' and 'Manage' functions of the NIST AI RMF explicitly require early identification of foreseeable misuse and safety risks.[121] Similarly, the EU AI Act requires risk management systems to be established and maintained for high-risk AI systems.[122] These systems comprise both risk assessment as well as risk management. Risk assessment may be carried out at the inception and planning stage. However, under the EU AI Act, the risks articulated correspond to those which may be mitigated during the development or design stage, or "when adequate technical information is provided."[123] As a result, this report discusses the operationalisation of the risk management system under the EU AI Act under the model design and training stage.

The EU AI Act stipulates the steps involved in the implementation of the risk management system to include identifying and analysing risks which may either be known or reasonably foreseen, particularly in terms of impact on health, safety, or fundamental rights.[124] The Guidance for Risk Management of Artificial Intelligence systems ("Guidance") specifies that risk identification includes identifying the sources of risk, and potential areas which the deployment of the AI system may impact, in order to create a comprehensive risk register.[125] The Guidance states that risk analysis involves examining the nature, sources, likelihood, and potential consequences of the identified risks.[126]

Subsequently, the AI Act requires that these risks are to be estimated and evaluated[127] prior to adopting appropriate and targeted risk management measures to mitigate the risks identified above for this stage.[128] The Guidance states that risk evaluation involves comparing the identified risks with the risk criteria, risk appetite and tolerance to determine whether an identified risk is acceptable.[129]

Governance and Accountability Allocation

This process involves defining and allocating clear roles for stakeholders involved across the AI lifecycle, including the AI owner, deployers, project manager, risk officer, human-oversight authorities, etc. This is significant because systemic risk often occurs due to failures and ambiguity in accountability. When stakeholders are unclear about the responsibility allocated to them, this can give rise to unsafe decision-making, governance bottlenecks, and organisational blind spots.[130]

This has been recognised by the EU AI Act, which details the obligations which must be allocated among providers,[131] deployers,[132] importers,[133] distributors[134] and other stakeholders. While it is not necessary to fulfil all the obligations at the inception stage, it is critical to document and make the relevant stakeholders aware of their corresponding responsibilities at this stage to prevent safety harms arising subsequently.

Similarly, the 'Govern' function under the NIST AI RMF emphasises early allocation of accountability structures within AI systems.[135] This includes practices like establishing, documenting and communicating clear, well-defined roles, responsibilities, and communication pathways relating to mapping, measuring, and managing AI risks, in order to support the systematic identification, assessment, and mitigation of AI-related risks.[136]

iii. Considerations for the Global South

Multinational companies typically operate out of multiple and diverse social contexts, which can present distinct challenges in how they strategise in terms of critical decisions. This is particularly relevant for markets with "cultural distance," where there is a difference in terms of cultural values. In such instances, there is often greater transfer of practices from the "home country", which in most cases relates to the Global North.[137] This can often result in limited incentives for some of the biggest global foundational model developers and deployers in key sectors such as health, recruitment and education towards investing in understanding local languages, cultural contexts, demographies, and power structures in non-priority markets in the Global South. Consequently, this may result in an underrepresentation of diverse needs. For instance, techno-linguistic bias favouring dominant languages from the Global North being prioritised in a manner that hinders AI systems from being able to correctly represent concepts from other communities.[138]

Often, due to the above factors, there is limited participation of users from the Global South and marginalised communities within these jurisdictions in developing and co-designing AI models at the inception stage.[139] It is also critical to take into account the dominance of AI players established in the Global North in developing AI infrastructure, and its implications on data sovereignty,[140] individual privacy and communities' rights over their cultural and linguistic resources.[141] This hegemony also risks technological dependency of the Global South on the Global North, a loss of local autonomy, and may overlook local needs, languages, and social norms in the inception and planning of AI systems.[142]

These challenges are also often likely to be exacerbated due to several Global South jurisdictions lacking effective governance frameworks such as data protection and AI safety regulations and industry standards.[143]

B. Data Collection and Preparation

Data collection is the stage of the AI lifecycle where raw data is acquired for training, validating, and testing an AI model.[144] Data collection can involve data acquisition, where new datasets may be discovered or generated. Data discovery involves extrapolating insights from existing datasets by identifying patterns and trends in data, which can improve security and inform decision-making.[145] Data generation refers to the process where synthetic data is artificially generated through computer simulations or generative models to supplement "real-world data" in a cost-effective manner.[146]

Data preparation is the stage of the AI lifecycle where the raw data is processed and refined before further stages.[147] This process involves improving upon or augmenting datasets or making them more representative by improving the quality of existing data to be more consistent, accurate and reliable.[148] The quality of datasets is often improved by handling missing data or values[149] or providing additional context to raw data through data labelling and annotation.[150]

Datasets may also be made more representative by ensuring diversity of data, either by replacing existing data that may cause bias,[151] or increasing representation of underrepresented data samples in datasets.[152]

i. Identified Risks

Several risks can arise at the data collection and preparation stage.

Transparency and explainability

A significant risk that may arise is a lack of transparency and explainability for users and deployers.[153] This is due to limited documentation of how processes in the data collection and preparation stage are employed. These processes include techniques such as data augmentation (where the size of the dataset is increased using synthetically generated data), and data repurposing (where data collected for a certain purpose is reused to train AI models to identify patterns for a different purpose).[154] These limitations in transparency may also contribute to challenges for users and deployers in identifying threats relating to adversarial manipulation,[155] which presents challenges in terms of determining accountability for harms stemming from a model's outputs.[156] Moreover, a lack of transparency due to limited documentation of the sources and nature of data collection may also give rise to certain concerns. This may be due to the quality of data collected and used for the AI model being 'bad.' For instance, misinformation detection tools developed using datasets drawing upon misinformation and disinformation present on the Internet, may be less accurate. This is because these tools may assume the misinformation they are trained on to be true.[157] In certain cases, there is also the risk of the nature of data collection being bad, as opposed to the data itself; for instance, where data has been scraped in an unauthorised manner by web crawlers.[158]

Data privacy violations

There may also be concerns of data privacy violations. These may arise in the absence of informed consent, and limited adherence to data protection principles.[159]

Bias

AI models may sometimes exhibit biased behaviours and produce inequitable and unfair outcomes. One form of such bias could be statistical bias, which can arise due to particular demographic groups being underrepresented in training datasets.[160] For instance, demographic underrepresentation in datasets used for computer vision models can produce higher error rates for darker-skinned individuals as compared to lighter-skinned individuals.[161] Limited obligations for explainability and transparency on prominent AI systems with regard to information on datasets can also result in users being unaware of the implications of potential biases.[162]

Another form of bias could include societal bias, relating to structural inequalities in society, which are reflected through the data that has been collected. For instance, predictive AI models which have been trained to identify and reproduce existing patterns in datasets may replicate prevalent societal bias surrounding male and female representation in different fields like medicine, engineering, etc.[163]

Risks in data labelling

Data labelling processes can often give rise to fairness risks. Data labelling refers to identifying and providing context to raw data, such as images and videos for a machine learning model.[164] This process can often fail to account for complexities in data. For instance, emotion recognition systems can create overly simplistic representations of complex humans, failing to capture local contexts of expressing emotions, and other situational nuances.[165] This could consequently produce unfair outcomes where an individual is incorrectly profiled for mental health concerns, for instance.[166]

Additionally, data labelling may also present risks in allocative decisions due to underrepresentation of data, or can reinforce societal biases due to misrepresentation in labels. For instance, the UK ICO illustrates how common practices like uniform sampling from the training dataset, may result in unfairness for underrepresented communities. This is because uniform sampling, where a data point is randomly selected from a dataset, is less likely to present a data point of an underrepresented community, simply due to less data being available about that community.[167]

ii. Governance Measures

The above identified risks are fundamentally tied to the data that models are trained on,[168] and certain governance measures like record keeping and audits can be effective in accounting for these risks at the data collection and preparation stage.

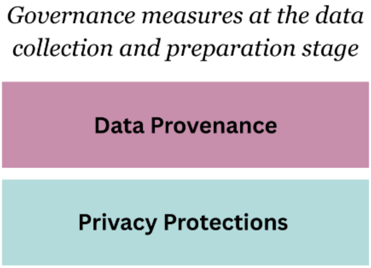

Governance measures at the data collection and preparation stage
Data Provenance
Privacy Protections

Figure 5: Collection and Preparation Stage of AI Model: Governance Measures

Data Provenance

Data provenance refers to the detailed record of the origin, history, and transformation of data inputted at the data collection stage.[169] This mechanism helps ensure that the data used to train or test AI systems is traceable and trustworthy - for instance, by verifying the data collected, the creators of the data, and records and histories of how the data has been modified. This can consequently help verify the authenticity and consent associated with the data.[170]

There are several mechanisms and technical measures through which data provenance can be operationalised. This includes auditable chains, which identify the origin, owner, validation, change points, and destination within the dataset. Some of the key artefacts underpinning these auditable chains include a registry of sources, end-to-end maps tracing data touchpoints from its origin across various processes and transformations, and detailed logs mapping outcomes to responsible decision-makers.[171] In the absence of uniform and universally accepted standards or frameworks for data provenance, several researchers have also proposed documentation standards like datasheets,[172] data statements,[173] and data cards.[174]

Regulatory provisions can also support AI model developers in adhering to principles of transparency, robustness and fairness through data provenance. For instance, the EU AI Act requires providers of general-purpose AI models to develop a "sufficiently detailed" summary of the data used to train the AI model. This summary must be in accordance with a template provided by the AI Office, and must be made publicly available.[175]

NIST AI Risk Management Framework 1.0 also underscored documentation, traceability, and data provenance as integral to trustworthy AI. Within its 'Map' and 'Govern' functions, the framework calls for systematic recording of information related to data collection, selection, system design, and testing. This documentation is intended to enable accountability, transparency, and oversight throughout the AI system lifecycle, including the data collection stage.[176]

Canada's Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems calls on model developers to publish training data descriptions. This includes data about the types of training data used to develop the AI system, as well as risk identification and mitigation measures.[177]

Moreover, a Checklist for AI Auditing, commissioned by the EU Data Protection Board, also recommended certain factors that could be accounted for while the training data is being audited.[178] This could include examining whether data sources have been documented, and whether input data sets, data use and intermediate data, and output data can be traced.[179]

Privacy Protections

Data provenance is a significant governance mechanism to support accountability and transparency within the data collection and preparation stage. However, it is imperative that such data provenance, as well as data collection and preparation more broadly, be done in accordance with data protection frameworks.

For instance, the General Data Protection Regulation[180] [181] [182] [183] [184] - the provision specifically articulates that producers should be encouraged to account for data protection rights while developing products and services which process personal data.

The Guidelines on the interplay between the EU Data Protection Regulation and the Artificial Intelligence Act[185] are slated to be released in 2026, according to the European Data Protection Supervisor.[186] However, literature already points to the existing impact of the GDPR on AI systems, including the significance of principles like data minimisation.[187] A study by the Panel for the Future of Science and Technology submitted to the European Parliament contended that AI systems can integrate data processing requirements and safeguards established by the GDPR.[188]

The study reiterated the need for mechanisms such as pseudonymisation with respect to data collected for developing the AI systems. The study also recommended identifying distinctions between the use of personal data for the purpose of developing a training dataset for AI systems, which involves broader correlations amongst individuals, as compared to personal data being used to profile and make decisions about individuals.[189] This can support the principle of purpose limitation provided under the GDPR by limiting data processing to the purpose for which it may have been collected. The study also recommended establishing deterrents against institutions that intentionally disregard the interests of data subjects and thereby exploit their trust.[190]

Research also points to consent registration as a technical measure that provides clarity regarding whether creators have consented to the collection and storage of data, and also allows people to grant, withdraw, or negotiate consent.[191]

The Checklist for AI Auditing, commissioned by the EU Data Protection Board, also recommended ascertaining whether data sources and the method of collection contravene data protection principles under the GDPR.[192] The recommendations also propose audits to examine whether the process of data collection followed the approach of privacy by design, as laid down under the data protection principles under the GDPR.[193]

Data governance and management best practices towards Fairness and Bias Mitigation

The EU AI Act requires that training, validation and testing data sets shall be relevant, sufficiently representative, and account for geographical and contextual characteristics.[194]

The UK Information Commissioner's Office ("UK ICO") provides implementational direction for this provision through its guidance on ensuring dataset fairness, which specifies certain factors to be considered and demonstrated to ensure the AI system is trained and tested on datasets that are representative, relevant, accurate, and generalisable.[195] These factors include designing datasets to be representative of relevant populations and communities by collecting sufficient data (both in terms of quantity and quality).[196] The guidance also acknowledges the significance of reliability and impartiality in data collection within the above factors by recommending that data be assessed, recorded and sourced through "sound collection methods" in an up-to-date and accurate manner.[197]

The UK ICO takes into consideration the various sources of bias in datasets and AI systems (including data collected through third parties).

Bias from external datasets

The UK ICO emphasises the need to assess risks of bias particularly when organisations procure datasets from other organisations through processes like data sharing.[198] For instance, the UK ICO highlights how selection biases in sampling and measurement may arise when non-representative external datasets are integrated with data collected directly by organisations themselves.[199]

Bias mitigation in data labelling/annotation

Harms can arise from misrepresentation or underrepresentation of certain data points in data labelling and annotation. This can be mitigated by implicit bias training to equip AI companies with an understanding of how underrepresentation and misrepresentation may impact the decision-making of the AI system. Participatory design through inclusion of underrepresented communities in the labelling process can be useful in helping organisations formulate inclusive labelling criteria and protocols.[200] For instance, the Uli dataset for automated detection of hate speech and gendered abuse in Hindi, Tamil and Indian English employs a participatory approach by selecting annotators who identify as women or members of the LGBTQIA+ community in South Asia.[201]

Bias mitigation through dataset modification

Bias mitigation at the pre-processing stage can also involve modifying datasets to either remove data that may result in biased outcomes and discrimination, or correct an imbalanced dataset by increasing representation of communities that were previously underrepresented.[202]

During the data collection process, every data point assumes an equal weight, which can lead to marginalised communities remaining underrepresented in datasets as well.[203] One way to address aggregation bias that may arise due to certain groups being underrepresented includes reweighting data points during data preparation to balance the representation. For instance, data was reweighted to avoid unfair outcomes in a clinical prediction model for postpartum depression (PPD). The model had been trained on potentially biased data and was diagnosing higher numbers of white women with PPD, contrary to medical literature pointing to evidence of higher prevalence of PPD among women from marginalised communities. In this case, reweighting data mitigated bias by reducing the effect of the biased data point (such as race) in prediction by removing it as a characteristic.[204]

Similarly, during the data preparation process, data labelling can often fail to account for structural and societal challenges and can perpetuate bias and discrimination.[205] These structural inequalities can be accounted for by changing data labels, wherein labels can be modified on the basis of algorithmic fairness metrics to mitigate risk against communities most vulnerable to such discrimination.[206]

Other bias mitigation strategies can include mechanisms such as data visualisations, which chart changes in training data's distribution patterns to indicate potential bias within the corresponding dataset.[207]

iii. Considerations for the Global South

Some of the risks identified in this stage may pose specific challenges when an AI system is developed or deployed in the Global South. Obtaining informed consent in data collection can often be a challenge due to linguistic diversity impeding the accessibility of transparency and explainability disclosures for individuals whose data is being collected.[208]

A study on low-resource government health systems in Zanzibar provides insights into how several patients were unaware about the individuals accessing their data or why it was being collected.[209] These issues arose primarily due to challenges in drafting data access management guidelines and translating policies into Swahili.[210] In any case, it is also pertinent to note that consent for data collection for AI systems in the Global South cannot be regarded to be fully meaningful due to power asymmetries between powerful multinational corporations and users in the Global South.[211]

Separately, in the course of our interactions with stakeholders, they highlighted methodologies and taxonomies under data justice initiatives such as SoberanIA, which may necessitate Big Tech companies having to pay license fees for using indigenous languages in the future fees.[212] Linguistic diversity can also pose challenges with respect to data preparation due to certain preprocessing methods being biased towards English.[213] Researchers have also identified how existing preprocessing methods such as tokenisation, normalisation, and embedding are biased towards English or other high-resource languages, leading to systematic errors.[214]

There are also challenges relating to traceability at the data collection and preparation stage, due to standards and methods of data collection not being uniform.[215] Non-standard data sources like paper records, verbal reports, and informal data, which may be more commonly used in the Global South,[216] may have higher measurement errors, inconsistent formats, and missing fields.[217] Moreover, case studies from the usage of a Nigerian health chat have revealed how a lack of traceability and explainability in sampling and weighting can lead to opacity and hinder transparency with respect to representation of communities while developing AI models.[218]

C. Model Design and Training

The model design stage includes selecting model algorithms, defining model architecture, and training techniques to address the objectives and problem statement identified at the inception stage.[219]

Subsequently, in the model training stage, the model is subjected to the data which was collected and prepared in the data preparation stage.[220] Thereafter, the model learns to recognise patterns within this data, which influences its capacity to generate outputs such as predictions or decisions. Appropriate training algorithms (which were selected during the model design stage), such as supervised, unsupervised learning, or reinforcement learning,[221] are applied and implemented for this purpose.[222]

i. Identified Risks

Misinformation

Certain AI systems, like LLMs, are designed to have the capability to generate synthetic media which is virtually indistinguishable from human-generated content. This can pose risks of misinformation and enable deception, fraud and impersonation.[223] These risks can arise when users are unaware they are interacting with AI chatbots, particularly in critical sectors such as finance and healthcare.[224] The usage of AI to generate "deepfakes" can also be misused by bad actors to create sexually explicit content,[225] impersonate celebrities,[226] carry out fraudulent activities such as identity theft and financial scams,[227] or even to influence electoral outcomes.[228] Misinformation is an indirect risk in the model design stage because this outcome is dependent on how AI systems are designed and safety measures for transparency of users are taken, as discussed subsequently.

Fairness and bias risks

In the data collection stage, the bias often occurs due to the training data being underrepresentative.[229] On the other hand, bias in the model training stage often occurs due to inherent biases in the algorithms used to train the machine learning model. These inherent biases refer to biased algorithm design, where debiasing techniques such as reweighting (where data is adjusted to be more representative) can remain ineffective due to biased assumptions by designers.[230] For instance, in an attempt to mitigate gender-based bias in hiring decisions, designers may fail to account for race-based bias within a gender that has been traditionally underrepresented.[231] These biases are subsequently reflected in the outputs produced by the AI system.[232] For instance, in the previous example, the AI system may subsequently prioritise white women in its hiring decisions over Black women.

Adversarial Vulnerabilities:

Adversarial evasion attacks occur when carefully designed, often imperceptible perturbations, known as adversarial examples, are introduced into an input by sophisticated attackers. This causes machine-learning models to produce highly inaccurate or otherwise arbitrary predictions, despite the input appearing unchanged to humans during oversight.[233] These attacks can undermine trustworthiness, particularly in high-risk AI systems like autonomous vehicles and critical infrastructure, where adversarial failures can have severe implications for AI safety.[234] For instance, a stop sign may be manipulated by adding imperceptible noise to the image, which may hinder an AI system from recognising the image as a stop sign.[235]

***Prompt Injections*:**

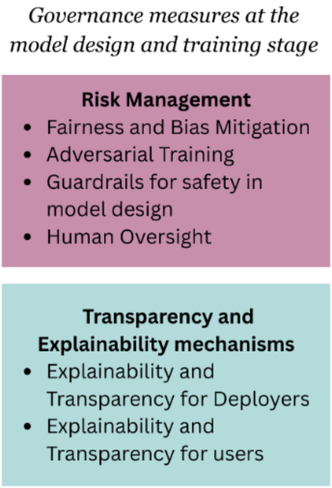
This refers to instances where bad actors, such as hackers, exploit large language models with malicious inputs disguised as legitimate prompts to produce outcomes containing misinformation, sensitive information, or information that contravenes human ethics.[236] For instance, in early versions of ChatGPT, users were able to craft prompts (such as under the guise of pretence) that enabled ChatGPT to produce concerning outputs such as racist and homophobic content.[237]

Misalignment:

This refers to instances where the AI model learns an indirect pattern, which differs from the human intent and the objectives outlined during inception.[238] Misalignment can give rise to unfavourable and unpredictable outcomes at the deployment stage.[239] For instance, social media recommender systems that have been trained to increase user engagement may prioritise political misinformation.[240]

ii. Governance Measures

Certain governance measures such as risk management and transparency mechanisms can be effective in accounting for these risks at the model design and training stage.



Governance measures at the model design and training stage

Risk Management

- Fairness and Bias Mitigation
- Adversarial Training
- Guardrails for safety in model design
- Human Oversight

Transparency and Explainability mechanisms

- Explainability and Transparency for Deployers
- Explainability and Transparency for users

Figure 6: Model Design and Training Stage of an AI Model: Governance Measures

Risk Management

The risk management system under the EU AI Act, discussed in the inception stage, is understood as a continuous iterative process across the AI lifecycle. While identification and evaluation take place at the inception stage, the techniques for risk management and mitigation are operationalised during model design and training. Although the EU AI Act does not stipulate any specific risk mitigation techniques, risk management techniques can include fairness and bias mitigation, adversarial training, safety guardrails, and human oversight.

- ***Fairness and Bias Mitigation:*** Bias mitigation may be implemented differently across the design and the training stage. With regard to the design stage, the AI Act requires that high-risk AI systems that continue to learn after market placement or first use, be developed in a manner that prevents bias arising due to feedback loops.[241] This may present itself as instances where biased outputs may be perpetuated as inputs for future iterations of the AI system.[242] For instance, AI systems trained on biased data may perpetuate biased hiring decisions, continuing to underrepresent marginalised communities. This could reinforce existing biases and inform biased datasets in the future, to be subsequently used to make hiring decisions.

  With respect to the training stage, bias and unfairness mitigation techniques involve in-processing methods. In-processing methods can often account for challenges of bias amplification that occur during the model training stage, where bias occurs due to unfair algorithms, irrespective of the representative nature of the dataset.[243] For instance, a study revealed that an algorithm over relied on past healthcare costs to predict future medical care needs, resulting in white patients being prioritised. This bias arose because the algorithm made an incorrect inference from data, failing to account for the fact that Black patients with similar health conditions as white patients were incurring lower healthcare costs due to socio-economic and access barriers.[244]

These in-processing methods often include fairness constraints and adversarial debiasing.

Fairness constraints refer to formal requirements or rules which are introduced during the training process to mitigate biases that may arise when AI predictions are correlated with sensitive attributes such as race or gender.[245]

Adversarial debiasing promotes fairness by training the primary AI model in opposition with an adversarial network (a secondary model that attempts to detect weaknesses, biases in the primary model) to reduce the influence of sensitive features.[246]

- ***Adversarial Training:*** Adversarial training involves training an AI system to improve robustness against adversarial attacks by training the model on a combination of inputs which are both clean as well as examples of adversarial perturbations.[247] As demonstrated in the risks above, adversarial perturbations refer to inputs that have been deliberately modified to produce unfavourable outcomes. Under this approach, AI systems improve robustness by learning how to accurately classify and differentiate between clean and perturbed inputs.[248]

The NIST AI 100-2 (2025) report on Adversarial Machine Learning provides detailed guidance on adversarial training, certified robustness techniques, and threat models tailored to different attack surfaces.[249]

Regulatory provisions also recognise the need for mitigation strategies for improving robustness against adversarial risks. For instance, the EU AI Act stipulates that high-risk AI systems must be designed and developed with appropriate robustness, cybersecurity, and resilience.[250] The provision specifies that security measures must include protection against adversarial inputs, confidentiality attacks, model flaws, and unauthorised alterations to datasets or pre-trained components through data poisoning and model poisoning.[251]

- ***Guardrails for safety in model design:*** The EU AI Act recommends taking technical and organisational measures to achieve robustness of high-risk AI systems against errors, faults or inconsistencies occurring within the AI system's operating environment.[252] These can include technical redundancy solutions, such as backup or fail-safe plans,[253] where multiple AI models run parallelly, allowing a model to ensure continuity or supervision in the event of another model failing or behaving unpredictably.[254]

  For instance, utilising two models, the primary model's outputs can be monitored for unsafe or misaligned behaviour, especially under adversarial queries. Under this approach of Two-Tier or Guardian-Based Models, a secondary or "guardian" model supervises the primary model.[255]

- ***Human Oversight:*** Human oversight refers to the ability of humans to exercise agency to monitor, intervene or override decisions made by AI systems.[256] Frameworks such as the OECD AI Principles acknowledge the potential of AI to negatively impact human rights and democratic values[257] (which may occur, for instance, due to misinformation and bias). This framework therefore recommended human agency and oversight as a safeguard against these risks.[258] Similarly, the United Nations Educational, Scientific and Cultural Organization (UNESCO) Recommendation on the Ethics of Artificial Intelligence proposes human oversight, for the purposes of ensuring both responsibility and accountability. The framework suggests that jurisdictions should seek to ensure that responsibility across the AI lifecycle can be attributed to specific physical persons or existing legal [259]☐

  The EU AI Act reiterates this need for human oversight, and embeds human oversight as an overarching governance mechanism across the AI lifecycle with specific ways corresponding to the model design and training stage, which are discussed subsequently.[260] The EU AI Act seeks to mitigate risks that may arise to health, safety or fundamental rights during the use or misuse of high-risk AI systems.[261] The EU AI Act specifies that high-risk AI systems be designed and developed in a manner allowing oversight by natural persons during deployment.[262] Under this provision, the natural persons assigned to carry out human oversight activities for this purpose must be equipped with the capacity to accurately interpret the merits and limitations of the relevant high-risk AI system.[263]

Researchers have noted that while Article 14 of the EU AI Act is a significant development towards human-centric AI, the provision does not provide specific guidance on the modes through which human oversight is to be operationalised.[264] Instead, the provision requires that oversight measures be proportionate to the potential risks, intended level of autonomy and context of deployment of the high-risk AI system.[265] This approach allows different sectors to contextualise human oversight measures to the needs of their sectors while maintaining accountability.[266]

The sub-guideline on Human autonomy and Oversight under the Inter-Parliamentary Union's Guidelines on Ethical Principles for AI Use in Parliaments recognises three primary modes of human oversight in AI.[267]

These include Human-in-the-loop (HITL), Human-on-the-loop (HOTL), and Human-in-command (HIC). Each of these modes corresponds to different levels of human intervention and autonomy of the AI system.[268] The HITL model involves human mediation of all decisions taken by an AI system. The HOTL model employs a balance between human oversight and AI autonomy wherein human intervention may be carried out during the project development phase, shifting to human monitoring of AI decisions and outputs during the operational phase. Through the HIC model, a more in-depth analysis of the implications of an AI system can be undertaken, supported by public feedback regarding the AI system.[269]

However, only the HITL mode comes into place during the design and training stage,[270] while the other two modes come into place primarily post training. The HITL mode of human oversight includes supervised learning, reinforcement learning from human feedback (RLHF) and active learning. HITL in supervised learning involves training AI models through labelled datasets which have been annotated by human data scientists (for instance, human labelling of text as “spam”).[271] Reinforcement learning refers to training an AI model through its interactions with its environment, and RLHF is a type of reinforcement learning where the AI model is trained with direct human feedback through reward functions.[272] In the case of active learning, human input is only sought for categories of predictions where the AI model presents lower confidence.[273]

Transparency and Explainability mechanisms

Explainable AI refers to processes and mechanisms providing human users with an understanding of the expected impact and potential risks and biases associated with an AI model.[274] Transparency is a key aspect of explainability,[275] and it can be embedded into the model design and training stage in a manner that allows both the deployers as well as the users to more meaningfully understand the AI system and its outputs.[276]

- **Explainability and Transparency for Deployers:** The EU AI Act acknowledges the need for transparency during the design and development of AI systems and lays down certain transparency requirements. For instance, the EU AI Act requires high-risk AI systems to be designed in a way that "its operation is sufficiently transparent" that allows deployers to appropriately use and interpret outputs generated by the AI system.[277] This includes providing documentation relating to the intended purpose, accuracy level, robustness, and potential risks of the AI system. Instructions for use should also specify human oversight measures, computational and hardware requirements, maintenance needs, and logging mechanisms which facilitate appropriate functionality.[278]

- **Explainability and Transparency for users:** Recent research has highlighted the need for explainable AI, which would ensure that users interacting with AI systems are able to understand its outcomes and decisions.[279] Explainability supports users in understanding, verifying, and ascribing accountability to decisions made by AI systems, facilitating greater transparency and trustworthiness for users.[280] For instance, it can help users understand any biases that may exist in the AI system and its impact on hiring decisions made with regard to a user.[281]

  More recently, research has acknowledged the need for "human-centred explainable AI,"[282] and supporting non-technical users in understanding AI systems and their outcomes or decisions.[283] Some of the mechanisms to operationalise such explainability for non-technical users include participatory design and non-conventional explainability techniques. For instance, research has underscored the significance of participatory and co-design practices in promoting accountability for the deployment of machine-learning interventions implemented in the context of local communities.[284] Similarly, participatory design practices can be leveraged towards collaboration between scientists and local communities for the purpose of co-designing AI systems towards more explainable systems.[285]

Moreover, the EU AI Act embeds transparency into the design stage by mandating transparency for users of the AI system, to prevent risks of impersonation and deception.[286] The EU AI Act places transparency obligations both with respect to interaction with AI systems, as well as engagement with outputs synthetically generated by the AI system.

For instance, the EU AI Act requires providers to ensure that AI systems designed to interact with natural persons are designed and developed to ensure that users are made aware of the fact that they are interacting with an AI system,[287] particularly in the deployment of emotion recognition systems or biometric categorisations.[288] Similarly, regulatory frameworks in the US also require the embedding of such transparency during deployment. For instance, California's law regulating AI companion chatbots requires platforms to explicitly inform child users that conversations are AI-generated.[289] Prior to this, the Utah Artificial Intelligence Policy Act was enacted, requiring entities to disclose to consumers in a clear and conspicuous manner that they were interacting with generative AI entities, and not humans.[290]

The EU AI Act also requires that AI providers mark synthetic output (such as text, audio, images and videos) generated by AI systems as artificially generated or manipulated.[291] This includes use cases where the AI system is deployed to generate content which constitutes a deep fake,[292] or to generate or manipulate text published in relation to matters of public interest.[293] Similarly, the Indian government released draft amendments to the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 ("IT Rules, 2021"), requiring Gen-AI and social media platforms to label 'synthetic information.'[294] It is important to note however, that there are concerns with watermarking and traceability, particularly privacy concerns surrounding the misuse of users' personal data in the absence of adequate safeguards.[295]

These requirements effectively become operative at the deployment stage when the AI provider makes the AI system available for use.[296] However, the First Draft Code of Practice on Transparency of AI-Generated Content ("Draft Code of Practice") published by the European Commission ("EC") suggests that certain mechanisms such as watermarking may be embedded at the training stage itself.[297]

Moreover, it is also imperative for AI systems to be designed in a manner to allow for compliance with transparency obligations at the deployment stage. The EC is currently developing Guidelines and a Code of Practice on Transparent AI systems to provide clarity on technical implementation means for operationalising Article 50.[298]

Once released, these technical implementation mechanisms can be embedded to adhere to transparency requirements at the model design and training stage. In this manner, transparency to the user in terms of notifying them when they are interacting with AI systems or AI-generated outputs, may help mitigate risks arising from deception and impersonation facilitated by interactions with AI systems.

iii. Considerations for the Global South

While AI systems are increasingly being deployed in Global South countries, technical governance mechanisms for AI safety deployed in these jurisdictions often fail to take local contexts into account.[299]

For instance, there are certain constraints with respect to adversarial robustness. Research has identified that developing adversarial robustness requires more training data.[300] This can have implications for marginalised communities in the Global South whose data may be underrepresented in datasets.[301] For instance, our expert stakeholders reiterated the challenge of limited existing digital data in the Global South. They highlighted the work of projects such as Masakhane in Africa which are working on creating datasets to increase access to low-resource African languages.[302] They also highlighted the need for the Global South to focus on curating higher quality datasets with less data in order to prioritise quality of data over quantity, the latter of which is currently limited.

Moreover, certain challenges in relation to human oversight also present themselves. Jurisdictions in the Global South have recently been recognising the need for human oversight as a governance mechanism to mitigate safety risks in the AI lifecycle. The recently released India AI Governance Guidelines also include human oversight as a safeguard to account for risks stemming from loss of control in sensitive sectors, within their People-First Approach under their Key Principles.[303] However, it is also imperative to consider the limitations of human oversight and not over-rely on this mechanism for AI safety.[304]

The India AI Governance Guidelines acknowledge the limitations of human oversight from a sectoral perspective (with respect to high-velocity sectors where direct human oversight may not be effective).[305] However, research on Article 14 of the EU AI Act points to broader challenges with human oversight, like humans' cognitive constraints, and how automation bias can pose constraints to the efficacy of the provision.[306] These concerns are particularly relevant considerations for the Global South as well. This is because the AI ecosystem in certain jurisdictions within the Global South faces significant constraints with respect to limited technical knowledge and experience operating AI tools.[307] This may present challenges with human-in-the-loop models currently envisioned.

It is also crucial to consider the concept of "explainability pitfalls."[308] Researchers contend that irrespective of explainability, users may disregard independent judgment and subordinate their decision-making to AI systems.[309] This is particularly relevant because studies have pointed to the heightened trust users in Global South jurisdictions such as China, Brazil and India have for algorithmic decisions and AI systems.[310] While this ties to the arguments of automation bias discussed by Fink, this heightened trust could also mean that transparency obligations such as those under Article 50 of the EU AI Act, requiring providers to indicate to the user that they are interacting with an AI system like a chatbot, may not be sufficient approaches for AI safety.

D. Verification and Validation:

The vocabulary of 'verification and validation' (henceforth, "V&V") of AI systems can be traced back to the need to ensure pre-deployment accuracy of the system. V&V represents both a pre-deployment procedural stage and ongoing operationalisation of technical measures throughout the AI lifecycle, focused on evaluating and testing an AI system. Concretely, the key purpose at this stage is to ensure that the system meets the requirements set at the inception stage and performs as intended.[311] It is, in essence, built into the lifecycle as a risk mitigation tool, seeking to address challenges that may arise ex-ante.

Verification ascertains whether the design and development of an AI system or its component is in line with pre-determined requirements at earlier stages, and validation checks if the AI system fulfils its objectives within the context of its purpose.[312] An example of verification methods includes testing a chatbot's factual accuracy against the knowledge base it was trained on.[313] Validation can involve testing the prediction abilities of the AI system just before it is deployed, by inputting example prompts in a virtual environment.[314] Rigorous testing and evaluation can determine risk and failure points throughout the lifecycle, and help mitigate these to ensure not only compliance but also build confidence in the ability of the system to handle real-world complexities.

i. Identified Risks

The risks previously identified in prior stages continue to exist, but they may present themselves differently. This is because risks associated with data collection, such as poisoned data, privacy leaks, bias in outputs, etc., must continue to be addressed at the V&V stage. Other specific challenges and risks unique to the nature of V&V are rooted in the technical operationalising of testing methods, especially when concerned with the novel features of AI systems.[315]

Lack of Specific Models and Criteria

A primary issue is the absence of clearly defined validation models and criteria, which hinders the design of test cases for AI functions. Scholars have identified attributes of AI systems that make it difficult to create clear models and criteria for V&V, such as the dependency of behaviour on training data and subsequent uncertain behaviour regarding untested data. They also discuss the oracle problem as a further V&V concern, as explained below, which makes it difficult to clearly define the accuracy criteria for the correct outputs for each individual input.[316]

Isolated Model Testing

Another concern is isolated model testing. V&V processes must focus on testing the AI systems within the complete socio-technical ecosystem, and not just individual models.[317] End-user and AI interactions determine real-world performance, and isolated model testing would fail to consider the overall impact of the system in the environment it is deployed into.[318] The full effect of an AI system can only be felt once it is released to the wider public, and can be influenced by individual attitudes and user decisions.[319] Such user behaviour can introduce variability, leading to emergent risks from entanglement with all components of the AI system.[320] This can be understood as a manifestation of the butterfly effect, where one component's interaction (here, the user behaviour) with all the other components of the AI system can lead to a compounded effect. For instance, real-world testing, where such user behaviour is incorporated, can present ethical risks and dilemmas on bias amplification from flawed data, and the overall system exhibits greater bias than a single component[321] Further, the opacity of an AI system, and the knowledge and communication gaps between experts (the designers, developers and deployers) and the end-users, is a hurdle when designing questions that can be used to test for outcomes that may be present in the real-world.[322] User decisions made based on outputs generated by an AI system have significant real-world impact. While the developers may be aware of the intended use of the AI system and design it accordingly, they may not give sufficient consideration to pre-existing biases that may be amplified when post-deployment user input is introduced to the system. Such knowledge and communication gaps can lead to instances where bias persists due to prediction errors. For example, predictive grades based on 'typical performance of students' may reinforce socio-economic inequalities, even though the algorithm was designed to avoid such inequalities in grading.[323]

The 'Oracle Problem'

The 'oracle problem' is another crucial risk for testing/validating AI systems.[324] If the AI system is expected to function autonomously for an extended duration, it is designed with the ability to adapt or change its behaviour and outputs according to its environment and input.[325] It is nearly impossible to certify what these circumstances can present as, and what the desired outcome to verify and validate must look like. A related risk is the impossibility of accurately verifying because of the sheer scale and diversity of the environments. For example, in the automotive industry, verifying and validating AI systems designed for autonomous vehicles requires testing the cars for the possibility of fatal accidents. According to researchers, it is infeasible given the breadth of testing required to conduct adequate testing that determines a safe AI application. This is because it would require driving 433 million kilometres and an 'impossible' amount of time to conduct complete testing.[326]

Additionally, there are very limited standards for V&V for different AI systems, including procedural guidance and metrics. Firstly, existing best practices designed for testing conventional function-based software are rooted in well-defined inputs and within specified behaviours. However, these have been stretched and adapted for AI systems that are much wider conceptually, based on machine learning models using data-driven training.[327] Secondly, the required tools and frameworks are rare, sector-specific or focused on niche domains, and where they are available, they are scattered and non-comprehensive.[328]

ii. Governance Measures

Notable governance and technical measures are relevant throughout the lifecycle, such as data provenance and adversarial training, which continue to apply and can be operationalised as key measures in addressing identified risks at the V&V stage as well. Interestingly, assessing the AI systems for potential harm can involve scientific and philosophical questioning, where AI is viewed as a method towards understanding its proximity to mimicking human intelligence.[329] Other testing and V&V looks at evaluating performance based on the AI system's success within real-world, practical applications.[330]

Governance measures at the verification and validation stage

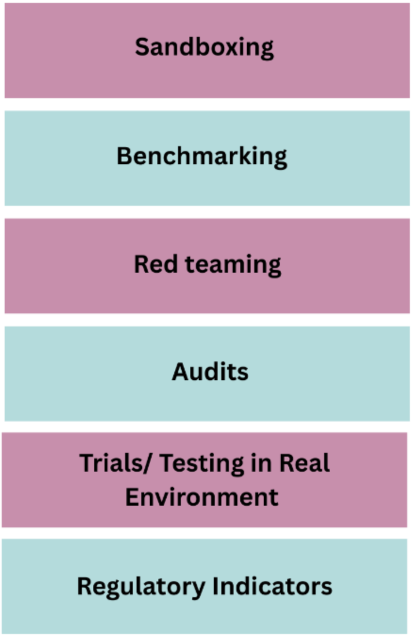


Figure 7: Verification and Validation Stage: Governance Measures

Sandboxing

AI sandboxes seek to foster innovation while safeguarding safety by testing within a controlled experimental environment in the pre-market stages.

Technical sandboxing is a form of contained testing, where the functions of an AI system can be executed under tightly controlled frameworks which can include runtime limits, restricted privileges such as limited access to the system's memory, etc.[331] This allows the developers to evaluate and test files or processes while mimicking real execution conditions, without exposing the system to compromise, such as malicious attempts like prompt injection.[332]

Regulatory sandboxing allows for this technical testing to take place in an environment with reduced regulatory constraints.[333] This creates a safe space for developers and legislators to work together, under regulatory supervision, with individual legal guidance.[334] As a governance mechanism, sandboxing can promote compliance with relevant regulatory obligations.[335] This is because conducting sandboxing testing during V&V encourages the improvement and implementation of legal certainty, best practices and evidence-based learning throughout the AI lifecycle.[336]

There are risks to regulatory sandboxing, and scholars argue that this may lead to developers evading responsibilities, lower safety standards and expose end-users to the very harms regulators seek to protect them from.[337] Careful and intentional operationalisation of sandboxing is necessary to ensure that the balance between innovation and regulation is maintained. To encourage this, the EU AI Act for example, calls for a sandbox plan - a document describing the objectives, timeline, methods and requirements for measures taken within the sandbox.[338]

Along similar lines of testing in a controlled environment, but distinct from sandboxing, are trials in a mock environment. Trials in a mock environment are a method of testing where the AI system is validated in an environment simulating an actual real-world environment. Such a trial can also be conducted in a direct or partial copy of the environment where the system will be deployed into.[339] This allows testers to look for malfunctions and correct them during pre-deployment. The potential damage is in a controlled environment and does not impact real-world performance negatively or cause harm to users.[340] Scholars present that trials can be cheaper in situations where the final environment is 'simple enough' to simulate.[341] For example, the trial of a music application was conducted by monitoring subjects' heartbeats while studying, thereby testing the ability of the system to recommend music that would improve efficiency. The system was used as intended, but the testing environment was 'mock' - in the way that the students were aware and studying in a replicated environment, and not when they would naturally be studying.[342] Given this simplified simulation of the real-world environment, there are limitations that follow. Scholars present that in a real-world environment, users' emotions and behaviour are influenced by various stimuli, and the whole breadth of these may not be represented in the mock environment.[343]

Benchmarking

Benchmarks are a particular combination of datasets that can include testing data, training data, and performance metrics.[344] Together, these represent specific tasks or abilities of the AI model and are selected as a shared framework to comparatively test the efficiency and safety of AI systems.[345] Conducting assessments using benchmarks operationalises AI safety efforts and increases standardisation due to the intentional selection of benchmarks and the decision-making process in developing benchmarking frameworks.[346] To further transparency and traceability, scholars opine that policymakers must ensure that applied and trusted benchmarks are well-documented, have clearly defined tasks and performance evaluation mechanisms.[347] Accordingly, the EU AI Act calls for the development of benchmarks and other measurement methods that can help assess accuracy and robustness levels of high-risk systems.[348]

Despite widespread reliance on benchmarks, researchers discuss concerns on the current use, such as ethical questions on what should be measured, according to which standards and what are the downstream effects.[349] As scholars have discussed, LLM benchmarks, for example, are hindered by limits of their creator's knowledge and may not be able to fully ascertain future and emerging AI capabilities, given they rely on the capacity of human understanding. As a result, these knowledge constraints of creators also impact how specialised and nuanced benchmarks can be, especially in more sensitive and contextual sectors like healthcare.[350] The performative and generative nature of benchmarks must also be considered, in the way that benchmarks not only describe and measure but also influence and actively shape the performance metrics that AI systems can be tested on.[351] Benchmarks are powerful and not neutral - they impact how AI systems are trained, tested, applied, and have economic, cultural and safety effects.[352] Large companies can shift their decision-making and business strategy, or raise massive capital based on the benchmark testing results.[353] Scholars also discuss how these large companies have a vested interest in the technology performing well, and can also exercise influence over the designing of the benchmarks. This can be a major conflict of interest in designing adequate benchmarks that can test for safe AI.[354] Specific concerns include ideas that mitigating AI safety risks that are continuously evolving and context-dependent are challenging for benchmarking as a V&V process.[355] Scholars recommend designing and applying benchmarks that are dynamic, continuously assessed for potential misuse, have rigorous evaluation protocols on validating and updating benchmark results, and can evaluate unintended consequences, alongside the performance of an AI system.[356]

Red teaming

Red teaming is a structured testing methodology that uses adversarial testing to assess the security and safety of an AI system.[357] The process involves simulated real-world attacks, misuse and abuse scenarios that target the weaknesses within a system.[358] These attacks can appear as prompt injection, data poisoning and model jailbreaks to test a system's resilience.[359] These 'threat models' are part of the interactive, iterative process and test the defences to gauge harmful behaviours such as violent content generation, and privacy issues like data leakage.[360] The verification processes can harness red teaming by engaging with external stakeholders like local civil society and advocacy groups, to incorporate risks of local malicious or authoritarian misuses that the AI system must be tested against.[361]

Audits

Audits as a governance mechanism are key in addressing risks in the development of AI systems and are typically concerned with investigating the systems and data, generating reports for both developers and users, and working towards transparency and explainability.[362] As part of V&V, internal audits allow for further accountability and traceability given the proximity to ongoing development procedures. Internal auditing can also address challenges that arise out of the 'blackbox' nature of algorithms, playing a key role in demystifying the opacity. By introducing structured oversight and verification methods, it promotes transparency through techniques like model documentation, where auditors require detailed logs of data inputs, processing logic, and output decisions.[363] This allows stakeholders to trace how algorithms arrive at conclusions, reducing reliance on unexplained predictions.[364] Adequate access to information during the V&V stage would allow for well-conducted internal audits.[365] Auditing helps accreditation with globally accepted standards, like the International Organisation for Standardisation.[366] There are also external audits, which are discussed later in the report as a governance mechanism for the Deployment and Post-Deployment stage.

Trials/ Testing in Real Environment

Real-world trials of AI systems are critical, since they are the sole non-hypothetical approach to testing whether an AI system is safe, trustworthy and effective under realistic conditions, beyond benchmarking and simulation-testing.[367] Real-world social contexts can pressurise AI systems, and analogous pressures are placed on the system in order to test it. Such trials can be conducted with increasing complexity, with small missions to the full scope of the AI system.[368] Rigorous trials can be laborious but they reveal possible issues in component collaboration, communication, sensors or usability of the system.[369]

The EU AI Act presents that the testing of high-risk systems must include real-world conditional testing and a comprehensive plan must be made, which is to be authorized by a market surveillance authority.[370] The real-world testing under V&V is required to last only as long as necessary, with informed consent, and suitable and qualified human oversight.[371]

Regulatory Indicators

While considerations around regulating AI can be traced in pre-existing and ongoing developments, industry practices and regulatory standards, there are limited obligations specific to V&V in the current regulatory climate. The EU AI Act indicates requirements where validation processes must be relevant, sufficiently representative, error-free and complete with regard to the intention of the testing mechanism.[372] Transparency of V&V methods and results is vital, as it serves as additional information for deployers, allowing them to interpret the AI system and use it appropriately.[373] The US NIST has also developed an AI Risk Management Framework.[374] NIST focuses on engagement with diverse internal teams, perspectives of stakeholders and external collaborators to improve the capacity for context understanding, assumptions of use, limitation grasp, etc. - which directly relate to ensuring robust testing mechanisms relevant to V&V processes are developed.[375] The framework emphasises a need for separation of duties, indicating that the developers and users of the AI system must be distinct from those conducting the V&V processes to ensure effective operationalisation of the risk mitigation efforts.[376]

Global South countries like Brazil have proposed new laws on governing artificial intelligence, which include guidance on developing sandboxes to foster innovation.[377] The law indicates requirements on validation for sandboxing and 'adoption of adequate parameters for the separation and organization of data for training, testing and validation of the results of the system.'[378] Further, the OECD AI Principles recommend testing and validation protocols, highlighting the iterative and not strictly sequential nature of V&V.[379] Continuous validation plays a key role in ensuring that risks associated with the oracle problem, for example, can be addressed in an ongoing manner, even post-deployment of AI systems.[380] It is only through continuous V&V that concerns around fault-tolerance, safety, or quality assurance can be resolved.

iii. Considerations for the Global South

As discussed, mechanisms under V&V critically depend on the quality of both the testing data and the team. Different people should bear the responsibility of training and testing; diversity should not be limited to the data, and must be present throughout the lifecycle. If the testing teams are not engaged locally, in regional contexts, V&V processes cannot adequately recognise risks specific to the local communities and the environment. AI models developed in the Global North are typically trained on data that underrepresents the diverse population, languages, culture and specific local context in the Global South. Testing of similar datasets as input directly without Global South context specific edits can lead to performance degradation and continuing bias when the system is deployed locally, because the system is designed and tested to perform at its highest ability in a Western Demographic and not the Global South.[381]

English-centric benchmarks are frequently poorly translated and are rife with errors.[382] They overlook local contexts like regional informal laws or traditions, and cannot capture the nuances required when using benchmarks to verify and validate an AI system.[383] Setting newer benchmarks that are Global South-sensitive with datasets on multiple categories ranging from literature to media, in both low and high-resource languages, require comprehensive efforts.

These benchmarks must capture both cultural breadth and depth to reflect local traditions and values.[384] One method to operationalise this is to include more substantive participatory approaches, and bringing in relevant stakeholders from local communities for consultations on inclusion, collaboration and ownership.[385] Participatory benchmarks enable performance measurement on broader dimensions, and can be designed and enabled to utilise human judgment by an individual with relevant knowledge and context.[386] For AI systems to be universally useful, they must meet the users where they are by respecting the regional contexts, and so must the verification and validation methodologies used to test them.

E Deployment & Post Deployment

The final stage of an AI system’s lifecycle can be further divided into specific stages: the initial deployment and post-deployment. Ongoing operation, continuous monitoring, revaluation and retirement are aspects of the post-deployment stage.[387]

Deployment is the stage at which the AI system is revealed to the target audience and can be put into use, provided it has been trained, tested, verified and validated. The risks at this stage are more pronounced than previously identified risks, since data that was not part of the training cycle, as well as unforeseeable inputs from users, and real-world impact of outputs, are no longer hypothetical.

i. Identified Risks

Users and impacted communities are at a risk from direct and indirect harms from AI systems which they are exposed to during real-world use, revealing vulnerabilities missed in prior risk mitigation and uncovering new threats. The following discussion focuses on the latter. Upon deployment, an AI system can present social harms, such as biased outputs that reinforce existing discrimination in policing or healthcare.[388] With newer inputs from end-users, privacy risks can be revealed.[389] Misuse of an AI system is a major risk post-deployment, where an AI system not intended for surveillance could be used to do so.[390] A further risk is the dehumanisation of personal healthcare decisions.[391] For example, the latter is part of a larger concern on the erasure of humanised healthcare systems, with increasing automated healthcare resulting in reduced doctorpatient relationships and individuals' medical autonomy.[392]

Data Drifting and Model Degradation

Post-deployment, AI systems bear the risk of losing accuracy due to data drifting. This phenomenon can occur in certain scenarios, where data collected and trained on begins to gradually differ from the data during the operation; it can in some cases have negative implications, leading to incorrect predictions or outcomes. When data drifts, it can lead to model degradation - where the AI system's performance declines with changes in the data or with changes in relationships between the variables on which it was trained and tested.[393]

Data drifting can be caused by the data sources exhibiting seasonal variations or changes in the user behaviour with evolving preferences, impacting input distribution.[394] A type of data drift is concept drifting, where the statistical relationship between input data and output/target variable changes with time. The risks associated with this are a direct challenge to the training of AI systems based on stationary data.[395] The extent of risk differs based on the application of the system. For example, AI systems used in diagnostic medicine have an impact on patient outcomes since the diagnoses depend on the reliability of disease prediction models; hence the continued/ long-term precision of the system post-deployment must be ensured.[396] Data drift in healthcare applications can occur due to differences between medical practices, training versus clinical use, changes in patient populations, disease patterns, etc.[397] For instance, in cases where data drifting leads to model degradation, real-world social harms can occur. When AI systems are deployed in predictive policing to identify recidivism, it can lead to an innocent person being misidentified as a possible repeat offender.[398] Automatic decision-making abilities of an AI system for sensitive topics such as crime control is a risky and unstable process, and requires interrogating the necessity of AI application in such fields.[399]

Model Inversion

Model inversion is a privacy risk where attackers seek to obtain sensitive information from AI systems, specifically machine learning models. Once an AI system has been deployed, a hostile user can reverse-engineer the learning and training procedures to reveal information that was part of the original training data.[400] Strategic querying exploits the tendency of an AI system to retain knowledge of training data, and attackers analyse outputs to deduce or reconstruct personally identifiable information that could include biometric information, medical records or financial details.[401]

Additional Risks

Previously discussed risks like data poisoning embedded into training data that lead to malicious backdoors, which may have been missed during the initial testing, can impact the accuracy and user experience once an AI system has been deployed.[402] Malicious backdoors, in particular, can weaken the system and reduce its trustworthiness, and make it vulnerable to manipulation once novel data is fed into the system as user inputs.

There are on-ground and practical concerns with AI system governance that persist throughout the lifecycle, such as infrastructure instability and low digital literacy among local deployers, oversight personnel and end-users. There are also concerns with a lack of information provided by AI developers to deployers and users, including a lack of proper documentation on the purpose, possible use cases, training data, V&V undertaken, human oversight requirements, and incident reporting mechanisms. Defining clear and traceable selection processes for responsible and expert personnel are essential to mitigate concerns like oversight gaps and accountability failures. Existing risks include the exacerbation of inequalities and vulnerability of users due to weaker oversight. Notably, there are also safety vs functionality trade-offs due to a rapidly deployed AI system given the innovation mindset.[403] When faster deployment is prioritised over sufficient risk and harm assessments and oversight mechanisms, it can expose the deployers' organisations to fines and operational errors.[404] All these are specific concerns that continuous monitoring must account for prior to asserting the accuracy and quality of an AI system.[405]

iii. Governance Measures

The governance and technical measures for the risks that present at this stage are similar to those discussed previously, and have a continuous nature to them. Risk mitigation and governance measures during and after deployment should account for abuse and accidental misuse of an AI system, social harms, inadequate human monitoring, and reporting on the purpose and limitations of the AI system for users, etc.

Deployers and providers of high-risk AI systems applied in sensitive environments also necessitate broad mechanisms such as sharing instructions on use for end-users, conducting fundamental rights impact assessment and conformity assessments, logging and sharing of technical information with the relevant stakeholders including accuracy metrics and record keeping.[406] The specifications come from the EU AI Act, but are indicative of governance measures that are omnipresent, some of which overlap with mitigation techniques discussed below.

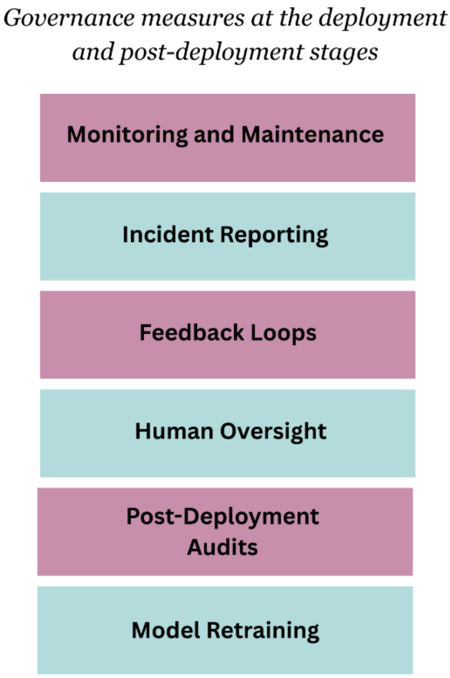

Governance measures at the deployment and post-deployment stages
Monitoring and Maintenance
Incident Reporting
Feedback Loops
Human Oversight
Post-Deployment Audits
Model Retraining

Figure 8: Deployment and Post Deployment Stage of an AI Model: Governance Measures

Monitoring and Maintenance

Monitoring and maintenance are necessary to ensure the AI system's robustness, ethical compliance, stakeholder trust, and safety of users and impacted communities. Specifically, to address data and concept drifting, drift detection frameworks are required.[407] Proactive approaches to drift detection require regular monitoring of the model performance based on pre-determined metrics, ongoing learning and collecting data. Tracking performance across different user groups are some ways to study the root cause.[408]

Behavioural pattern and anomaly detection is an element of ongoing monitoring which can help address model inversion risks.[409] By tracking and analysing repetitive, near identical prompts, possible model inversion can be detected early. For example, a single user submitting over a certain number of semantically similar queries can be flagged, like guardrails in a chatbot application that are triggered if a user input violates the safety policy of the deployed system.[410] Output screening and content filtering are important steps to be taken during the pre-deployment phase that can impact post-deployment risks, especially model inversion and malicious use.[411] Scrutinising what outputs the AI system creates can help govern the leakage of personally identifiable information during malicious attacks in a proactive manner.

Post-market monitoring is a requirement under the EU AI Act for high-risk AI systems, where systematic collection, documentation and analysis of relevant data allow the system providers to enable continuous evaluation.[412] The EU AI Act also indicates serious incident reporting requirements.[413] Similarly, the NIST AI RMF also recommends continuous vigilance, monitoring and human oversight of AI-systems.[414]

Logging (including automatic logs) and record keeping of the information collected as part of the monitoring is necessary to ensure traceability and foster accountability, and is also a requirement under the EU AI Act for high-risk systems.[415]

Incident Reporting

A necessary governance mechanism is the availability and encouragement of AI incident reporting. The OECD defines an 'AI incident' as an event where an AI system results in individual, environmental or societal harm due to malfunction, misuse of bias.[416] The availability to report an incident fosters trust amongst stakeholders, and creates space to have a risk or impact-based concern addressed.[417] Per the EU AI Act, for high-risk AI systems, incident reporting is a vital mechanism, enabling awareness and communication amongst deployers and end-users. The EU AI Act requires deployers to inform the providers of an AI system, distributors and relevant market surveillance authorities once a serious incident has been identified.[418] It is important that the process is straightforward, evidence is collected and the responsible individuals within the AI lifecycle are identified in order to investigate, address and adapt the system as required. Incident reporting mechanisms also act as a compliance measurement tool, enabling authorities to evaluate the shifting risks and associated compliance requirements, as the AI system continues to evolve once it has been deployed.[419]

Incident reporting can also be employed by the deployers themselves, where categories of internally identified incidents depending on the risk-level must be reported to regional authorities and other stakeholders. These can range from voluntary to mandatory reporting, based on the consequences of said incident, ensuring accountability.[420] The gravity of the incident determines whether it is mandatory to report or not. As an example, the new Brazilian artificial intelligence law indicates requirements for developers and deployers to report serious incidents to competent authorities.[421]

Per the OECD, complaint addressal forums and incident reporting are the first steps in creating a feedback loop, allowing for the communication and recording of any incident where an AI system may have resulted in individual, environmental or societal harm.[422] Such reporting can also help stakeholders such as policymakers assess the harms in light of building governance frameworks. The Ministry of Electronics and Information Technology, Government of India issued the India AI Governance Guidelines, where the creation of a national database for incident reporting has been suggested. This database is intended to act as a source of information for policymakers on real-world risks and harms, and also inform the development of risk frameworks. The recommendations encourage voluntary incident reporting by developers, and creates space for private organisations, sector-specific regulators and individuals to participate.[423]

Feedback Loops

Creating and implementing a feedback loop is key. A feedback loop allows the deployers to collect, analyse and learn from direct user data to operationalise a cycle of continuous improvement. Feedback can also be collected as consumer interactions and reviews, email surveys, etc.[424] Deployers can then engage with the incident reporting knowledge and feedback to maintain and improve the AI system.[425] In the India AI Governance Guidelines, the Ministry of Electronics and Information Technology, Government of India discusses the development of a structured feedback loop through reporting, which can guide policymakers on emerging risks, patterns of harm and their impact.[426]

Human Oversight

The 'human in the loop' mechanism, discussed in the model design and training stage, is also a key governance mechanism to prevent over-reliance on automated systems in the deployment stage. This is particularly significant in critical deployment sectors such as hiring practices, medical diagnostics, or health insurance.[427] Ongoing monitoring and control of an AI system requires human involvement and informed personnel capable of addressing issues that may arise during development and post- deployment of the system.[428] Human decision-making can be pre-defined, with designated tasks such as explicit approval before an AI system takes a certain step, or requiring approvals adhoc, such as intervention during irregular inputs leading to erroneous outputs .[429] The EU AI Act emphasises safeguards for individual rights, with transparency at its core. The EU AI Act mandates human oversight in high-risk AI decision-making, like healthcare, and requires that the personnel are competent, trained and authorised.[430] Human oversight is also required for operationalisation of other principles for safe AI, such as explainability, where the ability to understand and interpret AI systems underlines ethical deployment and post-deployment processes.[431] Techniques such as reinforcement learning from human feedback, are also part of the human oversight mechanism, and are employed to mitigate harmful outputs that may present themselves once an AI system has been deployed.[432]

Post-Deployment Audits

Conducting regular independent third-party security assessments and audits of the AI systems is necessary to determine vulnerabilities. This includes bias and fairness audits to ensure the overall safety of end-users. Auditability refers to the levels of access to information, quality of procedural documentation, and the ability to be systematically examined, analysed and understood.[433] Thoroughly conducted audits can indicate whether the principles of AI safety such as transparency, explainability and accuracy have been fulfilled. Third party audits should not just be permitted, but encouraged, including by civil society organisations, researchers and AI ethicists. The results and feedback from these post-deployment audits work alongside mechanisms such as maintenance and also work towards model retraining, as discussed next.

Model Retraining

Model retraining is an important governance measure - it indicates the willingness of developers and deployers to ensure continuous improvement of their AI system in order to maintain its accuracy and usefulness based on the initial purpose. Adaptive learning algorithms must be incorporated into the AI systems, along with the possibilities of human intervention and updating of data to retrain the model where it is unable to adapt and maintain accuracy. A regular retraining schedule depending on the environment and industry the AI system is deployed into, combined with a 'rolling window' of training - where the most recent data is regularly incorporated to maintain relevance can mitigate risks of the AI system's degradation.[434] Incorporating historical and new data requires a delicate balance between newer but less static trends and long-term patterns that have been previously recognised and verified in order to train the AI system.[435] Such retraining is an operationalised mechanism of a mandatory requirement under the EU AI Act, where providers of high-risk AI systems are required to take corrective actions if they believe the system no longer conforms with its requirements.[436]

iii. Considerations for the Global South

Various elements of heightened risks in the Global South are amplified post-deployment. Previously discussed considerations, such as low infrastructure availability, developing regional legal accountability mechanisms, inadequate reporting tools, etc. continue to present challenges.[437] Imbalanced development and staggered iterations of scientific advancements are unfortunate realities when deploying AI systems in the Global South, and must be approached carefully and intentionally in order to avoid exacerbating social harms.[438]

Historically disproportionate power dynamics continue to manifest even in technological deployment, and require differential application of AI systems, especially in sectors where racial and gendered information can impact the decision-making abilities of an AI system.[439] The sector-specific risks are necessary to highlight, for example, in hiring, where the AI systems risk reproducing systemic racism by relying on historically biased data and opaque algorithmic processes.[440] AI application for these purposes is a sensitive and risky field of application, especially with the multitude of human variables involved.[441] Predictive algorithms that are not adequately sensitised and contextually trained can continue to perpetuate discriminatory tones in studying linguistic patterns.[442] For example, hate speech detection algorithms identifying Black vernacular as 'toxic'.[443] With hate speech detection, there are further challenges in advancing research on South Asian languages, including the limited availability of data, data that is code-mixed, i.e. a mix of regional languages written in roman script and emojis.[444]

Consider the model inversion risks discussed earlier that can lead to biometric data leakage with targeted prompt injections. Such concerns around privacy are further exacerbated due to lower digital literacy and lack of awareness around sharing personally identifiable information online.[445] Users with limited understanding of digital privacy concerns are less likely to have foresight on potential harms they may face.[446] With drift detection, localised data may be of poor quality, non-standardised or scarce; and continuous vigilance as to the quality is complicated when seeking to prevent model degradation. This challenge arises because real-world data from diverse regions often includes variations like regional dialects, inconsistent collection methods, or limited samples that diverge from original training sets, requiring constant adaptation to avoid performance drops.

Models conceptualised in the Global North and embedded into AI systems are typically developed and trained on data from the Global North.[447] Their accuracy, once deployed in the Global South, can be impaired. With limited resources spent on post-deployment monitoring, lack of transparency, limited regional oversight and relaxed compliance and regulatory requirements, errors that may pose significant harms may not be immediately detected, let alone rectified.[448] Post deployment monitoring can be improved by documenting instances of negative impact, with examples on 'toxicity scoring' acting as reverse pedagogy - where impacted end-users can expose an AI system's pitfalls and limitations, eventually allowing for improvement in subsequent updated versions of the technology.[449] The flow of information must not be limited downstream, so to speak. Culturally rich and diverse digital ethics can help mitigate, to a certain extent, issues around lack of awareness amongst both deployers and end-users.[450]

The robustness and institutional resilience of an AI system is vital for continuous performance. Targeted localised bias auditing is necessary to ensure that an AI system is operationally robust based on the market it is deployed into, especially in the Global South. These audits and other monitoring systems must look at local categories like caste, socio-economic class and regional dialects in languages - categories that may not be extensively studied in the Global North. The monitoring can take place through impact assessments and audits, before and after deployment, and involve local communities and civil society organisations. In Brazil, the proposed Artificial Intelligence law also introduces requirements for deployer-led impact assessments for high-risk AI systems such as biometric identification.[451] Scholars have recommended a ‘Decoloniality Impact Assessment’, as an impact assessment methodology that is context-sensitive and evaluates AI systems in relation to the ‘inherent colonial legacies, global power asymmetries and epistemic injustices.’[452] Such assessments focus on questioning how the AI lifecycle and generic risk mitigation measures are not only insufficient in the Global South contexts, but in fact can reinforce structural inequalities, marginalise local knowledge pathways and continue to benefit from pre-existing exploitative systems.[453]

Endnotes:

[1] Dan Hendrycks, Mantas Mazeika and Thomas Woodside, 'An Overview of Catastrophic AI Risks' (arXiv, 21 June 2023) https://arxiv.org/abs/2306.12001 accessed 20 January 2026.

[2] Seth Lazar and Alondra Nelson, 'AI Safety on Whose Terms?' (2023) 381 Science 138 https://doi.org/10.1126/science.adi8982 accessed 20 January 2026.

[3] AI alignment is a sub-domain of AI safety working towards ensuring that the goals and characteristics of AI systems are in harmony with human values and objectives; Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (W W Norton & Company 2020).

[4] Department for Science, Innovation and Technology, *International AI Safety Report 2025* (2025) <https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf> accessed 21 January 2026.

[5] Norbert Wiener, 'Some Moral and Technical Consequences of Automation' (1960)131(3410) Science 1355 https://www.science.org/doi/10.1126/science.131.3410.1355 accessed 20 January 2026.

[6] Max Planck Institute for Biological Cybernetics, 'From Cybernetics to AI: The Pioneering Work of Norbert Wiener' (*Max Planck Neuroscience*, 25 April 2024) https://maxplanckneuroscience.org/from-cybernetics-to-ai-the-pioneering-work-of-norbert-wiener/ accessed 21 January 2026.

[7] Olle Häggström, 'AI Ethics and AI Safety' https://www.math.chalmers.se/~olleh/AIethicsVSAIsafety.pdf accessed 20 January 2026.

[8] This refers to the risk of autonomous goal-setting, where an AI system prioritizes self-generated motives over the objectives given to it by its human creators; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (OUP 2016)

[9] A form of AI capable of performing a vast majority of cognitive tasks emulating human-level efficacy; Arvind Narayanan and Sayash Kapoor, *AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference* (Princeton University Press 2024) https://press.princeton.edu/books/hardcover/9780691249131/ai-snake-oil?srsltid=AfmBOoqsA8dN2CTohNi_t8KIfy_M8Q8oBzKt9AWW44CIA5G1QffXuhVN accessed 1 February 2026.

[10] Shazeda Ahmed et al, 'Building the Epistemic Community of AI Safety' (SSRN, 1 December 2023) https://dx.doi.org/10.2139/ssrn.4641526 accessed 21 January 2026.

[11] Zhiqiang Lin, Huan Sun and Ness Shroff, 'AI Safety vs AI Security: Demystifying the Distinction and Boundaries' (arXiv, 21 June 2025) https://doi.org/10.48550/arXiv.2506.18932 accessed 20 January 2026.

[12] Will Henshall, 'When Might AI Outsmart Us? It Depends Who You Ask' (*TIME*, 20 January 2024) https://time.com/6556168/when-ai-outsmart-humans/ accessed 20 January 2026.

[13] John Bliss, 'Existential Advocacy: Lawyering for AI Safety and the Future of Humanity' (2024) *Georgetown Journal of Legal Ethics* https://www.law.georgetown.edu/legal-ethics-journal/wp-content/uploads/sites/24/2024/07/GT-GJLE240002.pdf accessed 20 January 2026.

[14] Davide Castelvecchi and Benjamin Thompson, '"It Keeps Me Awake at Night": Machine-Learning Pioneer on AI's Threat to Humanity' *Nature* (12 November 2025) https://www.nature.com/articles/d41586-025-03686-1 accessed 20 January 2026.

[15] Tech Desk, 'OpenAI Warns of "Potentially Catastrophic" Risks from Superintelligent AI, Outlines Global Safety Measures' *The*

Indian Express (New Delhi, 11 November 2025) Superintelligent AI risks: OpenAI warns of 'potentially catastrophic' risks from superintelligent AI, outlines global safety measures accessed 20 January 2026.

[16] John Bliss (n 13); Shazeda Ahmed et al, 'Building the Epistemic Community of AI Safety' (2023) SSRN https://dx.doi.org/10.2139/ssrn.4641526 accessed 21 January 2026 Shazeda Ahmed (n 10).

[17] John Bliss (n 13).

[18] Seth Lazar and Alondra Nelson, 'AI Safety on Whose Terms?' (2023) 381 Science 6654 https://www.science.org/doi/10.1126/science.adi8982 accessed 21 January 2026

[19] Timnit Gebru, 'Effective Altruism Is Pushing a Dangerous Brand of "AI Safety"' (*WIRED*, 30 November 2022) https://www.wired.com/story/effective-altruism-artificial-intelligence-sam-bankman-fried/ accessed 20 January 2026.

[20] Techno-utopianism supports the belief that technological competence can effectively mitigate and provide solutions to most of the human problems, including elimination of poverty and controlling diseases.

[21] Mark MacCarthy, 'AI Safety Met the Guillotine in Paris. Good Riddance' (*Tech Policy Press*, 24 February 2025) https://www.techpolicy.press/ai-safety-met-the-guillotine-in-paris-good-riddance/ accessed 20 January 2026.

[22] Ronen Bar, 'AI Moral Alignment: The Most Important Goal of Our Generation' (Effective Altruism Forum, 26 March 2025) https://forum.effectivealtruism.org/posts/4LimpA4pyLemxN4BF/ai-moral-alignment-the-most-important-goal-of-our-generation accessed 20 January 2026.

[23] Brian Christian (n 3).

[24] Dan Hendrycks, Nicholas Carlini, John Schulman and Jacob Steinhardt, 'Unsolved Problems in ML Safety' (arXiv, 16 June 2022) arXiv:2109.13916v5 https://arxiv.org/abs/2109.13916 accessed 20 January 2026.

[25] Mark MacCarthy, 'Are AI Existential Risks Real – and What Should We Do About Them?' (Brookings Institution, 11 July 2025) https://www.brookings.edu/articles/are-ai-existential-risks-real-and-what-should-we-do-about-them/ accessed 20 January 2026.

[26] In this context, achieving a reward corresponds to an agent having completed the task as directed by the developer; Leonard Dung, 'Current cases of AI misalignment and their implications for future risks' (2023) Synthese 202, 138 < https://doi.org/10.1007/s11229-023-04367-0 > accessed 31 January 2026.

[27] Brian Christian (n3).

[28] David Leslie, '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, June 2019) https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf accessed 20 January 2026.

[29] Gautam Misra and Supratik Mitra, 'Navigating AI Safety: A Socio-Technical and Risk-Based Approach to Policy Design' (*Tech Policy Press*, 19 December 2024) https://www.techpolicy.press/navigating-ai-safety-a-socio-technical-and-risk-based-approach-to-policy-design/ accessed 20 January 2026.

[30] Wissam Salhab et al, 'A Systematic Literature Review on AI Safety: Identifying Trends, Challenges and Future Directions' (2024) 12 IEEE Access 131762-131784 https://ieeexplore.ieee.org/document/10630784 accessed 20 January 2026.

[31] Mateusz Dolata, Stefan Feuerriegel and Gerhard Schwabe, 'A Sociotechnical View of Algorithmic Fairness' (arXiv Research Paper, 27 Sept 2021) arXiv:2110.09253 https://arxiv.org/abs/2110.09253 accessed 20 January 2026.

[32] Dan Hendrycks et al, 'Unsolved Problems in ML Safety' (arXiv Research Paper, 16 June 2022) arXiv:2109.13916v5 https://arxiv.org/abs/2109.13916 accessed 20 January 2026.

[33] Wissam Salhab (n 30).

[34] ibid.

[35] Renee Shelby et al, 'Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction' (arXiv Research Paper, October 2022) arXiv:2210.05791 https://arxiv.org/abs/2210.05791 accessed 20 January 2026.

[36] Samuel T Segun and others, 'Toward an African Agenda for AI Safety' (arXiv Research Paper, 12 August 2025) arXiv:2508.13179v1 https://arxiv.org/abs/2508.13179 accessed 20 January 2026.

[37] Renee Shelby (n 35).

[38] Peter Henderson and others, *Foundation Models and Copyright Questions* (HAI Policy & Society, November 2023) https://hai-production.s3.amazonaws.com/files/2023-11/Foundation-Models-Copyright.pdf accessed 20 January 2026.

[39] Nandana Sengupta and others, 'A Global South perspective for ethical algorithms and the State' (2023) *Nature Machine Intelligence* 5, 184-186 https://www.nature.com/articles/s42256-023-00621-9 accessed 20 January 2026.

[40] Sanchaita Hazra and others, AI Safety Should Prioritize the Future of Work (arXiv, 16 April 2025) arXiv:2504.13959 https://arxiv.org/abs/2504.13959 accessed 20 January 2026.

[41] Madhumita Murgia, *Code Dependent: Living in the Shadow of AI* (Picador 2024).

[42] Transformers are deep learning models with a neural network architecture employed for analysing sequential data and used in fields such as Natural Language Processing and Computer Vision; Tianyang Lin et al, 'A Survey of Transformers' (arXiv, 15 June 2021) arXiv:2106.04554v2 [cs.LG] https://arxiv.org/abs/2106.04554 accessed 20 January 2026.

[43] Emily M Bender and others, 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?' in *Proceedings of the 2021 ACM Conference on Fairness, Accountability and Transparency* (Association for Computing Machinery 2021)

610-623 https://doi.org/10.1145/3442188.3445922 accessed 20 January 2026.

[44] Emma Strubell and others, 'Energy and Policy Considerations for Deep Learning in NLP' (2019) in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* 3645-3650.

[45] Paul Mozur and others, 'From Mexico to Ireland, Fury Mounts Over a Global A.I. Frenzy' *The New York Times* (20 October 2025) https://www.nytimes.com/2025/10/20/technology/ai-data-center-backlash-mexico-ireland.html accessed 20 January 2026.

[46] Cecilia Marrinan, *Data Center Boom Risks Health of Already Vulnerable Communities* (TechPolicy.Press, 12 June 2025) https://www.techpolicy.press/data-center-boom-risks-health-of-already-vulnerable-communities/ accessed 20 January 2026.

[47] Arvind Narayanan and Sayash Kapoor, 'AI Safety Is Not a Model Property' (AI as Normal Technology, 12 March 2024) https://www.normaltech.ai/p/ai-safety-is-not-a-model-property accessed 20 January 2026.

[48] Roel I J Dobbe, 'System Safety and Artificial Intelligence' (arXiv, 18 February 2022) arXiv:2202.09292v1 [eess.SY] https://arxiv.org/pdf/2202.09292 accessed 20 January 2026.

[49] Brian J Chen and Jacob Metcalf, 'A Sociotechnical Approach to AI Policy' (Data & Society, 28 May 2024) https://datasociety.net/library/a-sociotechnical-approach-to-ai-policy/ accessed 20 January 2026.

[50] Seth Lazar and Alondra Nelson (n 3).

[51] Inioluwa Deborah Raji and Roel Dobbe, 'Concrete Problems in AI Safety, Revisited' (arXiv, 18 December 2023) arXiv:2401.10899 [cs.CY] https://arxiv.org/abs/2401.10899 accessed 20 January 2026.

[52] Jacqueline Harding and Cameron Domenico Kirk-Giannini, 'What Is AI Safety? What Do We Want It to Be?' (arXiv, 5 May 2025) arXiv:2505.02313 [cs.CY] https://arxiv.org/abs/2505.02313 accessed 20 January 2026.

[53] ibid.

[54] Dario Amodei and others, 'Concrete Problems in AI Safety' (arXiv, 21 June 2016) arXiv:1606.06565 [cs.AI] https://arxiv.org/pdf/1606.06565 accessed 20 January 2026.

[55] Distribution shift refers to the performance issues arising out of the mismatch between the training data on which the model was trained and what the model encounters post deployment.

[56] Jacqueline Harding and Cameron Domenico KirkG-Giannini (n 52).

[57] Victor Zhenyi Wang, 'Contesting AI Safety' (*TechPolicy.Press,* 12 September 2024) https://www.techpolicy.press/contesting-ai-safety/ accessed 20 January 2026.

[58] Jacqueline Harding and Cameron Domenico Kirk-Giannini (n 52).

[59] Tim G J Rudner and Helen Toner, 'Key Concepts in AI Safety: An Overview' (Center for Security and Emerging Technology, March 2021) https://doi.org/10.51593/20190040 accessed 20 January 2026.

[60] Dan Hendrycks and others, 'Unsolved Problems in ML Safety' (arXiv, 28 September 2021) arXiv:2109.13916 [cs.LG] https://arxiv.org/abs/2109.13916 accessed 20 January 2026

[61] Inioluwa Deborah Raji and Roel Dobbe, 'Concrete Problems in AI Safety, Revisited' (arXiv, 18 December 2023) arXiv:2401.10899 *[cs.CY]* https://arxiv.org/abs/2401.10899 accessed 20 January 2026.

[62] David Leslie, '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, June 2019) https://www.turing.ac.uk/sites/default/files/2019-06/understanding_artificial_intelligence_ethics_and_safety.pdf accessed 20 January 2026.

[63] Shazeda Ahmed , 'Building the Epistemic Community of AI Safety' (2023) SSRN https://dx.doi.org/10.2139/ssrn.4641526 accessed 21 January 2026.

[64] Brian Christian (n 3)

[65] ibid.

[66] Zhiqiang Lin, Huan Sun and Ness Shroff, *AI Safety vs. AI Security: Demystifying the Distinction and Boundaries* (arXiv, 21 June 2025) arXiv:2506.18932 [cs.CY] https://arxiv.org/abs/2506.18932 accessed 20 January 2026.

[67] ibid.

[68] Jacqueline Harding and Cameron Domenico Kirk-Giannini (n 52)

[69] Olle Häggström, 'On the Troubled Relation Between AI Ethics and AI Safety' (27 June 2024) https://www.math.chalmers.se/~olleh/AIethicsVSAIsafety.pdf accessed 20 January 2026.

[70] Billy Perrigo, 'U.K.'s AI Safety Summit Ends With Limited, but Meaningful, Progress' (*TIME*, 2 November 2023) https://time.com/6321502/uk-ai-safety-summit/ accessed 20 January 2026.

[71] Department for Science, Innovation and Technology, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023' *(GOV.UK,* 13 February 2025*)* https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 accessed 22 January 2026

[72] The UK AI Safety Institute was renamed as the AI Security Institute in February 2025.

[73] Gregory C Allen and Georgia Adamson, 'The AI Safety Institute International Network: Next Steps and Recommendations' (*Center for Strategic and International Studies*, 30 October 2024) https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations accessed 20 January 2026.

[74] Deepak P, 'AI Safety: Necessary, but Insufficient and Possibly Problematic' *AI & Society* 40(2) 1143-1145 (23 March 2024)

https://doi.org/10.1007/s00146-024-01899-y accessed 20 January 2026.

[75] Mark MacCarthy, 'Are AI Existential Risks Real – and What Should We Do About Them?' (Brookings Institution, 11 July 2025) https://www.brookings.edu/articles/are-ai-existential-risks-real-and-what-should-we-do-about-them/ accessed 20 January 2026

[76] 'The Paris Summit: Au Revoir, Global AI Safety?'

<https://www.epc.eu/publication/The-Paris-Summit-Au-Revoir-global-AI-Safety-61ea68//> accessed 28 January 2026

[77] 'The Wiretap: Trump Says Bye To The AI Safety Institute' <https://www.forbes.com/sites/thomasbrewster/2025/06/03/the-wiretap-trump-says-goodbye-to-the-ai-safety-institute/> accessed 28 January 2026

[78] Department for Science, Innovation and Technology and The Rt Hon Peter Kyle MP, 'Tackling AI Security Risks to Unleash Growth and Deliver Plan for Change' (*GOV.UK*, 14 February 2025) https://www.gov.uk/government/news/tackling-ai-security-risks-to-unleash-growth-and-deliver-plan-for-change accessed 20 January 2026.

[79] Tom Bristow, 'Britain dances to JD Vance's tune as it renames AI institute' (*Politico*, 14 February 2025) https://www.politico.eu/article/jd-vance-britain-ai-safety-institute-aisi-security/ accessed 22 January 2026.

[80] Danni Yu, Hannah Rosenfeld and Abhishek Gupta, 'The 'AI divide' between the Global North and Global South' (World Economic Forum, January 2023) https://www.weforum.org/stories/2023/01/davos23-ai-divide-global-north-global-south/ accessed 20 January 2026.

[81] Amba Kak, "The Global South is everywhere, but also always somewhere": National Policy Narratives & AI Justice (AI Now Institute, New York University, 2020) https://doi.org/10.1145/3375627.3375859 accessed 20 January 2026.

[82] Vidya Subramanian, Joanne D'Cunha and Angelina Dash, 'Exploring AISIs for the Global South' (Centre for Communication Governance, National Law University Delhi 2025) https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/exploring-aisis-for-the-global-south-805.pdf accessed 22 January 2026.

[83] 'What Is the AI Life Cycle?' (*Data Science PM*, 19 November 2024) https://datasciencepm.com/what-is-the-ai-life-cycle/ accessed 20 January 2026.

[84] Kazuaki Ishizaki, 'AI Model Lifecycle Management: Overview' (*IBM*) https://www.ibm.com/think/topics/ai-model-lifecycle-management accessed 20 January 2026.

[85] Daswin De Silva and Damminda Alahakoon, 'An Artificial Intelligence Life Cycle: From Conception to Production' (2022) 3(6) *Patterns* 100489 https://doi.org/10.1016/j.patter.2022.100489 accessed 20 January 2026.

[86] HM, 'Roles of Key Actors in the Lifecycle of Agentic AI Systems' (HM, 3 April 2024) https://hmstrategy.com/roles-of-key-actors-in-the-lifecycle-of-agentic-ai-systems/ accessed 20 January 2026.

[87] Katravath Rahul, 'AI Architect – Role, Responsibilities, Skills, Future' (GeeksforGeeks, 23 July 2025) https://www.geeksforgeeks.org/artificial-intelligence/ai-architect-role-responsibilities-skills-future/ accessed 22 January 2026.

[88] Katie Hewson, 'The roles of the provider and deployer in AI systems and models' (Stephenson Harwood, 12 September 2024) https://www.stephensonharwood.com/insights/the-roles-of-the-provider-and-deployer-in-ai-systems-and-models/ accessed 22 January 2026.

[89] Lumenova AI, 'AI Accountability: Stakeholders in Responsible AI Practices' (Lumenova AI Blog, 10 September 2024) <https://www.lumenova.ai/blog/ai-accountability-stakeholders-in-responsible-ai-practices> accessed 20 January 2026.

[90] ibid.

[91] Daswin De Silva and Damminda Alahakoon (n85)

[92] Giovanni Sartor et al, 'The impact of the General Data Protection Regulation (GDPR) on artificial intelligence' (European Parliament, June 2020) https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf accessed 19 January 2026; Chapman University Artificial Intelligence Hub, 'Bias in AI' (2025) https://www.chapman.edu/ai/bias-in-ai.aspx accessed 19 January 2026.

[93] Nathalie A Smuha, 'Beyond the individual: governing AI's societal harm' (2021) 10(3) *Internet Policy Review* https://policyreview.info/articles/analysis/beyond-individual-governing-ais-societal-harm accessed 19 January 2026.

[94] 'What is Sandboxing?' (Palo Alto Networks) https://www.paloaltonetworks.in/cyberpedia/sandboxing#sandboxing accessed 22 January 2026; 'AI lifecycle risk management' (VerifyWise) https://verifywise.ai/lexicon/ai-lifecycle-risk-management accessed 22 January 2026.

[95] Governance Framework: For Security Leaders' (Strobes Security, 20 June 2025) https://strobes.co/blog/ai-governance-framework-for-security-leaders/ accessed 22 January 2026.

[96] Chen Chen et al, 'Trustworthy, Responsible, and Safe AI: A Comprehensive Architectural Framework for AI Safety with Challenges and Mitigations' (arXiv Preprint, 23 August 2024) https://arxiv.org/abs/2408.12935 accessed 22 January 2026.

[97] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1, 2023) https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf accessed 19 January 2026.

[98] 'OECD AI Principle 1.4: Robustness, security and safety' (OECD.AI) https://oecd.ai/en/dashboards/ai-principles/P8 accessed 22 January 2026

[99] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 9.

[100] California SB 53 (2025–2026) https://legiscan.com/CA/text/SB53/id/3270002 accessed 22 January 2026.

[101] The Act also introduces the term "Critical safety incident", which translates to model behavior culminating in or nearly causing a fatality, hazards arising out of a catastrophic risk or a complete system loss, requiring immediate reporting; California Senate Bill 53.

[102] Roman Yampolskiy and Joshua Fox, 'Safety Engineering for Artificial General Intelligence' [2012] Topoi https://doi.org/10.1007/s11245-012-9128-9 accessed 2 February 2026

[103] Dan Hendrycks, 'Safe Design Principles | AI Safety, Ethics, and Society Textbook' https://www.aisafetybook.com/textbook/safe-design-principles accessed 2 February 2026

[104] Wissam Salhab (n 30).

[105] Roel Dobbe, 'AI Safety is Stuck in Technical Terms -- A System Safety Response to the International AI Safety Report' (arxiv, 5 February 2025) arXiv:2503.04743 [cs.CY] https://arxiv.org/abs/2503.04743 accessed 20 January 2026

[106] Dan Hendrycks, 'Introduction to AI Safety, Ethics and Society' (Taylor & Francis 2024) ch 4.4 https://www.aisafetybook.com/textbook/safe-design-principles accessed 22 January 2026.

[107] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)

[108] Ibid.

[109] Ibid.

[110] Ibid.

[111] National Academies of Sciences, Engineering, and Medicine, 'Widening Participation in the Design, Development, and Deployment of AI Tools' in *Human and Organizational Factors in AI Risk Management: Proceedings of a Workshop* (The National Academies Press 2025) https://doi.org/10.17226/29046 accessed 19 January 2026.

[112] In machine learning, failure modes can include both intentional failures (where the failure arises due to adversarial attacks by bad actors) as well as unintentional failures where the failure arises when the AI system produces unsafe outcomes which may be formally correct. *See* Ram Shankar Siva Kumar et al, 'Failure Modes in Machine Learning' (Microsoft, 12 March 2025) https://learn.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning accessed 19 January 2026.

[113] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)

[114] Ibid.

[115] Ibid.

[116] Ibid.

[117] Ibid.

[118] 'Widening Participation in the Design, Development, and Deployment of AI Tools' (n111)

[119] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)

[120] Ibid.

[121] Ibid.

[122] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689 art 9.

[123] Ibid.

[124] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689 art 9(2)(a).

[125] European Data Protection Supervisor, 'Guidelines on the management of AI risks' (Guidance, 11 November 2025) https://www.edps.europa.eu/system/files/2025-11/2025-11-11_ai_risks_management_guidance_en.pdf accessed 22 January 2026.

[126] Ibid.

[127] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689 art 9(2)(b).

[128] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689 art 9(2)(d).

[129] European Data Protection Supervisor, 'Guidelines on the management of AI risks' (Guidance, 11 November 2025) https://www.edps.europa.eu/system/files/2025-11/2025-11-11_ai_risks_management_guidance_en.pdf accessed 22 January 2026.

[130] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)

[131] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 16.

[132] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 26.

[133] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 23.

[134] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 24.

[135] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf accessed 19 January 2026.

[136] ibid.

[137] Sjoerd Beugelsdijk and others, 'Cultural Distance and Firm Internationalization: A Meta-Analytical Review and Theoretical Implications' (2018) 44(1) Journal of Management 89 https://journals.sagepub.com/doi/10.1177/0149206317729027 accessed 6th February 2026; Paula Helm and others, 'Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice' (arXiv PrePrint 2307.13714, 25 July 2023) https://arxiv.org/abs/2307.13714 accessed 6 February 2026.

[138] Paula Helm and others, 'Diversity and Language Technology: How Techno-Linguistic Bias Can Cause Epistemic Injustice' (2023) *arXiv:2307.13714* https://arxiv.org/abs/2307.13714 accessed 3 February 2026.

[139] Freya Smith and others, 'Codesigning AI with End-Users: An AI Literacy Toolkit for Nontechnical Audiences' (2025) 37(5) *Interacting with Computers* 444 https://doi.org/10.1093/iwc/iwae029 accessed 3 February 2026.

[140] Aníbal Monasterio Astobiza and others, 'Ethical Governance of AI in the Global South: A Human Rights Approach to Responsible Use of AI' (2022) 81 Proceedings 136 https://doi.org/10.3390/proceedings2022081136 accessed 19 January 2026.

[141] Hana Mesquita, Marina Garrote and Rafael Zanatta, ‘Regulating Artificial Intelligence in Brazil: the contributions of critical social theory to rethink principles’ (2024) 2024 Technology and Regulation 73 https://techreg.org/article/view/13251 accessed 19 January 2026; Damian Eke, Ricardo Chavarriaga and Bernd Stahl, ‘Decoloniality Impact Assessment for AI’ (2025) *AI & Society* https://doi.org/10.1007/s00146-025-02649-4 accessed 20 January 2026.

[142] Aníbal Monasterio Astobiza et al (n140)

[143] ibid.

[144] 'What is the AI Life Cycle?' (*Data Science PM*, 19 November 2024) https://www.datascience-pm.com/ai-lifecycle/ accessed 15 January 2026.

[145] Microsoft, ‘What is Data Discovery?’ (*Microsoft*) https://www.microsoft.com/en-us/security/business/security-101/what-is-data-discovery accessed 15 January 2026 and Palo Alto Networks, ‘What Is Data Discovery?’ (*Palo Alto Networks*) https://www.paloaltonetworks.in/cyberpedia/data-discovery accessed 15 January 2026.

[146] Rina Diane Caballar and Cole Stryker, 'Synthetic Data Generation' (IBM) https://www.ibm.com/think/insights/synthetic-data-generation accessed 15 January 2026.

[147] Some of the techniques involved in data preparation or the ‘pre-processing’ stage include data curation, entity resolution, and joining datasets. *See* Yuji Roh, Geon Heo and Steven Euijong Whang, ‘A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective’ (2021) 33 *IEEE Transactions on Knowledge and Data Engineering* 1328 https://doi.org/10.1109/TKDE.2020.3027479 accessed 20 January 2026.

[148] Sabrina Pochaba and others, 'Make Your Dataset Representative: Fill Data Gaps with Active Measurements' in 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (IEEE 2024) https://www.salzburgresearch.at/wp-

content/uploads/2025/03/MakeYourDatasetRepresentative.pdf accessed 19 January 2026.

[149] Louie Kangeter, *A Lifecycle Approach to AI Risk Reduction* (Institute for Security and Technology, October 2024) https://securityandtechnology.org/wp-content/uploads/2024/10/A-Lifecycle-Approach-to-AI-Risk-Reduction.pdf accessed 19 January 2026.; IBM, 'What is Data Augmentation?' (*IBM*) https://www.ibm.com/think/topics/data-augmentation accessed 19 January 2026.

[150] IBM, 'What Is Data Labeling?' (IBM) https://www.ibm.com/think/topics/data-labeling accessed 19 January 2026.

[151] Yuji Roh, Geon Heo and Steven Euijong Whang, 'A Survey on Data Collection for Machine Learning: A Big Data-AI Integration Perspective' (2021) 33 *IEEE Transactions on Knowledge and Data Engineering* 1328 https://doi.org/10.1109/TKDE.2020.3027479 accessed 20 January 2026.

[152] Sabrina Pochaba and others, 'Make Your Dataset Representative: Fill Data Gaps with Active Measurements' in 2024 20th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob) (IEEE 2024) https://www.salzburgresearch.at/wp-content/uploads/2025/03/MakeYourDatasetRepresentative.pdf accessed 19 January 2026.

[153] Rina Diane Caballar, '10 AI Dangers and Risks and How to Manage Them' (*IBM*) https://www.ibm.com/think/insights/10-ai-dangers-and-risks-and-how-to-manage-them accessed 19 January 2026.

[154] Karl Werder, Balasubramaniam Ramesh and Sophia Zhang, 'Establishing Data Provenance for Responsible Artificial Intelligence Systems' (2022) 13 ACM Transactions on Management Information Systems 1 https://dl.acm.org/doi/10.1145/3503488 accessed 18 January 2026.

[155] Ibid.

[156] Qingmei Joy Feng and others, 'The Risks of Artificial Intelligence: A Narrative Review and Ethical Reflection from an Oral Medicine Group' (2025) 31 Oral Diseases 348 https://pmc.ncbi.nlm.nih.gov/articles/PMC11976142/ accessed 18 January 2026.

[157] Camille Thibault and others, 'A Guide to Misinformation Detection Data and Evaluation' in Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Association for Computing Machinery 2025) https://doi.org/10.1145/3711896.3737437 accessed 20 January 2026.

[158] Daniel J Solove and Woodrow Hartzog, 'The Great Scrape: The Clash Between Scraping and Privacy' (2025) 113 Cal L Rev 1521 https://www.californialawreview.org/print/great-scrape accessed 19 January 2026

[159] Katharine Miller, 'Privacy in an AI Era: How Do We Protect Our Personal Information?' (*Stanford HAI*, 18 March 2024) https://hai.stanford.edu/news/privacy-ai-era-how-do-we-protect-our-personal-information accessed 19 January 2026.

[160] Information Commissioner's Office, 'What about fairness, bias and discrimination?' (Guidance on AI and data protection, 22 May 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/ accessed 22 January 2026.

[161] Charles Kinyua Gitonga, Dennis Murithi and Edna Chebet, 'Mitigating Demographic Bias in ImageNet: A Comprehensive Analysis of Disparities and Fairness in Deep Learning Models' (2025) 4 European Journal of Artificial Intelligence and Machine Learning 15 https://eu-opensci.org/index.php/ejai/article/view/1051 accessed 19 January 2026.

[162] Anshika Sharma, Shalli Rani and Mohammad Shabaz, 'A comprehensive review of explainable AI in cybersecurity: Decoding the black box' (2025) 11(6) *ICT Express* 1200 https://doi.org/10.1016/j.icte.2025.10.004 accessed 19 January 2026.

[163] Information Commissioner's Office, 'What about fairness, bias and discrimination?' (Guidance on AI and data protection, 22 May 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/ accessed 22 January 2026.

[164] Amazon Web Services, 'What is Data Labeling?' (*AWS*) https://aws.amazon.com/what-is/data-labeling/ accessed 19 January 2026.

[165] 'What about fairness, bias and discrimination?' (n163); Swati Punia et al, *Emerging Trends in Data Governance* (National Law University Delhi Press, 2022) ISBN 978-93-84272-33-3 https://ccgdelhi.s3.ap-south-1.amazonaws.com/uploads/ccg-edited-volume-emerging-trends-in-data-governance-343.pdf accessed 3 February 2026.

[166] Darlene Barker and others, 'Ethical Considerations in Emotion Recognition Research' (2026) 7(2) *Psychology International* https://www.mdpi.com/2813-9844/7/2/43 accessed 1 February 2026.

[167] Information Commissioner's Office, 'Guidance on AI and Data Protection: Annex A: Fairness in the AI Lifecycle' (15 March 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/annex-a-fairness-in-the-ai-lifecycle/ accessed 19 January 2026.

[168] Shayne Longpre and others, 'A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity' (2023) arXivLabs https://doi.org/10.48550/arXiv.2305.13169 accessed 19 January 2026.

[169] Tim Mucci, 'What is Data Provenance?' (*IBM Think*, 2025) https://www.ibm.com/think/topics/data-provenance accessed 19 January 2026.

[170] Ibid.

[171] Mark Sharron, 'Demonstrating Compliance With EU AI Act Article 10 Data and Data Governance Using ISO 42001 Governance Controls' (*ISMS.online*, 18 September 2025) https://www.isms.online/iso-42001/eu-ai-act/article-10/ accessed 19 January 2026.

[172] Gebru and others proposed that AI datasets should be accompanied by datasheets, similar to those used in the electronics industry, to document the motivation, composition, collection process, and recommended uses of the dataset. They contend that such documentation can enable transparency and accountability between dataset creators and dataset consumers. *See* Timnit Gebru and others, 'Datasheets for Datasets' (2021) 64 Communications of the ACM 86 https://arxiv.org/abs/1803.09010 accessed 19 January 2026.

[173] Bender and Friedman recommend providing data statements for datasets, through which developers and users would gain deeper insights into the deployment of AI systems and any potential biases that may be reflected in the outputs generated by the AI system. *See* Emily M Bender and Batya Friedman, 'Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science' (2018) 6 Transactions of the Association for Computational Linguistics 587 https://aclanthology.org/Q18-1041.pdf accessed 19 January 2026.

[174] Pushkarna and others propose data cards as a mechanism for documentation relating to datasets, which would summarise and explain key information about a dataset, including the sources and methods of data collection and annotation, and other details regarding the intended use cases of the AI system, etc. *See* Mahima Pushkarna, Andrew Zaldivar and Oddur Kjartansson, 'Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI' in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (ACM 2022) 1776 https://arxiv.org/abs/2204.01075 accessed 19 January 2026.

[175] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 53(1)(d).

[176] *Artificial Intelligence Risk Management Framework (AI RMF 1.0)* (n97)

[177] *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems* (Innovation, Science and Economic Development Canada, September 2023) https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems accessed 1 February 2026.

[178] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 5 and art 9.

[179] Ibid.

[180] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 5.

[181] Ibid.

[182] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 4 (7).

[183] Digital Personal Data Protection Act 2023, s 2(i).

[184] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 25; United Nations Development Programme (UNDP), Drafting Data Protection Legislation: A Study of Regional Frameworks

(March 2023) https://www.undp.org/sites/g/files/zskgke326/files/2023-04/UNDP%20Drafting%20Data%20Protection%20Legislation%20March%202023.pdf accessed 3 February 2026.

[185] These Guidelines are to be jointly developed by the European Data Protection Board and the European Commission. *See* Caitlin Andrews, 'Joint Guidelines on GDPR-AI Act Interplay to Come Soon, EDPS Says' (*IAPP*, 20 November 2025). https://iapp.org/news/a/edps-to-issue-joint-guidance-on-gdpr-ai-act-interplay-with-european-commission accessed 19 January 2026.

[186] Caitlin Andrews, 'Joint guidelines on GDPR-AI Act interplay to come soon, EDPS says' (*IAPP*, 20 November 2025) https://iapp.org/news/a/edps-to-issue-joint-guidance-on-gdpr-ai-act-interplay-with-european-commission accessed 19 January 2026.

[187] Giovanni Sartor and others, 'The impact of the General Data Protection Regulation (GDPR) on artificial intelligence' (European Parliament, June 2020) https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf accessed 19 January 2026.

[188] Ibid.

[189] Ibid.

[190] Ibid.

[191] It is imperative to note however, that traditional mechanisms of user consent often give rise to limitations such as consent fatigue and information asymmetry. Organisations like Spawning AI are attempting to develop the foundational infrastructure that enables a consent framework for AI data. This is being achieved through processes like collecting opt-in and opt-out information from data creators, and organising this information into searchable databases like their 'Do Not Train' Registry. *See* Shayne Longpre et al, 'Position: Data Authenticity, Consent, & Provenance for AI are all broken: what will it take to fix them?' (2024) 235 Proceedings of Machine Learning Research 32711

https://proceedings.mlr.press/v235/longpre24b.html accessed 19 January 2026.

[192] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 5 and art 9.

[193] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1, art 25.

[194] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 10.

[195] Information Commissioner's Office and The Alan Turing Institute, 'What Goes Into an Explanation?' in *Explaining Decisions Made with AI* (20 May 2020) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-1-the-basics-of-explaining-ai/what-goes-into-an-explanation/ accessed 19 January 2026.

[196] It represents the underlying population and the phenomenon you are modelling. *See* Information Commissioner's Office, 'Explaining decisions made with AI: Task 2: Collect and prepare your data' (ICO, 2020) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/part-2-explaining-ai-in-practice/task-2-collect/ accessed 19 January 2026.

[197] 'What Goes Into an Explanation? (n195)

[198] Information Commissioner's Office, 'Annex A: Fairness in the AI lifecycle' (15 March 2023) https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/annex-a-fairness-in-the-ai-lifecycle/ accessed 19 January 2026.

[199] Ibid.

[200] Ibid.

[201] Arnav Arora and others, 'The Uli Dataset: An Exercise in Experience Led Annotation of oGBV' in *Proceedings of the 8th Workshop on Online Abuse and Harms* (Association for Computational Linguistics 2024) 212. https://arxiv.org/pdf/2311.09086 accessed on 19 January 2025.

[202] 'Annex A: Fairness in the AI lifecycle' (n198)

[203] Matthew G Hanna and others, 'Ethical and Bias Considerations in Artificial Intelligence/Machine Learning' (2025) 38(3) *Modern Pathology* 100686 https://doi.org/10.1016/j.modpat.2024.100686 accessed 19 January 2026.

[204] Yoonyoung Park, 'IBM researchers investigate ways to help reduce bias in healthcare AI' (*IBM Research Blog*, 15 April 2021) https://research.ibm.com/blog/ibm-reduce-bias-in-healthcare-ai accessed 19 January 2026.

[205] Matthew G Hanna and others (n203)

[206] 'Annex A: Fairness in the AI lifecycle' (n198)

[207] 'Annex A: Fairness in the AI lifecycle' (n198)

[208] Ames Dhai and others, 'Understanding and processing informed consent during data-intensive health research in sub-Saharan Africa: challenges and opportunities from a multilingual perspective' (2024) 20(2) *Research Ethics* 145 https://pmc.ncbi.nlm.nih.gov/articles/PMC12346138/ accessed 19 January 2026.

[209] Tracey Li and others, 'Operationalizing Health Data Governance for AI Innovation in Low-Resource Government Health Systems: A Practical Implementation Perspective from Zanzibar' (2024) 6 Data & Policy e63 https://doi.org/10.1017/dap.2024.65, accessed January 2026.

[210] Ibid.

[211] Abeba Birhane, 'Algorithmic Colonization of Africa' (2020) 3 *Global Perspectives* 1200 10.1093/oso/9780192865366.003.0016 accessed 19 January 2026.

[212] SoberanIA, 'SoberanIA: A Inteligência Artificial que entende o Brasil' (*SoberanIA*) https://soberania.ai/ accessed 19 January 2026.

[213] Rachel Joseph, 'The Role of Linguistic Diversity in AI Development: Challenges and Opportunities in NLP' (2024) 44 Library Progress International 26106 https://doi.org/10.48165/bapas.2024.44.2.1 accessed 19 January 2026

[214] Due to diversity in structures of languages (for eg. word orders and code-switching, which is less prevalent in English as opposed to low-resource languages in the Global South), AI models for content moderation trained in English are often unable to detect words that connote sexual harassment. This can impact the detection of online harms in the Global South. *See*, Farhana Shahid, Mona Elswah and Aditya Vashistha, 'Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages' (2025) 8 Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 2331 https://doi.org/10.1609/aies.v8i3.36719 accessed 29 January 2026.

[215] DigiProd Pass, 'Supply Chain Traceability: Challenges and Proven Solution' (*DigiProd Pass*, 4 July 2025) https://digiprodpass.com/blogs/supply-chain-traceability-challenges-and-proven-solution accessed 19 January 2026.

[216] Ingmar Weber and others, 'Non-traditional data sources: providing insights into sustainable development' (2021) 64(4) *Communications of the ACM* 88 https://doi.org/10.1145/3447739 accessed 19 January 2026.

[217] Ibid.

[218] Genevieve Smith, 'How to Make AI Equitable in the Global South' (*Stanford Social Innovation Review*, 3 April 2024)

https://ssir.org/articles/entry/equitable-ai-in-the-global-south accessed 19 January 2026.

[219] Palo Alto Networks, 'What Is the AI Development Lifecycle?' (Cyberpedia) https://www.paloaltonetworks.in/cyberpedia/ai-development-lifecycle accessed 22 January 2026.

[220] Ibid.

[221] Supervised learning refers to training AI models using labelled datasets, whereas unsupervised learning involves training an AI model through analysing and clustering unlabelled datasets using machine learning algorithms. Reinforcement learning refers to training an AI model through a method of trial and error, based on its interactions with its environment. See Jacob Murel and Eda Kavlakoglu, *What is reinforcement learning?* (IBM Think, updated 2025) https://www.ibm.com/think/topics/reinforcement-learning accessed 7 February 2026.

[222]'What Is the AI Development Lifecycle?' (Cyberpedia) (n219)

[223] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, recital 133.

[224] Amplix, 'Deepfake AI Is Becoming A Rising Digital Menace' (*Amplix*, 3 July 2025) https://amplix.com/insights/deepfake-ai-is-becoming-a-rising-digital-menace/ accessed 19 January 2026.

[225] North Dakota Domestic & Sexual Violence Coalition, 'Generative AI & Sexually Explicit Deepfakes' (NDDSVC, 9 April 2025) https://nddsvc.org/generative-ai-sexually-explicit-deepfakes accessed 19 January 2026.

[226] BBC, 'Title Unknown' *BBC News (World Asia – India)* https://www.bbc.com/news/world-asia-india-67305557 accessed 1 February 2026.

[227] Don Hummer and Donald J Rebovich, *Seeing Isn't Believing: Addressing the Societal Impact of Deepfakes in Low-Tech Environments* (preprint, arXiv:2508.16618v1, 2025)

https://arxiv.org/html/2508.16618v1 accessed 19 January 2025. Hummer and J. Rebovich, 2023)

[228] Azmine Toushik Wasi and others**,** ‘Seeing Isn’t Believing: Addressing the Societal Impact of Deepfakes in Low-Tech Environments’ (2025) *arXiv* **2508.16618v1** https://arxiv.org/html/2508.16618v1 accessed 7 February 2026.

[229] Emilio Ferrara, *Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies* (2024) 6(1) Sci https://www.mdpi.com/2413-4155/6/1/3 accessed 1 February 2026.

[230] Alexandra Jonker and Julie Rogers, *What Is Algorithmic Bias?* (IBM Think, updated 2025) https://www.ibm.com/think/topics/algorithmic-bias accessed 1 February 2026.

[231] Angelina Wang, *Identities are not Interchangeable: The Problem of Overgeneralization in Fair Machine Learning* (2025) arXiv:2505.04038v2 https://arxiv.org/html/2505.04038v2 accessed 3 February 2026.

[232] Emilio Ferrara (n229)

[233] Ilmoi, ‘Evasion attacks on Machine Learning (or “Adversarial Examples”)’ (Towards Data Science, 14 July 2019) https://medium.com/data-science/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1 accessed 19 January 2026.

[234] Linyi Li, Tao Xie and Bo Li, *SoK: Certified Robustness for Deep Neural Networks* (arXiv:2009.04131v9, April 2023) https://arxiv.org/pdf/2009.04131 accessed 1 February 2026; AI safety frameworks emphasise *robust training* (e.g., adversarial training) to ensure models maintain safe behavior even under worst-case or adversarial inputs. However, current alignment and red-teaming practices are imperfect, and even aligned models may fail under cleverly optimised attacks (e.g., “adversarial suffixes” that force a model to violate its safety constraints). *See* Petr Spelda and Vit Stritecky, ‘Security Practices in AI Development’ (2025) 40 AI & Society 4869–4879

https://link.springer.com/article/10.1007/s00146-025-02247-4 accessed 1 February 2026.

[235] Palo Alto Networks, 'What Are Adversarial AI Attacks on Machine Learning?' (Cyberpedia) https://www.paloaltonetworks.in/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning accessed 22 January 2026.

[236] Annie Badman, *What is AI Risk Management* (IBM Think) https://www.ibm.com/think/insights/ai-risk-management accessed 1 February 2026.

[237] Leonard Dung, 'Current Cases of AI Misalignment and Their Implications for Future Risks' (2023) 202 *Synthese* 138 https://link.springer.com/article/10.1007/s11229-023-04367-0 accessed 1 February 2026; Zvi Mowshowitz, 'Jailbreaking ChatGPT on Release Day' (Don't Worry About the Vase, 2 December 2022) https://thezvi.substack.com/p/jailbreaking-the-chatgpt-on-release accessed 4 February 2026.

[238] Alexandra Jonker and Alice Gomstyn, 'What Is AI Alignment?' (*IBM*) https://www.ibm.com/think/topics/ai-alignment accessed 19 January 2026.

[239] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press 2014)Bostrom, 2014; Joseph Carlsmith, *Is Power-Seeking AI an Existential Risk?* (Open Philanthropy Project 2022) https://www.openphilanthropy.org/research/is-power-seeking-ai-an-existential-risk/ accessed 19 January 2026; Center for AI Safety, *Statement on AI Risk* (2023) https://www.safe.ai/statement-on-ai-risk accessed 19 January 2026; Leonard Dung, 'Current Cases of AI Misalignment and Their Implications for Future Risks' (2023) 202 *Synthese* 138; Richard Ngo et al, *The Alignment Problem from a Deep Learning Perspective* (arXiv:2209.00626, 2022) https://arxiv.org/abs/2209.00626 accessed 19 January 2026; Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury Publishing 2020)Ord, 2020; Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Viking 2019).

[240] Francisco Carvalho, 'Modelling the Recommender Alignment Problem' (arXiv:2208.12299v1, August 2022) https://arxiv.org/abs/2208.12299 accessed 1 February 2026.

[241] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 28 June 2024 on Artificial Intelligence (EU AI Act), art 15(4).

[242] Theodor Stoecker, Samed Bayer and Ingo Weber, 'Bias Mitigation for AI-Feedback Loops in Recommender Systems: A Systematic Literature Review and Taxonomy' (arXiv:2509.00109v1, August 2025) https://arxiv.org/abs/2509.00109 accessed 1 February 2026.

[243] Angelina Wang and Olga Russakovsky, *Directional Bias Amplification* (arXiv:2102.12594v1, 2021) https://arxiv.org/abs/2102.12594 accessed 1 February 2026.

[244] Charlotte Jee, 'A Biased Medical Algorithm Favored White People for Health-Care Programs' (MIT Technology Review, 25 October 2019) https://www.technologyreview.com/2019/10/25/132184/a-biased-medical-algorithm-favored-white-people-for-healthcare-programs/ accessed 4 February 2026.

[245] 'Fairness Constraint' (ScienceDirect) https://www.sciencedirect.com/topics/computer-science/fairness-constraint accessed 22 January 2026.

[246] Joshua Waithira, Ruth Chweya and Ratemo Makiya Cyprian, 'Adversarial Debiasing for Bias Mitigation in Healthcare AI Systems: A Literature Review' (2025) 12 Open Access Library Journal 1 https://doi.org/10.4236/oalib.1113340 accessed 22 January 2026.

[247] Georgios Ziras, Aristeidis Farao, Apostolis Zarras and Christos Xenakis, 'From Vulnerability to Resilience: Adversarial Training and Real-Time Detection for AI Security' (2025) 28 *Array* 100546 https://www.sciencedirect.com/science/article/pii/S2590005625001730 accessed 1 February 2026.

[248] Georgios Ziras, Aristeidis Farao, Apostolis Zarras and Christos Xenakis, 'From Vulnerability to Resilience: Adversarial

Training and Real-Time Detection for AI Security' (2025) 28 *Array* 100546 https://www.sciencedirect.com/science/article/pii/S2590005625001730 accessed 1 February 2026.

[249] Apostol Vassilev et al, 'Adversarial Machine Learning' (NIST AI 100-2e2025, National Institute of Standards and Technology 2025) https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf accessed 22 January 2026.

[250] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 15.

[251] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 15(5).

[252] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 15(4).

[253] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 15(4).

[254] Hanno Gottschalk, Matthias Rottmann and Maida Saltagic, 'Does Redundancy in AI Perception Systems Help to Test for Super-Human Automated Driving Performance?' in Hanno Gottschalk, Matthias Rottmann and Maida Saltagic (eds), *Deep Neural Networks and Data for Automated Driving* (Springer 2022) 81–106 https://link.springer.com/chapter/10.1007/978-3-031-01233-4_2 accessed 4 February 2026.

[255] Matthew Pisano et al, 'Bergeron: Combating Adversarial Attacks through a Conscience-Based Alignment Framework' (arXiv, 18 August 2024) https://arxiv.org/abs/2312.00029 accessed 4 February 2026.

[256] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689.

[257] OECD, 'Recommendation of the Council on Artificial Intelligence' (OECD/LEGAL/0449, 2019 [updated 2024]) https://www.oecd.org/en/topics/ai-principles.html accessed 22 January 2026.

[258] Ibid.

[259] UNESCO, 'Recommendation on the Ethics of Artificial Intelligence' (23 November 2021) https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence accessed 22 January 2026.

[260] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14.

[261] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14(2).

[262] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14(1).

[263] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14(4)(a); This includes awareness of potential propensity to overrely on AI-generated outputs, and to take decisions regarding the usage of AI systems including disregarding, overriding or reversing AI-generated outputs, or intervening system operations for safe shutdown procedures. *See* Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14(3).Article 14(4) of the EU AI Act.

[264] Jakub Szarmach, 'Human Oversight under Article 14 of the EU AI Act' (AI & Global Law Blog, 3 April 2025) https://www.aigl.blog/human-oversight-under-article-14-of-the-eu-al-act/ accessed 22 January 2026.

[265] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 14(3).Article 14(3) of the EU AI Act.

[266] Ana Zidarescu, 'Human Oversight under Article 14 of the EU AI Act' (AI Gouvernance, 3 April 2025) https://aigouvernance.com/human-oversight-under-article-14-of-the-eu-ai-act/amp/ accessed 22 January 2026.

[267] Inter-Parliamentary Union, Ethical principles: Human autonomy and oversight (Guidelines for AI in parliaments) https://www.ipu.org/ai-guidelines/ethical-principles-human-autonomy-and-oversight accessed 1 February 2026.

[268] Ibid.

[269] Ibid.

[270] Google Cloud, 'What is Human-in-the-Loop (HITL)?' (Google Cloud Discover) https://cloud.google.com/discover/human-in-the-loop accessed 22 January 2026.

[271] Cole Stryker, 'What is human-in-the-loop?' (IBM Think, 11 June 2024) https://www.ibm.com/think/topics/human-in-the-loop accessed 22 January 2026.

[272] Dave Bergmann, 'What is reinforcement learning from human feedback (RLHF)?' (IBM Think, 23 May 2024) https://www.ibm.com/think/topics/rlhf accessed 22 January 2026.

[273] Cole Stryker (n271)

[274] IBM, 'What is Explainable AI (XAI)?' (IBM) https://www.ibm.com/think/topics/explainable-ai accessed 19 January 2026.

[275] Martin Marzidovšek, 'What explainable AI is, why it matters and how we can achieve it' (OECD.AI Wonk, 23 May 2025) https://oecd.ai/en/wonk/what-explainable-ai-is-why-it-matters-and-how-we-can-achieve-it accessed 22 January 2026.

[276] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 13 and 50.

[277] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 13.

[278] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 13 and 50.

[279] Yao Rong and others, 'Towards Human-Centered Explainable AI: A Survey of User Studies for Model Explanations' (2024) 46 IEEE Transactions on Pattern Analysis and Machine Intelligence 2104 https://ieeexplore.ieee.org/document/10316181 accessed 19 January 2026.

[280] Martin Marzidovšek (n275)

[281] Alun Preece and others, *Stakeholders in Explainable AI* (arXiv:1810.00184v1, 2018) https://arxiv.org/abs/1810.00184 accessed 1 February 2026.

[282] Upol Ehsan and others, 'Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions' in *Proceedings of the 24th International Conference on Intelligent User Interfaces* (2019) 263–274 https://dl.acm.org/doi/10.1145/3301275.3302303 accessed 1 February 2026.

[283] ibid.

[284] Michael Katell and others, 'Toward Situated Interventions for Algorithmic Equity: Lessons from the Field' in *Proceedings of the 2020 Conference on Fairness, Accountability, and*

Transparency (2020) 45–55 https://dl.acm.org/doi/10.1145/3351095.3372850 accessed 1 February 2026.

[285] Yen-Chia Hsu and others, *Empowering Local Communities Using Artificial Intelligence* (arXiv:2110.02007v1, 2021) https://arxiv.org/abs/2110.02007 accessed 1 February 2026.

[286] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 50.

[287] Ibid.

[288] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 50(3); Recital 132 stipulates that implementation of Article 50 should take into consideration factors such as age and disability which can make users vulnerable. Recital 132 also requires that users (natural persons) should be notified when engaging with AI systems that may make inferences relating to emotions or intentions of the user through processing sensitive information such as their biometric data. Transparency obligations through these notifications should be made available to users with disabilities in accessible formats and should be clearly articulated no later than the first instance of interaction. *See* Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 50(5)and Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, recital 132.

[289] 'California introduces first AI chatbot safety law' (Digital Watch Observatory, 15 October 2025) https://dig.watch/updates/california-introduces-first-ai-chatbot-safety-law accessed 22 January 2026.

[290] Originally, all regulated businesses requiring state licenses or certification to practice were required to disclose the use of generative AI at the outset of consumer interaction. However,

the law has now been amended to place this obligation only in the context of "high-risk" AI interactions such as collection of sensitive personal information or the provision of personalised advice such as financial, legal, or medical advice. *See* 'Utah scales back reach of generative AI consumer protection law' *(Davis Polk, 4 April 2025)* https://www.davispolk.com/insights/client-update/utah-scales-back-reach-generative-ai-consumer-protection-law accessed 22 January 2026.

[291] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 50(2); The techniques and methods employed to carry out this marking must be designed to be effective, interoperable, robust and reliable.

[292] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 3(60); This mark can be done through techniques such as watermarks, metadata identifications, cryptographic methods for verification of authenticity and provenance, and fingerprinting. Where the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content in an appropriate manner that does not hamper the display or enjoyment of the work.

[293] Margaret Warthon, 'Restricting access to AI decision-making in the public interest: The justificatory role of proportionality and its balancing factors' (2024) 13(3) Internet Policy Review https://policyreview.info/articles/analysis/restricting-access-to-ai-decision-making accessed 19 January 2026.

[294] Government of India, Ministry of Electronics and Information Technology, *Draft Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Amendment Rules* https://www.meity.gov.in/static/uploads/2025/10/9de47fb06522b9e40a61e4731bc7de51.pdf accessed 7 February 2026.

[295] Asheef Iqubbal, *Limits of Labelling in India's Pursuit to Tackle Synthetic Information* (MediaNama, 13 November 2025) https://www.medianama.com/2025/11/223-limits-labelling-india-combat-synthetically-generated-information/ accessed 7 February 2026.

[296] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L 2024/1689, art 50(2).

[297] European Commission, First Draft Code of Practice on Transparency of AI-Generated Content (2025) https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content accessed 28 January 2026.

[298] European Commission, First Draft Code of Practice on Transparency of AI-Generated Content (2025) https://digital-strategy.ec.europa.eu/en/library/first-draft-code-practice-transparency-ai-generated-content accessed 28 January 2026.

[299] Chinasa T Okolo, Nicola Dell and Aditya Vashistha, 'Making AI Explainable in the Global South: A Systematic Review' in *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS 2022)* (ACM 2022) https://doi.org/10.1145/3530190.3534802 accessed 1 February 2026.

[300] Ludwig Schmidt and others, 'Adversarially Robust Generalization Requires More Data' in Proceedings of the 32nd International Conference on Neural Information Processing Systems (Curran Associates Inc 2018) https://dl.acm.org/doi/10.5555/3327345.3327409 accessed 28 January 2026.

[301] Genevieve Smith, 'How to Make AI Equitable in the Global South' (Stanford Social Innovation Review, 3 April 2024) https://ssir.org/articles/entry/equitable-ai-in-the-global-south/ accessed 19 January 2026.

[302] Masakhane, *Masakhane: A grassroots NLP community for Africa, by Africans* (Masakhane) https://www.masakhane.io/ accessed 7 February 2026;

Chinasa T Okolo, 'AI in the Global South: Opportunities and Challenges Towards More Inclusive Governance' (Brookings, 1 November 2023) https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/ accessed 7 February 2026.

[303] Government of India, Press Information Bureau, *India AI Governance Guidelines* (Ministry of Electronics and Information Technology, November 2025) https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf accessed 7 February 2026.

[304] Melanie Fink, *Human Oversight under Article 14 of the EU AI Act* (SSRN, 14 February 2025) https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5147196 accessed 7 February 2026.

[305] *India AI Governance Guidelines* (n303)

[306] Melanie Fink (n304)

[307] Chinasa T Okolo et al (n299)

[308] Upol Ehsan and Mark O Riedl, 'Explainability Pitfalls: Beyond Dark Patterns in Explainable AI' (2021) https://arxiv.org/abs/2109.12480 accessed 7 February 2026.

[309] Ibid.

[310] Shivani Kapania et al, '"Because AI Is 100% Right and Safe": User Attitudes and Sources of AI Authority in India', CHI Conference on Human Factors in Computing Systems (ACM 2022)

[311] Lalli Myllyaho et al, 'Systematic literature review of validation methods for AI systems' 181 Journal of Systems and Software 2021 https://doi.org/10.1016/j.jss.2021.111050 accessed 7 February 2026.

[312] Kristin McCann, 'How to use verification and validation for AI safety critical systems | Interview with MathWorks' Lucas Garcia' IOT Insider 4 April 2024 https://www.iotinsider.com/industries/ai/how-to-use-verification-and-validation-for-ai-safety-critical-systems-

interview-with-mathworks-lucas-garcia/ accessed 7 February 2026.

[313] Tellix AI, *How to Test AI Systems Before Full-Scale Deployment* (Tellix, 7 July 2025) https://tellix.ai/how-to-test-ai-systems-before-full-scale-deployment/ accessed 7 February 2026.

[314] Databricks, *Pre-deployment Validation for Model Serving* (Databricks Documentation, last updated 2 December 2025) https://docs.databricks.com/gcp/en/machine-learning/model-serving/model-serving-pre-deployment-validation accessed 7 February 2026.

[315] Lalli Myllyaho et al (n311)

[316] Fuyuki Ishikawa and Nobukazu Yoshioka, 'How do Engineers Perceive Difficulties in Engineering of Machine-Learning Systems? - Questionnaire Survey. 2019 IEEE/ACM Joint 7th International Workshop on Conducting Empirical Studies in Industry (CESI) and 6th International Workshop on Software Engineering Research and Industrial Practice (SER&IP) (2019); SEBoK, *Verification and Validation of Systems in Which AI is a Key Element* (Systems Engineering Body of Knowledge Wiki) https://sebokwiki.org/wiki/Verification_and_Validation_of_Systems_in_Which_AI_is_a_Key_Element accessed 7 February 2026.

[317] Ardi Janjeva, Anna Gausen and Tvesha Sippy, *Realising the Potential of Sociotechnical Approaches to AI Evaluation* (CETaS Expert Analysis, December 2024) https://cetas.turing.ac.uk/publications/realising-potential-sociotechnical-approaches-ai-evaluation accessed 7 February 2026.

[318] Megan E Salwei and Pascale Carayon, 'A Sociotechnical Systems Framework for the Application of Artificial Intelligence in Health Care Delivery' (2022) 16(4) *Journal of Cognitive Engineering and Decision Making* 194–206 https://pmc.ncbi.nlm.nih.gov/articles/PMC9873227/ accessed 7 February 2026.

[319] Daniel Kondor et al, 'Complex Systems Perspective in Assessing Risks in Artificial Intelligence' (2024) 382(2285)

Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 20240109 https://pmc.ncbi.nlm.nih.gov/articles/PMC11558246/ accessed 7 February 2026.

[320] Google, *Top Risks of Generative AI Systems – Secure AI Framework* (Secure AI Framework) https://saif.google/secure-ai-framework/risks accessed 7 February 2026.

[321] Daniel Kondor et al (n319); Emilio Ferrara, 'The Butterfly Effect in Artificial Intelligence Systems: Implications for AI Bias and Fairness' (2024) *Machine Learning with Applications* **15** 100525 https://doi.org/10.1016/j.mlwa.2024.100525 accessed 7 February 2026.

[322] Daniel Kondor et al (n319)

[323] ibid.

[324] Bill Dembski, *Artificial General Intelligence: The Oracle Problem* (Science and Culture Today, 27 February 2024) https://scienceandculture.com/2024/02/artificial-general-intelligence-the-oracle-problem/ accessed 7 February 2026.

[325] AI Hauptman, 'Adapting to the Human: A Systematic Review of a Decade of Research on Adaptive Autonomy' (2024) *Annual Review of Control, Robotics, and Autonomous Systems* (advance online publication) https://www.sciencedirect.com/science/article/pii/S0003687024001133 accessed 7 February 2026.

[326] Philip Koopman and Michael Wagner, 'Challenges in Autonomous Vehicle Testing and Validation' (SAE 2016 World Congress and Exhibition, SAE International, April 2016) https://users.ece.cmu.edu/~koopman/pubs/koopman16_sae_autonomous_validation.pdf accessed 7 February 2026.

[327] Daniel Kondor et al (n319)

[328] Lalli Myllyaho et al (n311)

[329] Inioluwa Deborah Raji et al, 'AI and the Everything in the Whole Wide World Benchmark' in *Proceedings of the 35th Conference on Neural Information Processing Systems*

(NeurIPS 2021) Track on Datasets and Benchmarks (J Vanschoren and S Yeung eds, Vol 1) (NeurIPS Datasets and Benchmarks 2021) https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf accessed 7 February 2026.

[330] ibid.

[331] Palo Alto Networks, *What Is Sandboxing?* (Cyberpedia) https://www.paloaltonetworks.in/cyberpedia/sandboxing#sandboxing accessed 7 February 2026.

[332] ibid.

[333] Wolf-Georg Ringe, 'Why We Need a Regulatory Sandbox For AI' (Oxford Law Blogs, 12 May 2023) https://blogs.law.ox.ac.uk/oblb/blog-post/2023/05/why-we-need-regulatory-sandbox-ai accessed 7 February 2026.

[334] Thomas Buocz, Sebastian Pfotenhauer and Iris Eisenberger, 'Regulatory Sandboxes in the AI Act: Reconciling Innovation and Safety?' (2023) 15(2) *Law, Innovation and Technology* 357–389 https://doi.org/10.1080/17579961.2023.2245678 accessed 7 February 2026.

[335] Jon Truby, 'A Sandbox Approach to Regulating High-Risk Artificial Intelligence Applications' 2022 (2) European Journal of Risk Regulation 13

[336] Douglas Arner, "Financial regulation, technology and the future of finance" in J Walker, A Pekmezovic and G Walker (eds), Sustainable Development Goals: Harnessing Business to Achieve the Sustainable Development Goals through Technology, Innovation and Financing 2019 Wiley

[337] Thomas Buocz et al (n334)

[338] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 57 and 58.

[339] Lalli Myllyaho et al (n311)

[340] ibid.

[341] ibid.

[342] ibid.

[343] ibid.

[344] Maria Eriksson and others, 'Can we trust AI benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation' https://arxiv.org/pdf/2502.06559 accessed 7 February 2026.

[345] Inioluwa Deborah Raji et al (n329)

[346] Maria Eriksson (n344)

[347] ibid.

[348] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 15(2).

[349] Maria Eriksson (n344)

[350] Timothy R McIntosh and others, 'Inadequacies of Large Language Model Benchmarks in the Era of Generative Artificial Intelligence' (2024) *arXiv* **2402.09880** https://arxiv.org/abs/2402.09880 accessed 7 February 2026.

[351] Maria Eriksson (n344)

[352] Maribeth Rauh, et al. Gaps in the Safety Evaluation of Generative AI. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7:1200–1217, October 2024. ISSN 3065-8365. doi:10.1609/aies.v7i1.31717

[353] Peter Douglas, *AI Systems Are Great at Tests. But How Do They Perform in Real Life?* (The Conversation, 25 August 2025) https://theconversation.com/ai-systems-are-great-at-tests-but-how-do-they-perform-in-real-life-260176 accessed 7 February 2026.

[354] Gabriel Grill, 'Constructing Capabilities: The Politics of Testing Infrastructures for Generative AI' in Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) (Association for Computing Machinery 2024) 1838–1849 https://doi.org/10.1145/3630106.3659009 accessed 1 February 2026.

[355] Maria Eriksson (n344)

[356] ibid.

[357] WitnessAI, 'AI Red Teaming: Strengthening AI Systems Against Real-World Threats' (WitnessAI, 13 August 2025) https://witness.ai/blog/ai-red-teaming/ accessed 28 January 2026.

[358] ibid.

[359] Evelyn Yee, 'AI Red-Teaming Design: Threat Models and Tools' (Center for Security and Emerging Technology, 24 October 2025) https://cset.georgetown.edu/article/ai-red-teaming-design-threat-models-and-tools/ accessed 28 January 2026; For model jailbreaks, read: Tom Krantz and Alexandra Jonker, 'AI Jailbreak: Rooting out an Evolving Threat' (IBM, 13 November 2024) https://www.ibm.com/think/insights/ai-jailbreak accessed 28 January 2026.

[360] Evelyn Yee (n359)

[361] Jamila Venturini, 'AI governance for and from the Global South' Medium 19 December 2023 https://medium.com/opendatacharter/ai-governance-for-and-from-the-global-south-07326645b053 accessed 28 January 2026.

[362] Rishiti Choudaha, 'On Third Party AI Audits: Access and Roadblocks' (*The CCG Blog*, 15 September 2025) https://ccgnludelhi.wordpress.com/2025/09/15/on-third-party-ai-audits-access-and-roadblocks/ accessed 28 January 2026.

[363] VerifyWise, 'Model Documentation Best Practices' (*VerifyWise*) https://verifywise.ai/lexicon/model-documentation-best-practices accessed 28 January 2026.

[364] Andrew Bell, Oded Nov and Julia Stoyanovich, 'Think about the Stakeholders First! Toward an Algorithmic Transparency Playbook for Regulatory Compliance' (2023) 5 Data & Policy e12 https://www.cambridge.org/core/journals/data-and-policy/article/think-about-the-stakeholders-first-toward-an-algorithmic-transparency-playbook-for-regulatory-compliance/10D7F194DB250DDF3A30471B5CEB9326 accessed 28 January 2026.

[365] Philip Adler et al. 'Auditing black-box models for indirect influence' 54 Knowl Inf Syst 2018 https://doi.org/10.1007/s10115-017-1116-3 accessed 28 January 2026.

[366] Jacqui Ayling, Adriane Chapman, 'Putting AI ethics to work: are the tools for purpose?' 2 AI and Ethics 2022 https://doi.org/10.1007/s43681-021-00084-x accessed 28 January 2026.

[367] Lalli Myllhalo et al (n311)

[368] ibid.

[369] ibid.

[370] Per the AIA, high-risk AI systems are identifiable as those that have significant harmful impact on the health, safety and fundamental rights of persons in the Union and such limitation should minimise any potential restriction to international trade. Market Surveillance Authorities are national authorities appointed by each Member State, and have intervening power if an AI system poses risks or fails to comply with the AI Act requirements. See more: https://www.aiact-info.eu/full-text-and-pdf-download/ ; Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 60.

[371] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on

artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 60; ECSS Working Group, 'Verification, Validation and Qualification of AI Systems' ECSS-E-HB-40-02A Machine Learning Qualification Handbook 28 September 2023 https://www.cosmos.esa.int/documents/10939403/13938157/SPAW23_ECSS_ML_final_MB%C3%A4_LM_v4.pdf/59a903d5-f111-5ae1-23b1-d8a753f1f70f?t=1695742845134 accessed 28 January 2026.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 10.

[373] Daniel Schnurr, 'Effective Implementation of Requirements for High-Risk AI Systems under AI Act: Transparency and Appropriate Accuracy' 2025 Centre on Regulation in Europe (https://cerre.eu/wp-content/uploads/2025/02/Effective-Implementation-of-Requirements-for-High-Risk-AI-Systems-Under-the-AI-Act_FINAL-1.pdf) accessed 21 February 2026; Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 13.

[374] National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (NIST AI 100-1, January 2023) https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf accessed 19 January.

[375] National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (NIST AI 100-1, 2023) 25 https://doi.org/10.6028/NIST.AI.100-1 accessed 28 January 2026.

[376] National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (NIST AI 100-1, 2023) 36 https://doi.org/10.6028/NIST.AI.100-1 accessed 28 January 2026.

[377] Rafael A F Zanatta and Mariana Rielli, 'The Artificial Intelligence Legislation in Brazil: Technical Analysis of the Text to Be Voted on in the Federal Senate Plenary'

(*Data Privacy Brasil*, 10 December 2024) https://www.dataprivacybr.org/en/the-artificial-intelligence-legislation-in-brazil-technical-analysis-of-the-text-to-be-voted-on-in-the-federal-senate-plenary/ accessed 28 January 2026.

[378] ibid.

[379] OECD, 'OECD Framework for the Classification of AI Systems' (OECD Digital Economy Papers No 323, 2022) https://www.oecd.org/content/dam/oecd/en/publications/reports/2022/02/oecd-framework-for-the-classification-of-ai-systems_336a8b57/cb6d9eca-en.pdf accessed 28 January 2026.

[380] Kristin McCann, 'How to Use Verification and Validation for AI Safety-Critical Systems | Interview with MathWorks' Lucas Garcia' *IoT Insider* (4 April 2024) https://www.iotinsider.com/industries/ai/how-to-use-verification-and-validation-for-ai-safety-critical-systems-interview-with-mathworks-lucas-garcia/ accessed 1 February 2026; Lalli Myllyaho et al (n311)

[381] Ibnu Fikri Ghozali, 'Decolonizing Algorithms: Artificial Intelligence Bias and Digital Colonialism in Global South AI Governance' (2025) 3 Jurnal Nawala Politika 77 https://doi.org/10.24843/jnp.v3i1.342 accessed 28 January 2026; Mélissa Anchisi, 'Beyond Translation: Multilingual Benchmark Makes AI Multicultural' (*Tech Xplore*, 2 June 2025) https://techxplore.com/news/2025-06-multilingual-benchmark-ai-multicultural.html accessed 28 January 2026.

[382] Yan Tao et al, Cultural bias and cultural alignment of large language models, *PNAS Nexus*, Volume 3, Issue 9, September 2024, https://doi.org/10.1093/pnasnexus/pgae346 accessed 28 January 2026.

[383] Yu Ying Chiu and others, 'CulturalBench: A Robust, Diverse, and Challenging Cultural Benchmark by Human-AI CulturalTeaming' (2024) arXiv preprint arXiv:2410.02677 https://arxiv.org/abs/2410.02677 accessed 28 January 2026.

[384] Mohamed bin Zayed University of Artificial Intelligence, 'Cultural inclusivity in AI: A new benchmark dataset on 100 languages' (*MBZUAI*, 15 January 2025) https://mbzuai.ac.ae/news/cultural-inclusivity-in-ai-a-new-benchmark-dataset-on-100-languages/ accessed 28 January 2026.

[385] Angelina Wang, Aaron Hertzmann and Olga Russakovsky, 'Benchmark suites instead of leaderboards for evaluating AI fairness' (2024) 5 Patterns 101080 https://doi.org/10.1016/j.patter.2024.101080 accessed 28 January 2026.

[386] Elizabeth Anne Watkins, '"It Doesn't Know Anything About My Work": Participatory Benchmarking and AI Evaluation in Applied Settings' https://openreview.net/pdf?id=lqqe0ANOoa accessed 28 January 2026.

[387] Rakesh Patel, 'AI Development Life Cycle: A Comprehensive Guide' (*Space-O AI*, 18 October 2025) https://www.spaceo.ai/blog/ai-development-life-cycle/ accessed 28 January 2026.

[388] Factually, 'Real-World AI Harm and Benefit Examples' (Factually) https://factually.co/fact-checks/technology/real-world-ai-harm-and-benefit-examples-140a61 accessed 28 January 2026.

[389] Zlatko Delev, 'AI Privacy Risks and Data Protection Challenges' (*GDPR Local*, 3 July 2025) https://gdprlocal.com/ai-privacy-risks/ accessed 28 January 2026.

[390] Nitin Vats, 'The Ethics of AI in Monitoring and Surveillance' (*NICE Actimize*, 1 January 2024) https://www.niceactimize.com/blog/fmc-the-ethics-of-ai-in-monitoring-and-surveillance accessed 28 January 2026

[391] Merlin Stein and Connor Dunlop, 'Safe beyond sale: post-deployment monitoring of AI' (*Ada Lovelace Institute*, 28 June 2024) https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/ accessed 28 January 2026; Adewunmi Akingbola and others, 'Artificial Intelligence and the Dehumanization of Patient Care' (2024) 3 Journal of Medicine,

Surgery, and Public Health 100138 https://doi.org/10.1016/j.glmedi.2024.100138 accessed 28 January 2026.

[392] Adewunmi Akingbola and others (n391

[393] Adam Stewart, 'Challenges in AI for Data Integrity Across Data Lifecycle: 6 Critical Issues' (*DialZara*, 5 June 2024) https://dialzara.com/blog/ai-data-lifecycle-management-6-key-challenges accessed 28 January 2026.

[394] Fabian Hinder et al, 'Model-based explanations of concept drift' 555 Neurocomputing 2023 (https://doi.org/10.1016/j.neucom.2023.126640)

[395] Nico Klingler, 'Concept Drift vs Data Drift: How AI Can Beat the Change' (*Viso.ai*, 4 April 2024) https://viso.ai/deep-learning/concept-drift-vs-data-drift/ accessed 28 January 2026.

[396] Ibid.

[397]Berkman Sahiner and others, 'Data drift in medical machine learning: implications and potential remedies' (2023) 96 The British Journal of Radiology 20220878 https://doi.org/10.1259/bjr.20220878 accessed 28 January 2026.

[398] Youngsub Lee et al 'The Effectiveness of Big Data-Driven Predictive Policing: Systematic Review' 7(2) Justice Evaluation Journal 2024

[399] Ibid.

[400] Wencheng Yang et al, 'Deep learning model inversion attacks and defenses: a comprehensive survey' 58:242 Artificial Intelligence Review 2025; Lior Romano, 'Model Inversion Attacks: A Growing Threat to AI Security' 14 March 2025 *Tillion* https://www.tillion.ai/blog/model-inversion-attacks-a-growing-threat-to-ai-security accessed 28 January 2026.

[401] Zhou et al, 'Model Inversion Attacks: A Survey of Approaches and Countermeasures' 2024 https://arxiv.org/html/2411.10023v1 accessed 28 January 2026.

[402] SentinelOne, 'Top 14 AI Security Risks in 2026' (7 January 2026) https://www.sentinelone.com/cybersecurity-101/data-and-ai/ai-security-risks/ accessed 2 February 2026.

[403] Lucinity, *Fast vs Slow AI Deployment: Finding the Right Balance in Compliance* (24 July 2025) https://lucinity.com/blog/fast-vs-slow-ai-deployment-finding-the-right-balance-in-compliance/ accessed 2 February 2026.

[404] Ibid.

[405] Tuhu Nugraha, 'AI Diplomacy: Why the Global South's Participation is Crucial for Global Stability' Modern Diplomacy 1 April 2025 https://moderndiplomacy.eu/2025/04/01/ai-diplomacy-why-the-global-souths-participation-is-crucial-for-global-stability/ accessed 2 February 2026.

[406] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 14, 18, 20 and 26.

[407] For concept drift, statistical tests like Drift Detection Method, Early Drift Detection Method, can be used to monitor the error rates of an AI system when it is operating. Data drifting detection can operationalise distribution comparison tests like the Chi-square test, which evaluates changes in the input distribution. See more: Nico Klingler, 'Concept Drift vs Data Drift: How AI Can Beat the Change' (*Viso.ai*, 4 April 2024) https://viso.ai/deep-learning/concept-drift-vs-data-drift/ accessed 28 January 2026.

[408] Jayita Gulati, 'Detecting & Handling Data Drift in Production' (*Machine Learning Mastery*, 17 April 2025) https://machinelearningmastery.com/detecting-handling-data-drift-in-production/ accessed 28 January 2026.

[409] Wencheng Yang et al (n400)

[410]Conor Bronsdon, 'How Attackers Extract Data Through "Innocent" Queries in Model Inversion Attacks' (*Galileo AI*, 1 August 2025) https://galileo.ai/blog/prevent-model-inversion-inference-attacks accessed 28 January 2026; Amazon Web Services, 'Detect and filter harmful content by using Amazon

Bedrock Guardrails' (*AWS Documentation*) https://docs.aws.amazon.com/bedrock/latest/userguide/guardrails.html accessed 28 January 2026.

[411] Content filtering in AI pre-deployment involves applying classifiers to screen model inputs and outputs for harmful categories like hate speech, violence, sexual content, or self-harm before release, blocking or annotating content exceeding severity thresholds. This reduces post-deployment risks such as model inversion attacks, where adversaries reconstruct sensitive training data from outputs, and malicious use like generating deceptive or exploitable responses. For more, read: Partnership on AI, 'PAI's Guidance for Safe Foundation Model Deployment' (*Partnership on AI*) https://partnershiponai.org/modeldeployment/ accessed 28 January 2026; Microsoft, 'Content filtering' (*Microsoft Learn*) https://learn.microsoft.com/en-us/azure/ai-foundry/concepts/content-filtering?view=foundry-classic accessed 28 January 2026.

[412] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 72.

[413] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 73.

[414] National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (NIST AI 100-1, January 2023) https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf accessed 19 January 2026.

[415] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 12.

[416] OECD, OECD Framework for the Classification of AI Systems (OECD Digital Economy Papers No 323, OECD Publishing, 22 February 2022) https://doi.org/10.1787/cb6d9eca-en accessed 2 February 2026

[417] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 12.

[418] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 20 and 26.

[419] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 28 June 2024 on Artificial Intelligence (EU AI Act) art 73 https://artificialintelligenceact.eu/article/73/ accessed 2 February 2026.

[420] Aditya Tiwari, 'AI Incident Reporting in Telecommunications Law' (*Jus Corpus*, 26 November 2025) https://www.juscorpus.com/ai-incident-reporting-in-telecommunications-law/ accessed 28 January 2026.

[421] Ciro Torres Freitas and André Zonaro Giacchetta, 'AI Watch: Global regulatory tracker - Brazil' (*White & Case*, 6 June 2025) https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-brazil accessed 28 January 2026; Brazil, Projeto de Lei nº 2338/2023 (dispõe sobre o uso da inteligência artificial) arts 25, 42 https://www.camara.leg.br/proposicoesWeb/prop_mostrarintegra?codteor=2868197&filename=PL%202338/2023 accessed 2 February 2026.

[422] Organisation for Economic Co-operation and Development, *OECD Framework for the Classification of AI Systems* (OECD Publishing 2022) <https://www.oecd.org/en/publications/oecd-framework-for-

the-classification-of-ai-systems_cb6d9eca-en.htmll> accessed 28 January 2026.

Ministry of Electronics and Information Technology, *India AI Governance Guidelines: Empowering Ethical and Responsible AI* (Press Information Bureau, 5 November 2025) https://static.pib.gov.in/WriteReadData/specificdocs/documents/2025/nov/doc2025115685601.pdf

[424] Manisha Khandelwal, 'Understanding AI Feedback Loop: What, How, & Why' (SurveySensum, 24 September 2025) https://www.surveysensum.com/blog/ai-feedback-loop accessed 28 January 2026.

[425] Sri Hari, 'Understanding Feedback Loops in Machine Learning Systems' 11 International Journal of Scientific Research in Computer Science, Engineering and Information Technology 2025

[426] n423

[427] Petar Radanliev, 'AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development' (2025) 39(1) Applied Artificial Intelligence 2463722 https://doi.org/10.1080/08839514.2025.2463722 accessed 28 January 2026.

[428] Andreas Tsamados, Luciano Floridi and Mariarosaria Taddeo, 'Human Control of AI Systems: From Supervision to Teaming' (2025) 5 AI and Ethics 1535 https://doi.org/10.1007/s43681-024-00489-4 accessed 28 January 2026.

[429] ibid.

[430] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 14 and 26.

[431] Peter Radenliev (n427)

[432] ibid; Reinforcement Learning from Human Feedback is an approach to train large language models with human values, by training them on direct feedback, such as rankings of generated

outputs. Rather than fixed rules or static labels, it leverages comparative human judgments to iteratively shape model behavior toward more helpful and safe responses.For more, see Reinforcement learning from Human Feedback' (GeeksforGeeks, 12 December 2025) https://www.geeksforgeeks.org/machine-learning/reinforcement-learning-from-human-feedback/ accessed 4 February 2026.

[433] Pinakin Ariwala, 'Why Auditability in AI Systems Matters More Than Ever in 2025' (*Maruti Techlabs*, 2025) https://marutitech.com/ai-auditability/ accessed 28 January 2026.

[434] Trang Tran Phuong, 'AI Model Drift Detection and Retraining: Maintenance Guide for Production ML Systems' (*SmartDev*, 24 November 2025) https://smartdev.com/ai-model-drift-retraining-a-guide-for-ml-system-maintenance/ accessed 28 January 2026; Trevor LaViale and Claire Longo, 'A Guide to Optimized Retraining' (Arize AI, 2023) https://arize.com/wp-content/uploads/2023/03/Arize-Guide-To-Optimized-Retraining.pdf accessed 28 January.

[435] Jayita Gulati (n408)

[436] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) [2024] OJ L/2024/1689, art 20.

437 Chinasa T. Okolo, 'AI in the Global South: Opportunities and Challenges towards More Inclusive Governance' (Brookings Institution, 20 August 2024) https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/ accessed 20 January 2026.

[438] Mishall Lallani, 'Decoloniality and AI: Possibilities and Pitfalls' in Christo El Morr, Yehya El-Lahib and Ricardo da Silveria Gorma (eds), *Beyond Tech Fixes* (Springer 2025) https://doi.org/10.1007/978-3-031-93022-5_14 accessed 22 January 2026.

[439] ibid.

[440] Vidushi Marda and Shivangi Narayan, 'Data in New Delhi's Predictive Policing System' (Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, January 2020) https://doi.org/10.1145/3351095.3372865 accessed 22 January 2026 ; Pragat Chauhan, 'AI and Human Rights: Global South Perspectives' (2025) 5(4) International Journal of Humanities Social Science and Management 563 https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf accessed 22 January 2026.

[441] Pragat Chauhan (n440)

[442] Shakir Mohamed, et al, 'Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence' Philosophy and Technology 33 (2020) 12 July 2020

[443] Maarten Sap et al, 'The Risk of Racial Bias in Hate Speech Detection' (Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence 2019) 1668.

[444] Deepawali Sharma and others, 'Hate Speech Detection Research in South Asian Languages: A Survey of Tasks, Datasets and Methods' (2025) 24(3) *ACM Transactions on Asian and Low-Resource Language Information Processing* 1–44 https://doi.org/10.1145/3711710 accessed 2 February 2026.

[445] Wencheng Yang (n400)

[446] Yang Hu and Yue Qian, 'Who Is Concerned about Digitalization? The Role of Digital Literacy and Exposure Across 30 Countries' (2025) *Information, Communication & Society* https://doi.org/10.1080/1369118X.2025.2592771 accessed 2 February 2026.

447 Danni Yu, Hannah Rosenfeld and Abhishek Gupta, *The AI Divide Between the Global North and Global South* (*World Economic Forum*, 23 January 2023) https://www.weforum.org/stories/2023/01/davos23-ai-divide-

global-north-global-south/ accessed 2 February 2026; Chinasa T Okolo, Nicola Dell and Aditya Vashistha, *Making AI Explainable in the Global South: A Systematic Review* in *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS '22)* (ACM 2022) https://doi.org/10.1145/3530190.3534802 accessed 2 February 2026.

[448] Syed Ali Hussain, Mary Bresnahan, and Jie Zhuang, 'Can artificial intelligence revolutionize healthcare in the Global South? A scoping review of opportunities and challenges' (2025) 11 Digital Health https://doi.org/10.1177/20552076251348024 accessed 22 January 2026.

[449] Shakir Mohamed et al n(442); Toxicity scoring assesses how harmful, offensive, or malicious content is, using categories like Race/Origin, Gender/Sex, Religion, Ability, and Violence. It flags content if the overall toxicity score—or any single category score—exceeds set thresholds. Read more: Brett Young, 'AI guardrails: Toxicity scorers' (Weights & Biases, 9 January 2025) https://wandb.ai/byyoung3/Generative-AI/reports/AI-guardrails-Toxicity-scorers--VmlldzoxMDg5Mzc5MA accessed 30 January 2026.

[450] Shakir Mohamed et al (n442)

[451] Rafael A F Zanatta (n377)

[452] Damian Eke, Ricardo Chavarriaga and Bernd Stahl, 'Decoloniality impact assessment for AI' (2025) AI & Society https://doi.org/10.1007/s00146-025-02649-4 accessed 22 January 2026.

[453] ibid.

॥ न्यायस्तत्र प्रमाणं स्यात् ॥
NLU
DELHI

CENTRE FOR
COMMUNICATION
GOVERNANCE

KONRAD
ADENAUER
STIFTUNG

Centre for Communication Governance
National Law University Delhi
Sector 14, Dwarka
New Delhi – 110078
011- 28031265

ccgdelhi.org | privacylibrary.ccgnlud.org |
ccg@nludelhi.ac.in | x.com/CCGNLUD | Instagram/ccgnlud