Report on AI Governance Guidelines Development

India's unique demographic and socio-economic landscape presents significant opportunities for AI-driven growth. However, to ensure inclusive progress and address the associated risks, it is crucial to establish robust governance mechanisms. Given the diversity of India’s needs, a whole-of-government approach is essential for effectively managing AI's potential and challenges.

Accordingly, the Government of India has approved the IndiaAI Mission on 7th March 2024, with a budgetary outlay of INR 10,371.92 crore. The IndiaAI mission will establish a comprehensive ecosystem catalysing AI innovation through strategic programs and partnerships across the public and private sectors. By democratizing computing access, improving data quality, developing indigenous AI capabilities, attracting top AI talent, enabling industry collaboration, providing startup risk capital, ensuring socially impactful AI projects and bolstering ethical AI, it will drive responsible, inclusive growth of India's AI ecosystem. The Mission will be implemented through its seven key pillars of IndiaAI Compute Capacity, IndiaAI Application Development Initiative, IndiaAI FutureSkill, Safe & Trusted AI, IndiaAI Innovation Centre, IndiaAI Datasets Platform, and IndiaAI Startup Financing.

The Safe & Trusted AI pillar aims to drive the responsible development, deployment and adoption of AI by creating indigenous tools, self-assessment checklists, and governance frameworks. To further this mandate, IndiaAI had issued an Expression of Interest (EoI) to support Safe & Trusted AI Projects across a range of themes, including Machine Unlearning, Synthetic Data Generation, AI Bias Mitigation, Privacy-Enhancing Strategy, Explainable AI Framework, AI Governance Testing Frameworks, AI Ethical Certification Framework and Algorithm Auditing Tools. After a comprehensive evaluation process, eight projects have been selected in the first round of the EoI and are under implementation. Following the success of this initiative, IndiaAI has also launched the 2nd round of the Expression of Interest (EoI), inviting organizations to submit proposals on themes, including Watermarking and Labelling, Ethical AI Frameworks, AI Risk Assessment & Management, Stress Testing Tools, and Deepfake Detection Tools.

In further recognition of the need for an India-specific approach to AI governance, under the chairmanship of the Principal Scientific Advisor (PSA) of India, a multi-stakeholder Advisory Group has been constituted to undertake development of an 'AI for India-Specific Regulatory Framework' including representatives from relevant ministries. The Advisory Group was tasked with providing guidance on AI governance and offering insights for the necessary regulatory oversight to enable sustainable and ethical development of AI technologies.

Under the guidance of the Advisory Group, a Subcommittee on 'AI Governance and Guidelines Development' was constituted to provide actionable recommendations for AI governance in India. The Subcommittee's mandate was to examine key issues related to AI governance in India, conduct a gap analysis of existing frameworks, and propose recommendations for a comprehensive approach to ensure the trustworthiness and accountability of AI systems in India.

The Subcommittee's report emphasizes the need for a coordinated, whole-of-government approach to ensure effective compliance and enforcement as India's AI landscape continues to evolve. Based on its extensive deliberations, the report outlines a series of recommendations that aim to shape the future of AI governance in India. These recommendations are based on a careful review of the current legal and regulatory context and reflect the Subcommittee's independent perspective on fostering AI-driven innovation while safeguarding public interests.

The Ministry of Electronics and IT acknowledges the valuable recommendations put forward by the committee to support its ongoing initiatives on AI governance as applicable. In evaluating these recommendations, the Ministry is publishing the report for wider public consultation to invite feedback to inform the development of a comprehensive AI governance mechanism that advances India's AI aspirations while protecting the interests of all citizens.

Report

I. BACKGROUND

The sub-committee was constituted by the Ministry of Electronics and Information Technology (**MeitY**) on November 9, 2023, to analyse gaps and offer recommendations for developing a comprehensive framework for governance of Artificial Intelligence (**AI**).

II. GOVERNANCE OF AI

AI refers to a range of technologies which can be used for both harm and good. Governing the use of AI is driven by the need to minimise risks and harms. While AI has been around for many decades, over the last decade, five major developments that have brought AI into households and, consequently, into the lexicon of policymaking, regulation, and governance:

- Significant technical progress in the field of machine learning;
- Access to much larger datasets for the purposes of training AI systems;
- Advancements in computation performance and scale;
- Advancement in natural language processing and capabilities; and
- Widespread availability of connected devices to deliver apps employing AI systems to users at scale.

The combination of these factors has given rise to a new paradigm for developing and deploying AI systems – the creation of '"foundation models" that, after being trained on a huge amount of broad data, can be adapted to many applications., including "generative AI" tools accessible to end-users to perform a variety of tasks.

AI systems today can perform complex tasks without active human control or supervision. This raises the concern that AI might generate outputs that humans may not expect or understand. It can be difficult to understand how different components within an AI system interact with each other and which specific component is responsible for any potential harm caused. Detecting design defects of an AI system can be challenging, especially for downstream users of the system.

A. **AI Governance Principles**

Since 2016, several organisations from government, industry, and civil society have published "principles" for "responsible and trustworthy AI (**RTAI**)".[1] These set out a vision for the development, deployment, and use of AI systems that should inform the design of regulation of such systems as well.

Much work has also already been done in India to put principles of AI governance into practice.[2] In India, the principles from the apex government think tank[3] and NASSCOM[4] represents a good baseline from government and industry, respectively.

---

[1] For examples of such frameworks, see: the UKRI, *Report on the Core Principles and Opportunities for Responsible and Trustworthy AI*, (2023); Center for Long-term Cybersecurity, *Decision Points in AI Governance*, (2020); Fjeld *et. al.*, *Principled Artificial Intelligence: Mapping Consensus in Ethical and Rights-based Approaches to Principles for AI*, Berkman Klein Center for Internet & Society, (2020).

[2] E.g., NITI Aayog Principles of Responsible AI (2021), Operationalising Principles (2021), and FRT Report (2022); Indian Council of Medical Research Ethics Guidelines for Application of AI in Biomedical Research and Healthcare (2023); Tamil Nadu Safe & Ethical AI Policy (2020); TEC Voluntary Standard for Fairness Assessment and Rating of AI systems (2023); TEC Voluntary Standard for Robustness of AI systems in Telecom Sector (under development); Telangana AI Procurement Guide (under development – 2023); Nasscom Responsible AI Resource Kit (2022) and Guidelines for Generative AI (2023).

[3] *See:* NITI Aayog Principles of Responsible AI (2021).

[4] *See:* Nasscom Responsible AI Resource Kit (2022) and Guidelines for Generative AI (2023).

The OECD AI Principles exemplifies attempts at global convergence.[5] These efforts are aligned in ***substance***.

A proposed list of AI Governance principles (with their explanations) is given below. This list is aligned with the OECD, NITI and NASSCOM efforts. The purpose of this list is twofold – to be a set of principles aligned with contemporaneous global and Indian work, and to set the stage for examining how to operationalise these principles in the Indian context:

1. ***Transparency:*** AI systems should be accompanied with meaningful information on their development, processes, capabilities & limitations, and should be interpretable and explainable, as appropriate[6]. Users should know when they are dealing with AI.
2. ***Accountability:*** Developers and deployers should take responsibility for the functioning and outcomes of AI systems and for the respect of user rights, the rule of law, & the above principles. Mechanisms should be in place to clarify accountability[6].
3. ***Safety, reliability & robustness:*** AI systems should be developed, deployed & used in a safe, reliable, and robust way so that they are resilient to risks, errors, or inconsistencies, the scope for misuse and inappropriate use is reduced, and unintended or unexpected adverse outcomes are identified and mitigated. AI systems should be regularly monitored to ensure that they operate in accordance with their specifications and perform their intended functions.
4. ***Privacy & security:*** AI systems should be developed, deployed & used in compliance with applicable data protection laws and in ways that respect users' privacy. Mechanisms should be in place to data quality, data integrity, and 'security-by-design'.
5. ***Fairness & non-discrimination:*** AI systems should be developed, deployed, & used in ways that are fair and inclusive to and for all and that do not discriminate or perpetuate biases or prejudices against, or preferences in favour of, individuals, communities, or groups.
6. ***Human-centred values & 'do no harm':*** AI systems should be subject to human oversight[6], judgment, and intervention, as appropriate, to prevent undue reliance on AI systems, and address complex ethical dilemmas that such systems may encounter. Mechanisms should be in place to respect the rule of law and mitigate adverse outcomes on society.
7. ***Inclusive & sustainable innovation:*** The development and deployment of AI systems should look to distribute the benefits of innovation equitably. AI systems should be used to pursue beneficial outcomes for all and to deliver on sustainable development goals.
8. ***Digital by design governance:*** The governance of AI systems should leverage digital technologies to rethink and re-engineer systems and processes for governance, regulation, and compliance to adopt appropriate technological and techno-legal measures, as may be necessary, to effectively operationalise these principles and to enable compliance with applicable law.

---

[5] *See:* OECD, AI Principles, (2019).

Efforts to operationalise the above principles would require commitment from both the government and the industry. Meaningful initiatives by the industry ecosystem to demonstrate self-governance can significantly enhance trust in the use of AI and complement government led governance initiatives.

B. Considerations to operationalise the principles

The sub-committee has identified three key concepts that can inform the operationalisation of the principles and underpin AI governance in India. [6]

1. Examining AI systems using a lifecycle approach

A ***lifecycle*** approach to the development, deployment, and use of AI systems is useful to examine putting the principles into practice effectively. This is because the ***risks*** of AI systems play out differently at different stages in a lifecycle of a given system. It is useful to think in terms of three broad stages:

- ***Development*** which involves examining the designing, training, and testing of a given system.
- ***Deployment*** which involves examining the putting of a given AI system into operation and use.
- ***Diffusion*** which involves taking a long-term view and examining the implications of multiple AI systems being widely deployed and used across multiple sectors and domains.

Governance efforts should consider the entire lifecycle when operationalising a set of principles.

2. Taking an ecosystem-view of AI actors

Multiple actors can be involved across the lifecycle of any AI system. Together, they create an ***ecosystem.*** For example, in the context of the lifecycle of a foundation model, multiple sets of actors can be involved, including:

- ***Data Principals***
- ***Data Providers***
- ***AI Developers*** (including Model Builders)
- ***AI Deployers*** (including App Builders and Distributors)
- ***End-users*** (including both businesses and citizens)

Traditional governance approaches can be limited if they focus on one particular set of actors in isolation. By looking at governance across the ecosystem, better and holistic outcomes are obtained. An ecosystem-view of actors could also look to clarify the distribution of responsibilities and liabilities between different actors involved in the ecosystem.

3. Leveraging technology for governance

A complex ecosystem of AI models, systems, and actors is currently unfolding and expanding rapidly before us. This complexity and speed can be attributed to various

---

[6] A complex Adaptive System Framework to Regulate AI, https://eacpm.gov.in/wp-content/uploads/2023/10/EACPM-WP26-A-Complex-Adaptive-System-Framework-to-Regulate-AI.pdf

factors, including the increasing use of AI models for various applications, the speed at which the new AI models are being developed and deployed, and the improvements in the quality of the outputs being generated using AI systems (particularly generative AI tools). This ecosystem cuts across sectors and domains, constituting a wide and evolving regulatory space.

Given this, a conventional "command-and-control" governance strategy may not be able to adequately monitor, oversee or promote the growth and expansion of this space. There is value in integrating a "techno-legal" approach into the governance strategy, where legal and regulatory regimes are supplemented with appropriate technology layers (e.g., of governance technology tools along with adequate human oversight) across actors and systems.[7] Such a strategy would recognise the value of using technology to mitigate risks, scale and automate compliance across large ecosystems, and enhance the monitoring capabilities of regulators and market actors.

Using technologies to identify the allocation of regulatory obligations/liability across the value chain can also help embed directives and guidelines on how different ecosystem players are to work together and collaborate. This could allow players to manage liabilities in the chain leading to lightweight, but gradually scalable, regulatory control.

There can be several different components to make up such a strategy. As a starting point, there is merit in examining how technology artefacts, similar to the concept of "consent artefacts" already proposed by MeitY in their Electronic Consent Framework, can perhaps leveraged to assign immutable and unique identities to participants, so that their activities can be tracked and recorded to establish liability chains between them

Such artefacts, combined with the contracts between the participants, may allow for liability to be spread and distributed between them. Such a chain could allow for each member of the chain to enforce or require good behaviour on their own part and the part of the chain they are connected to such as their suppliers. This could potentially enable successful self-regulation across the ecosystem.

A techno-legal approach could also be optimally used by the Government/ regulatory body, or an appropriate technical assistance may be extended by the developers and deployers of AI systems to identify or trace unlawful information and actors upon receipt of a valid request from the Government on specified grounds such as prevention, detection, investigation or prosecution of harms, crimes, and security incidents. However, it is equally important to note that any such automated tools using for automating compliance, monitoring, or regulation would need to be periodically reviewed themselves, keeping in mind their security, accuracy, fairness, and impact on fundamental rights (e.g., like speech and privacy).

There is a need to strengthen efforts to operationalise the proposed AI Governance principles through a well-defined regulatory framework, which leverages technology and incorporates a "digital by design" approach to achieve the desired outcomes.

---

[7] Examples of governance technology tools, also known as "RegTech tools", include codified-regulations, AI compliance systems, blockchain tracking and smart contracting. *See,* World Economic Forum, *Regulatory Technology for the 21st Century*, (2022).

While considering techno legal approaches, it is important to consider that laws encoded in technology will be enforced. This may be required in certain context and be desirable in some but may not be appropriate in all. Such an approach should not shut the door on a level of flexibility that will offer us the freedom to innovate and improve.[8]

III. GAP ANALYSIS

AI systems do not have agency, except to the extent we afford them. In undertaking a gap analysis, the sub-committee kept in perspective the fact that existing laws and regulations continue to apply to the use of AI. The principles of responsible AI, including with respect to safety, equality, inclusivity, non-discrimination, and privacy are grounded in the fundamental rights enshrined in our constitution.

Given this, there is a need to examine the suitability of existing laws to deal with risks and harms in the context of AI systems. This can provide a meaningful direction to strengthen the governance framework. Such an analysis should be anchored in areas where concerns already exist and where they are likely.

An informed view of the AI ecosystem is likely to improve efforts for effective compliance and enforcement of existing laws and address gaps. Given that use of AI is not confined to certain sectors or use cases, implementation of an appropriate AI governance framework would benefit from a cohesive and co-ordinated effort or a whole-of-government-approach.

Accordingly, the sub-committee focused on identifying gaps taking into consideration that:

(1) Where AI systems can be used to ***exacerbate*** well-understood harms, there is a need to prioritise efforts that enable the effective compliance and enforcement of existing laws;
(2) To govern effectively, regulators will first need ***access to adequate information*** about the dynamics of the overall AI ecosystem of data, models, apps, actors, users, AI systems, etc;
(3) Since the field of AI is rapidly ***evolving***, and AI systems are increasingly ***cross-cutting*** tools, there is a need for a whole-of-government-approach to deal with emerging risks of harm.

The sub-committee discussed a few other topics which may require policy position, or no action, or further observation of the developments before appropriate governance mechanisms may be considered. ***These are listed in the Annexure.***

The three considerations mentioned above are discussed in detail below.

A. <u>The need to enable effective compliance and enforcement of existing laws.</u>

1. Deepfakes/ fakes/ malicious content

There are existing legal safeguards/instruments to protect against misuse of foundation models for creating malicious synthetic media (i.e., malicious 'deepfakes'). In this case, depending upon the context and negative effect of the malicious synthetic media in question, multiple laws can apply. For example:

[8] See: Rahul Matthan, The Zone of Mischief (2024)

- ***Information Technology Act, 2000 (IT Act):*** Section 66D of the IT Act criminalises the use of computer resources for cheating by personation. Section 66E prescribes the punishment for capturing and publishing or transmitting the image of a private area of any person without his or her consent. Publishing or transmitting obscene material for instance, which could be generated by using deepfake technology is a punishable offence under Section 67A and 67B of the IT Act.

- ***Indian Penal Code (IPC):*** In addition to the IT Act, certain harms/cybercrimes perpetuated by AI could also fall under the IPC. For instance, identity theft and cheating by personation are offences under Section 419 (cheating by personation), section 463 and 465 (forgery for the purpose of cheating), section 292 and 294 (selling/circulating/distributing obscene objects), and section 499 (causing harm to reputation/defamation). It is to be noted that the IPC has been recently replaced by the Bharatiya Nyaya (Second) Sanhita (**BNS2**), and the BNS2 retains these offences.

- **Other laws:** In addition to the IT Act and IPC / BNS2, there could be more laws depending on the nature of crime or cause of actions involved, like Prevention of Children from Sexual Offences Act, 2012 (section 12) in the event of sexual harassment of children, Juvenile Justice (Care and Protection of Children) Act, 2015 (section 75) for causing harm to the children, the Copyrights Act (section 51), if synthetic content infringes copyrighted work.

It is worth noting that existing laws can also require specific measures from platforms and online service providers to detect and remove malicious synthetic media. Under the ***Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021*:**

- Rule 3(1)(b) requires intermediaries to inform its rules and regulations, privacy policy and user agreement to a user and to make reasonable efforts to prevent the dissemination of specific types of content, that may cause harm to its users – including information that may violate bodily privacy, cause harm to a child, is deceptive, among other things.
- Rule 3(1)(c) requires intermediaries to periodically inform their users about the effects of non-compliance with the rules and regulations, privacy policy, or user agreement of such intermediary.
- Rule 3(2)(b) requires the intermediary to, within 24 hours of receiving user complaint of content which is 'impersonation in electronic form, including artificially morphed images of such individual' remove or disable its access. Further, the grievance officer appointed by the intermediary should acknowledge user complaints within 24 hours.

The above shows that the legal framework may be adequate for the purposes of detecting, preventing, removing, and prosecuting the creation and distribution of malicious synthetic media. However, for this legal framework to be effective, it needs to be backed by requisite capabilities to enable stakeholders to effectively comply and for the authorities to enforce the legal framework.

This area points to possible gaps and opportunities for using technological measures for enabling effective compliance, so that malicious deepfakes are detected in time and/ or are removed before they cause serious harm. For example, as suggested above, traceability may be established by assigning unique and immutable identities to different participants, such as content creators, publishers, social media platforms, etc. These may then be used to watermark inputs to, and outputs from, generative AI tools. These may be used to track and analyse the lifecycle, from creation to use, of a deepfake – and to determine when they have been created without consent or in violation of a law (e.g., cheating by personation).

2. Cyber security

Existing laws including the provisions of the IT Act to deal with cyber security are equally applicable when AI is used to compromise cybersecurity.

The primary national law dealing with cybersecurity is the IT Act. It establishes the Indian Computer Emergency Response Team (**CERT-IN**) (section 70B) and the National Critical Information Infrastructure Protection Centre (**NCIIPC**) (sections 70, 70A). It also defines various types of cybercrimes and establishes both a right for individuals impacted by cybercrime to seek compensation, as well as criminal penalties for such crimes (if criminal intent is established). Further, under the IT Act:

- The CERT-IN Rules, as well as Cybersecurity Directions 2021, require all body corporates to report cybersecurity incidents as well as adhere to certain requirements concerning synchronisation of ICT systems clocks as well as maintaining security logs. Security incident reporting is also required from intermediaries under the IT Rules 2021.
- The NCIIPC Rules govern operators of “protected systems” and then requires them to impose various information security procedures and practices, including ensuring they have a Chief information security officer, establish a Cyber Security Operation Centre, conduct various risk management activities, etc. The Central Government has notified different protected systems from time to time.

The second national law that governs cybersecurity is the Digital Personal Data Protection Act (**DPDPA**), which requires data fiduciaries to put in place appropriate security safeguards to protect personal data against breaches.

Beyond the above national laws, sectoral regulators have introduced separate cybersecurity guidelines.

- The Reserve Bank of India (**RBI**) has prescribed comprehensive cybersecurity standards and guidelines for banks, non-banking financial companies, payment system operators, and payment aggregators.
- The Securities and Exchange Board of India (**SEBI**) has released circulars on cybersecurity for stock market participants.
- The Insurance Regulatory and Development Authority (**IRDAI**) also has guidelines on cybersecurity for all insurers and insurance intermediaries.
- The Department of Telecom requires telecom licensees to report security incidents under the licensing framework.

Given that AI enables non-technical specialists to carry out sophisticated measures, there may be a need to focus on ways to strengthen the application of the above cybersecurity framework in the context of use of AI systems. This could require suitably upgrading compliance and enforcement capabilities to deal with the rapid development of AI and related emerging threats. For example, there may be a need for guidance to help providers of AI systems build systems that work as intended and are "secure by design".

3. Intellectual property rights

A question that is currently the subject of significant analysis, and (in the global context) also a matter of litigation, is accounting, under existing laws, for specific intellectual property rights (**IPR**) violations that occur through use of AI, particularly generative AI.

The sub-committee examined two areas under the Indian copyright law:

a. Training models on copyrighted data and liability in case of infringement

Given that copyright law grants the copyright holder an exclusive right to store, copy etc., creation of datasets using copyrighted works for training foundation models, without the approval of the right holder, can lead to infringement.[9]

The Indian law permits a very closed list of activities in using copyrighted data without permission that do not constitute an infringement. Accordingly, it is clear that the scope of the exception under Section 52(1)(a)(i) of the Copyright Act, 1957 is extremely narrow. Commercial research is not exempted; [10] not-for-profit institutional research is not exempted. Not-for-profit research for personal or private use, not with the intention of gaining profit and which does not compete with the existing copyrighted work is exempted.

Therefore, while the law protects the rights of the copyright holder, do we have the capabilities to enforce compliance to the existing law? Do we need to identify and agree on steps that the entities training on data need to put in place so as to demonstrate compliance with the law?

Despite some guardrails, there might still be infringements of existing works, given that multiple persons (e.g., the end-user giving prompts or the model developer) may be involved in determining the output generated by a foundational model. Who would be held liable in case such an output is found to be infringing upon an existing copyright?

The above aspects may require a review to ensure compliance with the law as it stands.

There are also policy level questions – for example, should AI systems be allowed to train on bulk datasets that may include copyrighted data, without taking approval from each copyright holder? If so, under what circumstances this

---

[9] Section 14(c), Copyright Act, 1957.

[10] See, Rupendra Kashyap v. Jiwan Publishing House, 1996 (38) DRJ 81 at para 21. It says, "if a publisher publishes a book for commercial exploitation and in doing so infringes a copyright, the defence under section 52(1)(a)(i) would not be available to such a publisher, even though the book published by him may be used or be meant for use in research or private study".

may be considered so that rights of the copyright holders are not infringed? Do we need to interpret or clarify the scope of rights that should exist with the copyright holder? What guardrails must be introduced, if we are able to address the questions above? The answers can help improve legal certainty and clarify the way forward for a lawful use of AI systems.

b. Copyrightability of work generated by using foundation models

Given the requirement of 'human authorship' for copyright protection, the eligibility and scope of granting copyright for works generated by using foundation models is untested under existing law.

In the case of AI, while human interference is required to initiate the process of creating an output (writing the algorithm, giving a prompt, etc.), there may be not enough interpretation or guidance available, for example, on the following:

- How much human input is necessary to qualify the user or developer of an AI system as an 'author' of a generated work.
- Whether work generated by using AI models can be termed as a type of 'computer generated work' eligible for copyright protection under the Copyright Act.

By proactively creating appropriate guidance, the relevant authorities (Copyright Office, Ministry of Commerce & Industry) can provide certainty and clarity to the users as well as to other government authorities who may otherwise adopt inconsistent practices.

A consultation of what would be appropriate guidance to clarify whether and to what extent creative works generated by using foundation models can be eligible for copyright protection, might be useful.

After answering these policy level questions, we can examine opportunities for leveraging *techno-legal measures*, including those that enable tracing the use of copyrighted data in training of AI models.

4. AI led bias and discrimination

Issues of AI entrenched biases are cross cutting. Biases in a non-AI context are more likely to be in pockets and less co-ordinated. When they creep into AI systems, the effects would be based on system deployment and the concerns would be based on context and scale of deployment. It is also important to understand that in many decisions there may be a notion of bias, however, only biases that are legally or socially prohibited need to be protected against.

At a fundamental level, it is important to examine how our existing laws deal with harm like biases/ discrimination. For example, our employment laws, laws protecting minorities and consumer protection law provide protection against discrimination. However, how consumer/ user rights are adequately protected when AI is used in decision-making is an evolving subject and requires a whole of the government approach to assess and where needed provide clarity or appropriate mechanisms to deal with AI entrenched biases.

For example, in a non-AI context, individual consumers would be expected to complain about discriminatory conduct. In an AI context, it is possible individuals

may not easily understand discrimination, or even if they see it, they may just critique the AI logic and it might be more difficult to establish intent. AI systems may reinforce pre-existing biases and without transparency, this may go undetected.

Consider another scenario, deployers may use tools (unknowingly or unforeseeably) that frustrate complying with current law. E.g., if businesses use a biased recruitment tool, leading to violation of the Equal Remuneration Act (as subsumed in new Codes), they may not even realise there is a problem. In these kinds of scenarios, there may not be important gaps in the legal provisions, but the need for entities to understand the risks and risk mitigation while deploying "black box" models would become important.

B. <u>The need for transparency and responsibility across the AI ecosystem in India</u>

To govern effectively, regulators will first require adequate information from two perspectives –

- ***traceability*** of data, models, systems, and actors throughout the lifecycle of AI systems and
- ***transparency*** from actors regarding the allocation of liabilities and risk management responsibilities between each other through contracts.

This would be necessary to design effective, targeted, and appropriate governance mechanisms. An ecosystem-view is especially relevant to understand which AI systems are being developed and deployed in India (and by whom) that are high capability and/or likely to be deployed widely and/or deployed in sensitive use-cases. The risks posed by a system depends not just on their capability, but on the context of deployment as well. ***The categorisations of systems purely based on computational capacity or data parameters may not be effective*.**

For systems deployed in tightly regulated sectors, they would need to be assessed under existing sectoral laws before we evaluate the need for additional or fresh laws. The testing of such sectoral laws should, in particular, examine how existing rules on assigning liability for non-compliances (e.g., in health, banking, financial services and insurance, energy, etc.) can be applied to AI systems prone to high-risk.

However, there may well be situations where a sectoral view is limiting, since we may not fully understand (i) the risks and/or (ii) the possibility for risks to spillover across sectors. Therefore, a view that "high risk scenarios" are likely only in tightly regulated sectors may not be correct. Given this, as well as the fact that many governance concerns may be common or cross-cutting across sectors, it might be useful to start examining a baseline framework to ensure transparency and responsibility across the overall AI ecosystem.

C. <u>The need for a whole-of-government approach.</u>

There are many laws and regulators, departments to deal with many of the harms that may arise due to use of AI. Given the sectoral specialisation, this is indeed desirable. However, given the rapid technology developments and the

cross-cutting usage of AI, there is an inefficiency and possibility of gaps due to a ***fragmented approach.*** Existing departments and regulators are likely to be examining AI systems in silos. This prevents a common understanding from being built, especially on cross-cutting issues. This gap can make it difficult for the government to organise multiple initiatives around a common roadmap.

IV. Recommendations

The committee recommends the following:

1. **To implement a whole-of-government approach to AI Governance, MeitY and the Principal Scientific Adviser should establish an empowered mechanism to coordinate AI Governance.**

The empowered mechanism should be in the form of an Inter-Ministerial AI Coordination Committee or Governance Group (**Committee/ Group**). It should bring together the various authorities and institutions that deal with AI Governance at the national level. The Committee/ Group should have an ongoing status and should not be a limited duration mechanism.

The overall ***purpose*** of this Committee/ Group should be to bring the key institutions around a common roadmap and to coordinate their efforts to implement a whole-of-government approach. A collaborative and co-ordinated approach by various regulators can enable them to be more efficient and effective, given the complexity likely to be involved in dealing with AI systems at scale, especially when we take a long-term view of the diffusion stage of their lifecycle. This can be especially necessary in domains and areas where multiple authorities may be concerned (e.g., consumer protection, food, transportation, agriculture, health care, etc.).

The Committee/ Group should enable a whole of government approach to the AI ecosystem. Currently, regulators and government departments may have some visibility on the AI systems developed or deployed by entities who are under sectoral regulation (e.g., finance or health) or where the market is concentrated (e.g., ecommerce, social media, aggregators). However, the level of visibility would need to be adequate to assess potential risks associated with such entities in the context of AI. Further, there are likely to be AI systems developed or being developed and/ or deployed by entities who may not have an interface with the government/ regulators from a perspective of affording suitable visibility to enable a risk assessment in relation to AI.

A pre-requisite of governance would be for the government and the regulators to have a credible understanding the AI ecosystem in the country so that governance measures are rooted to the realities of existing and likely risks. The Committee/ Group should facilitate this task.

This would require a conversation-led approach with a view to develop an understanding of the ecosystem which can both serve as feedback for strengthening governance and enable understanding of possible challenges and gaps in complying

and enforcing existing laws. It is important to emphasise that such a mapping exercise should not result in regulatory overreach through at scale registration and reporting requirements.

With the above context, the Committee/ Group should meet at a regular basis to suggest measures to **catalyse collaboration between departments and regulators,** so that they can:

- apply and strengthen existing laws to minimise risk of harm due to use of AI;
- provide legal clarity and certainty around development and use of AI by issuing joint guidance;
- harmonise existing efforts and initiatives around common terminologies and risk inventories;
- enable demonstrable self-regulation to operationalise the responsible AI principles;
- take coordinated steps to respond to identified gaps with the benefit of multi-regulatory support;
- create a policy environment which enables the use of AI for beneficial use-cases; and
- promote the development and deployment of responsible AI applications in their domains/sectors.

In order to enable appropriate measurement of fairness, accountability and transparency in the Indian context, it is an essential prerequisite to have access to the right datasets, relevant to the Indian context, which allows users to assess the fairness and bias of their models across standard datasets. The creation of better datasets for the Indian context should be stimulated, and sector-specific datasets should be identified to enable creation and evaluation of fair models. These initiatives may be encouraged by the Committee/ Group.

The Committee/ Group should have a mix of both official and non-official members, because such a forum focused on coordinating AI governance must also bring in external expertise from industry and academia, given their central role in operationalising responsible AI principles in practice.

It may be headed by the Principal Scientific Adviser. Official members could include representatives deputed from MeitY, the NITI Aayog, the Telecommunication Engineering Centre, Bureau of Indian Standards, other departments of the Central Government, as well as sectoral regulators (e.g., RBI, Indian Council of Medical Research, SEBI, IRDAI, Telecom Regulatory Authority of India, etc.). Non-official members could include persons capable of representing the interests of AI developers, AI deployers, data providers, data principals, and end-users – so that the perspectives of the overall ecosystem can be considered. The Committee/ Group should invite external experts for discussions to understand and take on board diverse perspectives.

2. **To develop a systems-level understanding of India's AI ecosystem, MeitY should establish, and administratively house, a Technical Secretariat to serve as a technical advisory body and coordination focal point for the Committee/ Group.**

MeitY should establish and host a ***technical secretariat*** that brings in officers on deputation from departments and regulators participating in the Committee/ Group as well as experts from academia and industry. As the Committee/ Group's technical advisory body and coordination focal point, the Secretariat should:

- pool together multi-disciplinary expertise (tech, law, policy, economics, etc.) from existing institutions in academia and industry to strengthen capacity across departments and regulators;
- create a map of the stakeholders and actors involved in India's AI ecosystem and conducts regular horizon-scanning exercises of the AI field;
- assess risks to consumers & society across applications and domains (incl. cross-cutting issues like antitrust, online safety, security, data governance, public services, employment, etc.);
- facilitate the development of metrics (e.g., measurement standards for assessing environmental impact of AI in India) & common frameworks (e.g., on data provenance, system cards, security, evaluation data sets, use of open source, transparency reports etc.); and
- engage with industry to co-examine novel solutions (e.g., labelling of synthetic media, privacy-enhancing technologies, etc.) to operationalise responsible use of AI and enable development of appropriate guardrails[6]; and
- identify gaps, which may not be adequately addressable through delegated legislation and/ or existing State capacities (such as where a guardrail is required and how it can be implemented, or where existing adjudicatory or compliance capacities need to be strengthened to deal with issues and disputes arising from emerging technology led activities including the use of AI).

It is to be noted that a similar advisory body was recommended by NITI Aayog in 2021. The functions envisaged by the NITI Aayog for that body may be allocated appropriately between the Committee/ Group under recommendation 1 (which should focus on coordination of policymaking and regulatory functions) and its Secretariat (which should focus on horizon-scanning, risk assessment, gap analysis, standardisation, and technical advisory).

The NITI Aayog had noted that any such body must be an "*independent technology wheelhouse advising relevant Government agencies*" and "*should be autonomous to work with individual regulators and Ministries*".[11] While these are undoubtedly relevant considerations, the sub-committee notes that, at this stage, it is not recommended to establish such a Committee/ Group or its Secretariat as statutory

---

11 *See:* NITI Aayog, *Operationalizing the Responsible AI Principles*, (2021).

authorities, as making such a decision requires significant analysis of gaps, requirements, and possible unintended outcomes. Instead, existing pathways should be used to set up the Committee/ Group (such as has been done for existing mechanisms like the National Startup Advisory Council) and its Secretariat.

The proposed secretariat could be staffed by existing MeitY officials as well as lateral hires, young professionals, and consultants. MeitY may form an **AI Sub-Group** to suggest the form and structure of the proposed secretariat along with a detailed term of reference.

3. **To build evidence on actual risks and to inform harm mitigation, the Technical Secretariat should establish, house, and operate an AI incident database as a repository of problems experienced in the real world that should guide responses to mitigate or avoid repeated bad outcomes.**

To understand the actual incidence of AI-related risks in India, the Technical Secretariat should establish an AI incident database and nurture reporting to it. In the initial stages, the database should receive reports from public sector organisations deploying AI systems (whether directly or through public-private partnerships). Private entities should also be encouraged to voluntarily report AI incidents to the database. The focus should be on defining reporting protocols to ensure confidentiality and to focus on harm mitigation, not fault finding.

It is important to note that an “AI incident” may include a “cyber incident” or a “cyber security incident” under the IT Act, but the ambit of "AI incident" may also go beyond the scope of a cyber incident. While cyber incident relating to AI systems may refer to suspected or potential attack or vulnerability against AI systems, AI incidents may also refer to adverse or dangerous outcome, resulting from the use of AI that can disadvantage or harm individuals, businesses, and societies.[12] AI incidents may include malfunctions, unauthorised outcomes, discriminatory outcomes, unforeseeable outcomes and unexpected emergent behaviour, system failures, privacy violations, physical safety problems, etc.

Typically, these issues became known when they are reported in media or when there is an ad-hoc escalation by a victim or an identification by a regulator. There is merit in the development of an AI incidence reporting database to serve as a more systematically collected evidence base to inform governance initiatives. Over time, this can help to show patterns and establish a collective understanding of AI incidents and their multifaceted nature. The *OECD AI Incidents Monitor* is a useful reference – it documents AI incidents to help policymakers, AI practitioners, and all stakeholders worldwide gain valuable information about the real-world risks and harms posed by AI systems.[13]

Setting up such a database should be taken up as a fresh exercise and not lapsed into existing cybersecurity incident reporting regimes, since the concept of an “AI incidents” goes beyond cybersecurity.

---

12 *See:* AI Incident Database, *What is an incident?*

13 *See:* OECD AI Incidents Monitor, *Methodology and disclosures*

It is a given that any unlawful activity will be appropriately dealt with through the legal framework. However, the AI incident database should not be started as an enforcement tool. Its objective should not be to penalise people who report AI incidents. Instead, the objective should be to encourage reporting and the learnings should flow back into the ecosystem. Given this, the suitability of CERT-IN taking on the mandate of maintaining an AI incident repository, under the guidance of the Technical Secretariat, may be examined.

4. **To enhance transparency and governance across the AI ecosystem, the Technical Secretariat should engage the industry to drive voluntary commitments on transparency across the overall AI ecosystem and on baseline commitments for high capability/widely deployed systems.**

Efforts to operationalise the AI Governance principles would require commitment from both the government and the industry. In terms of transparency, this can start by encouraging demonstrable industry self-regulation through examining the adequacy of existing voluntary reports and disclosures being released by current AI developers and deployers (e.g., transparency reports, model cards, etc.).

Further, existing laws empower regulators to encourage and, where needed, mandate the relevant entities to implement necessary measures to address risks and mitigate harms. Through the Committee/ Group, regulators can collaborate to design and implement efficient and effective responses[6]. That said, there may be a need for a baseline framework that applies to the development and deployment of AI systems that are considered medium-to-high risk across domains and sectors.

Consequently, the Technical Secretariat could start such work by anchoring a collaboration with industry to build consensus around voluntary commitments. Such commitments can include elements such as:

- disclosures of the ***intended purposes*** of AI systems and applications;
- commitments to release regular ***transparency reports*** by AI developers and deployers;
- commitments to internal and external red-teaming of models or systems in areas;
- processes to test and monitor data quality, model robustness, and outcomes;
- processes to validate data quality and governance measures;
- processes to ensure peer review by third-party qualified experts;
- processes to ensure conformity assessments with accepted responsible AI principles; and
- security, vulnerability assessment, and business continuity requirements.

In addition to the private sector adoption of AI, government is also adopting AI for citizen welfare as well as for law enforcement purposes. Therefore, the above recommendations of the sub-committee, including transparency and governance measures listed above, may also be adopted by the government and their technology providers, wherever relevant.

Part of such commitments can include use of technological solutions to mitigate the risks. Scalable and reliable technological artefacts and design elements can be encouraged from actors involved in each lifecycle stage, and be embedded across

data gathering processes, model building and retraining processes, as well as the final apps and user interactions as the manifestation of the traceability and liability chain.

Industry could also be encouraged to create standardised risk assessment protocols that developers and organisations could adhere to during the design and development of AI systems for their respective domains.

The specific voluntary commitments are likely to vary for different industry segments and should be based on understanding of real risks. These should therefore result from deeper engagement through technical workshops and collaborative discussions with industry. These efforts may be led by relevant regulators and government departments.

It is expected that industry would be able to demonstrate their adherence to the voluntary commitments. The voluntary commitments are expected to complement the legal framework and should minimise the need for prescriptive/ onerous regulations. The voluntary commitments should provide the requisite flexibility to the industry to commit to measures which are meaningful and implementable while providing the much-needed visibility to the regulators and government to the governance measures being implemented.

The role of the Technical Secretariat should be to assist these efforts and bring in cross sectoral expertise and a baseline maturity into these commitments.

5. **The Technical Secretariat should examine the suitability of technological measures to address AI related risks.** Such measures may aid with establishing a systems-level approach. For example, technology artefacts may be used to model the interactions between datasets, models, applications and users in different domains and sectors (e.g., healthcare, finance, etc.) and to identify and track negative outcomes in real time.

As noted earlier, the current ***legal framework*** has provisions to deal with malicious synthetic media and is being refined from time to time. To complement the legal framework, there is a need to focus on examining the viability of technological solutions to strengthen compliance and enforcement tools across a range of AI related risks and in-fact across governance and regulatory risks more generally.

One example, where such viability can be examined, is the proliferation of malicious synthetic media (referred to as “deepfakes”). Consequently, one of the first areas of work, where the Technical Secretariat should undertake research, is to examine the viability of such technological solutions, including watermarking, platform labelling, and other fact-checking tools. The Secretariat should engage with industry and governments globally to evaluate the merit and feasibility of standards and mechanisms to enable a chain of content provenance even as the content is modified across different tools and the modifications are watermarked using different technologies. These can then feed into user awareness programs that can be rolled out nation-wide through different channels of communications that can be driven by the Committee/ Council members.

In parallel with the efforts to examine technological solutions, the Technical Secretariat should also undertake a deeper analysis to identify any specific gaps that

may be required to be addressed to better account for the prevention, detection, reporting and awareness of malicious synthetic media.

6. **Form a sub-group to work with MEITY to suggest specific measures that may be considered under the proposed legislation like Digital India Act (DIA) to strengthen and harmonise the legal framework, regulatory and technical capacity and the adjudicatory set-up for the digital industries to ensure effective grievance redressal and ease of doing business.**

Given the rapid development of digital technologies and newer business models, there is a need for the proposed legislation (DIA) to be suitably equipped. It is important for the proposed DIA to empower the government with appropriate regulatory and technical capacity and capability to minimise risks of harm from malicious use of emerging technologies, including AI.

For example, there is a need for the Government to review and strengthen the mechanisms for redress and adjudication of matters concerning digital technologies (including the risks posed by AI applications) keeping in mind the rapid growth in the digital ecosystem and the adoption of digital technologies at scale. This should extend to the appellate level. The strengthening should focus on adequate capacity in terms of human resourcing, qualifications and expertise, use of technology, and the avoidance of unnecessary overlaps between different forums. MeitY should consider introducing requirements for Grievance Appellate Committees (**GACs**) and Adjudicating Officers (**AOs**) to be "digital by design" and to employ online dispute resolution systems in the discharge of their functions. Further, currently, the IT Secretaries in State Governments serve as AOs. These are senior officers with several other responsibilities. This mechanism should be reviewed so that there is appropriate capacity from a future perspective. The qualifications and eligibility criteria to serve as an AO may be reviewed to enable officers to serve full-time on a dedicated basis and to also enable the entry of external experts (e.g., lawyers, cybersecurity professionals, etc.).

In order to holistically review possible areas in the legal framework that can be suitably strengthened through the DIA, the **AI Sub-Group** may be tasked to work closely with MeitY to make detailed recommendations.

Conclusion

Regulation should aim to minimise the risk of harm. Even enabling innovation is a minimisation of harm as people may not be able to innovate due to lack of clarity or gap in law. Therefore, harm mitigation should be *the* core regulatory principle while operationalising the seven principles discussed in this report.

The question - to regulate the “technology” or an “application,” can only be answered in the context of the above core principle and the answer will continue to evolve based what is needed to be done to minimise the risk of harm.

Regulation controls the behaviour of people by calling out what is permissible and/ or not permissible and penalising deviation from the desired behaviour. Regulation imposes costs on everyone. Therefore, the risk of harm has to be real and specific for a discussion on regulation and for regulation to be useful.

Regulation ranges from:

- some form of licensing/ authorisation (*entity-based regulation* – banking, health, telecom, car manufacturers) or
- a set of applicable regulation (*activity-based regulation* – taxation, online safety, consumer protection, data protection, anti-trust, copyright, patent, employment, contracting etc.) or
- a *combination approach* (threshold-based identification of entity and application of certain regulations for specified activities).

Given that the AI development industry is in the nascent stage, the starting point of regulation should be “activity-based regulation” through which people are motivated to minimise the risk of harms. Given the horizontal nature of AI use, when needed, this approach can evolve into a combination approach.

Given the broad and cross cutting use of AI, Government should:

- invest in strengthening existing regulatory capabilities and implementing a whole of government approach to effectively govern its use; and
- adopt a “digital by design” approach to regulation that encourages the participants in the AI ecosystem to self-regulate each other and develop lightweight but effective outcomes-focused regulations for timely intervention on the part of the regulator which allow the liability to be attributed to the defaulting parties; and
- leverage industry ecosystem efforts, including technology-based measures, which help demonstrate responsible development and use of AI.

***

Annexure.

Discussions notes of the sub-committee not directly leading to the recommendations.

Training of AI models on copyrighted data

The Indian Copyright Act provides protection to the copyright holder and does not allow AI systems to train on copyrighted content without the approval of the copyright holder. However, if AI systems were to be allowed to train on copyrighted content without an approval from each of the right holder – what should be the scope of such training? What guardrails would be required to be mandated? How will the rights of copyright holder be adequately protected? How will compliance and enforcement be implemented? Similar questions arise of eligibility of AI generated work for copyright. All this would need to be examined and based on the answers, the legal framework adapted or left as is.

Antitrust

Unfair trade practices as a result of abuse of dominance by a few entities owning large AI systems may be a concern given the concentration of computation capability, or data clouds. However, this is not a new concern for the competition law, in India or elsewhere. Abuse of dominance, vertical integration etc. can also be examined by the competition commission of india. Further, technology is rapidly evolving, a shortage of GPUs pushes the competition to look at models that do not need GPUs and would actually be beneficial to the advancement of the field.

Given the development of AI, new scenarios may emerge which might test the regulatory systems. For example, there could be scope for algorithmic collusion where without explicit communication between entities providing or deploying the algorithms. Therefore, the regulators would be best placed to keenly observe the developments and engage with industry proactively to understand possible risks.

Assessing “market dominance” and “abuse of such dominance” requires regulators to monitor and analyse the different dynamics across the AI ecosystem. There is value in examining how techno-legal measures can assist regulators in modelling the ecosystem impact of the models and apps.

Definition of AI

While some countries have defined AI, Indian laws have generally taken a technology agnostic position and focussed on harms and effects. A similar approach may be considered at this stage, given the evolving nature of AI technologies. Most definitions attempt to be future ready but are unlikely to capture how the technology may evolve. Other definitions tend to go too broad thereby creating uncertainty as traditional software could also be interpreted to be in scope. Definitions are probably useful when they are used to pinpoint certain kinds of technologies for which specific regulatory provisions are to be mandated. However, both the definitions and the manner of identifying systems for regulatory purposes is evolving and requires deeper evaluation.

Safe harbour

The safe harbour provision (Section 79) in the IT Act provides legal protection to intermediaries that host or transmit third-party content online. One of the conditions for availing safe harbour is that the intermediary does not “select or modify the content”. In case AI models, this condition would not be met in many scenarios. It is quite evident that AI systems providers or deployers cannot claim safe harbour as a default and where they do, they would need to demonstrate that they have satisfied the conditions under law.