AI KNOWLEDGE CONSORTIUM

AI Knowledge Consortium's Recommendations on the MeitY's Report on AI Governance Guidelines Development

MARCH 2025

Preface

Our comments on the AI Governance Guidelines Development report are a result of deliberations that took place during a AIKC workshop on the Report on AI Governance Guidelines Development convened on January 22, 2025.[1] This representation does not represent the specific institutional position of any of these organisations, and the discussion took place under Chatham House Rules.

Additionally, while the recommendations below engage with all facets of the AI Governance Guidelines Development Report, we would be remiss to not also mention the notes of caution sounded on the necessity of this exercise within our workshop. Several participants pointed towards the adequacy of existing laws and regulations to accommodate the development, deployment and diffusion of AI in specific sectoral and thematic contexts at the present time. That is, not all participants were of the view that the extant scope of the report, or indeed a horizontal governance/ institutional framework, is appropriate at the present stage.

The AIKC Secretariat has sought to synthesise the various comments and feedback received in our workshop, staying as true to the discussion as possible.

Other focal points from our submission, which are delineated in detail in the subsequent sections, are as follows:

- There is a need for clear, usable definitions for AI, ML, and foundation models
- AI is too heterogeneous to be lumped under a single expansive governance framework, because risks and harms are varied
- The application of existing laws should be seen in the context of the heterogeneity and technical complexities in each AI system; treating AI as a monolith is a recipe for failure
- Connecting between high-level principles and operational guidance is necessary; not all principles are relevant or appropriate to all AI systems
- India should steer away from overly subjective approaches to risk-classification
- Accountability constructs should not be confined to developers or deployers alone, an ecosystem approach is necessary
- Public sector deployments of AI require greater and perhaps more urgent oversight and audit; state-level involvement is also necessary
- A socio-technical approach, rather than a techno-legal one, is the need of the hour
- All legal aspects discussed in the report require greater context-specificity and nuance, including on aspects of liability on which AIKC members have produced research
- Equally, there a need to focus on standards design, development and implementation; as well as an impact assessment on unexplored dimensions such as digital trade
- There is also a need to reduce duplications in areas such as incident reporting, and in the case of overlapping institutional mandates or mechanisms
- Employing multistakeholderism, both on paper and in spirit throughout the development process of AI policies and institutional frameworks can help bridge gaps in state-capacity, aiding in a world-class policy design.

The following submission is structured in alignment with the framework of the AI Governance Guidelines Development Report.

Part I: Definition and Scope

A The AI Governance Guidelines Report fails to define key terms like Artificial Intelligence (AI), Machine Learning (ML), and foundation models. Precise definitions are a prerequisite for scoping the contours of the products and services to be governed. The Annexure section explains that a 'technology-agnostic' approach aims to maintain flexibility by avoiding overly broad definitions which may unnecessarily target low-risk technologies. But this approach undermines the framework's ability to offer clear guidance on the legal and regulatory path forward.

Recommendation:

We recommend the inclusion of at least 'working definitions' for 'AI' or 'AI Systems'.[2] These should be sufficiently broad to accommodate techniques and applications likely to emerge within the fold of AI. Additionally, definitions should also align with international standards to enhance global cooperation and legal certainty.[3] To this end, India may consider adopting the Organisation for Economic Cooperation and Development's (OECD) definition.[4] OECD defines AI systems as, 'a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment'. The definition is inclusive and encompasses simple to complex AI systems. It balances flexibility and specificity, ensuring future regulations are robust enough to handle dynamic developments while still being relevant for AI markets today.[5] The European Union (EU) AI Act is closely aligned with this definition.[6] Alternatively, International Standards Organisation- ISO/IEC 22989 defines AI systems as 'engineered systems that generate outputs such as forecasts, recommendations or decisions for a given set of human-defined objectives'.[7] The United States[8], and Australia[9] rely on this definition.

Similar approaches may be applied to define other key terms such as foundation models (also understood as general-purpose AI models (GPAI) or dual-use foundation models). AI models have garnered interest as highly capable systems that can perform a variety of tasks such as text synthesis, image manipulation and audio generation.[10] Given their adaptability, foundational models can be used in unforeseen ways, posing unique risks that may necessitate additional regulatory scrutiny to address their unique characteristics. This merits a shared understanding of the terminology among all relevant stakeholders.[11] In addition to defining these models, some countries have also introduced specific thresholds based on computational power or data scale to assist regulators in categorising AI systems more effectively, streamlining monitoring and enforcement efforts. For instance, the EU AI Act describes a "general-purpose AI model" as one trained with substantial data through self-supervision to achieve broad functionality, capable of performing a wide array of tasks. Such models are considered to have "high impact capabilities" when the computational effort for training exceeds 10^25 floating-point operations, as stipulated in Article 3(63).[12]

However, a growing body of evidence suggests the exercise of caution while using compute thresholds as a governance tool for AI risk mitigation,[13] as it oversimplifies the relationship between compute and AI risk. It assumes that greater computing power directly correlates with increased harm. The relationship between compute and AI capabilities is rapidly evolving, and non-linear. This is evident as smaller models are increasingly outperforming larger ones due to optimization techniques, making compute an unreliable sole risk indicator. This approach also neglects other critical factors influencing AI performance and risk, such as data quality, model architecture, inference-time enhancements, and system-level interactions. The current thresholds (10^25 FLOP in the EU AI Act and 10^26 FLOP in the now rescinded US Executive Order) may not be the most effective risk indicators. A more composite index that includes factors such as model performance benchmarks, security vulnerabilities, and real-world deployment impact, could be more effective.[14]

B The report does not emphasise the heterogeneity of AI technologies, which is critical for framing the risks or harms associated with these technologies accurately. AI systems vary widely in their capabilities, applications, and potential impacts. The report's failure to differentiate between types of AI technologies may limit its ability to address specific risks associated with various AI systems, ranging from simple automated decision-making systems to more complex and advanced foundational models.

Recommendation:

India should consider developing a taxonomy of AI models. A taxonomy is useful in categorising AI systems and assigning key regulatory objectives or principles to each category, enabling a clearer assessment of the risks and harms associated with these systems.

For the purpose of this document, we take the liberty of citing from existing research, a broad taxonomy of AI models. Dr Sangchul Park provides a useful framework for categorizing AI systems based on their lifecycle and interaction with humans. Based on their usage and level of human interaction, Dr Park classifies AI systems into five types: autonomous AI, generative AI, and discriminative AI. Discriminative AI is further divided into allocative AI, punitive AI, and cognitive AI.[15] The approach is premised on the understanding that the diverse array of AI systems available today induces a range of societal harms, each warranting distinct regulatory responses. Notably, an overlap exists in the use of each type of AI. For instance, the sensors of autonomous AI fall under cognitive AI. Transformers can be fine-tuned to implement downstream tasks that fall under allocative, punitive, or cognitive AI. Humanoids can classify people, speak, or paint. In such cases, regulations should be applied on the basis of specific tasks carried out by the system.

Autonomous AI

This includes robots and other autonomous systems—such as self-driving vehicles, automated facilities, surgical robotics, and pricing algorithms—that sense their environment, make reasoned decisions, and control operations. These systems may employ various forms of AI (discriminative, punitive, or generative) for decision-making. They should be distinctly regulated due to their potential safety implications, particularly in applications like autonomous vehicles.

Discriminative AI

Designed to score or classify individuals, this category involves models that assign benefits or detriments, or identify people, their conditions, or objects from datasets. Discriminative AI is typically developed using supervised learning from labelled data and is subdivided into:

- *Allocative AI*: Used for distributing limited resources (e.g., in recruitment, admissions, credit scoring, or insurance underwriting), this type must be closely monitored for potential discrimination and issues of inexplicability.
- *Punitive AI*: Employed to assign adverse sanctions (e.g., in criminal sentencing or fraud detection), this type necessitates rigorous scrutiny for inaccuracies and a lack of transparency.
- *Cognitive AI*: Covering applications such as computer vision, diagnostic imaging, and biometric identification, this form augments or replaces human cognitive functions and requires careful regulation regarding service quality, safety, efficacy, and privacy.

Generative AI

This category processes unstructured data, like text or images, into latent representations, which it decodes into creative outputs such as written content, artworks, or translations. Technologies like variational autoencoders, generative adversarial networks (GANs), diffusion models, and Transformers (e.g., OpenAI's GPT and Google's Gemini) exemplify this approach. Generative AI depends on unsupervised or self-supervised learning. Its regulation should consider its powerful ability to generate content, while also recognising its overlap with other categories, such as autonomous AI sensors or cognitive applications.

C The Report lacks a clear articulation of risks and harms associated with AI which is essential to ensure governance guidelines are precisely targeted to mitigate specific issues, facilitating effective and proportional regulatory interventions.

Recommendation:

AI systems differ significantly from traditional software systems in their development and operation. In the case of traditional software, developers explicitly define all the logic governing the system's behaviour, making it controllable, predictable, and easy to understand. In contrast, AI systems are developed by specifying objectives and constraints, selecting datasets, and employing machine learning algorithms. The approach makes them inherently less transparent and interpretable, as well as more complex to test and verify. This often introduces a broader range of risks and potential harms, affecting individuals, groups, communities, society, and even the environment.[16] Keeping these key differences in mind, we recommend the government to initiate discussions on the benefits and tradeoffs of a risk-based vs. rule-based approach to AI regulation. While a comprehensive risk assessment may be beyond the scope of a single ministry or committee, it is crucial to evaluate regulatory approaches that balance innovation, while ensuring accountability, and harm mitigation. A risk-based approach provides the flexibility to adjust regulations based on the potential risks of different systems. This allows for the allocation of resources to high-risk applications while fostering innovation in low-risk areas.

However, the subjectivity involved in defining and measuring risk can make this approach challenging to implement, especially given the rapid evolution of AI capabilities. On the other hand, rule-based approaches provide clear, standardised regulations that ensure consistency and legal certainty across AI applications, making enforcement more straightforward. But its rigidity in applying uniform requirements to both high and low-risk systems can stifle innovation and hinder technological advancement. A structured discussion on risk-tiered regulations vs. uniform compliance frameworks would help stakeholders assess which approach is best suited for India's AI governance landscape.

Countries are adopting distinct approaches to understand and mitigate these risks. For instance, in the United States, AI-related risk is considered as a function of the likelihood of an event and the magnitude of the harm if it occurs. The National Institute of Standards and Technology (NIST) has released voluntary AI Risk Management Frameworks (AI RMF) to guide individuals, organisations and others to manage these risks.[17] The European Union, too, through its EU AI Act, categorises AI systems into unacceptable, high, and limited risk. AI systems that are incompatible with EU values and fundamental rights are categorised as unacceptable risk.[18] AI systems that can negatively affect the health and safety of people are considered as high-risk AI systems. Additionally, AI systems that pose risk of manipulation or deceit are identified to be posing limited risk.[19] Meanwhile, Australia recognises the known as well as new and emerging AI risks. This is covered under a spectrum including technical risks, unpredictability and opacity, domain-specific risks, systemic risks, and unforeseen risks. Technical risks refer to design limitations or biases in training data, resulting in inaccurate or unfair outcomes. Unpredictability and opacity points at the opaque nature of AI systems and difficulty in identifying harms, predicting errors, and establishing accountability. Domain-specific risks highlight that AI can amplify existing harms or create new ones within particular sectors, such as spreading misinformation or generating harmful deepfakes. Systemic risks describe the potential for advanced AI models to cause large-scale, rapid, and unpredictable harms due to their high capabilities and increased accessibility. Unforeseen risks acknowledge that the rapid evolution and complexity of AI may lead to unexpected challenges that require agile, adaptable regulatory responses.[20]

While these frameworks demonstrate proactive efforts to address AI risks, it is important to exercise caution when using subjective or vaguely defined risk classifications. Overly broad categories, such as "high-risk" or "unforeseen risks," can be open to interpretation, often leaving critical decisions to courts rather than established regulatory bodies. This subjectivity may undermine regulatory clarity and lead to inconsistent enforcement across sectors. To mitigate this, clear, well-defined parameters replying on measurable performance benchmarks and context-specific risk indicators must accompany risk-based classifications. This ensures that classifications are transparent, consistent, and enforceable. In terms of choosing an appropriate governance model, developing countries like India face unique challenges. State capacity constraints, particularly in terms of technical expertise and enforcement resources, may make rule-based approaches more appealing due to their clarity, simplicity, and ease of enforcement. Uniform rules can serve as a baseline safety net, ensuring minimum standards across all AI applications. However, rule-based systems risk stifling innovation and may not adequately address the context-specific risks posed by emerging AI technologies.

It is also essential to note that risk-based approaches require building multistakeholder capacity, with active participation from government, industry, academia, and civil society. Collaborative efforts such as public-private partnerships and stakeholder consultations can help bridge capacity gaps and ensure informed, context-sensitive classifications. In this regard, the India AI Knowledge Consortium (AIKC) stands ready to play a role in facilitating cross-sectoral collaboration and developing practical, adaptable risk frameworks tailored to India's evolving AI landscape.

Part II: AI Governance Principles

D The Report lacks a clear operational framework linking high-level AI principles to the specific risks and harms these systems may generate. While the high-level principles are well-intentioned, they do not specify how to tackle the unique challenges that stems from AI models' opaque development processes, inherent unpredictability, and complex interactions with real-world environments.

Recommendation:

First-principles need to be contextualised based on the heterogeneity of AI models. For instance, in the case of autonomous AI, which includes self-driving cars or drones, the potential harms include safety deficiencies and security vulnerabilities. These harms necessitate implementing rigorous pre-incident testing, continuous tracking, deploying real-time safety alerts during imminent incidents, and maintaining detailed event data records post-incident to support investigations and accountability. In allocative AI, which affects critical sectors like public services, employment, and credit evaluation, it is essential to define clear, quantifiable fairness standards that can be integrated directly into these systems to prevent bias and discrimination. In generative AI, adequate disclosure of limitations combined with the use of system cards and model cards is helpful in enhancing transparency and guide users, helpful in mitigating risks associated with misuse or misinterpretation. A similar approach is adopted under the US NIST AI Risk Management Framework which outlines trustworthiness characteristics of AI. The RMF bases the trustworthiness of AI models on principles of validity, safety, security, accountability, transparency, explainability, and interpretability. In addition to stating these principles, it also provides concrete operational guidance by referencing established ISO standards and government initiatives (such as the NHTSA's transparency program).[21]

E The Report's high-level AI principles are at times disconnected from the current realities of AI. For example, the principle of reliability which typically demands deterministic accuracy can contradict the value proposition of generative AI and foundational models. These systems are designed around probabilistic modelling, where the focus is on generating creative and diverse outputs rather than ensuring consistent, error-free performance. The report seems to borrow many of its core principles such as reliability, robustness, accuracy and transparency from traditional software development without accounting for AI's unique nature where opaque learned models, unpredictable real-world data, and statistical accuracy replace deterministic code and predictable errors. The principles outlined in the report overlook the failure of even advanced probing methods to explain emergent behaviours such as unanticipated mathematical computations in large language models.

Recommendation:

The report would benefit from fine tuning the high-level principles on transparency, accountability, reliability, and robustness keeping in mind their inherent limitations. For instance, it highlights transparency as a key principle of AI governance, advocating that AI systems should be accompanied by meaningful information on their development, processes, capabilities, and limitations. It also stresses the importance of interpretability and explainability, where appropriate. However, excessive transparency requirements could risk compelling firms to disclose commercially sensitive information, including source code which has already sparked criticism under the EU AI Act due to its implications for trade secrets.[22]

India has historically supported source code disclosure,[23] but this stance could have unintended consequences for Indian AI firms operating globally. Many Indian companies are now at the cutting edge of AI innovation, and overburdening them with disclosure obligations could undermine their competitive edge when expanding abroad. Moreover, stringent transparency requirements may compromise system integrity and security, as disclosing technical details could expose AI systems to manipulation or exploitation by bad actors.[24] To this end, transparency mandates should be designed thoughtfully, ensuring that AI systems remain accountable and explainable without forcing firms to divulge proprietary information.

The scope of some principles such as accountability, is currently limited to developers and deployers and does not include all AI actors under AI lifecycle approach. India should also consider the roles and responsibilities of data providers, model trainers, regulators, and even end users.

Additionally, the principle on inclusive and sustainable innovation indicates an apparent conflation between regulating the technology itself and regulating the companies that develop and deploy it. A clearer demarcation is also required to separate regulatory strategies targeting technological risks from those addressing corporate practices and market dynamics. India should explore how to balance the equitable distribution of innovation benefits with the need to protect commercial interests. This could include examining mechanisms like Fair, Reasonable, and Non-Discriminatory (FRAND) licensing principles to safeguard intellectual property rights.

F The report does not adequately scrutinise the principles against the backdrop of AI technologies currently deployed by central and state governments, including various Digital Public Goods (DPGs). While it briefly acknowledges that operationalising these principles requires a joint commitment from both government and industry, it stops short of establishing a robust framework for evaluation. As a result, the gap analysis focuses primarily on existing laws rather than critically examining the core AI initiatives and technologies in practice. This oversight fails to provide actionable insights into how well the principles work when applied to real-world deployments.

Recommendation:

We recommend that India develop a comprehensive framework that scrutinises the current AI tools and technologies being deployed by central and state governments against relevant first-principles. This assessment should extend the gap analysis beyond existing legal frameworks to include a critical evaluation of the core AI initiatives in practice, thereby bridging the disconnect between regulatory intent and on-ground implementation.

The role of the government as a deployer requires much greater scrutiny. Citizens interacting with government-deployed AI systems may have little to no agency in opting out unlike in the private sector where there is often some degree of choice in engaging with AI-driven services The lack of choice raises concerns about transparency, accountability, and safeguards against potential harms. It is thus essential that all government-led AI deployments align with clear metrics of transparency and accountability.

Further, we also recommend India should encourage multi-stakeholder engagements in scrutinising government-led AI deployments. Academia, industry bodies, civil society organisations (CSOs), and technical experts should be actively involved in assessing whether AI tools deployed by central and state governments align with principles of transparency, fairness, and accountability. A structured consultation process through public advisory councils, multi-stakeholder roundtables, and expert review panels can help evaluate risks, identify gaps, and recommend safeguards tailored to the Indian governance landscape. Establishing formal redressal mechanisms with independent oversight from diverse stakeholders would help operationalise accountability measures. It would ensure inclusivity, transparency and accountability for government-deployed AI systems.

II. Considerations to Operationalise AI Principles

G The current AI lifecycle approach and ecosystem view of AI actors outlined in the report is narrow, focusing only on development, deployment, and diffusion stages, and terminating at the end user. This limited view fails to capture the wide range of affected parties, from individuals whose lives are indirectly impacted by AI-driven screening in job applications, parole decisions, and medical diagnostics, to name a few.

Recommendation:

We recommend the adoption of a more expansive AI lifecycle approach including stakeholders such as data providers, model trainers, deployers, regulators, end users as well as those who are indirectly impacted by AI-driven decisions. This may include job applicants, individuals involved in parole processes, and patients subject to medical AI systems. Additionally, the labour ecosystem of data workers, validators, and other contributors who sustain the "data provider and model trainer" roles needs to be further explored, with a focus on key risks and vulnerabilities they face.

H The report advocates for adoption of a techno-legal approach into its governance strategy where its legal and regulatory regime are supplemented with technology layers with some human oversight. However, this articulation of techno-legal approach is abstract and non-specific, failing to clearly connect regulatory technology (RegTech) and supervisory technology (SupTech) with specific AI governance principles such as accountability, transparency, and safety. In its current form, the report briefly mentions that technology can identify liability across value chains but offers little detail on applying liability constructs across the diverse range of AI systems and use cases.

Recommendation:

We recommend a more detailed and concrete operational framework linking regulatory technology tools to specific AI governance principles. India should have a more comprehensive policy discussion on how liability laws can be considered across the diverse range of AI systems, as they currently stand. In our AIKC Report on Crafting a Liability Regime for AI Systems in India, we explored the complex questions surrounding the determination of liability in the context of AI technologies.[25] Through a series of detailed case studies, we demonstrated that AI liability is highly contextual and varies significantly across different applications such as deepfakes, medical diagnostics, and autonomous vehicles. Liability regimes should be tailored to the specific context of AI deployment and the degree of control exerted over each system. Members’ research suggests that strict liability should not be uniformly applied across all AI systems. Instead, regulatory frameworks should be context-specific and focus on addressing identified risks without stifling innovation. The sectoral regulation and standards could play a crucial role in effectively governing these diverse systems.[26]

The government should also prioritise establishing robust oversight mechanisms for technologies used in regulation and supervision, ensuring transparency and accountability for entities controlling and deploying these tools.

We suggest that India should prioritise a socio-technical approach to AI governance as provided under NIST guidelines identifying and managing bias in AI. Such an approach recognises that technical fixes alone are insufficient to address issues like bias. It integrates technical, organisational, and societal factors to understand how AI systems interact within larger social contexts.[27] This approach is broader and superior to the techno-legal approach. This approach considers both technical and human factors, and offers a holistic understanding of AI's societal impacts.

III. Gap Analysis

I

The current Indian legal framework, as outlined in the report, relies on the Information Technology Act, Bhartiya Nyaya Sanhita, POSCO Act, Juvenile Justice Act, and Copyright Act to address deepfakes primarily as issues of obscenity, impersonation, and IP infringement. The report presumes that existing laws, with the support of techno-solutionistic measures such as watermarking and traceability protocols, are sufficient to tackle deepfake disinformation. However, this perspective neglects a nuance: not all deepfakes are inherently malicious—many have legitimate, even beneficial, applications. The current framework fails to clarify the governance scope of deepfakes, risking over-enforcement that could lead to wrongful arrests (e.g., of parodies or artistic expressions) and uncertainty within the industry. Moreover, without coordinated guidance, enforcement agencies may lack predictability and consistency in handling deepfake cases, undermining transparency and accountability. (Deepfakes)

Recommendation:

We note that it is imperative to first define and clarify the scope of deepfake governance. The Ministry of Home Affairs (MHA) should develop comprehensive guidance manuals differentiating harmful deepfake uses requiring enforcement from the legitimate ones. This framework could also include operational protocols such as assigning unique, immutable identities to content creators, publishers, and social media platforms to enable effective traceability and watermarking of both inputs and outputs of generative AI tools.[28] These measures will facilitate tracking the lifecycle of a deepfake, ensuring that non-consensual or illegal uses are detected and addressed promptly. However, the government should exercise caution when prescribing these tools. There is growing evidence suggesting that watermarking can be easily manipulated, so any requirement must be flexible and not impose overly severe penalties if detection or watermarking fails. Additionally, any provision targeting deepfakes should be part of a broader reform package aimed at combating misinformation. The Election Commission of India issued an advisory ahead of the Lok Sabha elections in May 2024[29] and in January 2025,[30] which represents a step in this direction. However, these measures are piecemeal, limited in scope and do not encompass the broader regulatory measures needed to counter systemic disinformation.

We recommend the introduction of bot disclosure requirements for social media platforms. Bots are automated programs designed to interact with users on social media platforms heavily deployed in disinformation campaigns as evidenced by the 2016 US election where over 50,000 bots were used to spread false narratives on Twitter.[31] Social media platforms can implement disclosures by requiring automated accounts to self-identify,[32] employing AI-driven detection systems to flag and label suspicious bot activity,[33] and transparency for accounts engaged in automated content amplification.[34] Countries like the United States (California B.O.T (Bolstering Online Transparency) Disclosure Act), and the EU AI Act[35] have already introduced regulatory frameworks requiring some level of bot transparency to counter disinformation and enhance accountability. This recommendation aligns with our proposal in the AIKC report on Crafting a Liability Regime for AI Systems[36] in India (explained above).

Additionally, collecting and analysing National Crime Records Bureau (NCRB) data on deepfake-related incidents are other small but operational actions that the government may take to ensure transparency and accountability in its approach. A well-defined governance framework, combined with coordinated enforcement guidelines, will provide predictability and certainty for all ecosystem players, limit wrongful arrests, and promote a balanced approach to managing deepfakes.

J

The report outlines India's existing cybersecurity framework by referencing the IT Act, CERT-IN, NCIIPC, DPDPA, and various sectoral guidelines (e.g., RBI, SEBI, IRDAI); it falls short in emphasising the critical need for standard-setting and rigorous testing tailored specifically to AI systems. It leans on the assumption that existing legal and cybersecurity processes are sufficient, overlooking the dynamic evolution of global best practices in cyber regulation and the need for a more nuanced, multi-stakeholder approach. The report's recommendations are limited to setting up a ministerial committee and a technical secretariat without a robust gap analysis or detailed guidance on integrating sectoral regulations and standards. The current framework primarily supports incident reporting and basic cybersecurity controls but lacks mechanisms to establish and enforce uniform standards for accuracy, robustness, and cyber resilience in AI. In contrast, international benchmarks such as the European Union's AI Act,[37] the US NIST AI Risk Management Framework,[38] and the UK's voluntary AI Cybersecurity Code of Practice[39] demonstrate a proactive approach including adversarial testing, continuous risk assessments, and cross-jurisdictional standards.

Recommendation:

India should transition from mere principle articulation to discussing operational methodologies, drawing on global frameworks like the Common Vulnerabilities and Exposures systems (CVE) for cybersecurity and mainstreaming regulatory and technical standards addressing both centralised oversight and sector-specific nuances. Discussions on specific standards and testing regimes aligning with international best practices in AI cybersecurity are most necessary. This includes an in-depth analysis of EU AI Act, NIST and CISA's cybersecurity standards, which require high-risk AI systems to undergo rigorous adversarial testing and continuous systemic risk assessments. India should also seek to align with cross-jurisdictional standards such as Quad Joint Principles[40] and assess feasibility of developing a voluntary AI Cybersecurity Code of Practice.[41]

K

The report's treatment of intellectual property in the AI context is incomplete. The report focuses predominantly on the input side specifically, training dataset without adequately addressing the nuances of how AI outputs interact with copyright law. It does not differentiate between various stages of AI operation like the creation of training datasets, the development of foundational models, and the generation of outputs. The report lacks clarity on how copyright applies differently at each stage, and how metadata or public domain materials should be treated compared to copyrighted content. This approach fails to capture the complexities of copyright dynamics, such as the need for fair use and Text and Data Mining (TDM) exceptions, which are essential for mitigating barriers to data access, particularly for smaller companies. Further, broader intellectual property issues, including personality rights and patent laws, trade secret protections, are not adequately addressed.

Recommendation:

India requires a more sophisticated and nuanced conversation on intellectual property governance in AI. On copyright, the report focuses solely on copyright concerns at the input stage, ignoring potential issues at the output stage. It also does not propose exemptions for text and data mining (TDM) despite acknowledging that India's Copyright Act, 1957 provides a limited fair dealing framework. This interpretation restricts the use of publicly available data for AI model training, which could disproportionately impact Indian AI startups that lack the resources to license data, putting them at a disadvantage against well-funded competitors. Countries like Japan and Singapore have introduced TDM exemptions within their copyright laws to facilitate AI innovation.[42] Japan allows copyrighted works to be used for data analysis under a "non-enjoyment" exemption, while Singapore's Section 244 of the Copyright Act (2021) explicitly permits computational data analysis. Further, even though Section 3(c)(ii) of the Digital Personal Data Protection Act, 2023 (DPDPA) exempts publicly available data under certain conditions, the requirement for entities to verify whether the data principal made the data public makes compliance difficult.

Additionally, India should explore mechanisms to protect broader IP rights, such as patent laws, personality rights and trade secrets, and seek to ensure that regulations are adaptable to the evolving technological environment. While Indian courts recognise personality rights, there is no codified law preventing the unauthorised use of an individual's identity, voice, or likeness, making AI-generated deepfakes and content replication a growing concern.[43] Similarly, India also lacks a dedicated trade secret law, leaving AI developers vulnerable to misappropriation of proprietary algorithms, training data, and model architectures.[44] Strengthening trade secret protections would help balance innovation and transparency, ensuring that startups and enterprises can safeguard their AI-driven innovations while maintaining fair competition.

A tailored, case-by-case approach would provide clarity and predictability for creators and AI developers alike, ensuring that intellectual property laws both protect innovation and foster a competitive, inclusive digital ecosystem.

L The report raises important issues regarding the cross-cutting nature of bias in AI systems, noting that biases which might be isolated in non-AI contexts can become amplified when embedded in AI-driven decision-making processes. Given that complete elimination of bias is not possible, proactive steps may be taken to manage it.

Recommendation:

India should aim for AI stakeholders to be bias-aware as completely eradicating bias in AI is impractical.[45] Instead, acknowledging bias enables the adoption of proactive measures to mitigate its harms.

Regular audits involving a diverse range of stakeholders are essential. These audits should prioritise safeguarding against human harms by incorporating feedback from various stakeholders, ensuring that the system's performance aligns with ethical standards. AI audits should adopt structured methodologies to assess functionality, transparency, ethical behaviour, and compliance. Examples of established frameworks include the NIST AI Risk Management Framework, ISO/IEC 42001, and Oxford University's CapAI tool. Additionally, best practices for AI audits include defining clear audit scopes, assessing data provenance and preprocessing, analysing user impact, and ensuring adherence to industry-specific regulations.

India should work towards establishing guidelines balancing transparency in design with intellectual property protection, ensuring that developers remain incentivised to open their systems for independent audits.

Additionally, incorporating socio-cultural context into the design process is critical for operationalising safety-by-design principles. Continuous feedback from end users help in identifying and mitigating bias throughout the system's lifecycle and ensure that AI tools remain responsive to real-world challenges. Finally, embedding the safety by design principles into the curricula of institutions like Indian Institute of Technology (IITs) and Indian Institute of Information Technology (IIITs) will further equip future AI engineers with the multidisciplinary expertise needed to develop bias-aware AI systems.

The report also considers discussion of international standards, such as ISO/IEC TR 24027:2021[46] for bias detection and the forthcoming ISO/IEC 12791 for bias mitigation techniques to provide a structured methodology to identify and address unwanted bias throughout an AI system's lifecycle.[47] Additionally, drawing on the UK Information Commissioner's Office's practical guidance on fairness and bias, could help anchor future developments in globally recognised best practices.[48]

IV. The Report's Recommendations

M The proposal to establish an Inter-Ministerial AI Coordination Committee is commendable for its ambition to create a whole-of-government approach. However its success will depend significantly on the clarity of its mission and the specificity of its mandate. Experience from past initiatives indicates that committees without a sharply defined purpose, such as the Parliamentary Standing Committee on Commerce's report on Promotion and Regulation of E-Commerce in India (2022) have often struggled to deliver tangible outcomes. In contrast, mission-specific groups such as Open Network for Digital Commerce (ONDC) have demonstrated greater effectiveness in coordinating actions and achieving policy objectives.

Recommendation:

The committee should proceed with a clearly defined purpose, structured mandate, and well-defined responsibilities. The lessons from previous committees with broad mandates should guide the framing of its objectives to prevent inefficiencies and ensure measurable policy outcomes. We support the suggestion on inclusion of a mix of official and non-official members to capture diverse perspectives. Additionally, we recommend an active engagement with state-level authorities. The inclusion of state governments would foster a more cohesive and responsive regulatory ecosystem as they play a critical role in implementing and enforcing AI governance measures at the grassroots level.

Furthermore, at AIKC, we stand ready to support this work by contributing research, policy insights, and strategic recommendations to aid the committee in developing a robust and well-coordinated AI governance framework.

N Establishment of a technical secretariat under MeitY is a promising step, several critical issues require further clarification. The accountability framework for the secretariat is not defined. It is important to specify how the secretariat's performance will be monitored, who it reports to, and what mechanisms are in place to ensure it remains independent and effective. The focus on scanning for risks without equally addressing the identification and mitigation of harms is another significant shortcoming. The secretariat should incorporate mechanisms to assess both risks and harms, ensuring a comprehensive evaluation of the AI ecosystem. Coordination with regulators, State Security Bodies (SSBs) and Civil Society Organisations, is also a concern. Given the limited state capacity and the busy schedules of key officials, the proposal must clarify how the secretariat will secure collaboration and draw on its mandated powers.

Recommendation:

The technical secretariat should have a well-defined accountability framework clearly explaining who it reports to, how its performance is assessed, and the mechanisms ensuring its independence. The secretariat's scope should be expanded beyond risk scanning to include mechanisms for identifying and mitigating AI-related harms, ensuring a holistic approach to AI governance. Stronger coordination mechanisms with regulators, State Security Bodies (SSBs), and Civil Society Organizations (CSOs), should be established. The proposal should clarify how the secretariat will leverage its mandate to coordinate efforts across stakeholders and drive AI governance at both national and state levels, given the limited state capacity.

Furthermore, it is important to clarify how the Technical Secretariat and the Coordination Committee mentioned in the report could converge with the recently announced India's AI Safety Institute.[49] The delineation of responsibilities between these entities should be made explicit to prevent redundancy and ensure efficient collaboration. Will the Safety Institute focus solely on research and standard-setting, while the Secretariat plays a more operational and regulatory role? How will insights from the Safety Institute feed into national AI governance decisions, and what mechanisms will ensure seamless coordination between these bodies?

O The proposal to establish an AI incident database is a promising initiative for systematically collecting evidence on the risks and harms arising from AI systems. However, it requires careful refinement in several areas including clarity on its scope, inclusivity of contributions, incentives for reporting, and alignment with global best practices.

Recommendation:

First, such a database must clearly define its scope to avoid duplicating existing reporting mechanisms by Cert-In and other cybersecurity systems, focusing instead on incidents that extend beyond traditional cybersecurity breaches. Second, it should broaden its inclusivity by enabling contributions from government, private sector entities, open source communities and independent developers, potentially through trusted flagger programs. Third, it should also establish incentive in the form of legal protections, recognition, or financial rewards to encourage widespread reporting while ensuring confidentiality. Finally, by drawing on successful global models like the Partnership on AI's AI Incident Database,[50] the OECD AI Incidents Monitor,[51] and the EU AI Act's incident reporting obligations,[52] the framework can be better structured to categorise incidents based on severity and impact, integrating continuous feedback into the regulatory process.

Endnotes

[1] The workshop was attended by Aapti Institute, Cyber Saathi Foundation, Esya Centre, Institute for Governance, Policies and Politics (IGPP), New Indian Consumer Initiative (NICI), Newschecker, Pahle India Foundation, Social Media Matters, Transitions Research, DeepStrat, and Internet Freedom Foundation (IFF).

[2] The term 'AI System' preferred over AI as it is more tangible and actionable

[3] https://www.lawsociety.ie/gazette/top-stories/2024/august/lack-of-ai-definition-gives-lawmakers-difficult-task/

[4] https://oecd.ai/en/wonk/definition

[5] Explanatory Memorandum on the updated OECD definition of AI system.

[6] Article 3: Definitions | EU Artificial Intelligence Act' <https://artificialintelligenceact.eu/article/3/> accessed 23 January 2025.

[7] https://cdn.standards.iteh.ai/samples/74296/c4efbadbf1a146d4af6d62fcad09438f/ISO-IEC-22989-2022.pdf

[8] https://www.law.cornell.edu/uscode/text/15/9401#3. A more evolved and recent definition of AI system appears in the National Institute of Standards and Technology (NIST). See here: https://airc.nist.gov/AI_RMF_Knowledge_Base/Glossary

[9] Department of Industry Science and Resources, 'Terms and Definitions | Voluntary AI Safety Standard | Department of Industry Science and Resources' (*https://www.industry.gov.au/node/93857*, 5 September 2024) <https://www.industry.gov.au/publications/voluntary-ai-safety-standard/terms-and-definitions > accessed 21 January 2025.

[10] https://www.techtarget.com/whatis/feature/Foundation-models-explained-Everything-you-need-to-know

[11] https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/

[12] General-Purpose AI Models in the AI Act – Questions & Answers | Shaping Europe's Digital Future' <https://digital-strategy.ec.europa.eu/en/faqs/general-purpose-ai-models-ai-act-questions-answers> accessed 23 January 2025.

[13] https://www.fenwick.com/insights/publications/interesting-developments-for-regulatory-thresholds-of-ai-compute

[14] https://arxiv.org/html/2407.05694v1

[15] Park, Sangchul. "Bridging the Global Divide in AI Regulation: A Proposal for a Contextual, Coherent, and Commensurable Framework." Washington International Law Journal 33, no. 2 (2024). https://digitalcommons.law.uw.edu/cgi/viewcontent.cgi?article=1937&context=wilj#page=29.09

[16] https://www.industry.gov.au/publications/voluntary-ai-safety-standard/risks-and-harms

[17] https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf

[18] https://www.nist.gov/system/files/documents/2024/10/07/09-24-about-the-ai-rmf-for-distro-9-25_508-edit.pdf

[19] https://www.trail-ml.com/blog/eu-ai-act-how-risk-is-classified#:~:text=%E2%80%8D-,Risk%2DClassifications%20according%20to%20the%20EU%20AI%20Act,the%20highest%20level%20of%20risk

[20] https://storage.googleapis.com/converlens-au-industry/industry/p/prj2452c8e24d7a400c72429/public_assets/safe-and-responsible-ai-in-australia-governments-interim-response.pdf

[21] https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[22] https://datainnovation.org/2024/03/the-eus-ai-act-creates-regulatory-complexity-for-open-source-ai/

[23] https://www.thedatasphere.org/news/negotiating-competing-interests-in-the-datasphere-unpacking-source-code-disclosure-provisions-in-trade-agreements/

[24] https://www.techpolicy.press/trade-pacts-should-not-have-special-secrecy-guarantees-for-source-code-algorithms/

[25] https://aiknowledgeconsortium.com/wp-content/uploads/2024/10/ReportESYACentreReport-CraftingaLiabilityRegimeforAISystemsinIndia.pdf

[26] Ibid.

[27] https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[28] These measures are reflected in international approaches. For instance, in the European Union, under the EU AI Act mandates both the technical marking by the provider and labelling of AI output by the deployer to ensure that AI generated/ manipulated content is readily identifiable. See here: https://technologyquotient.freshfields.com/post/102jb19/eu-ai-act-unpacked-8-new-rules-on-deepfakes; In the US, around twenty one states have enacted some form of regulation on non-consensual intimate deepfakes. In September 2024, California passed the Defending Democracy from Deepfake Deception Act, which mandates that large online platforms block and identify materially deceptive election-related content during specific periods before and after California elections. Additionally, these platforms must label certain content as inauthentic, fake, or false within 72 hours of receiving notice during those designated periods. Additionally, China has also introduced labelling requirements content generated using deep synthesis technology under its Regulations on the Administration of Deep Synthesis of Internet Information Services (Regulations) See here: https://www.aoshearman.com/en/insights/ao-shearman-on-data/china-brings-into-force-regulations-on-the-administration-of-deep-synthesis-of-internet-technology

[29] https://elections24.eci.gov.in/docs/2eJLyv9x2w.pdf

[30] https://www.eci.gov.in/eci-backend/public/api/download?url=LMAhAK6sOPBp%2FNFF0iRfXbEB1EVSLT41NNLRjYNJJP1KivrUxbfqkDatmHy12e%2FzGjJMI0%2FjETs7fjrM8lYn4ipTqYtDEvVosG8Bae5QB8%2Fj5TBF9Esc2hlzORgYtkmzyKzGsKzKlbBW8rJeM%2FfYFA%3D%3D

[31] https://aiknowledgeconsortium.com/wp-content/uploads/2024/10/ReportESYACentreReport-CraftingaLiabilityRegimeforAISystemsinIndia.pdf

[32] https://brandequity.economictimes.indiatimes.com/news/digital/twitters-new-update-allows-good-bots-to-self-identify/89633817

[33] https://arxiv.org/html/2407.15688v1

[34] https://academic.oup.com/cybersecurity/article/9/1/tyac015/6972135?login=false

[35] https://www.dentons.com/en/insights/articles/2022/july/21/the-regulatory-landscape-surrounding-the-use-of-bot-technologies

[36] Ibid

[37] Article 15 (1) EU AI Act. See also: Recital 115

[38] 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (National Institute of Standards and Technology (US) 2023) NIST AI 100-1 <http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf> accessed 23 January 2025.

[39] 'A Call for Views on the Cyber Security of AI' (*GOV.UK*) <https://www.gov.uk/government/calls-for-evidence/cyber-security-of-ai-a-call-for-views/a-call-for-views-on-the-cyber-security-of-ai> accessed 22 January 2025.

[40] 'Joint Principles of Quad Cybersecurity Partnership' <https://www.homeaffairs.gov.au/cyber-security-subsite/files/qscg-joint-principles.pdf> accessed 23 January 2025.

[41] 'Quad Principles on Critical and Emerging Technology Standards'.

[42] Hays, Seth. "AI Training and Copyright Infringement: Solutions from Asia." Tech Policy Press (blog), October 30, 2024. https://www.techpolicy.press/ai-training-and-copyright-infringement-solutions-from-asia/.

[43] https://www.ipandlegalfilings.com/advent-of-ai-voice-generation-and-threat-to-personality-rights/

[44] https://corporate.cyrilamarchandblogs.com/2024/05/the-22nd-law-commission-report-on-trade-secrets-call-for-a-balancing-act/

[45] https://itif.org/publications/2022/04/25/ai-bias-correctable-human-bias-not-so-much/. See also: https://sloanreview.mit.edu/article/manage-ai-bias-instead-of-trying-to-eliminate-it/; https://theconversation.com/eliminating-bias-in-ai-may-be-impossible-a-computer-scientist-explains-how-to-tame-it-instead-208611

[46] 'ISO/IEC TR 24027:2021' (*ISO*) <https://www.iso.org/standard/77607.html> accessed 23 January 2025.

[47] 'ISO/IEC TS 12791:2024' (*ISO*) <https://www.iso.org/standard/84110.html> accessed 23 January 2025.

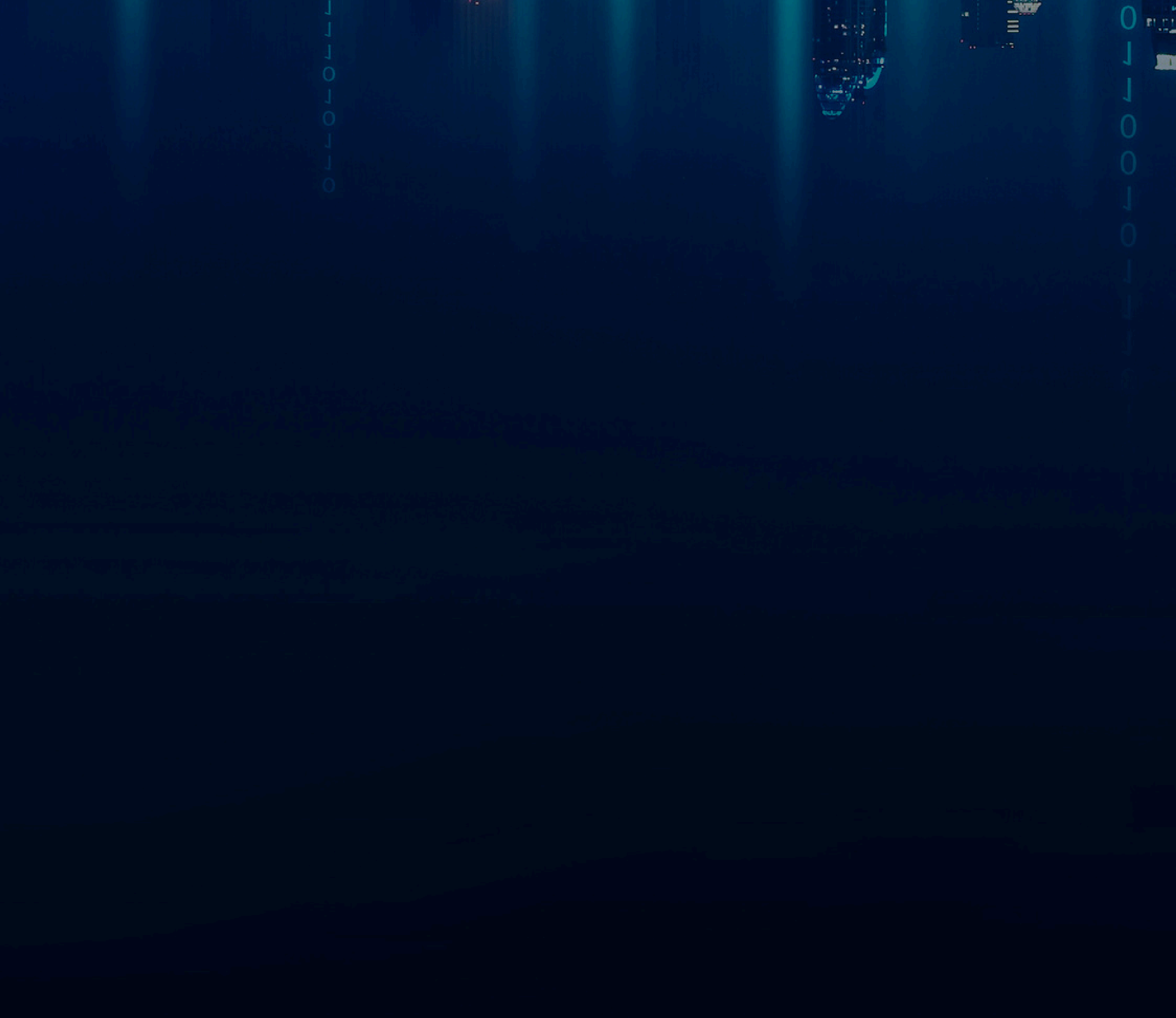
[48] Section 2(f-g) and Section 5(a-f). See: What about Fairness, Bias and Discrimination? (ICO) <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/how-do-we-ensure-fairness-in-ai/what-about-fairness-bias-and-discrimination/> accessed 23 January 2025.

[49] https://www.techpolicy.press/trade-pacts-should-not-have-special-secrecy-guarantees-for-source-code-algorithms/

[50] Welcome to the Artificial Intelligence Incident Database' <https://incidentdatabase.ai/> accessed 22 January 2025.

[51] 'Overview and Methodology of the OECD AI Incidents Monitor' <https://oecd.ai/en/incidents-methodology> accessed 22 January 2025.

[52] The EU AI Act under Article 60(7) mandates that both providers and deployers of AI systems and GPAI models are required to report incidents, even during testing phases. Under Article 55(1)(c); 73; 26(5), the providers and deployers of high-risk AI systems and GPAI models that pose systemic risks must notify appropriate governmental authorities—and, in certain cases, other relevant parties within the AI chain—about any serious incidents.



Secretariat

Koan Advisory Group, a New Delhi-based public policy consulting firm provides secretarial support to the AIKC

Contact Us

For inquiries, partnerships, or to learn more about our work with the AIKC, please write to:
Secretariat@aiknowledgeconsortium.com

Address: B40, Soami Nagar, New Delhi, 110017