

Responsible AI in Practice: Ensuring Fairness and Trust



This is a work of independent research produced by **Aapti Institute** and commissioned by [Google.org](https://www.google.org/) as a part of the Digital Futures Project to facilitate dialogue and inquiry into emerging technologies such as AI.

ACKNOWLEDGEMENTS

In addition to contributions from the wider Aapti team, this report also draws upon the expertise of numerous academic and industry experts, practitioners. We are grateful for their input during interviews and feedback sessions.

Illustrator: Ayesha Punjabi | **Designer:** Kartik Lav, Vasudha Varadarajan & Antara Madavane

Responsible AI in Practice: Ensuring Fairness and Trust

AUTHORS

GAUTAM MISRA, RATTANMEEK KAUR,
SUPRATIK MITRA, POORVI YERRAPUREDDY

THEMATIC GUIDANCE

SOUJANYA SRIDHARAN, DR. SARAYU NATARAJAN



Table of Contents

Glossary	6
Introduction	12
Methodology	16
Systems thinking and design thinking	17
The Community of practice approach	22
MODULE I: ALL ABOUT THE BIAS	25
Anchoring the Research	26
The AI value chain	26
Foundation models	33
Sectoral considerations	36
Bias: Meaning and implication	41
Sources of bias in AI	46
The Bias Framework	49
Methodology of the framework	49
How to read the framework	51
Mitigation Strategies	57
MODULE II: AI FOR DIGITAL INTEGRITY AND CYBERSECURITY: THINKING THROUGH THE LENS OF TRUSTWORTHINESS	66
Introduction	67
Anchoring the research	69

Deep Dives: Landscaping roles, harms and opportunities	74
Digital Integrity	74
Cybersecurity	77
Connecting the Dots: Trustworthy AI for digital integrity and cybersecurity	83
Levers of trustworthy AI	88
Trustworthy AI for Digital Integrity	89
Trustworthy AI for Cybersecurity	90
Implementation strategies and recommendations	91
Mitigation strategies for Digital Integrity	91
Mitigation strategies for Cybersecurity	96
<hr/>	
Way Forward	100
<hr/>	
Annexures	103
Endnotes	104
List of Experts	113
Community of Practice members	114
Policy Instruments	116

Glossary



Glossary

Abbreviations

AI	Artificial intelligence
API	Application Programming Interface
CoP	Community of Practice
CS	Cybersecurity
CSAM	Child Sexual Abuse Material
EU	European Union
FAIR	Findable, Accessible, Interoperable, and Reusable
FRT	Facial Recognition Technology
GDPR	General Data Protection Regulation
HELM	Holistic Evaluation of Language Models
IA	Information Assurance
IEC	International Electrotechnical Commission
IIA	Institute of Internal Auditors
ISO	International Organization for Standardization
LLM	Large language models
ML	Machine Learning
OBGV	Online Gender-Based Violence
RAI	Responsible Artificial intelligence
REVISE	Revealing Visual biases
RHO	Roles, Harms, and Opportunities
TTP	Tactics, techniques and procedures
UNESCO	United Nations Educational, Scientific and Cultural Organization

Definitions

Key terms related to frameworks presented in this report have been defined in the guides accompanying them.

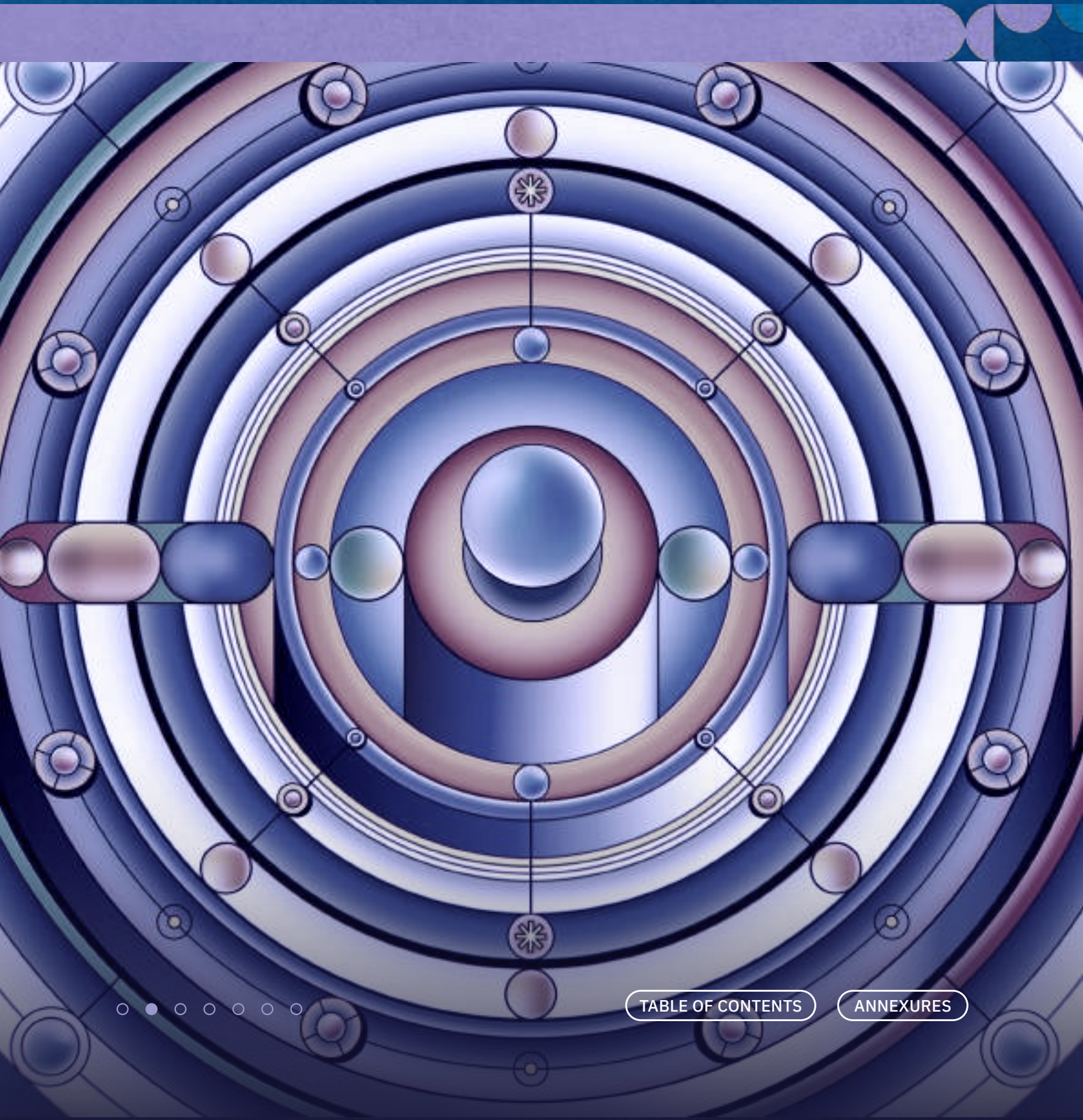
Adversary	Person, group, organization, or government that conducts or has the intent to conduct detrimental activities. (NIST)
AI value chain	The AI value chain is a network of interconnected processes and actors involved in the creation, deployment, and utilization of AI systems. (Attard-Frost & Widder, 2023b)
Anomaly	Condition that deviates from expectations based on requirements specifications, design documents, user documents, or standards, or from someone's perceptions or experiences. (NIST)
Application Programming Interface	The intermediary interface between the client and the application.
Artificial Intelligence system/tool	An engineered or machine-based system that can, for a given set of objectives, generate outputs such as predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy. (NIST AI Risk Management Framework)
Audit	Systematic, independent, documented process for obtaining records, statements of fact, or other relevant information and assessing them objectively, to determine the extent to which specified requirements are fulfilled. (NIST)
Benchmark	Standard against which results can be measured or assessed; Procedure, problem, or test that can be used to compare systems or components to each other or to a standard. (IEEE)
Bias	An inclination or prejudice, produced by computational or human cognitive biases, of a decision made by an AI system that is for or against one person or group, especially in a way considered to be morally or legally unfair. (Eirini Ntoutsi et al.)
Blue Team	The group responsible for defending an enterprise's use of information systems by maintaining its security posture against a group of mock attackers (i.e., the Red Team). (NIST)
Compute or computational power	A stack composed of hardware or software that allows for floating point operations, which is a mathematical operation that enables the representation of extremely large numbers with greater precision. (Jai Vipra & Sarah Myers West, Computational Power and AI)

Critical Infrastructure	System and assets, whether physical or virtual, so vital to a country, enterprise or individual that the incapacity or destruction of such systems and assets would have a debilitating impact on security, national economic security, national public health or safety, or any combination of those matters. (Aapti Analysis)
Cyber Threat	Any circumstance or event with the potential to adversely impact organizational operations (including mission, functions, image, or reputation), organizational assets, or individuals through an information system via unauthorized access, destruction, disclosure, modification of information, and/or denial of service. Also, the potential for a threat-source to successfully exploit a particular information system vulnerability. (NIST)
Cybersecurity	Prevention of damage to, protection of, and restoration of computers, electronic communications systems, electronic communications services, wire communication, and electronic communication, including information contained therein, to ensure its availability, integrity, authentication, confidentiality, and nonrepudiation. (NIST)
Deepfake	AI-generated or manipulated image, audio, or video content that resembles existing persons, objects, places, entities, or events and would falsely appear to a person to be authentic or truthful (EU AI Act)
Digital Integrity	Principles that govern the digital structures which directly impact human experiences online to ensure that an individual is free from interferences with her body or mind. (For a detailed discussion please refer to pg. 75)
Disinformation	False information which is deliberately intended to mislead– intentionally misstating the facts. (American Psychological Association)
Facial Recognition	is a way of using software to determine the similarity between two face images in order to evaluate a claim. Facial recognition uses computer-generated filters to transform face images into numerical expressions that can be compared to determine their similarity. These filters are usually generated by using deep “learning,” which uses artificial neural networks to process data. (Center for Strategic and International Studies)
Foundation Models	A foundation model is a model pre-trained on a large amount of data, capable of a range of general tasks such as interpret and mimic human language or images, in some cases AI models multi-modal having the ability to do both. (Bommasani & Liang, 2021)

Human cognitive bias	Human:cognitive biases relate to how an individual or group perceives AI system information to make a decision or fill in missing information, or how humans think about purposes and functions of an AI system. Human biases are omnipresent in decision-making processes across the AI lifecycle and system use, including the design, implementation, operation, and maintenance of AI. (NIST AI RMF 1.0)
Information Assurance	Measures that protect and defend information and information systems by ensuring their availability, integrity, authentication, confidentiality, and non-repudiation. These measures include providing for restoration of information systems by incorporating protection, detection, and reaction capabilities. (NIST)
Intrusion Detection	The process of monitoring the events occurring in a computer system or network and analyzing them for signs of possible incidents. (NIST)
Intrusion	A security event, or a combination of multiple security events, that constitutes a security incident in which an intruder gains, or attempts to gain, access to a system or system resource without having authorization to do so. (NIST)
Large Language Model	A class of language models that use deep-learning algorithms and are trained on extremely large textual datasets that can be multiple terabytes in size.
Machine Learning	A branch of Artificial Intelligence that focuses on the development of systems capable of learning from data to perform a task without being explicitly programmed to perform that task. Learning refers to the process of optimizing model parameters through computational techniques such that the model's behaviour is optimized for the training task. (NIST)
Misinformation	False or inaccurate information– getting the facts wrong. (American Psychological Association)
Natural Language Processing	A field concerned with machines capable of processing, analysing, and generating human language, either spoken, written, or signed.(NIST)
Ontology	A set of concepts and categories in a subject area or knowledge domain that shows their properties and the relationships among them to enable interoperability among disparate elements and systems and specify interfaces to independent, knowledge-based services for the purpose of enabling certain kinds of automated reasoning. (IEEE Guide IPA)
Open-source	Open source refers to something, historically software, that people can modify, share, and re-use because its design or “source code” is made publicly accessible. Opensource products provide universal access through an open-source licence that legally enables it. (Digital Public Goods Alliance)

Pentesting	A penetration test, or "pen test," is a security test that launches a mock <u>cyberattack</u> to find vulnerabilities in a computer system. (IBM)
Purple Team	Purple teaming is a cybersecurity exercise that combines red and blue teams to assess an organization's security. Red teams simulate attacks, while blue teams defend against them. The goal is to identify vulnerabilities, improve security, and develop better responses to cyberattacks. (LRQA)
Red Team	A group of people authorized and organized to emulate a potential adversary's attack or exploitation capabilities against an enterprise's security posture. (NIST)
Resilience	The ability to maintain required capability in the face of adversity. (NIST)
Response	Processes and procedures that are executed and maintained, to ensure timely response to detected cybersecurity incidents. (NIST)
Robust	The ability of an information assurance (IA) entity to operate correctly and reliably across a wide range of operational conditions, and to fail gracefully outside of that operational range. (NIST)
Socio-technical system	An approach to look at a technical system as an amalgamation of both the technical and social systems (Data Society)
Source of bias	Biases differentiated by the nature of incidence in AI systems. (Aapti Analysis)
Statistical Bias	A systematic tendency for estimates or measurements to be above or below their true values. Statistical biases arise from systematic as opposed to random error. Statistical bias can occur in the absence of prejudice, partiality, or discriminatory intent. (NIST)
Synthetic content	Also called AI generated content, refers to content—including text, audio, video, or other media—that has been created or "significantly altered" by algorithms. [NIST, also adopted by Future of Privacy Forum]
Tactics, Techniques and Procedures (TTP)	The behavior of an actor. A tactic is the highest-level description of this behavior, while techniques give a more detailed description of behavior in the context of a tactic, and procedures an even lower-level, highly detailed description in the context of a technique. (NIST)
Threat Actor	An individual or a group posing a threat. (NIST)
Zero day attack	An attack that exploits a previously unknown hardware, firmware, or software vulnerability. (NIST)

Introduction



Introduction

The growing influence of Artificial Intelligence (AI) across sectors has opened up unprecedented opportunities, captivating 91% of tech executives and 84% of the general public with its promise of transformation.¹ By 2030, the global economy stands to gain a staggering \$15.7 trillion from AI-driven innovations,² underscoring its role as a catalyst for unprecedented growth.

Research predicts that spending on AI could surge to \$100 billion in the U.S. and an impressive \$200 billion globally over 2025.³ As AI systems become more complex, challenges like trust, security, digital integrity and fairness must be addressed to fully realise the true potential of AI.

One aspect of AI adoption that has remained concerningly consistent is the level of risk mitigation organisations engage in to bolster trustworthiness. From 2019 to 2022, it was reported that there were no substantial increases in reported mitigation of any AI-related risks.⁴ While AI adoption has surged, there is a growing recognition of its associated risks, prompting organizations to take more proactive measures. A 2024 survey revealed that 78% of organizations now actively track AI as an emerging risk, a clear indication of heightened awareness and anxiety. Additionally, companies are 2.5 times more likely to be in advanced stages of digital risk maturity compared to the previous year, demonstrating a significant shift toward strategic risk management. Furthermore, 80% of organizations worldwide have adopted AI-driven security solutions to detect and prevent cyberattacks, underscoring the increasing focus on mitigating AI-related risks in critical operations.⁵ These developments mark a significant evolution in how organizations approach AI risk, reflecting a growing commitment to ensuring the safe and responsible deployment of AI technologies.

Our report is divided into two core pillars: Module 1 focuses on bias in artificial intelligence systems, which is not merely an issue of flawed algorithms or datasets, it is a systemic challenge, manifesting at various points across the AI value chain. From data collection and model training to deployment and application, biases can emerge due to complex interdependencies among

technical processes, societal dynamics, and stakeholder actions. These biases, if left unchecked, can perpetuate inequities, erode trust, and limit the transformative potential of AI. This part of the report delves into systems thinking and design methodologies to examine the sources, persistence, and mitigation of bias in AI systems. It identifies key inflection points in the AI value chain where bias takes root, explores the roles of various stakeholders, and proposes actionable strategies to foster fairness, accountability, and inclusivity in AI. This interdisciplinary approach highlights the need for systemic interventions and collaborative efforts to create AI systems that align with ethical principles and societal values. The findings presented here are grounded in extensive research, including insights from domain experts, community convenings, and documented examples of biased AI applications. While challenges such as data asymmetries and opaque decision-making remain, this report underscores the potential of systems and design thinking to drive impactful, context-aware solutions to mitigate bias and foster trust in AI.

For Module 2, we focus on how digital integrity and cybersecurity need to be approached from a trustworthiness lens. Digital integrity encompasses the protection of individuals' digital identities, data, and online experiences. This integrity is increasingly threatened by AI-driven misinformation and disinformation. The automation and amplification of harmful content, including hate speech and gender-based violence highlights the necessity of ethical oversight and governance of AI systems. Similarly, AI has emerged as a pivotal force in cybersecurity, acting both as a vital tool and a potential threat. Initially employed for anomaly detection and intrusion prevention, AI has evolved into a sophisticated instrument for defenders and adversaries alike in the digital landscape. On one hand, AI enhances threat detection, accelerates response times, and bolsters system resilience. On the other hand, threat actors exploit AI to uncover vulnerabilities in security frameworks, rendering cybersecurity a double-edged sword. Given these complexities, the significance of trustworthiness in AI cannot be overstated. Trustworthy AI systems must be robust, reliable, and aligned with ethical principles such as fairness, transparency, human oversight, and accountability. Trustworthiness serves as the cornerstone for safeguarding the digital integrity and security of individuals and

institutions. This report examines trustworthiness as a central governance principle, essential for mitigating harms and harnessing the opportunities presented by AI. Our approach entails defining and exploring the intersections between AI, digital integrity, and cybersecurity, identifying the dual capabilities of AI to either enhance or compromise these domains. By mapping out key challenges and opportunities, we assess how trustworthiness can function as a governance mechanism to underpin AI safety, responsibility, and accountability.

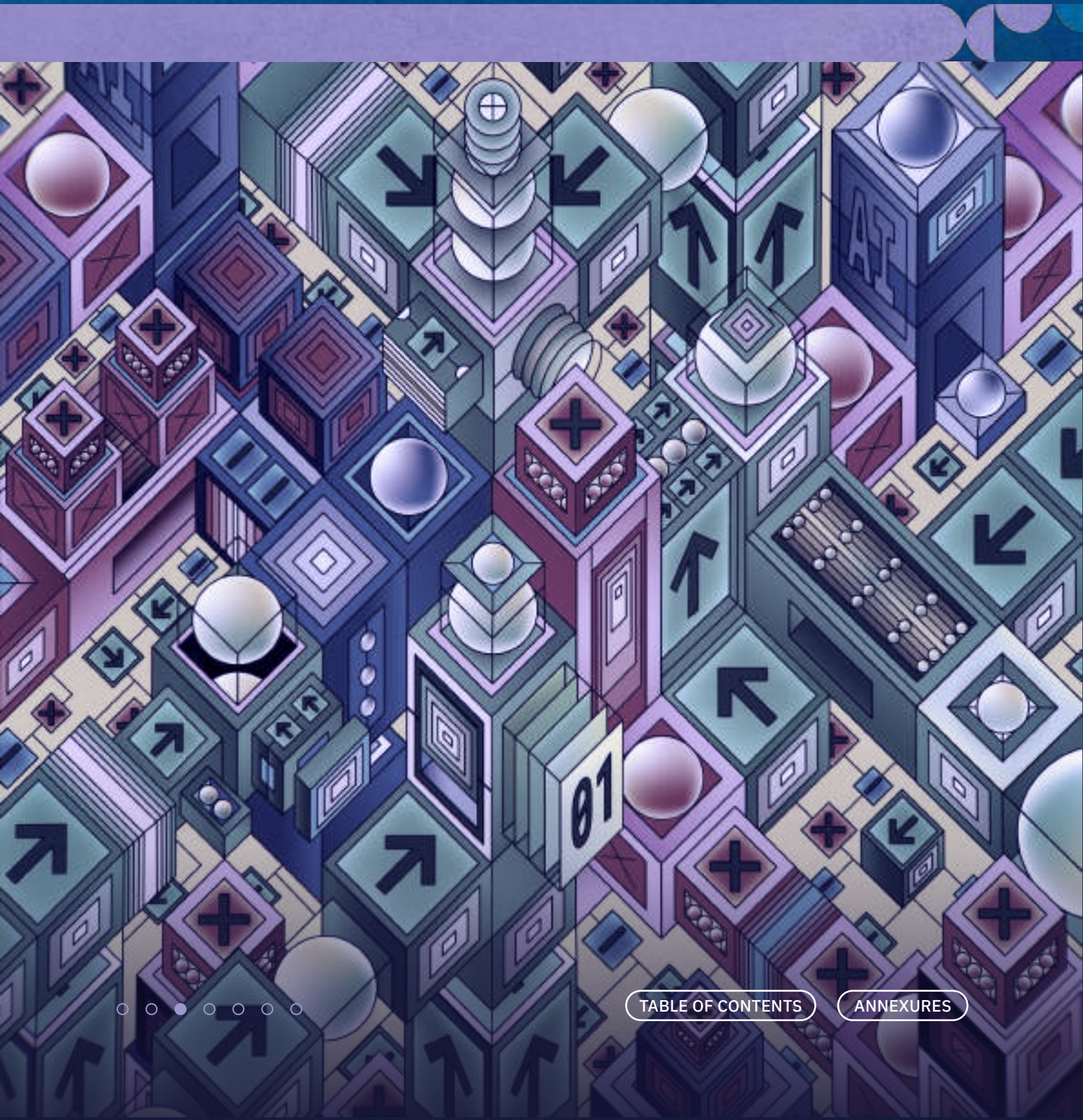
Grounded in both technical and societal considerations, this report bridges the gap between technological innovation and social responsibility. We argue that the integration of AI into cybersecurity and digital integrity efforts must be governed by a framework prioritizing trust as the foundational value. By proposing strategies for building trustworthy AI systems that align with global standards and accommodate diverse regulatory environments, we aim to foster an AI ecosystem that promotes digital safety, enhances security, and upholds the ethical principles essential for a thriving digital society.

Our analysis extends beyond module-specific inquiries to propose targeted mitigation strategies categorised into three core approaches: *Situation*, *Problematisation*, and *Resolution*. *Situation* focuses on macro-level concerns, exploring the broader contexts where governance issues emerge, such as the manifestation of challenges across the AI value chain in phases like pre-development, development, and adaptation (*Module A*), as well as at the intersection of AI with Digital Integrity and Cybersecurity (*Module B*). *Problematisation* delves into identifying vulnerabilities like biased datasets, lack of transparency in models and datasets, and insufficient benchmarking in *Module A*, alongside specific risks in *Module B*, such as the misuse of facial recognition technology (FRT), exploitation of synthetic data, adversarial AI, advanced phishing tactics, and lowered entry barriers for cybercriminals who are leveraging generative AI. Finally, *Resolution* offers tailored mitigation strategies to address these identified harms, drawing on insights from both modules to ensure actionable solutions that mitigate risks while maximizing AI's transformative potential. This structured approach forms the foundation of our work, enabling a nuanced understanding of AI governance and its associated challenges.

NOTE TO READERS

Executive Order 14110 by the Biden Administration, referred to in this report was repealed by the Trump Administration on 20th January 2025. An archived version has been added for the readers' reference.

Methodology



Methodology

AI can reflect the fullness of human innovation and potential, but it also highlights the collective challenges and ethical dilemmas we face. It embodies the synthesis of vast scientific and creative knowledge, while challenging stakeholders to think deeply about ethics and societal structures. These systems with unparalleled capabilities are not merely automating tasks or solving problems — they are redefining the very perception of knowledge, creativity and agency. The choices made in shaping and deploying AI will fundamentally define the legacy of industrial innovation and progress.

Our research on AI governance was grounded in a socio-technical approach, aiming to holistically examine the interplay between technological systems and societal impact. To achieve this, we conducted extensive secondary research and engaged with domain experts through detailed interviews. **Additionally, we employed two innovative methodologies to deepen our inquiry: a systems and design thinking approach and the establishment of a Community of Practice.** The CoP facilitated collaborative exploration among diverse stakeholders, while systems thinking enabled us to uncover how biases are embedded within AI models and systems and to identify the actors responsible for these biases. By pinpointing critical points of inflection, we devised multiple mitigation strategies to address biased algorithmic outcomes throughout the AI development and deployment pipeline. Furthermore, we incorporated design thinking principles to craft practical, user-centric solutions that promote safe and equitable AI systems.

Systems thinking and design thinking

In our attempt to examine responsible artificial intelligence, more specifically fairness in AI systems, we sought to look beneath biased algorithmic outcomes and outputs. To do so, the research unravelled critical sources underpinning bias that may occur through the process of AI design, development and deployment.

We approached this predicament with a system's thinking mindset. Such an approach allowed us to uncover what and how biases are entrenched in AI models and systems and who is responsible for them. By identifying such points of inflection, we were able to devise multiple mitigation strategies across the system of development and deployment of AI that could aid in limiting biased algorithmic outcomes, with a design thinking approach. The coalescence of both approaches was beneficial to see the problem of bias in AI systems as a “systems issue”, while grounding its mitigation to the needs of users and the feasibility of technology.

Systems thinking: Scope and application

Systems thinking is a methodological approach which by looking at the constituted elements of a system, tries to decipher the essence of the system.⁶

In doing so the question that a system thinking framework hopes to answer is how a system works as it does.⁷ It is an attempt to untangle the complex interrelationship across several elements, the approach postulates six core concepts.⁸ At its foundation is **interconnectedness**, which moves away from a linear mindset to a circular one, recognising that all elements rely on one another to persist. Closely tied to this is **synthesis**, which focuses on combining multiple components to create something new by simultaneously understanding the individual parts and the whole, emphasising relationships and dynamics rather than breaking complexity into isolated pieces, as in analysis. **Emergence** further highlights the unique outcomes or properties that arise when different parts of a system interact, producing results greater than the sum of their parts. **Feedback loops** are another vital concept, as they underscore the continuous and reciprocal influence elements have on one another, making it crucial to understand their types and dynamics to effectively analyze, interpret, and intervene. Central to feedback loops is **causality**, which involves identifying how actions or events lead to changes within an evolving system. Finally, **systems mapping** provides a powerful tool for visually representing a system's components and their interactions, uncovering insights that guide decisions and enable meaningful change. Embracing a system's thinking approach can be essential to understanding AI systems, especially how biases get entrenched

and flows through the development and deployment of AI systems. Allowing us to look at complex interplays and interconnections with which we can recognise how and when biases embedded into AI, creating reinforcing cycles, perpetuating inequalities.⁹

While having identified systems thinking as a methodological approach to investigating fairness in AI, we are also required to define what the system is. **To that end, we adopted a socio-technical ontology which defines technological systems, as an amalgamation of both the technical and social systems.**¹⁰ **The argument is that any technological system while having technical infrastructure and parts, is embedded and operated by humans and their social systems.**¹¹ Thus it is implied that, to effectively decode any technical system, we need to look at both the subsystems. We see artificial intelligence is also one such socio-technical system, which is developed and deployed through various actors and stakeholders taking part in and informing various elements which come together to formulate AI systems.¹²

Defining artificial intelligence as a socio-technical system, we investigate the AI system using a system's thinking approach. To identify the elements of the AI system, we borrow from the value chain ontology,¹³ in line with our social-technical definition of AI systems. The value chain approach helps us identify various processes which add value to the various elements of AI (both technical and non-technical) and actors who are responsible for these processes (both active and latent). Mapping these elements, we establish how the development of AI is shaped by the interconnected synthesis of active and latent processes and actors across the AI lifecycle. This interconnectedness produces feedback loops and casualties, which can showcase the emergence and propagation of bias across various stages of the AI value chain, between the different elements of the AI system. These insights enable us to build a comprehensive bias framework, which answers three key questions: How do such biases originate and persist in AI systems? *Where* in the AI lifecycle do they get entrenched? And *who* is responsible for it? Mapping these elements along with their interrelationships which produces biased AI outputs helps us arrive at a deeper understanding of bias in AI, allowing us to have more process-informed strategies for bias mitigation.

While we recognized the importance of adopting a systems perspective to understand how bias is integrated into AI systems, for its mitigation, we found that broad, aggregated approaches to reducing risks are often less effective.¹⁴ Instead, a disaggregated approach offers greater contextual awareness, enabling targeted interventions at the design level. This allows for more precise mitigation strategies that can address specific issues within their unique contexts, ultimately leading to stronger impact.¹⁵ To implement disaggregated bias mitigation effectively, we apply design thinking. The combination of these methodologies has already demonstrated its potential in addressing complex real-world challenges in other spaces,¹⁶ suggesting its suitability for tackling bias in AI systems.

Design thinking: Scope and application

Different from a system thinking approach, **design thinking is an approach anchored to a human-centred design perspective, where problem-solving is grounded in a solution-oriented and user-centric point of view.**¹⁷ Design thinking exists at the intersection of three key considerations: *desirability*, *feasibility*, and *viability*.¹⁸ ***Desirability*** focuses on understanding human needs to ensure solutions are meaningful, involving empathizing with stakeholders, identifying pain points, and prioritizing ideas that align with their expectations. By centering the human experience, desirability ensures the solution resonates with those it serves. ***Feasibility*** examines the practicality of implementing a solution, considering technical capabilities, resources, and constraints, evaluating whether an idea can be developed and deployed using attainable technology. ***Viability*** assesses the long-term sustainability and impact of a solution, ensuring it aligns with organizational goals and economic realities, analyzing scalability and the potential to be maintained over time. Together, these considerations form the foundation of a balanced and effective design thinking approach. The fruitfulness of design thinking for AI development has been well established to improve user-centricity, comprehending user wants, and inclusivity.¹⁹ In this research we try to take advantage of design thinking to understand the landscape of fairness and bias mitigation in the AI value chain, and gauge the need of all the actors involved in the development and deployment of AI. Thus to formulate our mitigation strategies we had to answer

what is desirable, and what is feasible and viable for actors at different phases of the AI value chain. To answer these questions, we used a mix of both primary and secondary qualitative research.

We conducted two community of practice convening's and one roundtable discussion, with developers, policymakers, multilateral and government institutions and civil society organisations. These discussions were specifically directed towards first analysing how the problem of bias in AI manifests itself in the real world, and then to understand what possible approaches to solving them would look like. With the presence of a diverse group of stakeholders in our convening we were able to capture multiple perspectives and approaches to deal with the risks of bias. Additionally, we also conducted interviews with both academics and practitioners working in the field of artificial intelligence and its governance, with a sector-specific lens. While we were unable to conduct user-centric research,²⁰ we used documented instances of biased AI systems to understand the impact it has on the end-users.

In distilling these perspectives and information, we were able to answer what is needed to make AI systems fair and keep users safe, while also understanding what is technologically possible, and how fairer and responsible development of AI can be made viable for the long-term. Based on our findings, we have devised mitigation strategies, which speak to how the current AI lifecycle operates while being grounded in the socio-technical realities of what is tenable and sustainable.



LIMITATIONS

While we do apply a design thinking approach to devise mitigation strategies, we were not able to conduct extensive user-centric research to capture how users of artificial intelligence approached and interacted with biased AI systems, and what are the challenges, opportunities and needs that emerge therein. Such an exploration in particular was difficult to formulate and conduct given that the deployment and user interaction with AI remains opaque and unexplainable,²¹ often due to the inherent information asymmetry,²² making it difficult to identify a comprehensive user base.

The Community of practice approach

The emergence of artificial intelligence (AI) as a powerful force in today's world has brought with it complex governance challenges. These challenges encompass issues of safety, accountability, fairness, and ethics. While regulatory frameworks have been evolving to establish accountability mechanisms and protect individual rights, these approaches are often limited by their focus on compliance, obfuscating complex concerns around stakeholder participation in AI governance in ways that devolve agency to a variety of actors. This is where the Community of Practice (CoP) approach becomes indispensable as a methodology.

Beyond Compliance: Reimagining AI Governance with CoPs

Traditional AI governance structures, largely shaped by regulatory bodies, have understandably leaned toward compliance-driven models. These frameworks aim to establish legal parameters and set standards that organisations must follow, ensuring AI systems are aligned with societal norms, legal requirements, and ethical principles. However, compliance is often reactive and does not always address the need for proactive engagement, especially at the community level.

The Community of Practice approach moves beyond mere compliance to conceptualise AI governance as a more dynamic and inclusive process. It positions stakeholders including technologists, civil society, regulators, and impacted communities as active participants in shaping AI ecosystems. By embedding individual and collective empowerment at its core, CoPs provide a space where these diverse groups can not only exchange ideas but also collaborate in shaping AI systems that reflect a range of needs, values, and aspirations.

Key Elements

CoPs as a methodology are built around a few essential elements, which contribute to the transformative role they can play in AI governance:



1. DOMAIN

The domain is the central, unifying focus that brings members of a Community of Practice (CoP) together. It represents the specific area of interest or concern that all participants are passionate about and committed to addressing. In the context of AI, this domain could range from critical issues like responsible AI, AI safety, or the ethics of AI deployment. By focusing on a common domain, the CoP aligns members' efforts and fosters collaboration around a shared goal, ensuring that discussions are relevant, purposeful, and geared toward addressing real-world challenges within the AI ecosystem.



2. DESIGN

In the context of a CoP, design involves creating structured yet flexible frameworks that allow for dynamic problem-solving while accommodating the diverse needs of stakeholders. The design of a CoP goes beyond superficial collaboration; it is about establishing intentional processes that guide participants toward innovative solutions. This can involve curating unique methodologies for inquiry, setting up collaborative environments for multidisciplinary thinking, and designing iterative feedback loops that refine ideas through collective input.



3. DIALOGUE

AI governance demands expertise across technical, social, economic, and ethical domains, making multidisciplinary collaboration essential. Communities of Practice (CoPs) excel at integrating diverse perspectives, bringing together stakeholders such as data scientists, engineers, human rights advocates, policymakers, and impacted communities. By fostering trust and mutual engagement, CoPs create a space for cross-disciplinary dialogue where knowledge is exchanged, and solutions are co-created. CoPs build shared practices—a dynamic repository of tools, strategies, and frameworks for safe, fair, and transparent AI deployment. These include ethical guidelines, governance principles, and methods for addressing challenges like algorithmic bias.



4. DISSEMINATION

One of the primary roles of CoPs is to document and disseminate the knowledge generated through their collaborative efforts. This collective knowledge forms a living repository of best practices, which can be adapted and used by other stakeholders in the AI ecosystem. This process of knowledge sharing contributes to the standardisation of governance approaches while allowing for customization based on specific contexts.

Future of COPs on AI regulation

We organised three convenings for our Community of Practice (CoP) to foster dialogue and co-create solutions for responsible AI governance. The first convening served as a launch event, where we outlined our agenda and scoped the inquiry, setting the stage for collaborative exploration. Our second workshop delved into AI's relationship with digital integrity and cybersecurity using a Roles, Harms, and Opportunities (RHO) framework. This structured approach enabled our diverse group of stakeholders including developers, policymakers, multilateral organisations, government institutions, and civil society organisations—to actively engage in the conversation and share insights and examples from their experiences in AI safety and governance.

Our third convening was a full-day session dedicated to examining how bias in AI manifests in real-world applications and identifying potential approaches to address these challenges. These discussions, enriched by the perspectives of a wide array of stakeholders, provided a holistic understanding of the risks associated with biased AI systems and strategies for mitigation. Additionally, we conducted interviews with academics and practitioners specialising in artificial intelligence governance, with a particular focus on its intersection with cybersecurity and sector-specific applications. While we were unable to conduct direct user-centric research, we relied on documented examples of biased AI systems to evaluate their impact on end users.

Through these convenings and supplementary interviews, the CoP has demonstrated the value of leveraging collective intelligence to address pressing challenges in AI governance. By maintaining an inclusive and participatory approach, we have taken significant steps toward fostering equitable, transparent, and trustworthy AI systems that respond to the diverse needs of communities.

MODULE I

All About the Bias



All About the Bias

Anchoring the research

The AI value chain

Any discussion on the responsible use and development of artificial intelligence (AI) will be remiss if the technology being discussed is looked at as a whole and not a result of multiple stages and processes. **An AI model is a result of multiple layers of processes, making it imperative to break down these stages in order to unpack the sources that give rise to bias within such models.** AI technologies in several myriad ways reflect and replicate the fault lines that exist in our society.²³ These errors may creep in at various stages of development of an AI model, be it a lack of representation in a dataset or stereotypes being embedded in an algorithm.²⁴ These fault lines once embedded in AI technology cause several societal harms, bias and discrimination being one of them. To further complicate the issue, the stages involved in the development of an AI model are complex, interconnected and exist in social, political, and economic contexts. Hence, breaking down AI technologies into their stages provides an integrative approach for researchers and policymakers to understand and intervene in AI technologies across different models, contexts, and sectors.

Approaches to breaking down AI systems

There are two significant approaches²⁵ which provide the theoretical framework to break down an AI model into the stages and resources required to build the final output. First, the value chain approach looks at the structured network of various processes for resource input and resource output received from AI systems. These processes are situated in a context and are interconnected. Furthermore, the value chain approach takes into consideration actors participating in these processes and their interdependencies. The value chain approach has a service-

dominant logic that focuses on intangible and tangible activities and necessarily takes into account the interconnectedness of these activities. Additionally, the value chain approach holds relationships between actors and co-creation of value at different stages as central, coupled with a focus on social, political and economic context of these activities and relationships.

Second, the supply chain approach refers to the stages and processes involved in making an AI model usable and consumable by end users. It follows a good dominant logic focusing on tangible activities and holds the final output as the focal point of all activities.²⁶ Moreover, the supply chain approach is linear in nature which means it only recognises one way movement of resources for the production of an AI system/ model. For the purpose of this research, we have adopted a value chain ontology as the empirical foundation to unpack sources of bias in AI systems.

Significance of the value chain approach

The value chain approach takes into account the interrelationships between stages and actors in AI development, as well as the patterned nature of the transactions taking place between various actors, to provide a comprehensive analytical framework.²⁷ For example, a supply chain ontology will be limited to the downstream flow of datasets from data owners, to model developers to application developers to end users. However, the value chain ontology will be able to take a step further and take into account the upstream flow of data and further training of AI models based on user interaction.

Additionally, tracing actors at each stage, ranging from chip manufacturers, cloud service providers, data collectors and annotators to end users, allows for clear attribution of sources of bias and allocation of responsibility for mitigation strategies.²⁸ For example, in a value chain ontology different actors at the stage of data collection can be traced, such as a platform collecting data from its users. In this case, the actors will be users as well as the platform. The value chain approach further provides for the framework to take into account the pattern of data flowing between these two actors, the business model and organisational structure, its impact on the downstream activities, and other social

and cultural factors. However, a supply chain approach will only take into account the transaction of the data being collected, as a disjointed part of a larger supply chain. This demonstrates that a value chain approach provides a more well rounded framework for the allocation of accountability within the matrix of transactions, relationships, and actors for techno-social problem-solving.

The value chain approach further enriches the discourse by taking into account the social, economic, political, and technical contexts. Crucially, in the context of AI technologies, geopolitics and tensions between global and domestic interests,²⁹ the technological gap between global north and south,³⁰ market competition and regulation,³¹ the concentration of AI capabilities with the private actors as compared to the nation states,³² social hierarchies and power imbalances,³³ economic resources, governmental impetus and political environment, international and domestic legal landscape play a key role.

For example, compute is a critical foundational resource for large AI models, however, at the same time it is highly scarce and controlled by a few big players such as, Taiwan Semiconductor Manufacturing Company, a chip fabricator, Nvidia,³⁴ a chip designer and cloud infrastructure providers like Google, Amazon and Microsoft. As countries attempt to develop AI technologies faster than ever, most of these players are exposed to these geopolitical tensions in forms like export restrictions. This significantly impacts the accessibility of computational capabilities and who gets to develop AI models. Another key contextual factor directly contributing to bias in AI technology is the lack of diversity among AI developers, which is a result of social hierarchies and power imbalances.

A critical analysis framework must provide room to consider these contextual factors to be able to meaningfully engage with the challenges at hand. The value chain approach provides an integrative approach to not only identify and unpack the sources of bias but also provides a framework to comprehensively think about the allocation of responsibilities and by extension of mitigation strategies.

Landscaping the AI value chain

The value chain framing for critical analysis of AI technologies and allocation of accountability is fairly new in the responsible AI discourse.³⁵ However, there is significant literature focusing on breaking down the relevant stages of development of AI models, important players participating at each stage, and other relevant factors impacting each stage. This section aims to landscape the existing literature and lay ground for the value chain framework adopted for this report.

Typically, the AI value chain is broadly divided into three stages: pre-development, development, and adaptation. Each of the stages are distinct phase of the AI systems lifecycle, encompassing various processes and actors that are part of it;



Pre-production: This stage extends from the design and planning of models to the pre-training of foundation models, which are used in the next phase of the development cycle as base models for further training.

- i. **Processes:** The first stage of the AI lifecycle begins with the identification, elucidation, and formulation of the problem.³⁶ This involves **planning and designing** the AI system, by setting out the objectives, and the typology of the strategy to develop the AI system. Further within this stage, the security risks, and ethical and legal conformity of the selected approach are also tested. This process is followed by **data sourcing**, a foundation step in the AI lifecycle. Collecting or sourcing data from diverse sources is essential for training, validating, and testing AI models. The accuracy and effectiveness of an AI system are directly impacted by the quality and quantity of the data gathered.³⁷ With the advent of foundation models, most AI systems built use one or more such foundation models.³⁸ These models are generally pre-trained models, on large datasets and can be further fine-tuned to perform specific tasks.³⁹ **Pre-training** of foundation models thus becomes a key part of the AI value chain. Further, these **foundation models are released** for use by downstream actors. The release strategy of the model (open-source or API), determines not only how foundation model providers monetise their

models but also determines the nature of power distribution between them and subsequent downstream actors.⁴⁰

- ii. **Actors:** The pre-development phase includes multiple upstream actors who inform and make decisions on how the system is built at each process. One key group actor is **enterprises** that commission AI systems. Depending on the business typology and sector, enterprises are embedded differently within the AI value chain. **Data providers** are a key part of the AI value chain. They provide datasets, which are either intended or not intended for model training and validation.⁴¹ Lastly, foundation model providers have become key actors in pre-development, with the foundation model becoming the base for most AI systems being developed. **Foundation model providers** are again embedded in the AI value chain depending on their business typology, while some model providers have restricted access allowing others to use their models with APIs, or others provide open source models.⁴²



Development: This stage extends from the selection of foundation models, which are then further refined with additional data and training methods to perform some specific tasks. These models post fine-tuning are verified and validated for their accuracy and functionality.

- i. **Processes:** The next phase of the AI lifecycle entails the creation of an AI model to perform a specific task. The phase begins with downstream actors **selecting relevant foundation models** given their use case. The choice in the foundation model is key, governed by not only the functionalities that are required but also needs to be contextualised to the data that it is trained on and its geographic and demographic contexts of deployment. In addition to selecting foundation models, this phase also involves curating, cleaning, and labelling the collected data for training an AI model. This process is key in refining the data, to make the AI system better equipped to learn, predict, and make decisions, ensuring a higher level of accuracy and reliability.⁴³ It is also a key point in the AI

value chain to remove biases and inconsistencies in the dataset. **Model training and fine tuning** is the next process in the AI lifecycle, which involves creating an algorithm or layers of algorithms that best suit the task at hand. The resultant model is exposed to prepared datasets, to perform specific tasks or make judgments. The model identifies patterns and relationships within the data, which enhances its ability to make accurate predictions or decisions based on new information.⁴⁴ Once the training phase is complete, the AI model's performance must be **validated and verified** using a separate validation dataset. This dataset, which the model has not encountered before, is used to test its ability to make accurate predictions.⁴⁵

- ii. **Actors: Model adapters and optimizers** are key actors in the development phase who perform multiple tasks to operationalize the AI model. They are generally AI developers, who design the model, train, and validate the model, by designing the architecture of the model, deciding on the mode of training, fine-tuning internal parameters for better accuracy, and analyzing the accuracy of the model.⁴⁶ Another key actors in the development phase are **data labelers and data annotators**. As models are trained on large datasets, their role becomes increasingly vital. They ensure that the data provided to models is well-organized and accurately labelled, which contributes to creating more effective and reliable models.



Adaptation: In this last phase of the AI lifecycle, where AI models are either integrated into existing digital systems or new digital systems are made to be deployed for users. Subsequently, post-deployment they are monitored and evaluated to test their real-world performance.

- i. **Processes:** This stage typically entails deploying a model into a user-facing application. Once the model is verified for its accuracy, **model integration** links various models to each other to perform a particular set of tasks. One or more models are then further integrated into the product environment, interacting with new data, making predictions, and delivering results in real time.⁴⁷ At this stage, the AI system is deployed for **user interaction**.

Following the deployment of an AI model, ongoing **monitoring and evaluation** are essential to ensure its continued optimal performance. This involves regularly assessing the model's predictions using relevant metrics and feedback. If there is a noticeable drop in accuracy or effectiveness, it indicates that the model needs to be refined or retrained, thus closing the lifecycle loop.⁴⁸ However, evaluation can take place across the AI value chain.

- ii. **Actors: Model integrators** are key actors in this stage of the value chain, who use different AI models to build AI tools to provide it to the end-user. The end-users within the AI pipeline are different depending on the objective and purpose of the AI model. **Enterprises** can be end-users of the AI models, when they use it within their workflow,⁴⁹ for example, a hospital using AI tools for administration and management. AI models can also be directly available to **consumers** where a consumer is an individual interacting with an AI product or service,⁵⁰ for example, AI tools for assisted learning are directly provided to students. Further consumers can also interact with AI models via interacting with **platforms** that host AI-generated content⁵¹ (Eg. AI-generated content on social media) or use AI as part of the service they provide (Eg. Suggestive AI models on YouTube). Lastly, **MLOps and evaluators** are actors providing tools and services for performance evaluation, auditing, safety assessments, etc, across multiple stages in the value chain.⁵²

KEY CONSIDERATIONS FOR THE VALUE CHAIN ONTOLOGY

1. The value chain may differ based on the typologies of the business models- for example, an AI model may be developed and used by the same entity or an AI model may be developed by one entity, bought by another but used by a third party.⁵³ In such a case the stages and the actors at each stage may vary.
2. The value chain may also differ based on the specific AI model under consideration. To elucidate, a generative AI model may have a different value chain than an expert systems value chain.⁵⁴
3. The value chain may also differ based on the specific sector or use case. That being said, a broad value chain ontology with clearly defined stages and actors can be dynamic and modular enough to be used for different typologies, AI models, and sectors.

Foundation models

A foundation model is a model pre-trained on a large amount of data, capable of a range of general tasks such as interpret and mimic human language or images, in some cases AI models multi-modal having the ability to do both.⁵⁵ It can be further trained through fine-tuning and other training methods to perform a wide variety of downstream tasks and applications.⁵⁶ Some of the key characteristics that define them are,

- Foundation models are trained using large datasets with many parameters, which allows them to capture intricate patterns in the data and easy scalability to perform a wide range of tasks
- Along with the need for large datasets, training foundation models also require subnational computational resources for both training and inferences
- Foundation models generally go through two steps of training, initially, they are trained on broad datasets to general features and patterns and subsequently can be fine-tuned on specific tasks using relatively smaller datasets
- Foundation models can be versatile in application as they can handle multiple tasks without needing retraining from scratch. Knowledge gained from one task can be easily applied to another task

Depending on the type of foundation model, it can be capable of a variety of tasks and applications, having the ability to interpret different modalities of data inputs, like texts, images, videos, and even audio.

Foundation models differ from other narrow models of AI, given that the latter are trained and can only perform specific tasks, trained on specific datasets, and are not designed to be used beyond their original purpose.⁵⁷ In contrast, foundation models provide a generalist architecture that can be adapted to a wide range of tasks.

MODEL TYPE	FUNCTION	EXAMPLES
Large Language models	LLMs can perform a variety of natural language processing (NLP) tasks. They are designed to understand and generate text that mimics human responses	GPT-4 BERT LLaMA
Vision models	Vision models use architectures like neural networks and methods like deep learning, allowing AI to understand interpret and respond to visual inputs	CLIP
Multimodal models	Multimodal AI are models capable of processing and integrating information from different modalities. These modalities can include text, images, audio, video and other forms of sensory input	DALL-E 3 Flamingo
Domain-specific models	Some foundation models are specialized for domains like healthcare, finance, or law, pre-trained on relevant data to support developers and researchers in those fields	MedPaLM FinBERT LegalBERT

PERILS OF THE CHANGING LANDSCAPE OF AI DEVELOPMENT

The rapid evolution in capabilities of large-scale foundation models, in their abilities to perform a variety of tasks and their capacity to be adapted to perform highly specific and complex tasks, has not only driven AI adoption across sectors but also shifted the AI value chain to a new trajectory.⁵⁸

Monopolising tendencies: The large-scale resources needed to build foundation models, have concentrated the market in the hands of few players.⁵⁹ While the fixed cost of developing foundation models remains high, the marginal cost of deploying them remains low.⁶⁰ This has given first movers an advantage in easily deploying their models, with relatively lower cost, as there is an uptake in their usage while disincentivising new players to build foundation models at scale.⁶¹ This is aided by the fact that there remain significant entry barriers to developing foundation models, in terms of the large-scale investments needed for computational infrastructure, talent and data.⁶² These tendencies

of market concentrations also produce a distinct power relationship between upstream and downstream actors across the AI value chain.⁶³ With foundation models becoming a key for further AI innovation, foundation model providers can dictate access through various business typologies (API access or open-source models).⁶⁴ These tendencies of monopolisation along with the power relationships have been pointed out to be a major source of concern.

Amplifying bias and societal risks: Like any AI system, the foundation model can yield inequitable outcomes, compounding historical existing inequities. While these border questions of algorithmic fairness and AI ethics need to be addressed across AI applications and AI models, the harms of an unfair foundation model have the propensity to intensify at a larger scale, given their pervasive usage in the development of downstream AI applications. This results in two types of harm, intrinsic harms, which are biases of foundation models which affect downstream applications, but also extrinsic harms, which are harms arising out of contexts of specific downstream applications, when foundation models are adapted. These risks are further heightened when we look back at their market concentration tendencies. Given the centrality of the foundation model in AI development along with the homogeneity of foundation models, where few models are reused for many applications, foundation models become a singular point of failure, where issues in these models can spread harm across numerous downstream applications.⁶⁵

Our Approach and its limitations

While there are concerns of a monopolising market, what is more important for our investigation of bias are the societal risks that foundation models can accentuate, given their market trajectory. The central role of foundation models in AI development, along with their tendency toward homogeneity, inadvertently makes them epistemically critical in our exploration of bias in AI systems and in developing effective strategies to mitigate its impact.

It is also important to note that there are other harms that are being currently exacerbated by the availability and access to foundation models like misusing the capabilities of foundation models to produce high-quality content for harmful and malicious purposes for cheap. Furthermore, there also remain important concerns regarding data rights and the environment, which have been flagged to be in considerable danger, as we try to build larger foundation models. These risks, while not situated in our conversation about bias in AI, are important questions for further research and policy regulation.

Sectoral considerations

The integration of AI across various sectors has introduced a complex landscape of bias manifestations, necessitating sector-specific analyses and tailored mitigation strategies. Adopting the sectoral approach acknowledges the differential impacts on individuals, contingent upon the type of AI technology deployed within each sector.

AI technologies are proliferating across sectors, however, for our research, we are limiting the scope to four sectors. The selection of these four sectors was based on certain criteria. Priority was accorded to sectors under heightened regulatory scrutiny, indicating perceived high-risk areas requiring AI governance. Additionally, we surveyed sectors exhibiting an increased adoption of AI and where there is a predicted expansion of the market. Understanding which sectors are more susceptible to bias allows for proactive intervention strategies, mitigating adverse impacts on affected populations. This comprehensive approach enables targeted interventions to preemptively address sources of bias and develop strategies for mitigation. By being cognizant of these sector-specific biases and their implications, our study aims to foster responsible AI deployment and mitigate societal risks, facilitating the ethical advancement of AI technologies across various sectors.



Healthcare

Artificial intelligence holds immense promise in revolutionising healthcare, particularly in enhancing patient outcomes and streamlining clinical processes. By employing algorithms to analyse vast amounts of data, AI can offer insights that facilitate quicker and more accurate diagnoses and treatment plans.⁶⁶

There has been an increase in AI integration in healthcare in India due to a shortage of qualified healthcare professionals and unequal accessibility to healthcare across India.⁶⁷

This potential is illustrated through IBM's utilisation of machine learning to detect diabetic eye disease at an early stage, showcasing how AI can augment diagnostic capabilities.⁶⁸ AI in

Healthcare Market is predicted to grow from \$14.6 billion in 2023 to \$102.7 billion by 2028.⁶⁹

The integration of AI in healthcare is not without its challenges. One significant concern is the potential for bias in AI algorithms, which can exacerbate existing health disparities among different demographic groups.⁷⁰ If left unaddressed, AI systems may inadvertently perpetuate historical patterns of discrimination based on factors such as race, ethnicity, gender, age, or disability status.⁷¹ This emphasises the importance of ensuring that AI technologies are developed and implemented with a keen awareness of these potential biases.⁷²

Research has revealed inherent biases in simplistic prediction rules for heart disease, historically employed in routine medical settings across industrialised nations.⁷³ Notably, the Framingham Heart Study's cardiovascular risk score exhibited robust performance among Caucasian patients but was unable to perform well when applied to African American individuals.⁷⁴ This disparity implies the potential for unequal distribution of care and inaccuracies in diagnosis and treatment within healthcare systems.⁷⁵ Consequently, there has been a surge in efforts aimed at governing the use of AI in healthcare.⁷⁶ These regulations seek to mitigate the risks associated with AI bias and discrimination,⁷⁷ while also promoting transparency and accountability⁷⁸ in AI-driven healthcare systems.^{79w}



Finance

The financial sector's investment in AI is poised for substantial growth, with the International Monetary Fund projecting a more than doubling of spending to \$97 billion by 2027, reflecting a compound annual growth rate (CAGR) of 29%. While fintech-enabled financial services have already had a transformative impact on financial institutions, AI is revolutionising how financial institutions operate.⁸⁰ AI models execute trades swiftly and precisely, leveraging real-time market data to uncover insights and guide investments. By analysing complex transaction patterns, AI enhances risk management in areas like security, fraud detection, and compliance. It also transforms customer engagement by predicting behaviour,

refining credit scoring, and enabling personalised interactions, leading to faster support and innovative financial products.⁸¹

One significant application of AI in banking lies in the assessment of creditworthiness. Traditional credit scoring models often rely solely on credit history, which may not provide a comprehensive view of an individual's financial reliability. AI algorithms leverage alternative data sources such as utility payments and consumption patterns to assess creditworthiness. This approach not only expands financial inclusion by catering to individuals with limited credit histories but also enhances risk assessment capabilities for banks.

Despite AI's benefits, concerns regarding bias in the finance sector persist. The complexities of AI algorithms make it challenging to fully understand and mitigate biases effectively. The risk of unintended bias, discrimination, and financial exclusion is particularly pertinent for consumers with protected characteristics or vulnerabilities. This highlights the importance of ongoing research, transparency, and regulatory oversight to address and mitigate potential biases in AI-driven banking applications.

The regulatory landscape surrounding AI adoption in insurance has witnessed heightened scrutiny.⁸² For instance, the reserve bank of India in 2024 established a committee to develop a Framework for Responsible and Ethical Enablement of Artificial Intelligence (FREE-AI) in the financial Sector for India.⁸³ This is part of a larger regulatory focus on the use of AI in this sector in India. Agencies like the securities and exchange board of India (SEBI) have also called for more responsibility on SEBI-regulated entities for the use Artificial Intelligence (AI) tools developed internally or third-party.⁸⁴ Regulators are increasingly focusing on ensuring AI systems' accountability, and bias mitigation mechanisms in AI-driven risk modeling, rigorous data validation, algorithm transparency, and ongoing ethical evaluations to ensure fair and accurate risk assessments.⁸⁵

If AI algorithms are not carefully calibrated and monitored they will inadvertently perpetuate biases present in historical data, leading to unfair outcomes for certain demographic groups. As a result, there is a growing interest in regulating AI's use in banking to ensure fair and ethical AI deployment, promote consumer trust, and mitigate potential risks associated with bias and discrimination.



Education

The integration of technology, particularly AI, has sparked a transformative wave in the education sector, revolutionizing traditional teaching methodologies and expanding the reach of educational resources. This new era of improved accessibility, enhanced execution of educational programs, and personalized learning experiences was amplified through the advent of classes being taken online post-COVID-19.




There has been a rapid uptake of AI-enabled learning management tools, with 47% of such tools expected to be AI-enabled by 2024.⁸⁶ Moreover, the AI market in education is projected to witness a remarkable Compound Annual Growth Rate (CAGR) of 40.3% between 2019 and 2025.⁸⁷

The widespread adoption of AI in education also raises critical concerns regarding its potential adverse impacts. The deployment of AI-driven systems in educational settings carries inherent risks of bias, leading to discriminatory outcomes and exacerbating inequalities.⁸⁸ For example, AI-driven essay grading systems can inherit biases from their training data, potentially reflecting the subjective judgments of human evaluators.⁸⁹

One significant risk stems from AI's ability to adapt learning experiences by adjusting the pace of the curriculum for students. If these adaptations are based on incomplete or biased data, erroneous assumptions about learning, or inadequate theories, they could perpetuate existing achievement gaps or even widen them. This highlights the imperative of ensuring that AI algorithms used in education are transparent, accountable, and free from biases to promote equitable learning opportunities for all students.

In response to these challenges, there has been a notable uptick in regulatory efforts aimed at governing the use of AI in education. UNESCO has issued press releases urging governments to implement regulations for AI in schools, emphasizing the need for ethical AI deployment and safeguarding student rights. Additionally, the Council of Europe has convened conferences specifically focused on regulating artificial intelligence in education, signalling a concerted global effort to address the complex ethical and regulatory dimensions of AI integration in educational settings.

Regulatory Focus, market spread and risk of bias across sectors

SECTOR	REGULATORY FOCUS	MARKET SPREAD	RISK OF BIAS
Finance 	<p>India is actively developing regulatory frameworks for AI in finance, focusing on data privacy, security, and ethical considerations. RBI has expressed <u>concerns</u> about the potential risks AI poses to financial stability, emphasizing the need for proper risk mitigation practices by banks</p>	<p>The financial sector in India has rapidly embraced AI to improve customer experiences, enhance risk management, and streamline operations. <u>Reports</u> have indicated significant AI adoption among financial services</p>	<p>AI in finance carries risks of bias, particularly in <u>lending and credit scoring</u>, where flawed data can <u>exacerbate inequalities for low-income and minority borrowers</u>. These biases can lead to discriminatory practices that <u>challenge</u> existing legal and ethical frameworks for fair lending</p>
Health 	<p>While there are not many regulatory frameworks in place to govern the use of artificial intelligence (AI) in healthcare right now. There is an emphasis on focusing on <u>data privacy, security, and ethical considerations</u> of AI in health. Both the draft DISHA act and the DPDP act, emphasise the need for patient data privacy, while the NHS stack and the Niti Aayog national strategy reflect on AI's ethical use in health</p>	<p><u>AI solutions for healthcare in India</u> are at an early stage, with most use cases still in development or testing, particularly in clinical interventions. Current applications focus on decision support systems, process optimization, and virtual assistants, with advancements in areas like disease detection, diagnostics, and patient-facing applications</p>	<p>The <u>risks of bias in AI for healthcare</u>, include unequal representation, flawed training data, systemic inequalities, and algorithmic misinterpretations, which can lead to disparate healthcare outcomes and amplify existing disparities</p>
Education 	<p>AI's role in education is recognized, but regulatory frameworks are still evolving. The focus has been on integrating AI to enhance learning outcomes, with <u>less immediate regulatory emphasis</u> than health and finance</p>	<p>The integration of artificial intelligence (AI) in India's education sector is <u>rapidly expanding</u>, with <u>applications</u> such as personalized learning, AI-powered tutoring, and content creation gaining traction</p>	<p>AI applications in education <u>risk reinforcing existing biases</u>, potentially affecting student assessments and resource access. Without careful design and oversight, these systems may disadvantage certain groups, impacting educational equity</p>

Bias: Meaning and implication

Fairness and Bias in AI

Conversations around fair AI or fairness in AI have focused on bias. While fairness and bias are not necessarily the same concepts, what has been repeatedly highlighted is that bias is a major source of unfairness and discrimination in AI decision-making.⁹⁰ While the correlation between them is yet to be effectively explored, one can look at other domains to understand how bias can produce unfairness and discrimination. Specifically, looking at approaches to bias within jurisprudence and legal theory, we can see how unfairness and discrimination link to biases.

The rule against bias, considered a central tenet of natural justice, across the world's legal systems, argues for a fair procedure of decision-making but also an unbiased decision-maker.⁹¹ In articulating the need for decision-makers to approach matters with an open mind and free of prejudice,⁹² the latter asserts that bias can impede fair and just decision-making. This rule, while originating in the judicial process, has within democratic frameworks been extended to a vast range of public decision-makers.⁹³

Justice and human rights frameworks have further drawn a very comprehensive understanding of how social biases can produce structural discrimination leading to injustices. This covalence of societal biases stemming from race, gender, ethnicity, etc, and discrimination stems from the understanding that such prejudices are not only arbitrary considerations but also morally and ethically unjust.⁹⁴ The argument is that arbitrary and irrational distinctions are tantamount to discrimination since they fail to treat individuals with the equal concern and respect they deserve as autonomous human beings.⁹⁵

Furthermore, research studying bias has produced evidence on how incorrect and arbitrary biases can engender systemic social harm. Research looking at the labor market has found that employer bias has a direct effect on wages, job assignments, and promotions.⁹⁶ Gender biases in teachers have been seen to have a direct correlation to the performance of female students,⁹⁷ while societal gender biases have resulted in lower enrollment of girls over boys.⁹⁸ Such investigations assert how biases, specifically negative stereotypes and prejudices, can systematically create unfair outcomes, reinforcing existing disparities among social groups.⁹⁹

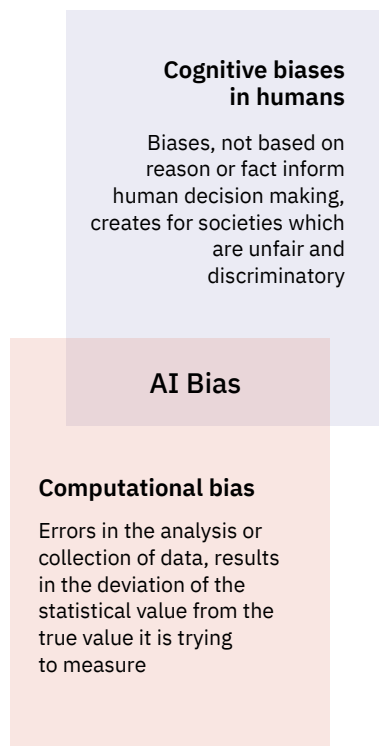
Understanding Bias

Bias as a subject of inquiry has received multi-disciplinary attention, from psychology, social psychology, statistics jurisprudence, and law. Thus, before looking at what we mean by bias within AI systems, it can be valuable to look at how other disciplines have approached the question of bias. The aim here is to gain a holistic understanding of bias across differing themes of literature.

In answering what drives human thinking and decision-making, psychology has identified that people often take ‘shortcuts’ over reasons or facts. These ‘shortcuts’ are known as cognitive heuristics or biases.¹⁰⁰ While the study of bias within psychology has categorized many such cognitive biases, at its core, all these biases influence human decision-making, leading to judgments, decisions, or actions that are not necessarily based on scientific and logical reasoning.¹⁰¹

Social psychology has further explored how such cognitive biases, including prejudices and stereotypes, impact society at large. Prejudices are unjustified negative attitudes towards individuals, generally directed based on membership to social groups along the lines of gender, ethnicity, race, and other characteristics.¹⁰² Stereotypes, on the other hand, are “overgeneralized beliefs” about particular groups.¹⁰³ While stereotyping might be value-neutral in some cases, negative stereotypes combined with prejudicial attitudes can produce fear and hostility, leading to discrimination against certain groups. Thus, cognitive biases implicitly impact not only how we interact with the world and make decisions and judgments but also dictate how others experience it, often creating a society that is unfair and discriminatory towards some.¹⁰⁴

Beyond human bias, which we look at above, we also see how biases are present in statistical computation and their effect on outcomes and inferences. Statistical biases refer to statistical values that systematically differ from the real values that they are trying to measure. Generally caused due to systematic issues in how the data is collected or in its analysis, leading to wrong conclusions and inferences drawn from such statistics.¹⁰⁵



Source: Aapti Analysis

Literature examining bias in AI has generally treated it as a technical issue and has often restricted the understanding of bias as a function of statistical phenomenon. Within this understanding, bias in AI systems is seen as a systematic inaccurate behaviour of the system resulting from computational errors. However, research that looks at the AI value chain has refocused attention on how human cognitive biases can find their way into AI systems.¹⁰⁶ It is argued that AI products which are built within social, cultural, and political situations, are influenced by human cognitive biases of actors and institutions that interact with the AI system within its value chain, which often results in cognitive biases like prejudices and stereotypes becoming encoded into the AI system.¹⁰⁷ What's more is that cognitive heuristics can also be used in heuristic algorithms to make decisions and judgments in the same ways humans use them, in solving complex computational problems with efficient solutions.¹⁰⁸ Thus it becomes necessary that when looking at AI bias, we look at both cognitive biases and computational biases, encompassing a broad range of sources, typologies, and risks posed by biased AI systems.

Most literature in defining AI bias has looked at prejudice present in AI decision-making, which can be harmful or be claimed as unfair for an individual or a group, locating bias within the decision made by the AI system.¹⁰⁹ While others have focused on defining bias in systematic technical processes which produces erroneous or unfair outcomes.¹¹⁰ In both cases, it becomes clear that in defining bias, literature has looked at the decisions produced by the AI system to find prejudices and unfairness. Espousing a similar approach, Ntoutsis et al's definition of bias, as **an inclination or prejudice of a decision made by an AI system which is for or against one person or group, especially in a way considered to be unfair.**¹¹¹ However we believe to effectively be able to tackle bias in artificially intelligent systems, the key is to look at sources of bias and further, define what constitutes unfairness. To that end, we augment over current definitions of bias, to be considered **an inclination or prejudice, produced by computational or human cognitive biases, of a decision made by an AI system that is for or against one person or group, especially in a way considered to be morally or legally unfair.**

Current AI fairness frameworks

The cross-disciplinary focus on fairness in AI has ensued in multiple frameworks that have tried to solve for bias within AI systems. While each has a distinct strategy to address bias, they can be broadly categorized into three distinct approaches:

- 1. Software toolkits:** Many stakeholders have developed toolkits to facilitate the adoption of fairness metrics to mitigate bias in AI systems. They are generally algorithmic methods to assist AI practitioners and AI fairness experts to audit and mitigate bias in models.¹¹² These toolkits, often developed by multilateral technology organizations like [Microsoft](#) and [IBM](#), as well as those independently created by researchers and tech think tanks—such as [TextAttack](#) and [XAI](#)—have been deemed beneficial for AI developers, aiding developers recognize and moderate bias within their models. However, there remain conceptual limitations. A primary issue with toolkits has been their inability to represent the complexities of distributive unfairness in the world.¹¹³ Creating a metric for fairness requires a deep understanding of the problem domain and social context, making it difficult to develop mathematical underpinnings of fairness.¹¹⁴ Furthermore, the inability of such tools to metrify abstract notions of fairness has been seen to perpetuate a narrow conception of fairness in AI developers interacting with it for the first time.¹¹⁵
- 2. Fairness principles:** Ethics principles specifically dealing with bias in AI systems are being developed by multiple stakeholders. Ethical frameworks like [UNESCO's AI ethics principles](#) or [NASSCOM's responsible AI principles](#), are generally aimed at guiding and embedding ethical thinking within the AI pipeline. However, what has been realized is that they can be hard to operationalize in practice. Principle-based frameworks have only been able to give a list of considerations, rather than a decision-making tool.¹¹⁶ Without having the ability to be operationalized, it does not help AI practitioners closely interact with the system and think through ethical dilemmas.¹¹⁷ Additionally, what has also been noted is that such frameworks tend to focus solely on the design and development of AI models, which overlook processes like deployment that can be pertinent pain points.¹¹⁸
- 3. Regulatory frameworks:** Legal and policy frameworks in the last few years have been trying to play catch up to the growing developments in AI and its application. In trying to mitigate the harmful outcomes of AI decision-making, actors have either tried to develop new regulations or have tried to extend existing statutes to encompass such harms. Some regulatory approaches show promise, such as the EU AI Act's risk-based approach to governing AI usage and efforts to interpret algorithmic decision bias under existing anti-discrimination laws.¹¹⁹ However, criticism has emerged regarding regulatory frameworks' effectiveness in addressing AI system biases. Research looking at the EU AI Act notes that there remains an overarching technocratic approach in dealing with AI harms.¹²⁰ Approaches to bias mitigation have relied on debiasing models and data, where fairness metrics are used to improve the model's performance and representation within datasets.¹²¹ Such an approach fails to accommodate other system design decisions that can be potential sources of bias.¹²²

POLICY

BIAS

India



In the latest [advisory](#) by the Ministry of electronics and information technology (MeitY), bias is referred to as unfair or prejudiced behaviour that AI models might exhibit, leading to unequal treatment or discrimination. It emphasizes that AI systems should not perpetuate bias or compromise fairness. In addition to the advisory, NITI Aayog's [national strategy for artificial intelligence](#) postulates a reactive, use-case-based approach to mitigate the harms of biased AI decision-making.

Singapore



The [Proposed Model AI Governance Framework for GenAI](#) (2024) highlights the importance of transparency in AI systems, asking for safety disclosures, including steps to mitigate bias. It also emphasizes the need for standardizing bias correction techniques and safety measures.

EU



The [EU AI Act](#) emphasizes mitigating bias in AI systems, particularly in high-risk applications. Article 10(2)(f) highlights the need for datasets to be free from bias and specifies that bias detection and correction cannot be achieved through synthetic or anonymized data. It also mandates that AI systems reduce bias risks and address feedback loops in future operations.

USA



The [Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence](#)¹²³ emphasizes addressing bias as a critical issue in AI development, particularly in sectors like housing, finance, healthcare, and consumer protection. It highlights the need for AI systems to be evaluated for bias and discrimination, ensuring fairness and equity, while protecting vulnerable populations from harm due to AI misuse. Further, the NIST's [AI Risk Management Framework](#) also talks about fairness and management of AI bias as a key lever for trustworthy AI and also has come out with a [framework](#) to set standards for identifying and managing bias.

Brazil



Brazil's [proposed bill](#) on AI places significant emphasis on bias mitigation through regular public impact assessment for all AI models and systems. The proposed bill also suggests the adoption of data management measures for the elimination of discriminatory biases and highlights the importance of up-to-date and representative datasets for training AI models for high-risk applications. The proposed bill also gives people the right to non-discrimination and corrections of any direct, indirect, illegal, or abusive discriminatory biases within AI systems.

Consequently, while fairness frameworks can chip away at the problem of bias, current approaches fail to look at bias holistically. From the survey above, we can see two key vulnerabilities, one being the inability to encompass the entire AI value chain, and secondly, a technocratic perspective that fails to comprehensively understand a predominantly socio-technical problem. In our research, we hope to address these challenges by adopting a techno-societal perspective to understand bias, by looking at sources of biases within the AI value chain.

Sources of bias in AI

Bias can be incorporated into the AI system from multiple sources. Data that is used to train AI models, the modelling of the AI system itself, or even in user interaction with the AI models, can become points through which existing biases can be baked into the AI system or new kinds of bias can get entrenched.¹²⁴

Literature studying sources of AI bias has either had a data-oriented approach,¹²⁵ looking at how data at various points in the AI pipeline, or has looked at the source of bias from their places of origin, such as humans, computational, or systemic.¹²⁶ While such frameworks hold value in understanding a myriad of sources of bias, they do not locate these sources within the AI value chain.

Locating such sources within the AI Value chain becomes important for being able to not only effectively identify how and when different sources of bias impact the system within the AI pipeline, but also aid in thinking about strategies for mitigating these biases by allowing us to allocate responsibility and accountability across actors within the value chain.

Given the importance of looking at the sources of bias from a value chain ontology, we adopt a framework¹²⁷ that posits three broad sources of biases differentiated by their nature of incidence within the AI value chain.

- 1. Pre-existing bias:** Pre-existing biases exist independently of the technology and have roots in social institutions, practices, and attitudes. Such biases can become embedded in technological systems either through conscious and explicit efforts or implicitly and unconsciously, as part of the data or the model.
 - a. Individual bias:** Individual biases and human preferences of actors with decision-making power within the AI value chain can be embedded within the AI systems from the data or the model. (*Eg. Biases of data labellers embedded in training data, transferred into AI systems.*)
 - b. Societal bias:** Bias within society at large, such as in the organization, institutions, and culture, can also become embedded within systems. (*Eg. Biases in the U.S. criminal*

system lead to more African-American convictions, which perpetuates racial biases.)

- 2. Technical bias:** Technical biases arise from the technical design of the system itself. These biases can emerge from various aspects of the design process, such as the use of decontextualized algorithms, data processing and collection, and other system design decisions. Unlike pre-existing biases, technical biases are inherently linked to the system's design and do not exist independently from it.
 - a. Algorithmic bias:** These are biases that are not present in the data but rather present in the algorithm itself. Selection of features, models, and training procedures can introduce biases, due to inefficiencies of these systems in characterizing the data. *(Eg. Certain models of statistics like regression can fail to capture correlations between attributes and sub-groups)*
 - b. Data Bias:** These biases stem from selection methods of data sources or data collection procedures. It includes sampling biases arising from non-random sampling, where a population might be under or over-sampled. Representation bias arises when population sub-groups and outliers are not considered for and thus underrepresented in the data. *(Eg. Models for facial recognition trained primarily on Caucasian data fail when confronted with different racial identities)*
- 3. Emergent bias:** Emergent biases are biases that occur when an AI system is deployed and users interact with the system.
 - a. Population bias:** Bias can originate from the inability of the system to be able to effectively represent the population, that is using the system. *(Eg. Systems trained and modelled for men produce biased results when implemented for women.)*
 - b. User bias:** Bias can also be produced from the user's interaction with the AI system. *(Eg. Chatbots interacting with racist or misogynistic user inputs can learn to reproduce similar outputs)*

PRE-EXISTING BIAS	TECHNICAL BIAS	EMERGENT BIAS
Pre-existing biases have roots in individuals and social institutions that are baked into AI systems	Technical biases arise from the design of the system and data processing	Emergent biases occur when an AI system is deployed , and from user interaction
Individual bias Biases of individuals within the AI value chain <i>Biases of data labelers embedded in training data transfer into AI systems</i>	Algorithmic bias Bias introduced within the features, models and training procedures <i>Certain statistical models, like regression, fail to capture correlations between attributes and subgroups</i>	Population bias Biases can arise from the system's inability to represent the population <i>Systems trained and modeled for men produce biased results when implemented for women</i>
Societal bias Biases existing in institutions, cultures and organizations <i>Biases in the U.S. criminal system lead to more African-American convictions, which perpetuates racial biases</i>	Data bias Biases stemming from the selection and/or collection of data <i>Models for facial-recognition trained primarily on Caucasian data fail when confronted with different racial identities</i>	User bias Bias can also be produced from the user interaction <i>Chatbots interacting with racist or misogynistic user inputs can learn to reproduce similar outputs</i>

Source: Apti Analysis; Caton and Haas, "[Fairness in Machine Learning: A Survey](#)"; Friedman and Nissenbaum, "[Bias in Computer Systems](#)."

The Bias Framework

Our framework aims to demonstrate where and how biases get entrenched at the different stages of the AI value chain. To that end, **our framework maps the various processes and actors that are part of building AI systems across three stages: pre-development, development, and adaptation. Further, the framework divides biases according to their various sources into three broad categories: pre-existing, technical, and emergent, further mapping where in the value chain these biases get encoded into AI systems.**

Lastly, the framework also maps the differences in the value within three different sectors (*finance, health, and education*), and the manifestations of bias categorised according to its origin within the sector-specific value chains.

Methodology of the framework

The framework begins with using a value chain ontology to investigate the AI lifecycle. The ontology helps us identify the various processes and actors that are embedded in the building AI systems, across three different stages, pre-development, development, and adaptation. Each of these stages is further disaggregated into processes that are required to be performed for an AI system to be developed and deployed.

Our research into the AI value chain elucidated a disaggregation of these processes across many independent actors. That is to say, with the proliferation of foundation models, more and more AI applications are developed using such models, thus not requiring entities building AI applications to develop and train AI models from scratch. While inviting innovative AI applications, this disaggregation further complicates responsibility and accountability across the value chain. Thus to tackle such a predicament, we mapped actors who perform and are responsible for the various tasks across the AI value chain. This allowed us to answer two important questions, **who has the power to effect decision-making across the process in the AI value chain and subsequently who should be responsible for biases getting entrenched at each juncture of the AI value chain?**

We borrow from literature about biases in computer and AI systems, to classify biases according to their origins and sources underpinning them. While mapping manifestations of bias is important to be able to monitor and measure its harmful impacts on users of AI systems, manifestations of biases are sector and context specific, making it difficult to exhaustively map its various harmful manifestations. Further, mapping sources of biases instead of manifestation also helps us to plot where biases get encoded into AI models, across their lifecycle, aiding us to think about procedural and design-level mitigation strategies, moving beyond outcome-based regulations. Our mapping of sources of bias across the AI value chain, thus not only allows us to delineate where and how sector-specific examples of manifestation of bias are entrenched into systems, but also subsequently allows us to suggest stage-adaptive and actor-specific mitigation strategies.

How to read the framework

The framework is divided into two main parts, the general framework, which provides a holistic overview that demonstrates the general purpose AI value chain (processes and actors) and the sources of bias within it. The sector-specific frameworks illustrate how domain-relevant value chains evolve across finance, healthcare, and education, highlighting respective nuances and departures.

a. General Framework

The general framework is divided into two sections. The first section consists of rows and columns that represent key elements of the AI value chain. The columns correspond to different stages:

- i. *Pre-development*, which includes algorithm modelling and dataset creation;
- ii. *Development*, which involves testing, training, and validating the model;
- iii. *Adaptation*, where the model is integrated into user-facing applications and continuously learns from user interactions to enhance functionality.

The rows capture various elements of importance, including the process, which identifies the steps involved at each stage;

- i. *Actor*, highlighting the stakeholders responsible for specific processes;
- ii. *Sources of bias*, mapping how and where different types of bias become embedded in AI models.

b. Sector-specific frameworks

We focus on three sectors—finance, health, and education—each of which is mapped on a separate table. In these tables, the columns represent the various stages of the AI value chain, while the rows capture different elements of significance. The process element outlines how the steps and actions differ across sectors in the development of an AI system. The actor's element highlights how different stakeholders are involved at various stages of the sector-specific value chains. Lastly, the bias manifestations element maps examples of bias and indicates at which stage they are most likely to become embedded in the AI system.

Purpose of the framework

Using the framework, we have made pointed strategies to reduce tendencies of biases getting entrenched in AI models. These strategies based on the sources of biases at different stages are also consequently mapped to actors' processes and stages of the AI value chain. We hope that with this disaggregated approach to bias mitigation, we can adopt standards and strategies for the design and procedures of development and deployment of AI systems, which can attenuate the risks and harms arising from biased AI outputs.

Beyond its relevance to this project, the framework also hopes to help tech developers, civil societies, researchers, academicians, and policymakers to navigate the complexity of value chains and a specific type of risk under consideration, such as bias and its sources.

GLOSSARY

PROCESSES

Design and planning: Initial process of laying out the system’s architecture.

Data sourcing: Collecting data used in training AI models.

Pre-training: Training foundation AI models for general tasks.

Foundation model release: Foundation models are released in accordance to specific business typologies.*

Foundation model selection: Downstream actors according to their use case choose foundation models to further train and fine tune them for specific tasks.*

Model training and fine tuning: Training foundation models for specific tasks and objectives.

Verification and validation: Checks correctness, ensures alignment with the model's intended purpose

Model integration: Model integration is embedding one or more trained AI model into a system or application.

User interaction: Users engaging with the AI tool, giving it prompts and tasks to perform.

Monitoring and evaluation: Tracking performance and effectiveness, and safety.

** While the release and selection of foundation models do not account for significant sources of bias, they are key to tracing bias flow into downstream AI systems.*

ACTORS

Foundation model providers: Actors developing foundation models.

Dataset providers: Provide datasets for model training.

Model adapters and optimizers: Actors who carry out a varied range of tasks, like data processing, data labeling, fine tuning, and model training which operationalize AI models for specific applications.

Model integrators: Actors integrating AI models into AI systems and providing them as services or to end-users.

Platforms: Actors like social media and others where AI generated content can be shared and hosted.

User: Actors using AI systems as enterprises (B2B) or consumers (B2C).

- **Enterprises** are actors who integrate AI in their workflow or as a part of their service offered.
- **Consumers** are actors who directly use AI services and tools for their own purpose.

MLOps and evaluators: Actors providing tools and services for performance evaluation, auditing, safety assessments and etc, across multiple stages in the value chain.

SOURCES OF BIAS

A. Pre-existing Bias: Pre-existing biases have roots in individuals and social institutions that are baked into AI systems.

1. Individual Bias: Biases of individuals within the AI value chain.

2. Societal Bias: Biases existing in institutions, cultures and organizations.

B. Technical Bias: Technical biases arise from the design of the system and data processing.

1. Algorithmic Bias: Bias introduced within the features, models and training procedures.

2. Data Bias: Biases stemming from the selection and/or collection of data.

C. Emergent Bias: Emergent biases occur when an AI system is deployed, and from user interaction.

1. Population Bias: Bias can arise from the system's inability represent the population.

2. User Bias: Bias can also be produced from the user interaction.

GENERAL FRAMEWORK

Key

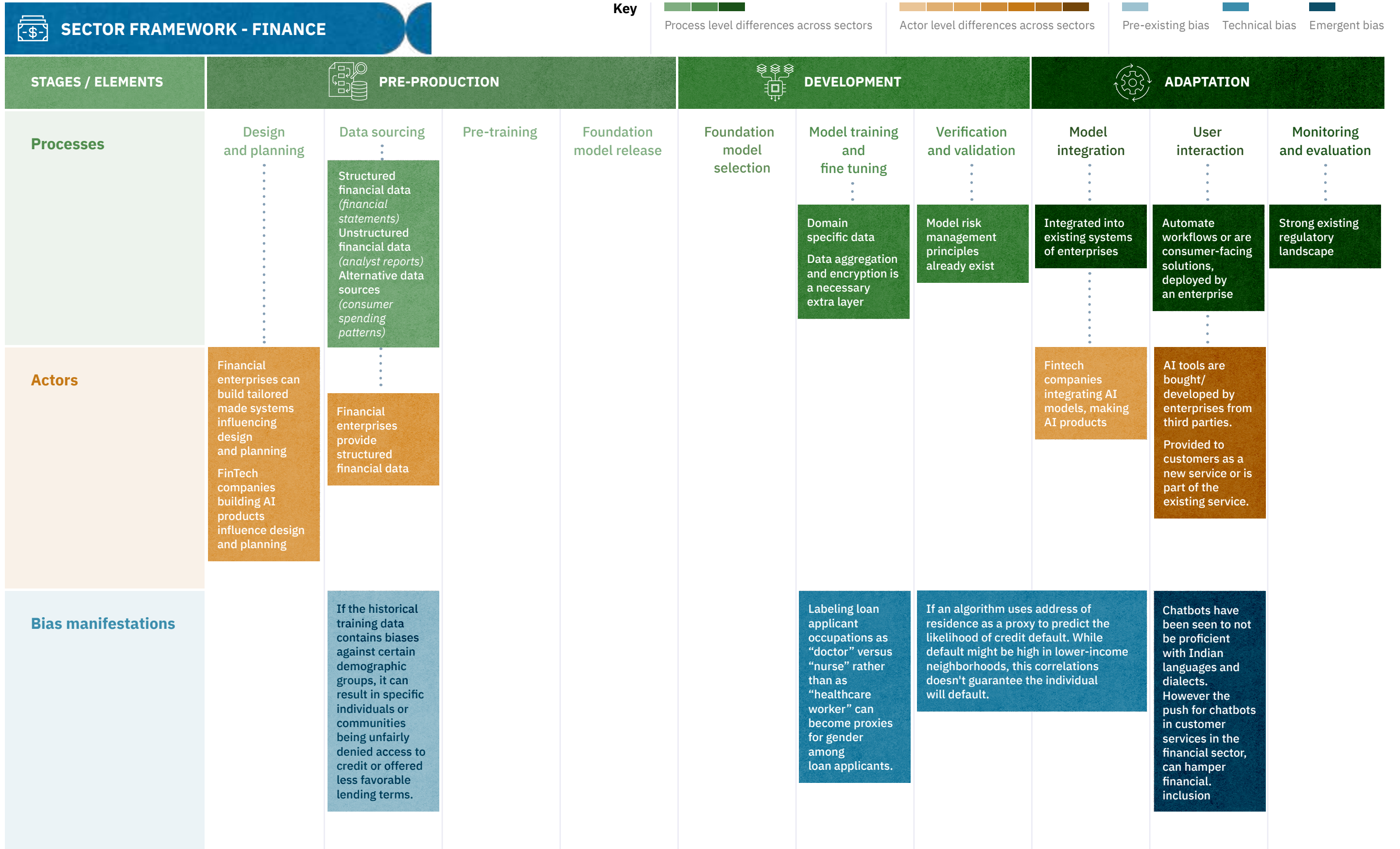


 The first part of the text is a blue square icon.

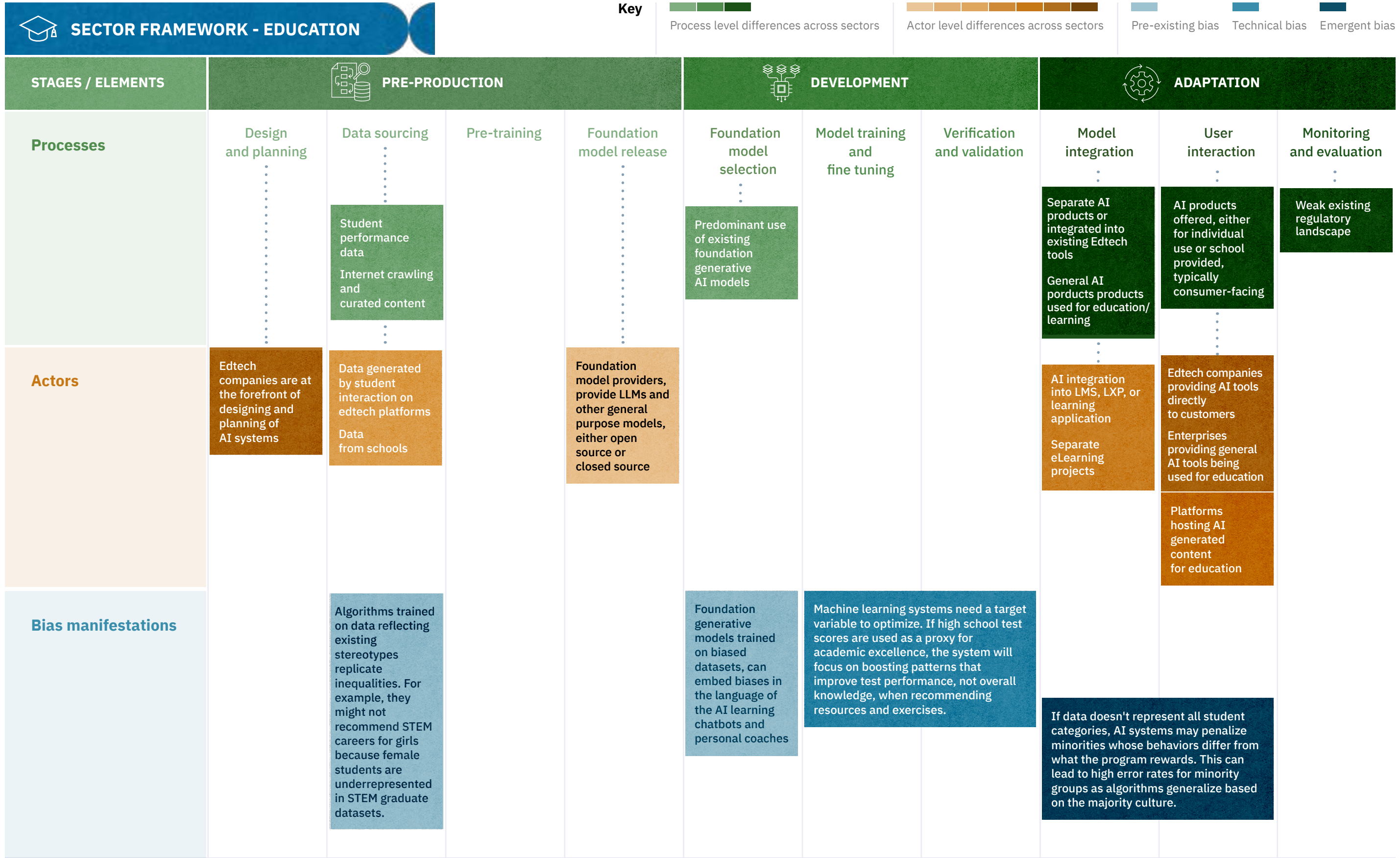
© 2015 Pearson Education, Inc. or its affiliate(s). All rights reserved.

 UNIVERSITY OF NORTH CAROLINA




STAGES / ELEMENTS	PRE-PRODUCTION				DEVELOPMENT			ADAPTATION		
Processes	Design and planning	Data sourcing	Pre-training	Foundation model release	Foundation model selection	Model training and fine tuning	Verification and validation	Model integration	User interaction	Monitoring and evaluation
Actors		Foundation model providers				Model adapters		Model Integrators	Platforms	
		Dataset Providers							Enterprises (B2B)	
		Data labelers							Consumers (B2C)	
	Enterprises (B2B)					Evaluators				
Sources of bias										
			Pre-existing							Pre-existing
		Technical				Technical	Technical			
				Emergent			Emergent			
Disaggregated sources of bias	• Individual Bias • Societal Bias	• Societal bias • Data bias	• Individual bias • Data bias • Algorithmic bias	• Population Bias	• Population Bias	• Individual bias • Algorithmic bias	• Algorithmic bias • Data bias	• Data or algorithmic bias • Population bias	• Population Bias • User Bias	• Societal Bias • Individual Bias







MITIGATION STRATEGIES

STAGES	STRATEGIES	PROCESSES	ACTORS	BIAS	MODALITIES
Pre-production 	Reducing bias in datasets	Data sourcing	<ul style="list-style-type: none"> Foundation model providers Dataset providers 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Fair data curation frameworks Bias measuring tools
	Fairness aware learning objectives	Pre-training	<ul style="list-style-type: none"> Foundation model providers 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Fairness checklists Fairness aware loss-functions
	Model documentation	Foundation model release	<ul style="list-style-type: none"> Foundation model providers 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Model cards
Development 	Bias evaluation for foundation models	Model selection	<ul style="list-style-type: none"> Model adaptors 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Audit frameworks for foundation models Bias testing tools
	Bias correction on foundation models	Model training and fine tuning	<ul style="list-style-type: none"> Model adaptors 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Post-processing debiasing methods Debiasing toolkits
	Comprehensive benchmarking of AI models	Verification and validation	<ul style="list-style-type: none"> Model adaptors Regulators Model integrators 	<ul style="list-style-type: none"> Pre-existing Technical 	<ul style="list-style-type: none"> Responsible AI benchmarking Standardised responsible AI benchmarks
Adoption 	Understanding contextualized bias	Model integration	<ul style="list-style-type: none"> Model integrators Enterprises Platforms 	<ul style="list-style-type: none"> Pre-existing Technical Emergent 	<ul style="list-style-type: none"> Impact assessment Participatory approaches to AI deployment
	Bias audits and monitoring	User interaction/ Monitoring and evaluation	<ul style="list-style-type: none"> Enterprises Platforms 	<ul style="list-style-type: none"> Technical Emergent 	<ul style="list-style-type: none"> Bias-audit tools & frameworks Third-party evaluators



The Pre-Production Phase

ACTORS

Foundation model providers, dataset providers

SOURCE OF BIAS

Pre-existing

Strategy: Reducing bias in datasets

One of the main ingredients of foundation models is large-scale datasets used for pretraining. Such datasets often mirror social prejudices, embedding pre-existing bias into such models.¹²⁸ Further, methods of curation of such large datasets have largely been based on ad-hoc processes and based on heuristics.¹²⁹ Such decisions about what data to include or exclude informed by subjective or context-dependent criteria, can also further introduce biases.¹³⁰

Modalities

- **Data curation frameworks:** The application of data curation practices, and frameworks, for the creation of datasets for model training, have already been recognised to be beneficial for the AI lifecycle, improving, fairness, accountability, and transparency in dataset development.¹³¹ Some frameworks have been adapted based on data curation theory and principles from other disciplines, like FAIR,¹³² to be applied to the data collection pipeline for AI. While such frameworks are a good checkpoint for pre-existing biases in datasets, they also become a guide for stakeholders to take decisions on datasets, limiting variances based on individual heuristics and biases.
- **Bias measuring tools:** Open-source tools like [REVISE](#) uses statistical methods to inspect a data set for potential biases or issues of underrepresentation along three dimensions: object-based, gender-based and geography-based. Similarly tools like [IBM's AI fairness toolkit](#) has been able to identify biases in large datasets, developing 70 fairness metrics and 10 bias mitigation algorithms. Google's [Vertex AI](#) toolkit provides metrics that allows one to check for biases within datasets. Such tools can also help actors look at existing biases in datasets.

The Pre-Production Phase

ACTORS

Foundation model providers

SOURCE OF BIAS

Pre-existing, technical

Strategy: Fairness-aware learning objectives

Matching AI learning objectives and our expected or desired behavior from AI models can be key in reducing bias. However, learning objectives have increasingly been defined by utility, losing sight of key aspects such as fairness. Incorporating fairness as a learning objective in the training process can help in being able to reach desired AI behavior, allowing developers to account for pre-existing (data) and technical (proxies) biases. Additionally, incorporating such practices also allows for AI training to be attuned to reducing bias, embedding fairness consideration within the training process.

Modalities

- **Fairness checklists:** Checklists have been used in many other domains to guide decision making and prompt critical conversations. In aviation and medical science, while checklists operate as memory aid, in domains like structural engineering, checklists aid discussion among stakeholders about probable risks. Thus, checklists can be value levers, which prompt conversation and discussion on ethics. Ensuring participatory co-designed checklists, can help build organisational infrastructures to formalise ad-hoc processes and precipitate fairness conversation in model development and training.

Some fairness checklists already exist, some developed by dominant technology corporation, and some checklists by civil society organisations and academic researchers which specifically cater to fairness. While such checklists can be good starting point, research show better overall impact of checklist when they are developed through participatory methods within organisations, aligned to existing workflows and organisational culture.¹³³

- **Fairness-aware learning objectives:** Machine learning algorithms are predominantly designed to reduce loss functions, which is defined as the cost of wrong output. The goal of machine learning algorithms is to reduce this loss function, indicating a closer approximation of the desired output. Incorporating ideas of fairness and bias in loss functions, thus can be a key measure to attune AI learning objectives to reduce bias. Some research to this end has seen beneficial results.

A modified loss function, Sensitive loss incorporating demographic information and discrimination-aware rules can guide the learning process of AI models to have more fairness and unbiased representations. Similar work, on developing loss function based on bias parity scores, which measures fairness as the average bias of all classes compared to the population, have also been fruitful.

The Pre-Production Phase

ACTORS

Foundation model providers

SOURCE OF BIAS

Pre-existing, technical

Strategy: Model documentation

Documentation has been an essential part of software development, it has allowed transparency between developer and other non-technical stakeholders and also for use by other actors in the software lifecycle. Similar standardised documentation is also present in other industries, like components of electronic hardware have detailed characterizations of components' performances under different test conditions. While ML models often are complex and harder to explain, information about model biases, underlying datasets, model performance and limitations are important information for subsequent actors to understand if a model fits the purpose and use case. Such information can help model adaptors choose more appropriate models and account for existing foundation model biases in their subsequent training.

Modalities

- **Model cards:** Model cards primarily serve the purpose of documenting and disclosing information about trained models. Including information on what datasets were used, assumptions made during training and others. Additionally model cards can also contain information in relation to model performances, across different demographics and tasks and intended use and limitations. Model cards can thus offer a standardized approach to presenting information about models, making it straightforward to comprehend, maintain, and utilize. Platforms like [Hugging Face](#) have developed a database of various models, along with standardised model cards.



The Development Phase

ACTORS

Model adaptors, Model integrators

SOURCE OF BIAS

Pre-existing, technical, emergent

Strategy: Bias evaluation for foundation models

Intrinsic biases encoded into existing foundation models through data or otherwise, can cause biases in downstream applications of the models. As foundation models become the starting point for many AI systems being developed by downstream actors, understanding existing biases in models can be key to remitting biases of its downstream application. Further, different foundation models may have different hidden biases, across demographic groups etc, it is key to understand such biases before making an appropriate model choice.

Modalities

- **Frameworks of audits for foundation models:** Auditing is one of the promising forms of governance mechanisms stakeholders can employ to ensure AI models are legal, ethical, and technically robust. Much like audits of financial transactions tests for correctness, completeness and legality, audits for foundation models can be designed to pit against pre-defined fairness principles. While there remain critiques to auditing foundation models existing frameworks for auditing foundation models¹³⁴ can have some efficiencies. IIA's AI audit framework which is aimed at helping organisations build their own internal audit frameworks is one example.
- **Bias testing tools:** Tools to map biases in large language models already exist. These tools can aid model adaptors and integrators run diagnostics to evaluate pre-existing and technical biases, by mapping outputs to predetermined sets of inputs. Some of these tools are also open-source frameworks, allowing stakeholders to add ethical concerns depending on the demographic and societal contexts of its users. Google's What-if tool, allows one to use counterfactuals to see how model prediction changes, this can be used to test biased model outputs across different demographics. Similarly Google's fairness indicator also allows users computation and visualization of commonly-identified fairness metrics for classification models, making it easy to compare performance.

The Development Phase	
ACTORS Model adaptors, Model integrators	SOURCE OF BIAS Pre-existing, technical
<p>Strategy: Bias correction on foundation models</p> <p>Foundation models absorb vast amounts of data to recognise patterns, allowing them to then produce outputs. It has already been established that the mountains of data that these foundation models absorb and learn biased patterns within the data, surmounting to biased outputs and treatments of individuals. As suggested, bias audits and evaluation of foundation models becomes key in identifying such biases. Subsequent debugging of foundation models is an essential second step, which can to some extent limit bias patterns in foundation models to seep into downstream AI systems, built over them.</p>	
<p>Modalities</p> <ul style="list-style-type: none"> • Post-processing debiasing methods: While retraining and finetuning models to meet fairness standards are some ways of debias foundation models, they can often be infeasible in practice for large-scale trained models due to large computational and storage costs, low data efficiency, and model privacy issues. However, debiasing techniques like FairReprogram and Fair infinitesimal jackknife (FairIJ) which have been seen to have successfully improved group fairness, without requiring the model to be retrained. Further techniques like SenSeI have been seen to improve individual fairness. • Debiasing toolkits: IBM's AI Fairness 360 toolkit provides a number of tools for bias mitigation, including tools for identifying bias in training data, training fairness-aware AI models, and evaluating the fairness of AI models. Similarly, Microsoft's Fairlearn toolkit, gives developers fairness metrics to identify biases and algorithms for their mitigation. 	

The Development Phase	
ACTORS Model Integrators, regulators, researchers	SOURCE OF BIAS Pre-existing, technical, emergent

Strategy: Comprehensive benchmarking of AI models

Benchmarks are essentially goals set for the AI models to achieve. While, currently AI models have scored high against such benchmarks, they have continued to produce biased outputs. Exploration around such benchmarks points to the source of this inconsistency to the lack of comprehensive benchmarks.¹³⁵ Researchers have pointed out that, current benchmarks are attuned to assessing only one goal, which cannot comprehensively capture multi-modalities of newer AI systems, suggesting a need to move towards benchmarks that help conceptualise trade-offs between accuracy and bias/toxicity.¹³⁶ Further, while responsible AI frameworks have proliferated the ecosystem, there remains inconsistency in reporting of responsible AI benchmarks.¹³⁷ Without standardised responsible AI benchmarks, it becomes hard for other actors to compare and assess different AI models.

Modalities

- **Responsible AI benchmarks:** In response to benchmark saturation, there has been movement away from traditional benchmarking towards more task specific benchmarks with higher standards against which AI models are tested, with a focus on bias aware benchmarking. Like [HELM](#) which is designed to evaluate LLMs across diverse scenarios, including reading comprehension, language understanding, and mathematical reasoning. Similarly, [TruthfulQA](#) is a benchmark designed to evaluate the truthfulness of LLMs in generating answers to questions. Similar benchmarks across various modalities have been developed to test models across diverse tasks.
- **Standardised responsible AI benchmarks:** While there has been uptake in the development of responsible AI benchmarking, there hasn't been any consensus on standardisation. The lack of standardisation has also led to selective reporting of benchmark performances. Making it important to build consensus around benchmarking standards, across AI models.



The Adaptation Phase

ACTORS

Model adaptors, Model integrators

SOURCE OF BIAS

Pre-existing, technical, emergent

Strategy: Understanding contextualised bias

Concepts of fairness can vary across different use cases of AI systems. It can also vary depending on the demography and geography of its user base. Because of such variance in fairness understandings and definitions, it is difficult for omnibus frameworks of fairness and bias to be impactful on the ground. Identifying and mapping stakeholders and their interrelationships can help reveal sources of emergent biases and pre-existing biases in the collected and used data. Having a good understanding of what cognitive and structural human biases are at play in a given context can be translated to mechanisms that limit their exasperation through AI systems.

Modalities

- **Participatory approaches to AI deployment:** Such approaches can be an efficient way to account for emergent biases in AI systems. By consulting affected stakeholders before the deployment of given AI systems, actors can be cognizant of stakeholders' interests and consideration are accounted for prior to the roll out of any AI-based system that will impact them. In doing so, a participatory approach not only allows deployers to account for specific biases and harms, but also make the process of deployment fairer. Approaches like human-centered AI, can be a starting point for participatory approaches. Additionally, civil society and academia are already working frameworks that operationalise a participatory approach for AI systems.
- **Impact Assessment:** AI impact assessment tools and frameworks, can help assessment of actual and real-time potential impacts of AI systems on individual and community users. A system-level AI Impact Assessment (AIIA), developed by the Responsible Artificial Intelligence Institute (RAI Institute), allows actors to identify risks through a set of defined controls across stages of the system lifecycle and impacts which are categorised in line with generally accepted principles for safe and trustworthy AI, in particular: accountability and transparency, fairness, safety, security and resilience, explainability and interpretability, validity and reliability and privacy. Similar standardised fairness assessment frameworks have also been developed by telecommunication engineering centres in India, looking at the entire lifecycle of AI systems, across data, model used and scenario testing to assess potential biases in AI systems.

The Adaptation Phase

ACTORS

Enterprises, Platforms, Model evaluators

SOURCE OF BIAS

Emergent

Strategy: Bias audits and monitoring

Bias and fairness are not absolute concepts rather are dependent on the application contexts. While bias mitigation at the upstream stages of the AI value chain can to some extent deter biased algorithmic outputs, the real-world accuracy and efficacy of a system can only be measured when they are deployed or piloted for real time user interaction. Further, given the nature of AI systems, research has also established how user interactions can also encode bias into AI models. Thus, auditing AI systems for bias before scaling deployment and adoption, along with regularised auditing of AI models post deployment are key to mitigate for emergent biases.

Modalities

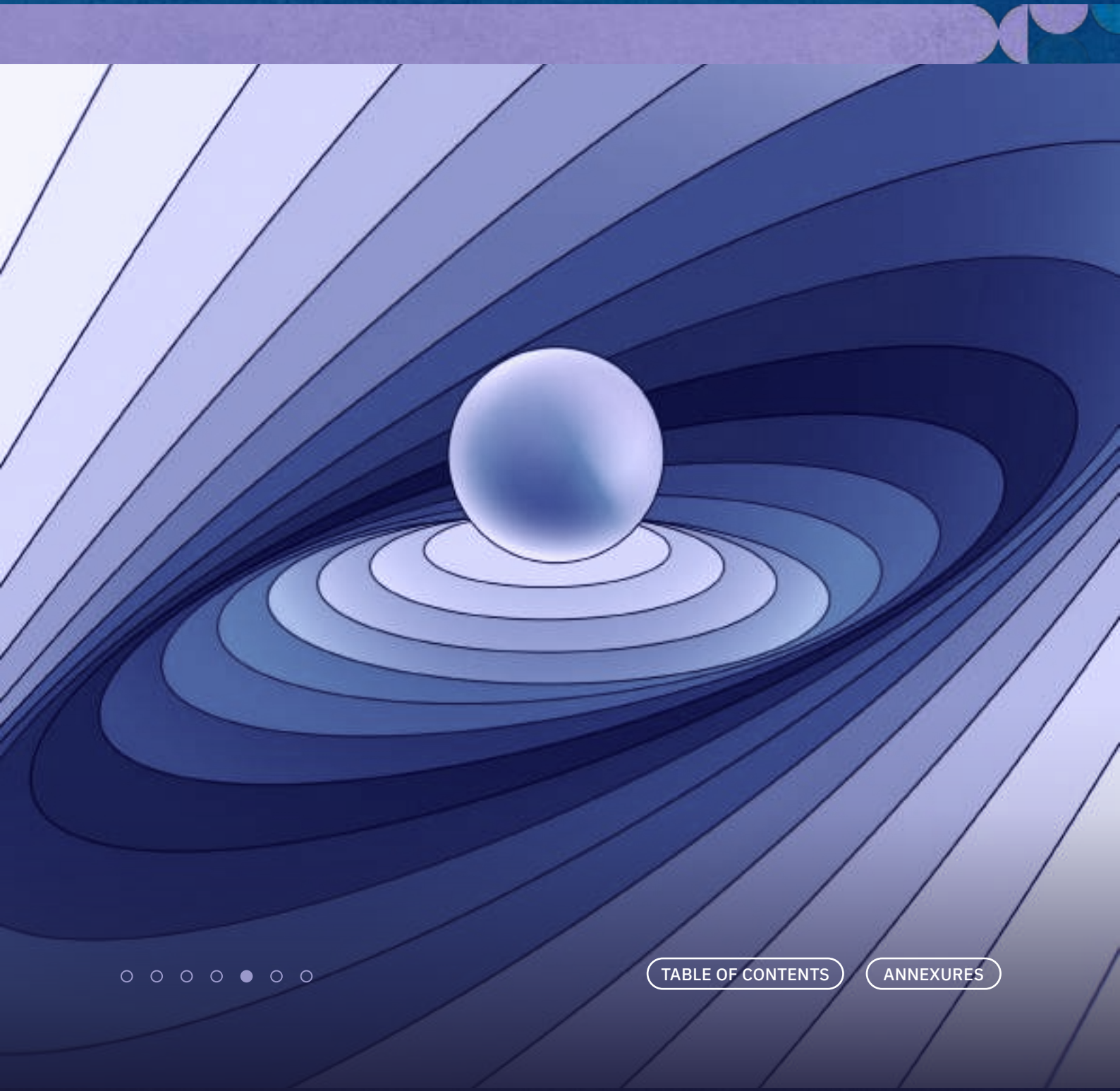
- **Bias audit processes and tools:** Open-source bias audit tools and monitoring frameworks also allows stakeholders to perform their own regularised internal audits of AI systems being adopted and deployed.

Monitoring guidances and standards by International Organization for Standardization (ISO) and the International Electrotechnical Commission (IEC) ([ISO/IEC 42001](#), [ISO/IEC 23894:2023](#), [ISO/IEC 23053:2022](#)) can help organisation establish AI risk management practices, and also can help guide organisations to setup robust processes and mechanisms for AI adoption. Open source tools like [Aequitas](#) can audit AI systems for discriminatory and biased predictive AI tools, across multiple bias and fairness criteria, for different types of interventions.

- **Third party evaluators:** Independent third-party evaluation of AI models, which looks at input data and output decisions, are essential for an exhaustive understanding of bias in AI systems; it also helps build transparency and trust for such systems.

MODULE II

AI for digital integrity and cybersecurity: thinking through the lens of trustworthiness



AI for digital integrity and cybersecurity: thinking through the lens of trustworthiness

Introduction

The growing influence of Artificial Intelligence (AI) across sectors has opened up unprecedented opportunities, particularly across domains such as cybersecurity and digital integrity. However, alongside these advancements come profound challenges.

As AI systems become more integral to the functioning of critical infrastructures—ranging from financial services¹³⁸ to healthcare¹³⁹ and education¹⁴⁰— they introduce new vulnerabilities that require urgent attention. The rapid pace at which AI technologies are developing has outstripped regulatory frameworks, leaving gaps in governance that could exacerbate risks such as data breaches, algorithmic bias, misinformation, and cyber threats.¹⁴¹

In the context of digital integrity, AI plays a dual role: it both bolsters and undermines the safety and security of digital environments. Digital integrity refers to the protection of individuals' digital identities, data, and online experiences.¹⁴² The erosion of trust, fueled by AI-driven misinformation and disinformation, has become a significant threat to this integrity. Moreover, AI's ability to automate and amplify harmful content,¹⁴³ such as hate speech or gender-based violence, raises urgent questions about how to ensure ethical oversight and governance of AI systems.

Similarly, AI has emerged as both a vital tool and a potential threat in cybersecurity. Initially used for anomaly detection and intrusion

prevention.¹⁴⁴ AI has evolved into a sophisticated weapon for both defenders and adversaries in the digital space.¹⁴⁵ On the one hand, AI enhances the detection of threats, accelerates response times, and strengthens system resilience. On the other hand, threat actors are increasingly using AI to exploit vulnerabilities in security frameworks, making cybersecurity a double-edged sword. The integration of AI into cybersecurity, while necessary, introduces ethical concerns about transparency, accountability, and bias, particularly in the realm of automated decision-making.¹⁴⁶

Given these complexities, the importance of trustworthiness in AI cannot be overstated. This report examines trustworthiness as a central governance principle to mitigate the harms and harness the opportunities presented by AI. A trustworthy AI system must not only be robust and reliable but must also respect ethical principles such as fairness, transparency, human oversight, and accountability. Trustworthy AI can serve as the foundation upon which the digital integrity and security of individuals and institutions are built.

Digital integrity refers to the safeguarding of individuals' digital identities, data, and interactions, ensuring that they remain free from corruption or exploitation. As our online presence grows, the importance of digital integrity becomes more pronounced, particularly in light of AI's rapid advancements. AI plays a pivotal role in both enhancing and undermining digital integrity. For instance, generative AI can amplify misinformation, while AI-driven moderation tools help combat disinformation and protect vulnerable users.¹⁴⁷

Similarly, AI's integration into cybersecurity has revolutionised how organisations detect and respond to threats. With its ability to process large-scale data and make real-time decisions, AI offers significant advantages in threat detection and response.¹⁴⁸ However, it also presents risks, as adversaries increasingly exploit AI to execute sophisticated attacks. Thus, a careful balance must be struck between leveraging AI for cybersecurity and mitigating the risks it introduces.

Anchoring the research

This inquiry is built upon a three-pronged framework designed to comprehensively address the intersections of AI trustworthiness, digital integrity, and cybersecurity. At its core, our approach is grounded in the principles of trustworthy AI, which serve as the foundation for linking these interconnected domains. Once this foundational alignment is established, we adopt a socio-technical lens to examine the broader ecosystem. This perspective enables us to analyse the interplay between technological systems and societal dynamics, ensuring a balanced understanding of the challenges and opportunities at hand.

In the final phase, we apply a Roles, Harms, and Opportunities (RHO) framework to contextualize digital integrity and cybersecurity. This structured approach allows us to identify the key stakeholders, assess potential risks and harms, and explore avenues for leveraging opportunities. By integrating these elements, the RHO framework provides a roadmap for crafting targeted and effective mitigation strategies that address both technical vulnerabilities and societal implications, paving the way for a more secure and trustworthy digital future.

Trustworthy AI

We adopt a trustworthiness approach to inform AI adoption strategies and propose effective mitigations for potential harms arising from the integration of AI across various domains. This approach is particularly suited to addressing the complex interplay of technical, societal, and ethical dimensions that define AI's impact on digital integrity and cybersecurity.

The rationale for adopting a trustworthy AI governance approach is twofold. First, trustworthy AI frameworks are embedded within academic discourse, and they crop up in enterprise practices and policy landscapes across multiple jurisdictions. Many countries have adopted these frameworks as guiding principles for AI governance, providing a strong alignment with global standards and policy priorities. This enables the development of strategies that are internationally cohesive while addressing local nuances. Second, the trustworthy AI framework complements a techno-

societal governance perspective by combining technical requirements such as robustness, reliability, and security with human-centric attributes of trust. Trust serves as a bridge between technical rigor and socio-legal principles, encompassing elements like transparency, accountability, and fairness.

Socio-technical lens

We employ a socio-technical approach to explore the multifaceted impact of AI on digital integrity and cybersecurity. This approach recognizes that AI operates at the intersection of technology and society, emphasizing the interconnectedness between technical systems and the social, ethical, and policy environments in which they function.

By positioning AI within this broader socio-technical framework, we aim to unpack the dual impact of AI both on the robustness of technical infrastructures and on human lives and societal dynamics. This perspective highlights that AI safety is not merely a technical challenge but also a societal imperative.

It underscores the need to move beyond isolated technical solutions and examine how AI systems interact with governance models, societal values, and cultural contexts. By embracing this lens, we delve into the co-evolution of human systems and technological advancements, emphasizing that effective mitigations must balance technological innovation, social responsibility, and inclusive policy design. This holistic understanding ensures that Responsible AI frameworks account for the broader societal implications of AI deployment, fostering solutions that are technically sound, ethically grounded, and contextually appropriate.

Roles, Harms and Opportunities (RHO) Framework

To further translate the socio-technical approach into practice, we adopt a three pronged framework including roles, harms and opportunities. This framework serves as a lens through which we can assess not only the technical challenges of AI systems but also the societal harms that arise from its deployment, as well as the

potential opportunities for innovation and benefit. By understanding the intersection of AI with Digital Integrity and Cybersecurity, the RHO framework allows us to map out the feedback loops between technology and society, especially as it pertains to safeguarding against threats while promoting responsible development.

AI's applications across various sectors present a paradox. In sectors like financial services and healthcare, where regulation is stringent due to the handling of highly sensitive personal data, AI has proven essential in upholding security protocols. The RHO framework is particularly useful in untangling these paradoxes. It helps us recognize the roles AI plays in fortifying security and enhancing user experience, the harms that may arise from exploitation or misuse, and the opportunities for creating a safer, more transparent digital ecosystem. Thus, as AI continues to evolve across sectors, a techno-societal approach supported by the RHO framework ensures that we are addressing both technical and societal concerns in a balanced, comprehensive manner.

Harms come second and primarily refer to the adverse impacts associated with the use or role of AI. Lastly, opportunities lie at the opposite end of the spectrum, representing the benefits or positive use cases that can be leveraged through the trustworthy development and application of AI.

UPON LANDSCAPING THE INSTANCES OF AI USE, WE DISTILLED AI ROLES INTO THREE CATEGORIES



AI-Powered

AI-powered refers to systems or tools where AI is the core driver, autonomously performing tasks, making decisions, and executing functions without direct human intervention.



AI-Augmented




AI-augmented refers to systems where AI supports or enhances human decision-making and actions by providing insights, tools, or automation for specific tasks.






AI-Generated

AI-generated refers to the creation of outputs, content, or data through AI algorithms, which can include text, images, videos, insights, or predictions.

DIGITAL INTEGRITY

ROLE	HARMS	OPPORTUNITY	MITIGATIONS	EXAMPLES
<div>AI Powered</div> <div></div>	Abuse of facial recognition technology	Opportunity for regulatory innovation by governments and international organisations	Translation of legal and ethical requirements of using FRTs and biometric technologies into technical standards for easy adoption and enforcement. (accountability, rights respecting)	ISO/IEC 2382-37:2022 standards established for systematic description of the concepts in the field of biometrics pertaining to the recognition of human beings and the adoption of basic safeguards within the design of facial recognition technologies.
	Scrapping personal/ sensitive personal information online	AI tools for anonymisation of big datasets	Adoption of differential privacy to train AI models (rights respecting)	Amazon Web Services provides AWS Clean Room Differential Privacy services which allow users such as researchers or model developers to use datasets where individual personal information is not revealed.
			Adopt scraping prevention tools (rights respecting)	Glaze or NightShade are easy-to-use tools developed by the University of Chicago, that make small changes to an image and make it extremely difficult to understand for an AI model.
			Embed privacy-centered licensing and contracting practices. (transparency and accountability)	X (formerly, Twitter) prohibits data scraping without prior permission in its terms of service.
<div>AI Augmented</div> <div></div>	Synthetic content for misinformation, disinformation, and other harmful and violent content.	AI software for the detection and removal of fake news or explicit imagery	Enhance transparency and robustness of AI tools (transparency and robustness)	Meta Oversight Board has recommended labeling AI-generated and manipulated media to be labelled on Meta’s social media platform. Accordingly, “AI Info” labels will be added to AI generated content on platforms such as Facebook and Instagram.
			Conduct robust testing, consistent auditing, and oversight. (accountability)	
<div>AI Generated</div> <div></div>	Algorithmic promotion of misinformation/ disinformation and hate speech	Using NLP and ML tools to real-time flag fake news and hate speech.	Conduct regular internal and third-party audits. (accountability)	Third-party sock puppet audits were conducted by researchers to understand the misinformation bubble on YouTube and how to revert the bubble enclosure.
			Collaborate on cross-platform moderation. (robustness)	Lantern by Tech Coalition is a significant cross-platform initiative supported by several large-scale online platforms resulting in meaningful cross-platform moderation for harmful content.

CYBERSECURITY

ROLE	HARMS	OPPORTUNITY	MITIGATIONS	EXAMPLES
AI Powered 	Threat actors leveraging adversarial ML to target ML models	An emerging machine learning (ML)-powered cybersecurity opportunity lies in the protection of ML models themselves through ML powered adversarial defence systems	Dynamic threat intelligence and self-updating models are AI-powered tools that autonomously identify and adapt to emerging risks (Secure and Reliable)	AI-Powered Phishing Detection by Darktrace incorporates explainability into its AI-based anomaly detection to help human analysts understand why specific emails or actions are flagged as threats. Equitable Incident Prioritization in Endpoint Detection AI systems by vendors like CrowdStrike Falcon prioritize incidents based on unbiased severity assessments rather than potentially skewed data sources. (Secure and Reliable)
			AI-driven fault-tolerant systems and adversarial training to handle errors and adversarial attacks fall under AI-powered solutions. (Robust and Resilient, Secure and Reliable)	IBMs Adversarial Robustness Toolbox (ART) is an open-source project, started by IBM, for machine learning security and has recently been donated to the Linux Foundation for AI (LFAI) by IBM as part of the Trustworthy AI tools. (Robust and Resilient)
			Automated audits and data protection mechanisms like homomorphic encryption, differential privacy and federated learning are powered by AI for secure data management. (Rights Respecting, Secure and reliable, Accountability)	Microsoft Smartnoise provides a mathematically measurable privacy guarantee to individuals by adding a carefully tuned amount of statistical noise to sensitive data or computations. Microsoft Presidio helps to ensure sensitive data is properly managed and governed. (Rights Respecting)
AI Augmented 	Threat actors using Generative AI to augment harmful adversarial code	Leveraging natural language processing (NLP) and deep learning to understand adversarial code generation patterns, AI-augmented detection systems can isolate malicious scripts before they are executed. This reduces the success of generative AI-driven attacks aimed at evading static signatures and rule-based defences	Real-time explanation features in AI systems assist human analysts, making this a prime example of AI-augmented decision-making. (Transparency and explainability)	IBM's Watson OpenScale to analyze your AI with trust and transparency and understand how your AI models make decisions. Detect and mitigate bias and drift. Increase the quality and accuracy of your predictions. Explain transactions and perform what-if analysis. (Fair, Transparent and Explainable)
			Logging mechanisms that support tracing and auditing with AI assistance augment the human ability to analyze malicious actions and incidents. (Transparency and Accountability, Secure and Reliable)	Splunk Phantom is a Security Orchestration, Automation, and Response (SOAR) system that logs and traces CS events using AI. (Accountability)
AI Generated 	Threat actors leveraging generative AI to create personalised phishing scams that are indistinguishable from enterprise emails	NLP, Anomaly Detection, AI-Enhanced user Behaviour Analytics are all methods that can scan and analyse abnormal requests or behaviours that result in process or successful phishing attacks	Feedback loops driven by AI-generated insights about bias or unfair outcomes ensure systems adjust autonomously based on real-world data. (Fair, Accountable)	Google's What-if Tool is an interactive visual interface designed to help visualise datasets and better understand the output of your TensorFlow models. generates insights about bias in data or models, ensuring fairness in cybersecurity decisions. (Fairness)
			AI systems generating predictive insights about potential cybersecurity threats or vulnerabilities contribute to safer operations through AI-generated strategies. (Safe, Secure and Reliable)	The MITRE ATT&CK framework (MITRE ATT&CK) is a universally accessible, continuously updated knowledge base for modeling, detecting, preventing and fighting cybersecurity threats based on cybercriminals' known adversarial behaviors. It helps predict and identifies potential vulnerabilities or attack vectors, providing insights for preemptive actions. (Safe)

Deep dives: Landscaping roles, harms and opportunities

Digital Integrity

Digital integrity is a polysemic phrase and finds varied conceptions in technical, ethical, and legal literature. In terms of technical understanding of digital integrity, it refers to the correctness of the digital object that has not been corrupted. Furthermore, digital integrity as an attribute of datasets refers to the accuracy, consistency, and completeness of data throughout its lifecycle.¹⁴⁹

Digital integrity as an ethical and legal concept is an evolving one. Recently, the Constitution of Switzerland was amended to include a fundamental right to digital integrity.¹⁵⁰ The article provides a non-exhaustive list of the rights covered under the right to digital integrity. These include the right against abusive processing of data related to digital life, the right to security, right to be forgotten among other things. Furthermore, it creates a positive obligation on the State to promote digital inclusion and awareness.¹⁵¹ While the Article covers a broad spectrum of rights and protections, it does not provide a foundational definition or meaning to the right to digital integrity itself.

Beyond this, there is limited literature on the scope and meaning of digital integrity in ethics and law,¹⁵² thus an inquiry into understanding integrity lends itself becomes imperative. In ethics, integrity as a human characteristic primarily focuses on keeping oneself intact and uncorrupted. While it is viewed as a relationship with oneself, it is also instrumental in establishing boundaries of relationships with other individuals or institutions.¹⁵³ Thereby, governing the social, political, or economic structures and their impact on an individual's integrity.

This may be observed when looking at the right to integrity under the international human rights jurisprudence. Typically, the right to personal integrity as a human right confers rights on individuals and gives rise to obligations that govern various actors and institutions.¹⁵⁴ For example, in Article 3 of the EU Charter of Fundamental Rights, the right to personal integrity protects individuals from the use of arbitrary or excessive physical force by state actors (for example, unlawful confinement) or unwarranted

use of force for certain actions (for example, political imprisonment or religious persecution).¹⁵⁵ Thus, this right governs the actions of various state institutions such as the police or judiciary. In this light, a rights-oriented approach to digital integrity allows the building of principles for governing digital structures or technologies that directly impact human experiences.

Once it is established that the right conception of digital integrity helps govern structures around individuals, it is also important to pinpoint what aspects of human experience are protected. Generally, the right to personal integrity protects individuals from arbitrary, excessive, or unwarranted physical force. However, a broader interpretation of the right to integrity¹⁵⁶ includes the right to be free from interferences to one's body and mind. This broader interpretation of the right is essentially significant in the context of digital integrity as it encompasses harms beyond the use of physical force or physical harm. **Therefore, the right to digital integrity as a concept primarily focuses on principles that govern the digital structures which directly impact human experiences online to ensure that an individual is free from interferences with her body or mind.**




Unpacking non-interference and corresponding rights

Digital integrity can be further disaggregated with a primary focus on the phrase “free from interferences to her body and mind” and will include:

- 1. Bodily and mental integrity:**¹⁵⁷ bodily integrity includes the protection of the physical well-being of an individual both online and offline. The principle is also extended to the body as data and protects personal data online. Further, mental integrity includes mental well-being and mental autonomy.
- 2. Informational self-determination:**¹⁵⁸ informational self-determination confers the right to every individual to control their information and to decide what information about them can be disclosed, to whom, and how it will be used.
- 3. Freedom:** freedom of thought, opinion, and expression is essential to protect an individual from interference to their mind. Thus, this principle covers aspects of informational integrity¹⁵⁹ and is instrumental in ensuring that information provided online is accurate, consistent, and reliable.

AI in Digital Integrity

Disaggregation of digital integrity is further instrumental in identifying the instances of when AI technologies interfere with digital integrity thus creating harms. Moreover, it provides an ethico-legal compass to navigate and generate a clear understanding of what demands to be done or what opportunities exist to mitigate these harms by leveraging AI technologies. Accordingly, the following table provides a deep dive into the harms and opportunities that arise across different AI roles:

HARMS		OPPORTUNITY	
 AI-POWERED			
Surveillance tools such as Facial Recognition Technology		Opportunity for regulatory innovation by governments and international organisations	
Scrapping personal/ sensitive personal information and non-consensual use of personal data for model training		AI tools for anonymisation of big datasets including tools	
 AI-GENERATED			
Generating synthetic content that violates privacy and bodily integrity including CSAM, deepfake pornography		AI software for detection and removal of explicit imagery	
Creating misinformation/ disinformation through synthetic content- hate speech and OBGV		AI models for detection of synthetic content	
 AI-AUGMENTED			
Algorithmic promotion of misinformation/ disinformation and hate speech		Using NLP and ML tools to real time flag fake news and hate speech.	

Cybersecurity

The integration of AI into cybersecurity dates back to 1987, when researchers first explored machine learning for anomaly detection in intrusion detection systems.¹⁶⁰ These early efforts paved the way for more advanced autonomous cybersecurity solutions, laying a foundation for today's evolving digital security landscape. The recent acceleration in AI's adoption in cybersecurity is largely driven by the global shift to digital infrastructures across sectors like education,¹⁶¹ finance¹⁶² and healthcare,¹⁶³ which was expedited by the COVID-19 pandemic. This transition has not only amplified the demand for robust cybersecurity measures but also highlighted the potential of AI to revolutionise defence mechanisms.

As both individuals and organisations become more reliant on online systems for critical operations, the need for airtight cybersecurity has never been more pressing. AI plays a transformative role in this domain by offering unprecedented processing power,¹⁶⁴ pattern recognition,¹⁶⁵ and decision-making accuracy.¹⁶⁶ These capabilities allow for enhanced threat detection, faster response times, and a stronger overall resilience to cyber threats.¹⁶⁷

However, the introduction of AI in cybersecurity also brings a paradoxical dynamic. While AI enhances security systems and deters cyber criminals, it also introduces vulnerabilities that can be exploited. Sophisticated attackers could manipulate AI algorithms¹⁶⁸ or feed them tainted data to bypass detection systems. This underscores the need for continuous vigilance, frequent updates, and refinement of cybersecurity strategies. AI's use also raises concerns about bias, transparency, and accountability, especially in decision-making processes that are becoming more automated.

The convergence of AI and cybersecurity is a double-edged sword: it offers significant potential for advancing cyber defences but simultaneously creates new risks that demand careful consideration. A multi-stakeholder dialogue is essential for innovation in this space, ensuring AI's full potential is harnessed responsibly while minimising inherent risks.¹⁶⁹

AI's application in cybersecurity has transformed how organisations defend against and respond to digital threats. Its scope spans several critical domains, offering cutting-edge solutions that enhance the detection, prevention, and mitigation of cyber attacks.

The significance lies in its ability to enhance defence mechanisms against the increasingly complex and sophisticated nature of cyber threats. Traditional rule-based systems struggle to keep up with the volume and velocity of attacks, but AI's advanced capabilities—such as real-time data analysis, anomaly detection, and predictive analytics—enable organisations to identify and respond to threats faster and more effectively. AI's capacity to continuously learn and adapt allows it to detect evolving tactics, techniques, and procedures (TTPs) used by cyber criminals, reducing the risk of zero-day attacks and human error.¹⁷⁰ Additionally, AI helps bridge the cybersecurity talent gap by automating routine tasks, optimising resource allocation, and enabling cybersecurity teams to focus on more strategic interventions. However, as AI strengthens defensive strategies, it is also employed by threat actors, heightening the need for vigilant and responsible AI deployment. The integration of AI into cybersecurity is crucial for ensuring the resilience and safety of digital infrastructures in an increasingly interconnected world.

AI for Blue Teaming: Defensive Security

As organisations face mounting cyber threats and a shortage of skilled professionals, AI's role in defence (commonly referred to as blue teaming) has become indispensable. AI excels in analysing vast and diverse data sets—ranging from logs, events, and user behaviour to network activity—across an entire organisation. This capacity allows for real-time detection of anomalies and suspicious behaviour, enabling proactive responses to potential breaches before they escalate.

AI-driven systems offer continuous monitoring of network traffic, reducing the risk of human error through automation. They can detect novel patterns and threats, including zero-day attacks, that traditional rule-based systems often miss.¹⁷¹ AI's ability to scale and adapt is crucial in today's environment, where vast amounts of data and connected devices constantly increase. Additionally, AI

streamlines compliance processes, generating reports and triggering necessary actions, thereby reducing the burden on cybersecurity teams.

Moreover, AI can identify and neutralise threats before they materialise into full-scale attacks, helping prevent costly data breaches. The automation of routine tasks, such as patch management and vulnerability scanning, allows cybersecurity teams to focus on higher-level strategic decisions, thus increasing the overall efficiency of cybersecurity operations.¹⁷²

AI for Red Teaming: Offensive Security

While AI significantly strengthens defence mechanisms, it is equally important to recognize its growing use in offensive security (red teaming) by cyber criminals. During the Bletchley AI Safety Summit in November 2023, global leaders addressed how AI is not only enhancing cybersecurity defences but also empowering threat actors.¹⁷³ Several key judgments highlighted the risks associated with AI's irresponsible deployment, noting that AI will inevitably amplify the frequency and severity of cyberattacks.¹⁷⁴

AI's role in offensive tactics is growing, with threat actors leveraging AI tools to evolve their tactics, techniques, and procedures (TTPs). Technologies like Generative AI (GenAI) streamline reconnaissance, social engineering, and attack planning,¹⁷⁵ making these malicious activities more efficient and harder to detect. The use of AI in analysing exfiltrated data allows cybercriminals to sift through stolen information rapidly, expediting subsequent attacks.

Furthermore, threat actors are training AI models using stolen data, which compounds the effectiveness of their operations. Techniques like model extraction, model inversion, and model stealing enable attackers to reveal sensitive information through public-facing APIs, relying on input data to bypass security. As AI continues to evolve, these threat actors can automate and scale attacks, reducing the technical barriers for novice hackers, hacktivists, and cybercriminals-for-hire.

The rise of AI in offensive security brings with it the threat of an increasing preponderance of cybercrime. AI lowers the entry

barriers for attackers, allowing even individuals with limited expertise to execute sophisticated cyber attacks. This further underscores the need for AI-enhanced defence mechanisms and highlights the escalating importance of AI in the cybersecurity arms race.¹⁷⁶

AI in Purple Teaming

AI plays a transformative role in enhancing purple teaming by optimising collaboration between red and blue teams. In purple teaming, AI can be used to simulate advanced attack scenarios, identify vulnerabilities, and continuously learn from red team activities. AI tools analyse vast amounts of data from simulated attacks, enabling the blue team to understand patterns, automate threat detection, and strengthen defences. By using machine learning models, AI can generate real-time insights, making it easier for both teams to adapt and refine their strategies. This integration of AI not only accelerates the detection of weaknesses but also automates remediation, leading to more dynamic and effective defence systems.¹⁷⁷

AI in Security Penetration Testing (Pentesting)

AI significantly enhances security penetration testing by automating the identification of vulnerabilities and potential attack vectors. This is done by experts who simulate an attack on a computer system to identify weaknesses in its defenses. AI-powered pentesting tools can mimic the behaviour of advanced persistent threats (APTs), using pattern recognition and predictive analytics to discover weak points that human testers might overlook. These tools can simulate a range of cyber attacks at scale, analyse network traffic, and detect misconfigurations or exposed endpoints in real-time. AI's ability to continuously learn from previous tests also helps improve the accuracy and depth of security assessments, allowing organisations to address critical issues faster and with greater precision. This AI-driven approach elevates traditional pentesting, making it more efficient and effective in protecting against ever-evolving cyber threats.

AI in Cybersecurity




The evolving relationship between AI and cybersecurity can be analysed through sector-specific lenses to understand its roles, harms, and opportunities. For instance, in financial services, AI plays a critical role in detecting fraud and insider threats, while in healthcare, AI is employed to safeguard patient data and ensure compliance with privacy regulations. In critical infrastructure, AI helps secure vital systems against state-sponsored attacks and advanced persistent threats (APTs).

By adopting a sector-specific approach, we can explore how AI intersects with cybersecurity across industries, identifying both opportunities for innovation and risks that must be mitigated. AI's growing influence in cybersecurity will undoubtedly shape the future of defense strategies, with its integration continuing to evolve in response to emerging threats.

We chose to incorporate a Roles, Harms, and Opportunities (RHO) framing to dissect the nuanced relationship between the risks and benefits arising from AI's intersection with cybersecurity. This approach helps to systematically evaluate how AI tools and technologies are deployed across different cybersecurity functions, whether for safeguarding systems (as seen in blue teams) or for testing their resilience through simulated attacks (as executed by red teams).

The lines between these uses may appear blurred, as both defenders and adversaries often leverage similar AI-driven methods, such as anomaly detection, adversarial techniques, or behavioral analysis. By drawing clear distinctions and exploring the multifaceted applications of AI, we can better identify gaps, overlaps, and areas of improvement within the cybersecurity landscape. This structured understanding allows us to propose effective mitigation strategies and adopt best practices, ensuring that as AI continues to permeate cybersecurity operations, it enhances security and resilience rather than introducing new vulnerabilities.

Across sectors, AI presents compelling cybersecurity opportunities that far outweigh its harms:

HARMS		OPPORTUNITY	
 AI-POWERED			
Threat actors leveraging adversarial ML to target ML models		An emerging machine learning (ML)-powered cybersecurity opportunity lies in the protection of ML models themselves through ML powered adversarial defence systems	
 AI-GENERATED			
Threat actors leveraging generative AI to create personalised phishing scams that are indistinguishable from enterprise emails		NLP, Anomaly Detection, AI-Enhanced user Behaviour Analytics are all methods that can scan and analyse abnormal requests or behaviours that result in process or successful phishing attacks.	
 AI-AUGMENTED			
Threat actors using Generative AI to augment harmful adversarial code		Leveraging natural language processing (NLP) and deep learning to understand adversarial code generation patterns, AI-augmented detection systems can isolate malicious scripts before they are executed. This reduces the success of generative AI-driven attacks aimed at evading static signatures and rule-based defences.	

Connecting the Dots: Trustworthy AI for digital integrity and cybersecurity

As highlighted above, digital integrity and cybersecurity are two essential areas of enquiry for achieving online safety. It is established that these two spheres are affected due to proliferation of AI tools. However, at the same time, these tools also present a gamut of opportunities for strengthening and bolstering of digital integrity and cybersecurity online. To this end, it is essential to formulate governance mechanisms geared towards developing human-centric and robust AI systems to promote integrity and security. These governance mechanisms may be a mix of regulatory and techno-societal mechanisms. Accordingly, we adopt a trustworthy AI governance approach that informs AI adoption, as well as proposes mitigation strategies to address harms that may potentially arise from AI integration across the domains of interest.

The reasons for adopting a trustworthy AI governance approach are twofold. First, trustworthy AI frameworks are well embedded within academic and regulatory efforts, and are even adopted by several countries for governing AI technologies.¹⁷⁸ This allows us to develop strategies which are aligned with policy perspectives and priorities of multiple jurisdictions and global standards. Second, trustworthy AI is well suited for techno-societal approach to governance, as it not only focuses on the technical requirements of AI systems, but layers those requirements with the human-centric characteristics of trust.¹⁷⁹ Trust thus provides a balanced mix of technical attributes of robustness, reliability, security¹⁸⁰ but also socio-legal principles such as transparency, accountability¹⁸¹ etc.

This section examines various policy frameworks for trustworthy AI adopted globally, theoretical underpinnings of trust, levers of trustworthy AI and relevant stakeholders. Furthermore, the section will justify the significance of adopting this framing for digital integrity and cybersecurity and the inter-relationship of trust with digital integrity and cybersecurity respectively.

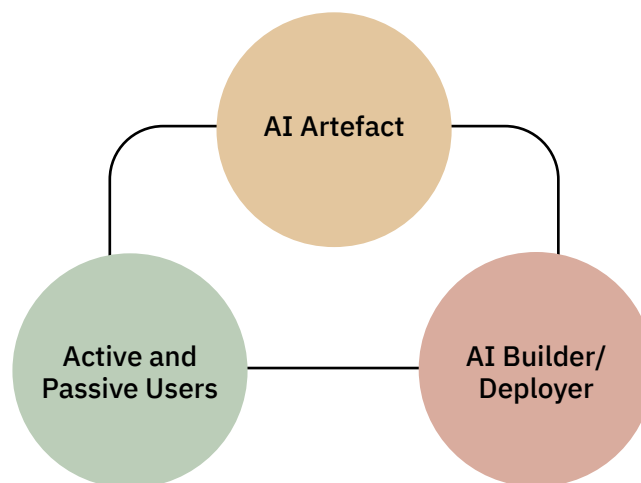
Meaning and scope of trustworthy AI

Trust is understood as a mental state. It may be intuitive and tacit or based on an explicit promise or commitment. Typically, trusting a person or an object requires that the trustor can be **vulnerable** to the trustee's actions, **rely on the trustee to be competent** to do what the trustor wishes them to do, and **rely on them to be willing** to do it.¹⁸² Trust may further be between two individuals, understood as **interpersonal trust**, or at the level of society or groups, understood as **social trust**.¹⁸³ Interpersonal trust is based on the characteristics of the two individuals, their relationship and familiarity, and the context in which the trust is being placed. Interpersonal trust is normally guided by personal morals, social norms, or legal tools such as contracts.

On the other hand, social trust deals with a much broader concept covering a range of trust relationships in society across individuals, groups, communities, and collectives. Additionally, it also deals with relationships among individuals, organisations, and institutions. Pertinently, social trust is increasingly influenced by social structures and positioning of actors, legal frameworks and structures such as the constitutional or criminal justice systems, and economic structures as well. It must also be noted that apart from rules and social mores, trust is also guided by the risks, benefits, and other trade-offs based on the context relationship.¹⁸⁴

Social trust can further be divided into **horizontal and vertical**, where horizontal trust is within individuals and groups; vertical trust encompasses relationships of trust between individuals and institutions, typically governmental and state institutions.¹⁸⁵ This distinction becomes all the more important when looking at the relationship of trust between individuals and artificial intelligence systems. Trust in artificial intelligence systems can be understood using a vertical trust framework to define relationships of trust among AI artefacts, AI users, and the AI builder.¹⁸⁶

Where, **“AI artefacts”** mean a wide range of AI models and technologies including foundation models, applications of foundation models, softwares or applications with AI integrations etc. Further, **“AI builder”** includes entities and organisations developing AI models and other players responsible for deployment and integration of AI models in other technologies. **“AI**



users” include **active and passive users** of AI technologies, where **active users** are characterised by direct use and interaction with the technology whereas **passive users** mean indirect or mediated use and interaction with the technology.

This vertical trust framework provides an important lens to understand the existing trustworthy AI frameworks for the following reasons:

1. **Actors and elements:** The vertical trust framework allows for the breaking down of trust as a relationship between various actors and identifies essential elements of the relationship of trust.¹⁸⁷ In this context, the relationship of trust and its elements can be identified as follows:

RELATIONSHIPS	LEVERS
Between AI tool and user	The relationship largely depends on the competence of the tool to perform the task in the expected manner consistently.
Between AI builder and user	This relationship is significantly layered by the power dynamic between the user and the developer of the tool. ¹⁸⁸ Thus, to create a relationship of trust, there is a need to adopt measures that rebalance the power dynamic. These measures may include organisational governance policies, transparency and accountability mechanisms.

- 2. *Functions:*** The vertical framework further puts emphasis on the need to further contextualise trust within existing structures and other relevant factors influencing the relationship of trust. These include legal, social and cultural structures within a local context, as well as, global considerations such as geopolitics of AI tech, global hegemonies and colonial histories.
- 3. *Reciprocity of relationship:*** Furthermore, a vertical setting acknowledges the positioning of various actors in the AI ecosystem, the resultant power dynamic and reciprocities between the users of an AI system, the AI tool, and the creators of the tool.

Thus, accordingly, the next section traverses trustworthy AI frameworks to identify the essential elements governing the identified relationships of trust. Further, the section will also identify key structural factors that affect these relationships of trust and the role of legal and social structures.

Landscaping: Frameworks and other policy mechanisms

Trustworthy AI frameworks have become ubiquitous in governance approaches to AI in multiple jurisdictions across the world as well as a key component of international policy for governing AI. Trustworthy AI frameworks, broadly, provide foundational principles for AI models to be trustworthy as well as risk management guidance. These principles largely rely on a mix of technical methods and ethico-legal principles for governance.

European Union; EU Ethics Guidelines for Trustworthy AI (2019),¹⁸⁹ EU AI Act (2023)¹⁹⁰

The Guidelines have two parts.

Part I: Essential elements/conditions of trust that need to be met throughout the lifecycle of AI -

- it should be **lawful**, complying with all applicable laws and regulations;
- it should be **ethical**, ensuring adherence to ethical principles and values; and
- it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

Part II: Principles for **realising trustworthy AI** -

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

OECD (Original version: 2019; Updated: May, 2024)¹⁹¹

Principles of trustworthy AI:

- Inclusive growth, sustainable development, and well-being
- Human rights and democratic values including privacy and fairness
- Transparency and explainability
- **Robustness, security, and safety**
- Accountability

US Executive Order on Safe, Secure, and Trustworthy Development of AI (2024)¹⁹² and NIST Risk Management Framework (2024)¹⁹

Characteristics of Trustworthy AI include:

- Valid and reliable
- Safe, secure, and resilient,
- Accountable and transparent,
- Explainable and interpretable,
- Privacy-enhanced,
- Fair with harmful bias managed.

Levers of trustworthy AI

Based on a review of emerging regulatory developments, policy frameworks and ethical guidelines released by various actors, including multilateral organisations, state institutions and dominant technology corporations, we have arrived at the following levers to guide our research. These levers influence the relationship of trust between the AI tool and user as well as the relationship of trust between the AI builder and user. The levers function as guiding principles for the development of mitigation strategies and scaffoldings around the opportunities. Moreover, these levers directly contribute towards promotion of digital integrity and cybersecurity as highlighted in the next section.



Rights respecting

As observed from the existing trustworthy AI frameworks and policies, there are multiple principles which require protection of or adhering to basic human rights including non-discrimination, privacy, liberty and dignity, human autonomy and agency. Thus, the first and foremost principle is rights respecting AI.



Secure and reliable

This principle requires AI systems or tools to be able to perform as expected under the given conditions and for a wide range of inputs. Reliability is essential for ensuring trustworthiness, as an AI system that produces inaccurate or inconsistent results under the same conditions gives rise to anticipated and unanticipated risks.



Safe

The safety of AI systems is directly related to the protection of several human rights, especially dignity and integrity. Accordingly, this principle requires AI systems to be built in such a manner that they do not endanger human life. Safety is crucial for promoting the adoption of AI, especially in high-risk or high-stake sectors such as healthcare or finance.



Robust and Resilient

Resilience principles entail that an AI tool must withstand unexpected or adverse events and be able to fall back to the status quo. The resilience principle is closely related to reliability and complements it. Furthermore, resilience contributes to the trustworthiness of AI by minimising the risks that may arise from adverse events.



Transparency and explainability

Transparency and explainability require adequate disclosure of relevant information about the AI models and datasets, including measures such as data provenance, model provenance, informed consent, etc.



Accountability

Accountability refers to meaningful opportunities to place responsibility for AI systems and their outcomes. Accountability can include measures such as auditing, reporting of risks, and other organisational policies.



Trustworthy AI for Digital integrity














Rights respecting	This lever promotes digital integrity as a human right of being free from unwarranted interferences online. Rights such as privacy, human autonomy and dignity directly feed into digital integrity.
Secure and reliable	Reliable and consistent AI tools play a critical role in maintaining information integrity and freedom of thought online. Additionally, secure AI systems ensure that user privacy is not compromised.
Safe	As this pillar promotes mitigation of AI risks to human safety both physical and online, it promotes protection of humans from abuse of personal information online and violent or abusive content online. Thus this principles acts as scaffolding around AI tools from being detrimental to physical and mental integrity.
Robust and resilient	Robustness of AI tools determines the capacity of AI tools to perform safely even in adverse events. Thus, robust AI tools will ensure that information provided online is correct and accurate. Moreover, robustness acts as guardrails preventing AI tools from creating hateful or violent content. Thus promoting informational and personal integrity online.
Transparent and explainable	Transparent and explainable AI models play a significant role in promoting integrity online by providing crucial information on the functioning of the AI models such as data provenance, model provenance. Thus, allowing for meaningful scrutiny over AI model function and pursuing redressal mechanisms.
Accountability	Lastly, this lever is critical in realising digital integrity by providing an opportunity to seek redressal for violation of integrity online.

Trustworthy AI for Cybersecurity


Rights respecting	Privacy enhancement through robust cybersecurity measures go in tandem to protect data from unauthorised access and misuse.
Secure and reliable	How critical infrastructure systems can fail and how they can be targeted coupled with the design of autonomous yet robust monitoring systems that detect and prevent attacks and failures.
Safe	AI models are being integrated into products for anomaly detection, for example. As the AI interventions increase in deployment, it is imperative to reduce the number of false positives that come along with it.
Robust and resilient	AI systems need to be able to withstand unexpected adverse events or unexpected changes in their environment or use. This is imperative in the face of internal and external adversity equipping systems to degrade safely and gracefully whenever applicable.
Transparent and explainable	Transparent AI cybersecurity systems can enhance the feedback loop towards cybersecurity by making the underlying decision-making processes understandable and verifiable for security teams. We are seeing an introduction of chatbots that assist cybersecurity specialists which contributes to explainability.
Accountability	Accountability in cybersecurity ensures that there are clear mechanisms for defining responsibilities and providing redress when these systems cause harm.

Implementation strategies and recommendations

Mitigation strategies for Digital integrity

HARMS	OPPORTUNITY	MITIGATION	
 AI-POWERED			
Abuse of facial recognition technology	Opportunity for regulatory innovation by governments and international organisations	Translation of legal and ethical requirements of using FRTs and biometric technologies into technical standards for easy adoption and enforcement	 
Scrapping personal/ sensitive personal information online	AI tools for anonymisation of big datasets	Adoption of differential privacy to train AI models	
		Adopt scraping prevention tools	
		Embed privacy-centered licensing and contracting practices	
 AI-GENERATED			
Synthetic content for misinformation, disinformation, and other harmful and violent content	AI software for detection and removal of fake news or explicit imagery	Enhance transparency and robustness of AI tools	
		Conduct robust testing, consistent auditing and oversight	 
 AI-AUGMENTED			
Algorithmic promotion of misinformation/ disinformation and hate speech	Using NLP and ML tools to real time flag fake news and hate speech	Conduct regular internal and third-party audits	
		Collaborate on cross platform moderation	

The mitigation strategies are discussed in detail in the following pages, and substantiated with examples from policy and industry practices to demonstrate how they function in action. Most crucially, we have attempted to map the examples to T-AI levers to show how the mitigation strategies when adopted could help adherence to the trustworthiness frameworks mentioned in the report.

	AI-POWERED	TRUST LEVERS: ACCOUNTABILITY, RIGHTS RESPECTING
<p>Harm 1:</p> <p>Use of AI-powered facial recognition technologies can result in arbitrary or excessive surveillance, breach of privacy, and discriminatory policing.</p>		
<p>Mitigation Strategy</p> <p>Adopt a techno-legal approach to FRT design: Embed ethical and legal principles within standards and design of FRT technologies. Principles such as de-identification of biometric data and interoperability of data and systems baked into the ISO standards have proven to be effective in ensuring adoption of privacy-enhancing technologies and accountability.¹⁹⁴ Moreover, principles of privacy by design and privacy by default as articulated under Article 25 of the GDPR, prioritise embedding privacy as a technical and design decision from the very start.¹⁹⁵ This ensures that principles including data minimisation, purpose limitation, and storage limitation are adopted not just for compliance but are an integral part of design choice. Lastly, to ensure that facial recognition technologies are not discriminatory or biased, it is essential to adopt fairness frameworks and standards across the value chain development.¹⁹⁶</p>		



Harm 2:

Scraping personal and sensitive personal information online. As AI technologies are data intensive and their efficacy is often associated with access to high volumes of data, the development of these models often relies on information and data scraped from the internet using crawlers or bots. This raises challenges to privacy online and creates vulnerabilities arising from non-consensual use of sensitive personal information.

Mitigation Strategies

- **Adopt differential privacy:** The differential privacy method guarantees the preservation of individual privacy by ensuring that sensitive personal information is not reflected or has minimal impact on the aggregate output.¹⁹⁷ In simple terms, differential privacy methods introduce enough noise, through a randomised algorithm, into two datasets with similar records so that the unique records in the two datasets are not reflected in the aggregated outcome. Adopting differential privacy helps in masking the identifiable information in the datasets and also prevents attack agents from finding identifiable information.
- **Adopt scraping prevention tools:** At present, there are several tools that prevent data scraping using different techniques such as firewalls or softwares that detect suspicious behaviour on the websites.¹⁹⁸ Website publishers can use these tools to detect a web crawler or a bot and automatically block them. Moreover, simple tools such as captcha have proved to be helpful in preventing bots from accessing the websites. Further, to prevent misuse of personal information, such as images from social media, using tools that make tiny changes to the pixels of an image that are not visible to the human eye can be used. For example, Glaze or NightShade are simple tools developed by University of Chicago, that make simple changes to an image and make it extremely difficult to understand for an AI model.¹⁹⁹ While these tools are originally made for protecting copyright interests of artists, these can possibly be used for masking sensitive personal information in our images online.
- **Embed privacy-centered licensing and contracting practices:** This strategy is two layered. First, websites can contain strict terms and conditions that prevent users from scraping data from their website. Moreover, while entering any licensing agreement for the use of data from a website, terms and conditions can be added to ensure that personal sensitive information is not used for AI training. Second, the websites that aim to use personal sensitive information online must provide clear notice to the users about the use of information for AI training. This ensures that the users get a real meaningful opportunity to give or revoke consent for their data use.

**Harm 1:**

Synthetic content creation using generative AI tools has significantly impacted the online information ecosystem. Generative AI is being used to produce a slew of misinformation and disinformation, disrupting the integrity of the informational ecosystem, online as well as offline. Moreover, synthetic content is being used to create hateful content targeting religious, ethnic or racial groups, gender and sexual minorities, etc. Further, sythetic content has intensified online gender based violence and privacy violations through creating deepfakes and child sexual abuse material (CSAM).

Mitigation Strategies

- **Enhance transparency and robustness of AI tools:** While AI powered tools for detection of false information will be critical for combating spread of misinformation or disinformation,²⁰⁰ there is also a need to embed ethico-legal standards of transparency in development and use of AI tools.²⁰¹ At present use of generative AI presents a significant risk of hallucination that can result in incorrect information or inferences being spread. Thus, media platforms using AI generated content must provide transparency disclosures to ensure that users are always well aware of interacting with AI generated content.²⁰² Moreover, in sensitive informational ecosystems such as health and finance, there must be transparency around the AI models being used and the training datasets for public scrutiny and auditing.²⁰³
- **Conduct robust testing, consistent auditing and oversight:** For highly capable generative AI tools there is a need to build scaffoldings to ensure that AI is not used to create content that may be violent, harmful or hateful, especially towards vulnerable demographics. This includes measures such as ensuring that datasets do not contain CSAM,²⁰⁴ hate speech or violent imagery. Hashing techniques and data provenance can be used for identifying and tagging CSAM or other harmful content within datasets. Further, outputs must be put under rigorous prompt testing and consistent monitoring post deployment. Allow users to flag content and provide feedback for harmful content and establish procedures for taking retrospective action.²⁰⁵











Harm 1:

Algorithmic recommendation systems on online media platforms can result in further amplification of harmful or hateful content online, often understood as a filter bubble or echo-chamber effect.²⁰⁶ This may have adverse impact on safety and privacy of groups and individuals, and informational integrity.

Mitigation Strategies

- **Conduct regular internal and third-party audits:** Internal auditing of recommender systems by platforms for risks and impact assessment is crucial for ensuring that the algorithms are not exacerbating or amplifying harmful content. With the legal recognition of internal and external audit requirements,²⁰⁷ it is all the more critical to formally adopt the practice. Third party auditing techniques can include code/ data audit where the code and datasets are made available to auditors, crowd source auditing where data is collected from user collectives, or API audits where user data is accessed by auditors through platform provided APIs.²⁰⁸
- **Collaborate on cross platform moderation:** A partnership between online media platforms to build shared datasets, identification and filtering systems that curb the spread of harmful online content through recommender systems.²⁰⁹ These cross platform initiatives²¹⁰ can include measures such as sharing email address or user name of offenders, digital hash of CSAM, shared databases of harmful/ violent language, metadata, etc. There have been several initiatives such as the Lantern by Tech Coalition, which are supported by several large-scale online platforms resulting in meaningful cross platform moderation for harmful content.

Mitigation strategies for Cybersecurity

HARMS	OPPORTUNITY	MITIGATION	
AI-POWERED			
Threat actors leveraging adversarial ML	AI powered interventions for blue teaming	Dynamic AI threat intelligence	
		AI driven fault tolerant systems	
AI-GENERATED			
Threat actors using AI to improve and encrypt malicious code	Advanced NLP and AI powered Anomaly Detection	AI systems with real time explanation features to support threat analysts	
		AI assisted logging mechanisms	
AI-AUGMENTED			
Threat actors leveraging generative AI to write phishing emails	Feedback loops driven by AI insights using real-time data	AI generated insights that bolster systems using real-time data	 *
			
	AI anomaly detection to identify abnormal requests and behaviour	Advanced AI for predictive analytics to scan for threats and vulnerabilities	 



* **Principles of fairness in AI** as represented by the 'All about the Bias' framework also has implications for cybersecurity. AI bias can cause false positives, mistakenly flagging harmless content as threats. Over-classification risks arise when AI detection systems fail to differentiate slang and code words from phishing, leading to unnecessary alerts.



- i** These strategies rely on AI systems to autonomously execute tasks, such as analyzing threats, mitigating risks, or ensuring compliance.

Harm 1:

Threat actors leveraging adversarial ML to target ML models

Mitigation Strategies

- **Optimise for automated threat intelligence and self updation:** To address the harm of threat actors leveraging adversarial machine learning (ML) to target models, the first strategy involves *optimising for threat intelligence and self-updating mechanisms*.²¹¹ This means AI systems can integrate real-time threat intelligence from global sources to anticipate evolving adversarial tactics. Additionally, self-updating ML models using continual learning techniques can dynamically adapt to new threats, reducing reliance on static updates and maintaining robust defenses. Lastly, collaborative threat-sharing networks can enhance communal defenses by pooling insights into adversarial patterns, ensuring that organizations remain proactive against emerging threats.²¹²
- **Leverage AI driven fault tolerant systems and adversarial training:** Fault-tolerant architectures, such as ensemble learning, ensure operational integrity even under partial failure caused by adversarial inputs. Adversarial training exposes models to maliciously crafted examples during development, enabling them to recognize and counteract similar attacks in real-world scenarios. Regular resilience testing, including red team simulations of adversarial attacks, further refines these defenses, ensuring that systems remain reliable even when faced with sophisticated threats.
- **Automate the audit process and data protection mechanism:** The third strategy emphasizes *automating audit processes and protecting data throughout its lifecycle*.²¹³ AI-powered auditing tools continuously monitor model inputs, outputs, and behavior to detect anomalies, verify compliance with security protocols, and identify early signs of adversarial manipulation. Data integrity can be safeguarded using mechanisms like blockchain for traceability,²¹⁴ differential privacy,²¹⁵ and homomorphic encryption,²¹⁶ ensuring that sensitive information remains secure from tampering or exposure. By combining automation and robust data protection, organizations can minimize vulnerabilities and strengthen their defenses against adversarial ML threats.



- i** These strategies involve AI generating outputs or insights that can inform decisions or actions but do not necessarily involve active human participation during the generation process.

Harm 1:

Threat actors leveraging generative AI to create personalised phishing scams that are indistinguishable from enterprise emails

Mitigation Strategies

- **Capitalise on Feedback Loops Driven by AI -Generated Insights:** Feedback loops powered by AI-generated insights play a crucial role in detecting and countering personalized phishing attempts. AI systems can continuously monitor phishing detection performance and analyze outcomes to identify biases or unfair practices, such as disproportionately flagging legitimate communications as threats or overlooking sophisticated scams. These systems adjust autonomously based on real-world data, refining their ability to discern subtle differences between legitimate and malicious emails. For example, AI can assess factors such as linguistic nuances, metadata, and sender authenticity in real time. If a phishing attempt initially bypasses detection, feedback loops ensure the system learns from this oversight, enhancing future detection capabilities. This iterative approach also reduces false positives, allowing organizations to streamline their email security processes while maintaining robust defenses against generative AI-driven phishing attacks.
- **Leverage Predictive AI Systems for Cybersecurity Threat Insights:** AI systems capable of generating predictive insights about potential phishing threats significantly enhance cybersecurity operations. By analyzing patterns in past phishing campaigns, generative AI can anticipate new attack vectors and provide actionable strategies for preemptive mitigation. For instance, these systems can identify vulnerabilities in enterprise email infrastructure or detect compromised employee accounts that could be exploited to launch phishing scams.

AI-generated predictive models can also prioritize threats based on their potential impact, helping organizations allocate resources effectively. For example, by recognizing a spike in phishing attempts targeting specific departments (e.g., finance or HR), AI can recommend tailored training sessions for employees in those areas or suggest updates to email filters and authentication protocols. Additionally, predictive AI insights enable the creation of adaptive security policies, such as dynamic email filtering rules that evolve in response to emerging phishing tactics.



- i** These strategies leverage AI to assist or enhance human decision-making, providing actionable insights or streamlining complex processes.

Harm 1:

Threat actors using Generative AI to augment harmful adversarial code

Mitigation Strategies

- **Enable AI-augmented decision making:** To address the harm of threat actors using generative AI to augment harmful adversarial code, the first strategy involves *real-time explanation features for cybersecurity analysts*. Generative AI has the capability to analyse vast datasets in real-time to detect patterns and anomalies that are indicative of malicious activity, such as dynamically generated adversarial code. By leveraging *explainable AI principles (XAI)*, these systems can break down complex decision making processes into comprehensible explanations for human analysts. For example, Palo Alto Networks Cortex XDR combines AI with endpoint and network telemetry to detect advanced threats.²¹⁷ These threats include malicious scripts. flagging a piece of code as malicious could highlight specific attributes or behaviors such as structural similarities to known malware, abnormal execution patterns, or evasion techniques that triggered the alert. This transparency not only helps analysts trust the AI's recommendations but also allows them to prioritize threats more accurately. Additionally, real-time explanation features can guide analysts in choosing the most effective countermeasures. For instance, if an AI system identifies adversarial code exploiting a specific vulnerability, it can provide a step-by-step explanation of the exploit mechanism and recommend targeted remediation strategies, such as patching affected systems or updating firewall specifications.²¹⁸ These explanations can also include risk assessments, helping analysts understand the potential impact of the threat if left unaddressed.²¹⁹
- **Offer AI-assisted logging mechanisms that support tracing and auditing for human analysts:** The second strategy focuses on AI-assisted logging mechanisms that support tracing and auditing. Generative AI can enhance logging systems by analyzing and contextualising log data, identifying anomalies, and correlating suspicious activities across distributed systems.²²⁰ These mechanisms enable the tracing of attack vectors back to their source, assisting in incident investigation and attribution. Advanced AI-driven auditing tools can also assess system compliance with security protocols in real time, ensuring vulnerabilities exploited by adversaries are identified and addressed promptly. By automating these processes, organizations can maintain comprehensive oversight of their systems, reducing the window of opportunity for generative AI-enabled threats.

Way Forward



Way Forward

AI is rapidly transforming and revolutionising multiple spheres of our lives, as people increasingly interact with AI tools, directly or indirectly. Governance of AI has hence become an important question for policymakers, AI developers and users alike. Moreover, the AI question demands a holistic interdisciplinary evaluation bringing together perspectives of all significant rightholders and stakeholders. Policymakers and regulators across the globe are not only thinking about governance of AI for protection of individuals from the harms but also to leverage AI tools for promotion of developmental goals. Thus, the primary focus of regulation is to foster a conducive environment for AI innovation. Accordingly, the trends reflect many countries opting for soft laws and policies such as regulatory sandboxes, governance frameworks or voluntary standards to promote responsible and sustainable development of AI.

Through this report we aim to supplement these regulatory efforts in three major ethical and functional realms of bias, digital integrity and cybersecurity. To this end, we have developed strategies for mitigating harms arising from different sources of bias, namely, pre-existing, technical and population bias. Using a design-centric approach to bias across AI development value chain, the recommendations provide holistic multi-stakeholder strategies for combating bias in AI. Similarly, through the 'Roles, Harms and Opportunities' (RHO) framework, our report highlights the duality of impact that AI has on digital integrity and cybersecurity. The opportunities and mitigations section presents ways in which trustworthy AI principles can be leveraged to use AI for protection and promotion of digital integrity and cybersecurity. Thus, in totality, the recommendations of the report present a gamut of pathways for AI governance that can complement existing or upcoming regulatory mechanisms.

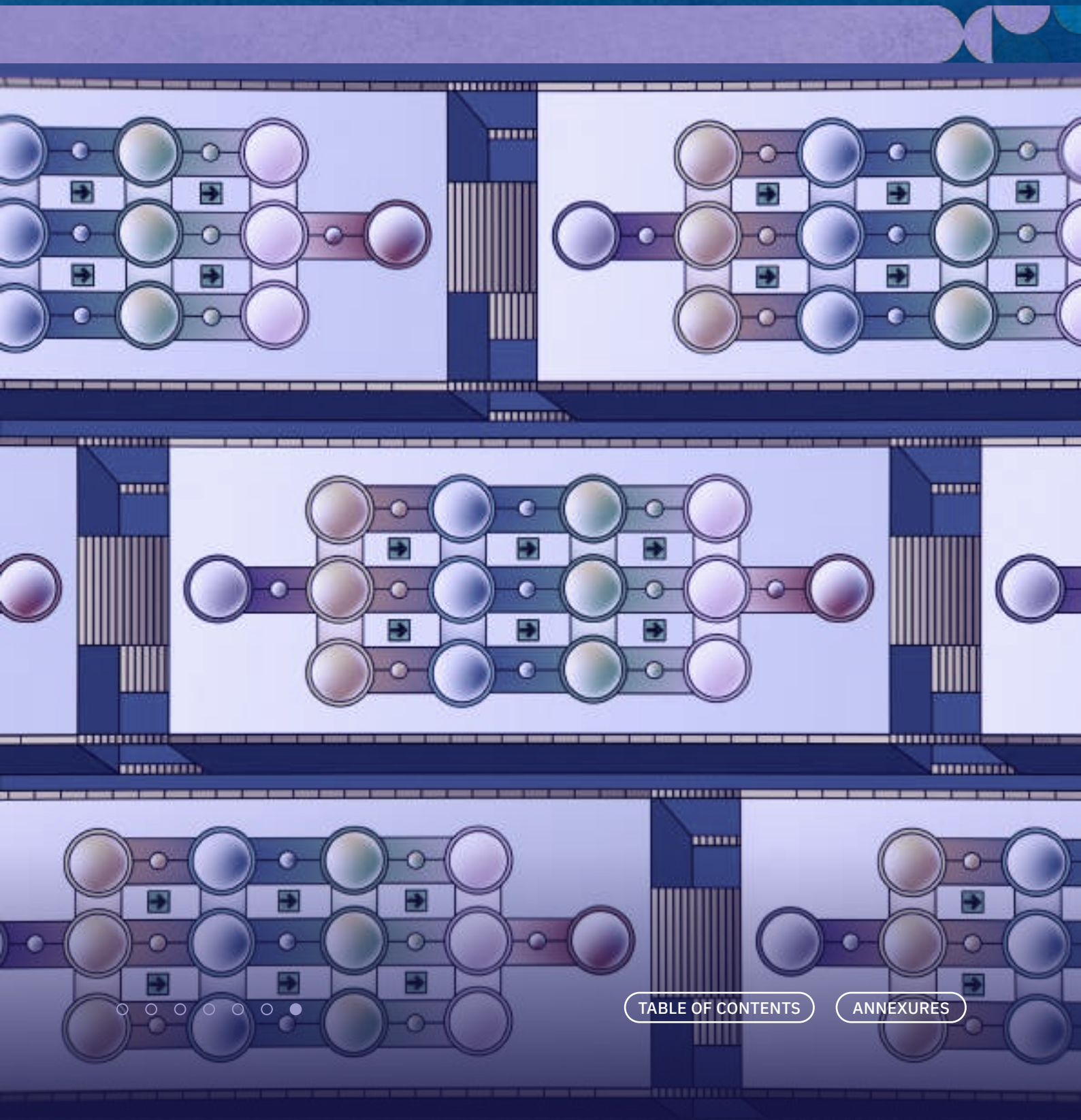
While the report engages with broad areas of bias, digital integrity and cybersecurity - the discussion and mitigation strategies are

limited to the processes of AI model development and do not deal with ancillary debates on computational infrastructure and data labour. The discourse on computational infrastructure, especially revolving around the concentration of compute capabilities with a small number of players, is a critical one.²²¹ Compute is proving to play a crucial role in AI development and deployment, as well as, regulatory considerations at a global scale.²²² It is also being predicted to be a critical factor in geopolitical power dynamics in 2025, where national governments will rely on private transnational corporations for the building blocks of AI technologies.²²³ Thus, any future steps on the governance of AI must bear in mind the compute considerations and widespread impact of the same. Similarly, discourse on data annotation work is another crucial area in the AI development value chain that is gaining momentum. As data annotators form the backbone of developing robust AI tools, several questions on fair working conditions and compensation demand urgent attention.²²⁴

Lastly, moving forward this year, there is a need to develop a global AI regulatory agenda where countries have a shared vision on responsible and trustworthy development of AI technologies. This must be coupled with national efforts to understand and mitigate AI-related risks in local contexts. While, multiple major jurisdictions established AI Safety Institutes last year to cover an institutional gap in AI policy,²²⁵ The Safety Institutes have not only been envisioned as crucial players in the international policy making arena, but also provide direction and build consensus on AI governance at the national level.²²⁶ However, it is yet to be seen how these goals pan out.

Lastly, with the increased reliance on foundational models and use of AI models in larger work streams, there is a need to focus on their governance.²²⁷ As more and more use-cases of foundational models permeate the market, it has become essential to develop clear rules for governance of foundational models as the infrastructural layer of AI development.²²⁸ Thus, going forward, foundational model governance is going to be key for responsible innovation of AI.

Annexures



Annexures

Endnotes

1. [PricewaterhouseCoopers. \(n.d.\). PwC's Global Artificial Intelligence Study: Sizing the prize. PwC.](#)
2. [Ibid](#)
3. [AI investment forecast to approach \\$200 billion globally by 2025. \(2023, August 1\). Goldman Sachs.](#)
4. [Chui, M., Hall, B., Mayhew, H., Singla, A., & Sukharevsky, A. \(2022, December 6\). The state of AI in 2022—and a half decade in review. McKinsey & Company.](#)
5. [2024 Digital Risk Report: Opportunities and Challenges of the AI Frontier. \(2024b, June 27\). AuditBoard](#)
6. [Arnold, R. D., & Wade, J. P. \(2015\). A Definition of Systems Thinking: A Systems Approach. Procedia Computer Science, 44, 669–678.](#)
7. [Ibid](#)
8. [Tools for Systems Thinkers: The 6 Fundamental Concepts of Systems Thinking - Interaction RBP. \(2022b, January 27\). Interaction RBP.](#)
9. [Bycsunpandian. \(2024, March 27\). The role of systems thinking in creating unbiased AI systems.](#)
10. [A Sociotechnical Approach to AI Policy. \(n.d.\). Data & Society.](#)
11. [Ibid](#)
12. [Ibid](#)
13. [Attard-Frost, B., & Widder, D. G. \(2023\). The Ethics of AI Value Chains. arXiv \(Cornell University\).](#)
14. [Schmitz, A., Mock, M., Görg, R., Cremers, A. B., & Poretschkin, M. \(2024b\). A global scale comparison of risk aggregation in AI assessment frameworks. AI And Ethics.](#)
15. [Ibid](#)
16. [Pourdehnad, J., Wexler, E. R., & Wilson, D. V. \(n.d.\). Systems & Design Thinking: A Conceptual Framework for Their Integration.](#)
17. [What Is Design Thinking & Why Is It Important? | HBS Online. \(2022b, January 18\). Business Insights Blog.](#)
18. [IDEO Design Thinking. \(n.d.\). IDEO | Design Thinking.](#)
19. [Sreenivasan, A., & Suresh, M. \(2024\). Design thinking and artificial intelligence: A systematic literature review exploring synergies. International Journal of Innovation Studies, 8\(3\), 297–312.](#)
20. See Limitations
21. [Liao, Q. V., & Vaughan, J. W. \(2024\). AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. Harvard Data Science Review, Special Issue 5.](#)
22. [Eliot, L. \(2023, October 5\). AI Ethics Wary About Worsening Of AI Asymmetry Amid Humans Getting The Short End Of The Stick. Forbes.](#)
23. [Zajko, M. \(2022\). Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. Sociology Compass, 16\(3\).](#)
24. [Davis, J. L., Williams, A., & Yang, M. W. \(2021\). Algorithmic reparation. Big Data & Society, 8\(2\).](#)
25. [Frost, B.A., Widder, D.G. \(2023\). The ethics of AI value chain: An approach for integrating and expanding AI ethics research, practice and governance.arXiv:2307.16787v2; Widder, D. G., & Nafus, D. \(2023\). Dislocated](#)

- accountabilities in the “AI supply chain”: Modularity and developers’ notions of responsibility. *Big Data & Society*, 10(1).
26. [Ibid](#)
 27. [Ibid](#)
 28. [Ibid](#)
 29. [Pnp. \(2023, October 11\). Computational power and AI. AI Now Institute.; The geopolitics of AI and the rise of digital sovereignty | Brookings. \(2023, June 24\). Brookings.; World Economic Forum. \(n.d.\). Strategic Intelligence | World Economic Forum. Strategic Intelligence.](#)
 30. [Muldoon, J., Wu, B.A. Artificial Intelligence in the Colonial Matrix of Power. *Philos. Technol.* 36, 80 \(2023\).](#)
 31. [Ch. 4, Competition and AI. OECD Business and Finance Outlook 2021. In OECD business and finance outlook. \(2021\); Tyler, E. \(2024, February 8\). ANALYSIS: Antitrust Bills Aim at AI Pricing Collusion. Bloomberg Law.; Begovic, B. \(2024, February 22\). EU antitrust chief calls for scrutiny of AI impact on merger control policy | Digital Watch Observatory. Digital Watch Observatory.](#)
 32. [Gray, J. E. \(2022, March 5\). Market concentration, democracy and artificial intelligence. . .a call to policymakers. Medium.](#)
 33. [Ulnicane, I., & Aden, A. S. \(2023\). Power and politics in framing bias in Artificial Intelligence policy. *Review of Policy Research*, 40\(5\), 665–687.](#)
 34. [Morris, Z. C. & B. \(2023, May 30\). Nvidia: The Chip Maker that became an AI superpower. BBC News.](#)
 35. [Brown, I. \(2023, June 29\). Expert explainer: Allocating accountability in AI supply chains. Ada Lovelace Institute.; Kuspert, S., Meos, N. & Dunlop, C., \(2023, February 10\). The value chain of general-purpose AI. Ada Lovelace Institute.; Engler, A.C., Renda, A., \(2023\). Reconciling the AI value chain with the EU’s artificial intelligence act. CEPS.; Responsible AI lifecycle, IndiaAI.; Introduction to AI assurance. \(2024, February 26\). Department for Science, Innovation and Technology.GOV.UK.](#)
 36. [Daswin De Silva and Daminda Alahakoon, “An artificial intelligence life cycle: From conception to production,” *Patterns* 3, no. 6 \(June 1, 2022\): 100489](#)
 37. [“What Is the AI Development Lifecycle?” Palo Alto Networks, n.d.](#)
 38. [“What is a foundation model?” n.d.](#)
 39. [“Pre Trained Model Definition | Encord,” n.d.](#)
 40. [Küspert, S., Moës, N., & Dunlop, C. \(2023, February 10\). The value chain of general-purpose AI. Ada Lovelace Institute.](#)
 41. [Spear Invest, “Diving Deep into the AI Value Chain,” Nasdaq, n.d.](#)
 42. [“What Is a Foundation Model?”](#)
 43. [“What Is the AI Development Lifecycle?”](#)
 44. [Ibid](#)
 45. [Ibid](#)
 46. [Madhulika Srikumar, Kasia Chmielinski, and Jiyou Chang, “Risk Mitigation Strategies for the Open Foundation Model Value Chain,” *Risk Mitigation Strategies for the Open Foundation Model Value Chain, 2024*](#)
 47. [“What Is the AI Development Lifecycle?”](#)
 48. [Ibid](#)
 49. [Spear Invest, “Diving Deep into the AI Value Chain.”](#)
 50. [Srikumar, Chmielinski, and Chang, “Risk Mitigation Strategies for the Open Foundation Model Value Chain.”](#)
 51. [Ibid](#)
 52. [Ibid](#)

53. [Ibid](#)
54. [Ibid](#)
55. [Reflections on Foundation Models. \(2021, October 18\). Stanford HAI.](#)
56. [Ibid](#)
57. [What is a foundation model? \(n.d.\).](#)
58. [The value chain of general-purpose AI. \(n.d.\).](#)
59. [Ibid](#)
60. [Vipra, J., & Korinek, A. \(2019\). Market concentration implications of foundation models: THE INVISIBLE HAND OF CHATGPT. Brookings.](#)
61. [Ibid](#)
62. [Ibid](#)
63. [The value chain of general-purpose AI. \(n.d.\).](#)
64. [Ibid](#)
65. [Vipra, J., & Korinek, A. \(2019\)](#)
66. [Shedding light on healthcare algorithmic and artificial intelligence bias. \(n.d.\). Office of Minority Health.](#)
67. [AI in healthcare is India's trillion-dollar opportunity. \(2024, September 10\). World Economic Forum.](#)
68. [Zoting, S. \(2024, August 26\). Artificial intelligence \(AI\) market size to reach USD 3,680.47 bN by 2034.](#)
69. [Stanly, M. \(2023, July 27\). AI in Indian healthcare sector: Promises and challenges. IndiaAI.](#)
70. [Krasniansky, A., & Krasniansky, A. \(2019, October 29\). Understanding racial bias in medical AI training data - Bill of Health. Bill of Health - The blog of the Petrie-Flom Center at Harvard Law School.](#)
71. [Lekadir, K., Quaglio, G., Garmendia, A. T., & Gallin, C. \(2022\). Artificial intelligence in healthcare. In Artificial Intelligence in Healthcare \(Report PE 729.512\). \(Original work published 2022\)](#)
72. [Mitchell, T. \(2024, December 6\). Algorithmic bias in health care exacerbates social Inequities—How to Prevent it | Harvard T.H. Chan School. Harvard T.H. Chan School of Public Health.](#)
73. [Bracic, A., Callier, S. L., & Price, W. N. \(2022\). Exclusion cycles: Reinforcing disparities in medicine. Science, 377\(6611\), 1158–1160.](#)
74. [Shanklin, R., Samorani, M., Harris, S., & Santoro, M. A. \(2022\). Ethical Redress of Racial Inequities in AI: Lessons from Decoupling Machine Learning from Optimization in Medical Appointment Scheduling. Philosophy & Technology, 35\(4\).](#)
75. [Mitchell, T. \(2024, December 6\). Algorithmic bias in health care exacerbates social Inequities—How to Prevent it | Harvard T.H. Chan School. Harvard T.H. Chan School of Public Health.](#)
76. [Muñoz Parry, C., & Aneja, U. \(2023\). Artificial intelligence for healthcare: Insights from India. In Chantam House.](#)
77. [Chin, M. H., Afsar-Manesh, N., Bierman, A. S., Chang, C., Colón-Rodríguez, C. J., Dullabh, P., Duran, D. G., Fair, M., Hernandez-Boussard, T., Hightower, M., Jain, A., Jordan, W. B., Konya, S., Moore, R. H., Moore, T. T., Rodriguez, R., Shaheen, G., Snyder, L. P., Srinivasan, M., . . . Ohno-Machado, L. \(2023\). Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. JAMA Network Open, 6\(12\), e2345050.](#)
78. [Roadmap for the NIST Artificial Intelligence Risk Management Framework \(AI RMF 1.0\) | NIST. \(2023, March 14\). NIST.](#)
79. [World Health Organization. \(2023\). Regulatory considerations on artificial intelligence for health.](#)
80. [“What Is Artificial Intelligence in Finance? | IBM,” n.d.](#)

81. [Ibid](#)
82. [Kestenbaum, J. \(2023, July 5\). New AI Bias Law Broadly Impacts Hiring and Requires Audits. Bloomberg Law.](#)
83. [Parasnis, S. \(2024, December 26\). RBI announces members of Committee to develop Responsible AI Use Framework. MEDIANAMA.](#)
84. [Parasnis, S. \(2024a, December 20\). Regulated entities fully responsible for data privacy in case of AI usage: SEBI. MEDIANAMA.](#)
85. [Sharma, V. \(2019, August 19\). AI's impact on the Indian insurance sector. IndiaAI.](#)
86. [Sg, R. \(2021, January 30\). Budget 2021: Digitalisation of education sector can help India become 'Atmanirbhar' Business Today.](#)
87. [Saxena, P. \(2022, January 10\). AI impact on India: AI in education is changing India's learning landscape. IndiaAI.](#)
88. [FairAIED: Navigating fairness, bias, and ethics in educational AI applications. \(n.d.\).](#)
89. [The Evolution of Education: How AI is Reshaping Grading | The Princeton Review. \(n.d.\).](#)
90. [Ferrara, Emilio. "Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies." arXiv \(Cornell University\), January 1, 2023.](#)
91. [Sumeet Kachwaha, "The Rule Against Bias and the Jurisprudence of Arbitrator's Independence and Impartiality," Asian International Arbitration Journal 17, no. Issue 2 \(October 1, 2021\): 103–34](#)
92. [Matthew Groves and H. P. Lee, "The rule against bias," in Cambridge University Press eBooks, 2007, 316–29](#)
93. [Ibid](#)
94. [Altman, Andrew, "Discrimination", The Stanford Encyclopedia of Philosophy \(Winter 2020 Edition\), Edward N. Zalta \(ed.\)](#)
95. [Meital Pinto, "Arbitrariness as Discrimination," Canadian Journal of Law and Jurisprudence 34, no. 2 \(July 12, 2021\): 391–415](#)
96. [Christian A. Ruzzier and Marcelo D. Woo, "Discrimination with inaccurate beliefs and confirmation bias," Journal of Economic Behavior & Organization 210 \(June 1, 2023\): 379–90](#)
97. [Megalokonomou, R., & Lavy, V. \(2023, September 19\). Teacher gender biases exist and have long-term effects. VOXEU CEPR.](#)
98. [Raina, S. \(2021\). GENDER BIAS IN EDUCATION. INTERNATIONAL JOURNAL OF RESEARCH PEDAGOGY AND TECHNOLOGY IN EDUCATION AND MOVEMENT SCIENCES, 1\(02\).](#)
99. [Banaji, Mahzarin R., Susan T. Fiske, and Douglas S. Massey. "Systemic racism: individuals and interactions, institutions and society." Cognitive Research 6, no. 1 \(December 20, 2021\).](#)
100. [Joe Hitchcock, "Cognitive Bias – Everything You Need to Know," InsideBE, October 22, 2023](#)
101. [Amos Tversky and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases," Science 185, no. 4157 \(September 27, 1974\): 1124–31](#)
102. [Allport, G. W. \(1954\). The nature of prejudice. Addison-Wesley.](#)
103. [Bordalo, P., Coffman, K., Gennaioli, N., & Shleifer, A. \(2016\). Stereotypes*. The Quarterly Journal of Economics, 131\(4\), 1753–1794.; Kahneman, D., & Tversky, A. \(1972\). Subjective probability: A judgment of representativeness. Cognitive Psychology, 3\(3\), 430–454.](#)
104. [Susan T. Fiske, "Prejudice, Discrimination, and Stereotyping," Noba, 2024](#)
105. ["5 Types of Statistical Biases to Avoid in Your Analyses," Business Insights Blog, June 13, 2017](#)

106. [Tomáš Kliegr, Štěpán Bahník, and Johannes Fürnkranz, “A review of possible effects of cognitive biases on interpretation of rule-based machine learning models,” Artificial Intelligence 295 \(June 1, 2021\): 103458](#)
107. [Kate Crawford and Trevor Paglen, “Excavating AI,” The AI Now Institute, NYU, September 19, 2019](#)
108. [Dan Pilat and Sekoul Krastev, “Heuristics,” The Decision Lab, 2021](#)
109. [Friedman, Batya, and Helen Nissenbaum. “Bias in computer systems.” ACM Transactions on Office Information Systems 14, no. 3 \(July 1, 1996\): 330–47.; Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. “Towards a standard for identifying and managing bias in artificial intelligence,” March 15, 2022.](#)
110. [Ferrara, Emilio. “Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies.” arXiv \(Cornell University\), January 1, 2023.; Shashkina, Victoria. “What is AI bias really, and how can you combat it?” ITReX, May 11, 2023. Accessed June 29, 2024.](#)
111. [Eirini Ntoutsi et al., “Bias in data-driven artificial intelligence systems—An introductory survey,” Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery/Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery 10, no. 3 \(February 3, 2020\)](#)
112. [“Industry Analysis: The AI Fairness Toolkits Landscape,” Borealis AI, May 6, 2022](#)
113. [Balayn, Agathe, Mireia Yurrita, Jie Yang, and Ujwal Gadiraju. ““Fairness Toolkits, A Checkbox Culture?’ On the Factors that Fragment Developer Practices in Handling Algorithmic Harms,” August 8, 2023.](#)
114. [“Industry Analysis: The AI Fairness Toolkits Landscape.”](#)
115. [Balayn et al., ““Fairness Toolkits, A Checkbox Culture?’ On the Factors That Fragment Developer Practices in Handling Algorithmic Harms](#)
116. [Cansu Canca, “Operationalizing AI ethics principles,” Communications of the ACM 63, no. 12 \(November 17, 2020\): 18–21](#)
117. [Ibid](#)
118. [Ben Gansky and Sean McDonald, “CounterFAccTual: How FAccT Undermines Its Organizing Principles,” 2022 ACM Conference on Fairness, Accountability, and Transparency, June 20, 2022](#)
119. [Divij Joshi, “Algorithmic Fairness and Anti-Discrimination Law - Centre for Law & Policy Research,” Centre for Law & Policy Research, November 5, 2020](#)
120. [European Digital Rights \(EDRi\). “If AI is the problem, is debiasing the solution? - European Digital Rights \(EDRi\),” September 20, 2021](#)
121. [Ibid](#)
122. [Ibid](#)
123. Note: Executive Order 14110 by the Biden Administration, referred to in this report was repealed by the Trump Administration on 20th January 2025
124. [Mary Reagan, “Understanding Bias and Fairness in AI Systems - Towards Data Science,” Medium, January 7, 2022](#)
125. [Ferrara, “Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies.”](#)
126. [Schwartz, Reva, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. “Towards a standard for identifying and managing bias in artificial intelligence,” March 15, 2022.](#)
127. [Friedman, Batya, and Helen Nissenbaum. “Bias in computer systems.” ACM Transactions on Office Information Systems 14, no. 3 \(July 1, 1996\): 330–47.; Caton, Simon, and Christian Haas. “Fairness in](#)

- [Machine Learning: A Survey.](#)” arXiv (Cornell University), January 1, 2020.
128. [Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. \(2021\). Data and its \(dis\)contents: A survey of dataset development and use in machine learning research. Patterns, 2\(11\), 100336.](#)
 129. [Bhardwaj, E., Gujral, H., Wu, S., Zogheib, C., Maharaj, T., & Becker, C. \(2024\). Machine learning data practices through a data curation lens: An evaluation framework. 2022 ACM Conference on Fairness, Accountability, and Transparency, 31, 1055–1067.](#)
 130. [Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. \(2021\). Data and its \(dis\)contents: A survey of dataset development and use in machine learning research.](#)
 131. [Ibid](#)
 132. [Huerta, E. A., Blaiszik, B., Brinson, L. C., Bouchard, K. E., Diaz, D., Doglioni, C., Duarte, J. M., Emani, M., Foster, I., Fox, G., Harris, P., Heinrich, L., Jha, S., Katz, D. S., Kindratenko, V., Kirkpatrick, C. R., Lassila-Perini, K., Madduri, R. K., Neubauer, M. S., . . . Zhu, R. \(2023\). FAIR for AI: An interdisciplinary and international community building perspective. Scientific Data, 10\(1\)](#)
 133. [Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. \(2020\). Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. Association for Computing Machinery.](#)
 134. [Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. \(2023\). Auditing large language models: a three-layered approach. AI And Ethics.](#)
 135. Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark, “The AI Index 2024 Annual Report,” AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.
 136. [Ibid](#)
 137. [Ibid](#)
 138. [How Artificial Intelligence is Transforming the Financial Services Industry. \(2023, August 11\). Deloitte.](#)
 139. [PricewaterhouseCoopers. \(n.d.-a\). No longer science fiction, AI and robotics are transforming healthcare. PwC.](#)
 140. [The future of learning: AI is revolutionizing education 4.0. \(2024, September 12\). World Economic Forum.](#)
 141. [Cyber security risks to artificial intelligence. \(2024, May 14\). GOV.UK.](#)
 142. [Aapti Analysis](#)
 143. [Neena. \(2024, September 17\). Content moderation in a new era for AI and automation | Oversight Board. The Oversight Board.](#)
 144. [OTIFYD. \(2023, April 10\). Intrusion & Anomaly Detection | OTIFYD - Safeguarding OT Networks.](#)
 145. [Learn about the double-edged sword of AI in cybersecurity. \(2024, September 10\). World Economic Forum.](#)
 146. [Raimondo, G. M., U.S. Department of Commerce, National Institute of Standards and Technology, & Locascio, L. E. \(2023\). Artificial Intelligence Risk Management Framework \(AI RMF 1.0\). In NIST AI 100-1.](#)
 147. [How AI can also be used to combat online disinformation. \(2024, September 10\). World Economic Forum.](#)
 148. [What is the role of AI in threat detection? \(n.d.\). Palo Alto Networks.](#)
 149. [CLIR. \(2021, February 14\). Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust • CLIR.; Authenticity, integrity, and security in a digital world. \(2019\). In National Academies Press eBooks.](#)
 150. [Kolbe-Guyot, M. \(2023, October 23\). “Big” News from Geneva: Making sense of the fundamental right to](#)

[digital integrity and its potential implications on digital sovereignty and beyond. C4DT.](#)

151. [Ibid](#)
152. [Integrity \(Stanford Encyclopedia of Philosophy\). \(2021, July 26\).](#)
153. [Ibid](#)
154. [Hill Jr, D. W. \(2012\). The right to personal integrity in international and domestic law \[The Florida State University\].](#)
155. [Article 3 - Right to integrity of the person. \(2024, November 19\). European Union Agency for Fundamental Rights.](#)
156. [Ibid](#)
157. [Rochel, J. \(2021\). Connecting the dots: Digital integrity as a human right. Human Rights Law Review, 21\(2\).](#)
158. [Breuer, J., Heyman, R., & Van Brakel, R. \(2022\). Data protection as privilege—Factors to increase meaning of GDPR in vulnerable groups. Frontiers in Sustainable Cities, 4.](#)
159. [United Nations. \(2021\). United Nations Global Principles for Information Integrity Recommendations for Multi-stakeholder Action. In United Nations Global Principles For Information Integrity Recommendations for Multi-stakeholder Action.](#)
160. [Nigro, P. \(2024, December 17\). The intersection of cybersecurity and artificial intelligence. Security Magazine.](#)
161. [Artificial Intelligence in Education. \(n.d.\). UNESCO.](#)
162. [Chlouverakis, Dr. K. \(n.d.\). How AI is Reshaping the Financial Services Industry. Ernst and Young.](#)
163. [Bajwa, J., Munir, U., Nori, A., & Williams, B. \(2021\). Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthcare Journal, 8\(2\), e188–e194.](#)
164. [What is the role of AI in threat detection? Palo Alto Networks. \(n.d.\).Schmelzer, R. \(2023, October 5\). Understanding the recognition pattern of AI. Forbes.](#)
165. [AI in cybersecurity: Revolutionizing threat detection, decision-making, and beyond. intwo. \(2024, July 11\). Barton, D., & Li, Dr. A. Z. \(2019, November 14\). A brief history of machine learning in Cybersecurity. Security Info Watch.](#)
166. [NIST identifies types of cyberattacks that manipulate behavior of AI systems. NIST. \(2024, January 4\).](#)
167. [Learn about the double-edged sword of AI in Cybersecurity. World Economic Forum. \(n.d.\).](#)
168. [NIST identifies types of cyberattacks that manipulate behavior of AI systems. NIST. \(2024, January 4\).](#)
169. [Learn about the double-edged sword of AI in Cybersecurity. World Economic Forum. \(n.d.\).](#)
170. [Bolen , S. \(2024, November 21\). AI: The New Frontier of Zero-day exploits. Medium.](#)
171. [Ibid](#)
172. [How AI is orchestrating blue Team success against advanced threats. \(n.d.\).](#)
173. [Baker, T. \(2023, October 24\). What does AI Red-Teaming actually mean? Center for Security and Emerging Technology.](#)
174. [Street, P. M. O. 1. D. \(2023, November 1\). The Bletchley Declaration by countries attending the AI Safety Summit, 1-2 November 2023. GOV.UK.](#)
175. [Generative AI in Cybersecurity. \(2024, June 10\). Centre for Emerging Technology and Security - the Alan Turing Institute.](#)
176. [Pearcy, S. \(2025, January 8\). A guide to AI red teaming. HiddenLayer | Security for AI.](#)
177. [Hay, A. \(2024, July 3\). Preparing for Tomorrow: Addressing Future Trends and Challenges with Purple Teaming. Lares.](#)

178. Standardization Trends on Safety and Trustworthiness Technology for advanced AI. (n.d.); How countries are implementing the OECD Principles for Trustworthy AI - OECD.AI. (n.d.)
179. Díaz-Rodríguez, N., Del Ser, J., Coeckelbergh, M., De Prado, M. L., Herrera-Viedma, E., & Herrera, F. (2023). Connecting the dots in trustworthy Artificial Intelligence: From AI principles, ethics, and key requirements to responsible AI systems and regulation. Information Fusion, 99, 101896.; Dehghani, F., Dibaji, M., Anzum, F., Dey, L., Basdemir, A., Bayat, S., Boucher, J., Drew, S., Eaton, S. E., Frayne, R., Ginde, G., Harris, A., Ioannou, Y., Lebel, C., Lysack, J., Arzuaga, L. S., Stanley, E., Souza, R., De Souza Santos, R., . . . Bento, M. (2024). Trustworthy and responsible AI for Human-Centric Autonomous Decision-Making systems. arXiv (Cornell University).
180. Ibid
181. Ibid
182. Trust (Stanford Encyclopedia of Philosophy). (2020, August 10).
183. Chin, I., Joerin, J., & Schubert, R. (2023). FRS Working Paper #7: Strengthening Social Resilience – The Importance of Trust (No. 7); AI and Trust. (2023, November 27). The Belfer Center for Science and International Affairs.
184. Ibid
185. Ibid
186. Conventionally, vertical trust is understood as a relationship between individuals and the state and works on the assumption that the state institutions work for public good. While this assumption does not apply for AI tools and the companies building them, it still provides a framework of thinking to explore the relationship of trust between individuals, AI artifact, technology.
187. This also allows us to deal with the existing criticism of literature on trust and AI. Lukyanenko, R., Maass, W., & Storey, V. C. (2022). Trust in artificial intelligence: From a Foundational Trust Framework to emerging research opportunities. Electronic Markets, 32(4), 1993–2020.
188. Choung, H., David, P., & Ling, T. (2024). Acceptance of AI-Powered Facial Recognition Technology in Surveillance scenarios: Role of trust, security, and privacy perceptions. Technology in Society, 102721. Trust in technology is associated with trust in the creators and owners of the technology, the technology companies, the infrastructure of connectivity, data storage in the cloud, security protocols, and authorities with access to the data
189. Publications Office of the European Union. (2019). Ethics guidelines for trustworthy AI. Publications Office of the EU.
190. The Act texts | EU Artificial Intelligence Act. (n.d.).
191. AI Principles Overview - OECD.AI. (n.d.); <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
192. House, W. (2023, October 30). FACT SHEET: President Biden issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence. The White House.; House, W. (2023a, October 30). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. The White House.
193. Team, N. A. (n.d.). NIST AIRC - AI Risks and Trustworthiness. NIST Trustworthy & Responsible AI Resource Center.; The United States works with domestic and international AI communities to establish frameworks that advance trustworthy AI for all - OECD.AI. (n.d.).
194. Grother, P., Shevtsov, D., Tabassi, E., & Wolf, A. (2014). Face recognition standards. In Encyclopedia of

- Biometrics (pp. 1–10).
195. Davida, Z., & Lubasz, D. (2021). Privacy by Design – searching for the balance between privacy, personal data protection and development of artificial intelligence systems. In Nomos Verlagsgesellschaft mbH & Co. KG eBooks (pp. 337–360).; Almeida, D., Shmarko, K., & Lomas, E. (2021). The ethics of facial recognition technologies, surveillance, and accountability in an age of artificial intelligence: a comparative analysis of US, EU, and UK regulatory frameworks. AI And Ethics, 2(3), 377–387.
 196. Refer to Section “Mitigation Strategies” in Module I.
 197. CSDL | IEEE Computer Society. (2022, June 9).
 198. Top strategies to prevent web scraping and protect your data.
 199. About the Glaze project. Night Shade.; Burgess, M., & Rogers, R. (2024, October 12). How to stop your data from being used to train AI. WIRED.
 200. Everything in moderation. New America.
 201. AbuJarour, S., Qarariah, A., Saadeh, N., & Salem, M. (2024). AI, Misinformation, and Fake News: A Literature Review of Ethical and Technical Approaches. Contributions to Finance and Accounting, 641–652.
 202. Transparent AI disclosure obligations: Who, what, when, where, why, how. (n.d.).
 203. Germani, F., Spitale, G., & Biller-Andorno, N. (2024b). The dual nature of AI in information dissemination: Ethical considerations. JMIR AI, 3, e53505.
 204. Gheorghiu, D. (2023, November 12). Explaining the technology for detecting child sexual abuse online — CRIN. CRIN.
 205. Thorn & All Tech Is Human. (2023). Reducing the Risk of Synthetic Content: Preventing generative AI from producing child sexual abuse material. In NIST.; Raimondo, G. M., Locascio, L. E., & National Institute of Standards and Technology. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. In NIST Trustworthy and Responsible AI (NIST AI 600-1). U.S. Department of Commerce.
 206. Narayanan, A. (2023, March 9). Understanding social media recommendation algorithms. Knight First Amendment Institute.
 207. Article 37 of the Digital Services Act in the EU requires mandatory auditing of recommender systems deployed by the online platforms and also provides for other third party audits
 208. Auditing recommender systems. (n.d.). Interface.
 209. A side of Whack-A-Mole (Part 1). (2024, March 16).
 210. What is Lantern? (n.d.).
 211. AI in Threat Detection. (n.d.). Palo Alto Networks.
 212. AI RMF. (n.d.). AI RMF Playbook. In AI RMF Playbook (pp. 4–142).
 213. AI RMF. (n.d.). AI RMF Playbook. In AI RMF Playbook (Govern 4.1).
 214. Gaur, V. (2020, April 14). Building a transparent supply chain. Harvard Business Review.
 215. Differential privacy. (n.d.). Harvard University Privacy Tools Project.
 216. What is homomorphic encryption? - IEEE Digital Privacy. (n.d.).
 217. Discover Cortex XDR for detection response. (n.d.). Palo Alto Networks.
 218. Freitas, S., Chen, S., Wang, Z. J., & Chau, D. H. (2022, April 4). UNMASK: Adversarial Detection and Defense through robust feature alignment - Microsoft Research. Microsoft Research.
 219. 8 Ways Generative AI Can Enhance Cybersecurity. (n.d.).
 220. Behere, G. (2025, January 13). Human-Friendly Observability with Generative AI - The AI Spectrum - Medium. Medium.

221. [Belfield, H., & Hua, S. \(2022\). Compute and antitrust. Verfassungsblog. https://doi.org/10.17176/20220819-181907-0](https://doi.org/10.17176/20220819-181907-0)
222. [Pavel, B., Ke, I., Spirtas, M., Ryseff, J., Sabbag, L., Smith, G., Scholl, K., & Lumpkin, D. \(2023, November 3\). AI and Geopolitics: How Might AI Affect the Rise and Fall of Nations? RAND. https://www.rand.org/pubs/perspectives/PEA3034-1.html](https://www.rand.org/pubs/perspectives/PEA3034-1.html)
223. [Ibid](#)
224. [Muldoon, J., Cant, C., Graham, M., & Spilda, F. U. \(2023\). The poverty of ethical AI: impact sourcing and AI supply chains. AI & Society. https://doi.org/10.1007/s00146-023-01824-9](https://doi.org/10.1007/s00146-023-01824-9); [Rowe, N. \(2023, November 15\). Underage workers are training AI. WIRED. https://www.wired.com/story/artificial-intelligence-data-labeling-children/](https://www.wired.com/story/artificial-intelligence-data-labeling-children/)
225. [AI Safety Institutes: Can countries meet the challenge? - OECD.AI. \(n.d.\). https://oecd.ai/en/wonk/ai-safety-institutes-challenge](https://oecd.ai/en/wonk/ai-safety-institutes-challenge)
226. [Ibid](#)
227. [Stroponiati, K. \(2025, January 2\). From No-Code to auto payments, critical infrastructure shifts will drive enterprise AI in 2025. Crunchbase News. https://news.crunchbase.com/ai/critical-infrastructure-shifts-2025-stroponiati-brilliant/](https://news.crunchbase.com/ai/critical-infrastructure-shifts-2025-stroponiati-brilliant/)
228. [Schneider, J., Meske, C., & Kuss, P. \(2024\). Foundation models. Business & Information Systems Engineering. 66\(2\), 221–231. https://doi.org/10.1007/s12599-024-00851-0](https://doi.org/10.1007/s12599-024-00851-0); [Wu, S. D. a. N. \(2024, May 21\). The remarkably rapid rollout of foundational AI Models at the Enterprise level: a Survey. Lightspeed Venture Partners. https://lsvp.com/stories/remarkably-rapid-rollout-of-foundational-ai-models-at-the-enterprise-level-a-survey/](https://lsvp.com/stories/remarkably-rapid-rollout-of-foundational-ai-models-at-the-enterprise-level-a-survey/)

List of Experts

Aditya Gopalan - Indian Institute of Science

Akshat Goel - Rocket Learning

Amrita Mahale - ARMAN

Anupam Guha - Center for Policy Studies, IIT Bombay

Blair Attard Frost - University of Toronto

Danish Pruthi - Indian Institute of Science

Dr. Pramod Verma - Center for Digital Public Infrastructure

Joanna Bryson - Hertie School

Madhulika Srikumar - Partnership on AI

Maitreya Shah - Berkman Klein Center for Internet and Society

Sachin Lodha - Tata Consultancy Services Research

Mani Shukla - Tata Consultancy Services Research

Ajeet Singh - Tata Consultancy Services Research

Bala Mallikarjuna - Tata Consultancy Services Research

Varun Hemachandaran - Agami

Community of Practice members

Akshat Goel - Rocket Learning

Aman Taneja - Ikigai Law

Angela Thomas - Software Freedom Law Center

Anwesha Sen - Takshashila

Arjun D'Souza - Software Freedom Law Center

David Menezes - People Plus AI

Deepa Padmar - Vidhi Center for Legal Policy

Devina S - Independent researcher

Dr. Sivaramakrishnan - CeRAI

Eunsong Kim - UNESCO

Gagesh Varma - Saraf Partners

Harleen Kaur - Digital Futures Lab

Isha Suri - Center for Internet and Society

Ishan - Independent researcher

Jagriti - OMI Foundation

Jian Xi Teng - UNESCO

Kailas Kartikeyan - Independent

Kamesh Shekhar - The Dialogue/Core-AI

Kaustubha Kaldindi - Tattle

Krishnan Narayanan - CeRAI

Maya Sherman - GPAI

Mihir Kulkarni - Wadhwani AI

Minesh Mathew - Wadhwani AI

Nandini Chami - IT for Change

Pallavi Bedi - Center for Internet and Society

Rashika Narain - Agami

Rijesh - Takshashila

Satya Shoova Sahu - Takshashila

Shivangi Narayan - IT for Change

Sidharth Deb - The Quantum Hub

Srishti Joshi - Koan

Sundar Narayanan - Independent

Tarunima Prabhakar - Tattle

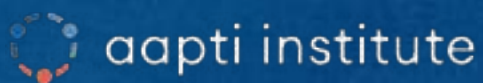
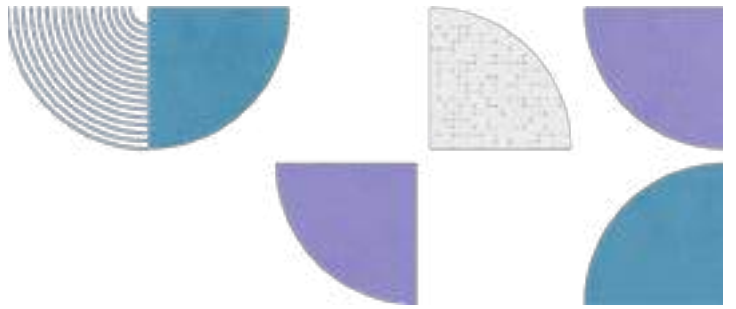
Varun Hemachandran - Agami

Policy Instruments

1. [Availability and use of Artificial Intelligence, Bill 2338/ 2023 \(Brazil\).](#)
2. [Charter of the Fundamental Rights of the European Union, 2000.](#)
3. [Ethics Guidelines for Trustworthy AI, 2019 \(EU\).](#)
4. [Executive Order 14110—Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 2023 \(USA\).](#)
5. [Ministry of Electronics and Information Technology, eNo. 2\(4\)/2023- Cyber Laws-3, Due Diligence by Intermediaries/ Platforms under the Information Technology Act, 2000 and Information Technology \(Intermediary Guidelines and Digital Media Ethics Code\) Rules, 2021.](#)
6. [Model AI Governance Framework for GenAI, Infocomm Media Development Authority, 2024 \(Singapore\).](#)
7. [NIST-AI-600-1, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile.](#)
8. [OECD/LEGAL/0449: Recommendation of the Council on Artificial Intelligence adopted on 22 May 2019.](#)
9. [Regulation \(EU\) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence.](#)

NOTE TO READERS

Executive Order 14110 by the Biden Administration, referred to in this report was repealed by the Trump Administration on 20th January 2025. An archived version has been added for the readers' reference.



Aapti is a public research institute that works at the intersection of technology and society. Aapti examines the ways in which people interact and negotiate with technology both offline and online.

contact@aapti.in | www.aapti.in

This work is licensed under the Creative Commons
Attribution-NonCommercial-ShareAlike 2.5 India License.

View a copy of this license at creativecommons.org/licenses/by-nc-sa/2.5/in/