MITIGATING BIAS IN AI

A HANDBOOK FOR STARTUPS

INDIAai

A MEITY, NEGD & NASSCOM INITIATIVE

Introduction

- AI is a broad term that describes machines that are programmed to think, work, and react like humans. Modern AI models can learn from experience and perform tasks that require human intelligence.
  - In a very simplified form, AI can be classified into three broad heads based on the degree to which the technology can conduct tasks:

Narrow AI

AI that is programmed to perform only a single task and has a narrow range of abilities. It is developed in an environment where the problem is front-and-centre. A common example of Narrow AI is chatbots and conversational assistants such as Siri and Alexa.

A

General AI

General AI or strong AI refers to machines that can mimic human intelligence and behaviours with the ability to learn, innovate and apply their intelligence to solve problems.

B

Super AI

An AI that can not only mimics human intelligence and understanding but surpasses them with higher self-awareness. Such AI, though not existent at present, keeps researchers and scientists worried as it might lead to the extinction of the human race.

C

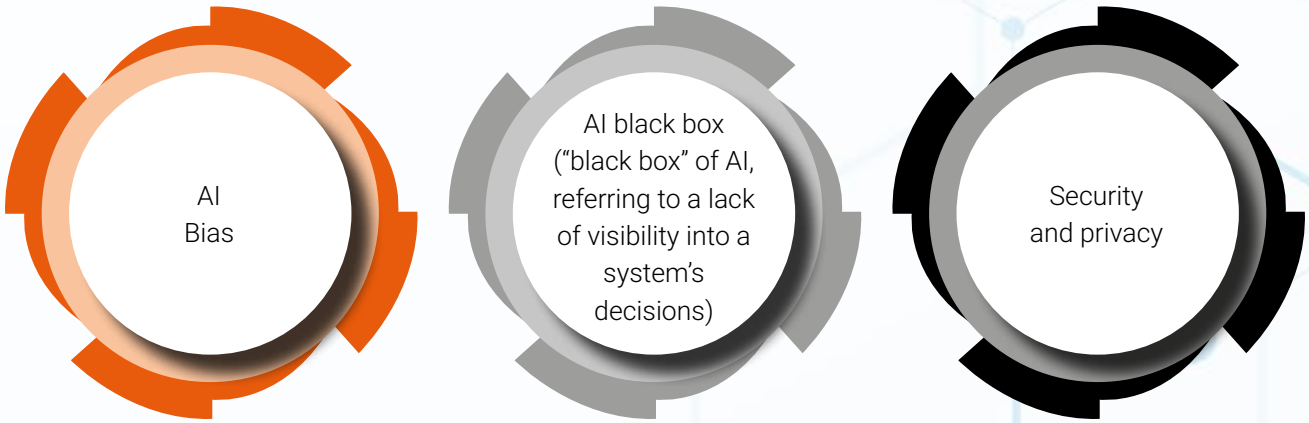
- Machine Learning (ML), a branch of AI, is the process of teaching a computer system how to make accurate predictions when data is fed into the system. It can train software to perform a task and improve its capability by ***learning and experience*** over time.
  - All ML is AI but not all AI is ML.
  - Deep learning is a subset of ML and is designed to function like a human brain.
- AI has changed the business landscape due to its ability to offer speed and reliability of an outcome at a lower cost compared to its human counterpart. It's because of this that AI has various applications in today's society.
  - **AI in Healthcare:** AI helps in making a better and faster diagnosis than humans through better predictability and consistency
  - **AI in Finance:** AI is enabling automation, algorithm trading, adaptive intelligence and chatbot interaction
  - **AI in Agriculture:** Application of AI in agriculture for crop monitoring and predictive analysis

- **AI in Retail:** AI-engine helps bring curated shopping journeys to customers. For retailers, AI has transformed inventory management, visual curation, conversational support, and outreach
- **AI in Education:** Automation of basic administrative activities like grading, offer personalized training and make learning universally accessible are some of the advantages of introducing AI in education
- **AI in Travel & Transport:** AI can assist in making travel arrangements to suggesting hotels, flights, and best routes to the customers.
- **AI in Logistics & Supply Chain:** Algorithms are today improving delivery times as well as reducing the costs, while autonomous vehicles and smart warehouses are improving efficiency.

- **AI in social media:** AI can organize and manage massive amounts of data generated from social media and analyse them to identify the latest trends, hashtags, and requirements of different users.

Major AI apprehensions in deployments

AI apprehensions mostly speak about AI's imprecise future and it can be tough to evaluate their validity in the short run. We indeed fear what we do not understand and the future of AI, with its possibility of multiple gloomy outcomes, gives us sleepless nights. Though this might not hold true for IT leaders and executives who are trying to build a practical AI strategy, many of the fears are well-founded. The premonitions' surrounding AI is relevant as the technology is capable of self-learning and improving exponentially. The most important apprehensions that are being discussed among academicians and technocrats are:



III. The Concept and Principles of Responsible AI

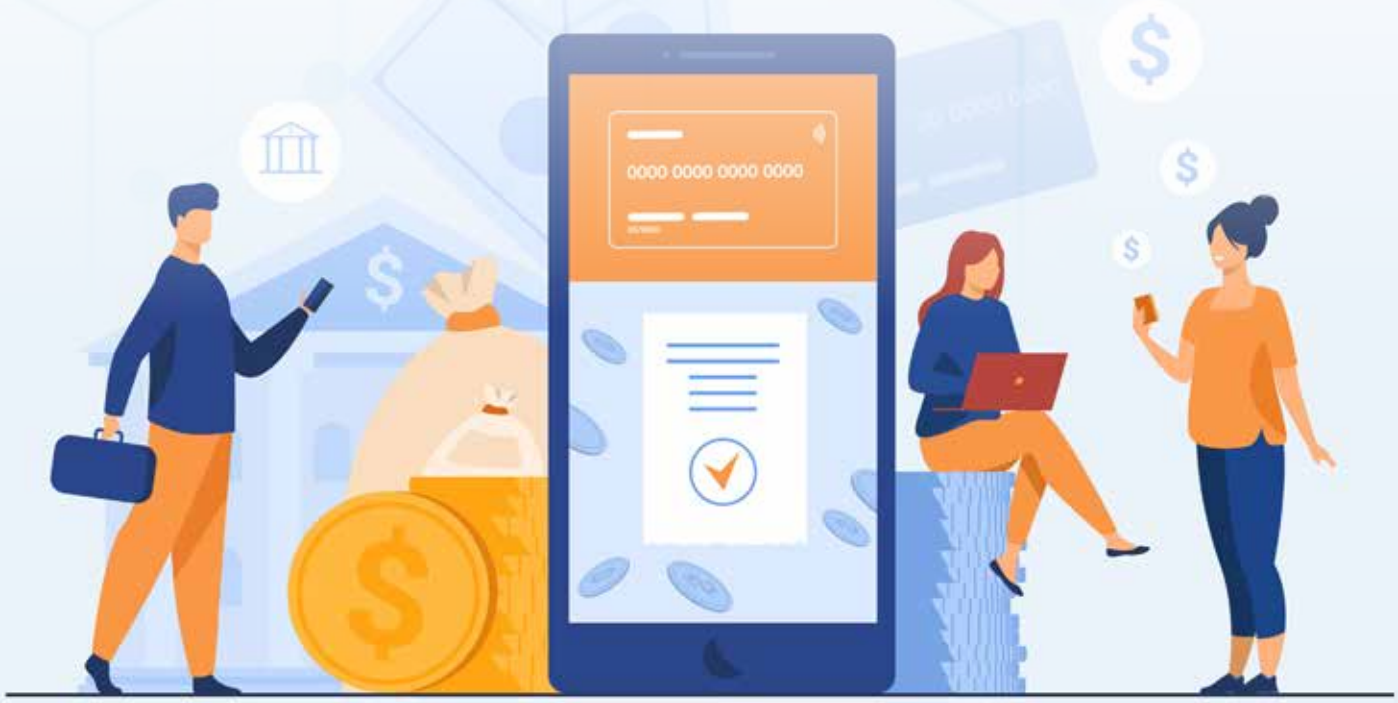
Responsible AI is a set of principles and frameworks that holds AI models responsible for the decision that they make. It is a practice that needs to be followed for designing, building and deploying AI in a manner that empowers people and businesses, yet, fairly impacts customers and society. The interest in designing responsible AI has emerged in response to growing misuse and abuse of the system, which often is unintended.

The Governing Principles of Responsible AI

With growing awareness of the risks associated with deploying AI systems that often violate legal and ethical norms, building a responsible AI has become an important aspect across all sectors. There are some of the main principles that provide a practical framework to keep in mind when designing, developing, or maintaining AI and ML systems.

- **Accountability**
  Accountability is one of the most important dimensions of decision-making and can be considered as the foundation of trust in society. In simple terms, it means an ethical, moral and answerable platform that guides organizations and allows them to reason out the logic behind any decision. To a great extent, accountability also ensures the expectation that organizations will function within the regulatory framework and be answerable for their decisions and actions. However, the intensity of the accountability obligation will vary according to the degree of autonomy and criticality of the AI system.

- **Transparency and Explainability**
  Transparency can be defined in multiple ways as several related concepts are often used synonymously such as 'explainability', 'understandability' and 'interpretability'. However, in most basic terms, it means the understanding of the system's algorithm and why is it giving a particular decision and how did it arrive at this conclusion. However, depending on what constitutes transparency and the level to which it is allowed, the meaning of transparency may vary.



- **Fairness and Non-discrimination**
  Fairness is a complex concept as the term is defined in different ways by different individuals. It is a social construct and is often subjected to biasedness. There are numerous mathematical definitions of fairness and when choosing one to build an algorithm, we automatically exclude the other. Nevertheless, fairness and parity imply achieving the same outcome across geographies, demographics and populations. In the AI context, fairness means decisions made by the system should be fair and non-discriminatory by ensuring that users are aware of the goals and aspirations of the AI system and the limitations that they have.

- **Safety and Reliability**
  The ability of AI to better human lives comes with a cost – the cost of safety. AI safety and security can be broadly defined as the attempt to ensure that no harm is caused to humanity. The safety and security of an AI system can range from the physical safety of an individual to the privacy of an individual, organization or nation and can vary from country to country. For example, while in Singapore, safety implies a "human-centric" approach which includes protecting the safety and security of people's data and wellbeing, in Europe, in addition to the physical safety and data security, stress is laid on the existence of a mechanism to test a system's adaptability and the capability to process and identify potential risk associated with AI services.

- **Privacy and Security**
  As AI becomes more prevalent, protecting the privacy and securing important personal and business information is becoming more critical and complex. With AI, privacy and data security issues require especially close attention because access to data is essential for AI systems to make accurate and informed predictions and decisions about people. AI systems must comply with privacy laws that require transparency about the collection, use, and storage of data and mandate that consumers have appropriate controls to choose how their data is used.

III. AI Bias: A Major Challenge to Responsible AI Adoption

In the pre-digital world, it was humans and organizations who were responsible for making decisions based on state laws that incorporated concepts of fairness, transparency and equity. Today, these decisions can be made entirely by machines based on AI algorithms developed by humans. And by their nature, algorithms can automate, amplify and hard code all sorts of biases as we have seen in the examples listed below.

Amazon's hiring algorithm favoured applicants who had the word 'executed' or 'captured' in their resumes and it was most visible in men's resumes as compared to women's. This automatically created exclusion through an unknown bias

AI used in one of the US health care systems to allocate care to millions of patients in the US was found to be biased wherein coloured patients (particularly black patients) received a lower standard of care despite having the same comorbid conditions as their white counterparts

Bias is an unwanted and disproportionate weightage given in favour or against an outcome or idea that results from prejudice or an unfair approach to the problem

- While AI decision-making was meant to solve human bias and unintentional errors, in reality just the opposite happened!! Biased AI systems are likely to become a widespread problem as AI algorithms start moving out of the data science lab and interacting with the outside world.
- The question is here is do the AI system get biased? The answer is, yes. Even if algorithms aren't explicitly codified to be biased, they can end up being so. Listed below are possible reasons for getting biased:

The primary factors for AI bias are:

- Human Evolutionary Problems:
  - Homo Sapiens are only 150,000 years old on this planet, and our brain is mostly animal/ beast level. This brings out various aspects of instinctual behaviour for the protection of territory etc. Every varying context our predecessors faced evolved into a bias to keep them secure/ alive.
  - Most of our observations and responses are subconsciously shaped by cultural wiring that often outlives the survival utility; for instance – racial segregation.
  - Human biases are extremely diverse and difficult to detect making it extremely difficult to pre-empt them
  - Tunnel vision problem: AI developers are more focused on achieving a specific goal rather than looking at the broader dimensions and social impact context. Insensitivity/ unawareness of real-world implications lead to a limited span of application at best and institutionalised discrimination or endangerment of life at worst.
- Digital Access Problems:
  - The great digital divide results in skewed representations and exclusions.
  - It also determines who codes the decision-making flows and what errors of omission and commission they bring in.
- Organisational Data Problems:
  - Sufficient data isn't getting generated: Sometimes, certain data points aren't captured because the business team couldn't visualise their relevance or usefulness.
  - Data architectures leading to AI bias through partial data loading: Most Data Lake architects, Business Analysts and CXO's tend to overlook this aspect and tend to load partial data into the ecosystem to save space, load and processing muscle.
  - Data Cleansing/Standardising problem: one of the main reasons that organisations are reluctant to take all the data (Lock-stock barrel) is usually one of the top three reasons that keep CXO's awake. It takes 80% of the time to cleanse legacy data for MVP, leading to major project delays.
  - Insufficient data to train algorithms: At the end of the day, it is the data that goes to the algorithm, for the algorithm to process that creates the field of play. Limited data availability which is non-representative of the real-world scenarios creates a bias in the decision output.
  - The pilot problem: What works in Pilot, may not work in Real life. A pilot environment would be very controlled and exacting to the requirement and chooses only limited,

cleansed, data sets. However, real-life interactions have to factor in a lot more complexities.

- Data collection: Often data may be collected in a way that leads to over-representation of a group thereby leading to bias
- Implicit bias in data: Many times data generated by humans have in-built biases and as AI systems learn from such data, the system automatically reflects those biases.
- Incorrect way of framing the problem and wrong choice of attributes: While building an AI algorithm, developers may incorrectly identify the problem that needs to be solved and the end result that they want to achieve. This leads to the wrong choice of attributes leading to the entire system getting biased.

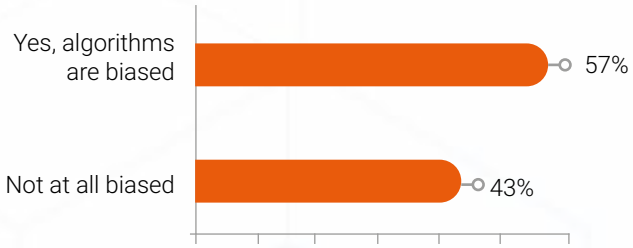
- Limited auditing facilities for AI applications:
  - The use of AI applications in a high-risk environment often calls for external audits to ensure unbiasedness in the system. However, due to data protection and privacy, such audits become impossible which leads to even more bias.

INDIAai Responsible AI Startup Survey

To have a better understanding of India's startups perspective on responsible AI and bias in the AI system, INDIAai undertook a survey of startups. The survey analysed how AI bias was viewed and dealt with by the startups. Based on these findings, two major challenges have been identified, which have been dealt with separately in the subsequent section.

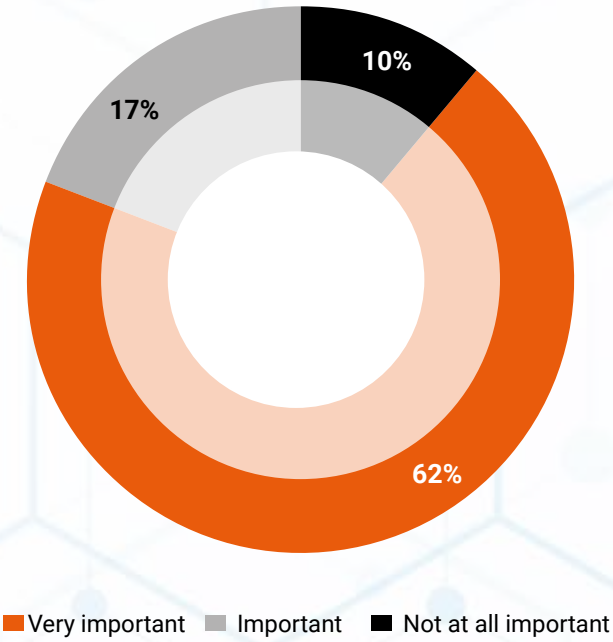
- 57% of the respondents agree to the presence of bias in their AI algorithms

Is there bias in your AI Algorithms



- 6 out of 10 respondents believe that an unbiased algorithm is extremely important for business

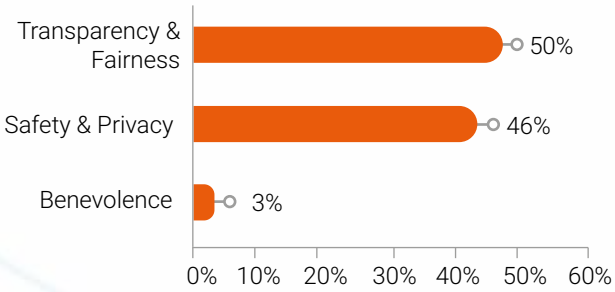
Importance of unbiased algorithms in business



- 50% of respondents opined that transparency and fairness are the 2 most important principles for designing a Responsible AI followed by safety and privacy

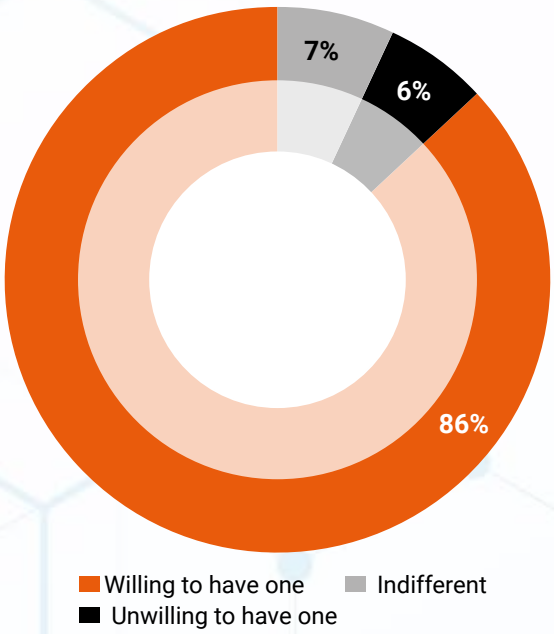
Not at all important

Important principles for Responsible AI



- Startups are also in favour of setting up a regulatory framework for creating responsible and unbiased AI

Willingness to welcome a regulatory framework to ensure AI products created are fair and unbiased



IV. Road to Responsible AI: Overcoming AI Bias

The practice of incorporating responsible AI in business is a tedious task as it includes concerns around transparency, explainability, accountability, fairness, security and privacy. The pursuit of responsible AI, is hence, a team effort that needs to be orchestrated amongst all stakeholders. Based on the INDIAai survey, the terms explainability and accountability needs further attention.

IV.a. Explainability

Explainable AI or XAI simply means making AI systems more explainable to humans as "Black Box" systems has shown to give predictions that are problematic due to inherent biasedness. In recent years, there has been an increasing popularity for XAI. The reasons are mostly the need to justify an outcome rather than just describing the approach towards the problem, the ability to control and improve so that future systems are well-versed and smart.

- IBM Explainability 360 is an open-source software toolkit unified API that helps users to gain an understanding of how AI systems make predictions, guides and tutorials in one interface
- DataRobot provides a model blueprint that outlines every processing step that the model uses to arrive at a conclusion. It also provides prediction explanations that show how each variable can impact the model's outcome

However, researchers and scientists have voiced the concern that it is extremely difficult and complex to make systems explain themselves.

Apart from this, often an algorithm is confidential which cannot be disclosed to the public or the complexity of it may be too difficult for a common man to understand which might lead to misinterpretations. Also, there is the challenge of debugging the incorrect output from a trained model. Researchers also fear that if clients have access to the reasoning and explainability of the model, they could use adversarial behaviours to change the decision-making process.

IV.a.i. XAI: The Technique

As algorithms are being used by society at large without an iota of understanding of its premise and inner workings, the question arises as to how to identify the problem of biasedness at every stage and act towards a more responsible path.

This is where XAI enters the scene. XAI is an emerging field that offers techniques to overcome the bias of algorithms and ML models and generate standard human explanations. The main reason behind XAI prominence is to build trust and detect the bias at an early stage so as to eliminate them. The explainable nature of AI is to ensure adaptability and interpretability to AI models which are further woven into contextual reasoning. There are multiple approaches in categorizing XAI which can range from applicability of a method to different models to the scope of the explanation. A more recent approach is to add a layer of explainability to an ML model that is already deployed. This is called a post-hoc method, where prediction is done first and then a layer of explainability is added to reason. However, the problem arises in terms of accountability if something goes wrong. In such a case, the question arises who should be held responsible: the team that built the machine learning model, or the team that built the explainability model. This is a question that needs to be analyzed and probed further.

IV.b. Accountability

Accountability is the first step towards creating an ethical and responsible AI as it makes the system responsible and answerable for an outcome. While responsibility has a touch of moral praise or blame and social approval, answerability relates to willingness to reveal the reasons behind an outcome. The big question is, who should be accountable for the misuse of data? Should it be data collectors, model developers or end-users?

The problem of distributed responsibility acts as a challenge to the process of identification of the responsibility. In this context, the INDIAai survey can give a clearer picture. As good as 70% of startups believe that developers should be accountable in some form or the other as they feel it is the reasonability of the developer to make the product safe and fair. Only 30% of them is of the view that others can be held responsible.

IV.c. Auditing: A Means to XAI?

Auditing data or algorithms simply means examining whether it is accurate and reliable along with the reliableness of the system to produce intended results. Algorithmic auditing has two roles: first, testing the algorithms periodically by internal teams to check if they perform as expected under variable conditions and that they are bias-free and second, testing those algorithms that have a significant role to play in people's lives. At present, where AI ecosystems lack regulations, the need of the hour is to have an external third party for the audit so as to ensure customers' trust and faith. However, the process may not be simple as organizations may resist the idea of external auditing due to concern over leakage of internal sensitive information.

- Ensuring auditing for AI applications: The use of AI applications calls for external audits to ensure the principles of Responsibility are not violated. However, due to a host of reasons including the black-box nature of algorithms, business secrets, IP rights, data protection and privacy, multidisciplinary skill requirements et al, such audits cannot be enforced easily.
  - Auditing Governance & Methodology
    - Auditing Boards
    - Stat financial Audit vs Algo Audit
    - Policies around Audit
    - IP protection
  - Auditing the Data:
    - The data ranges: Ensuring that the data set that has to be audited for is complete for known data fields (ex: age/gender: if the data set has a particular age group or gender frequency skew, that will alter the results).
    - Data Availability: Suppose, a logical data point that could provide more insight to the overall case is absent, ensuring that the same is made available or compensated with enough caveats in the output is important. (ex: quantifying domestic labour while underwriting)
    - Averaging: When there are null values, most developers, tend to average the previous values or overall indiscriminate values causes huge skew
    - Absence of data documentation: All adjustments made are to be audited and looked for biases.
  - Auditing the algorithm
    - Lift-Shift coding of algorithms: Where biases are transferred from analogue to digital.
    - Compensatory weights: Thorough evaluation of compensatory weightage might alter the data
    - Documentation of the algorithm
    - Understanding the algorithms business purposes

V. AI Regulation in India: A Quick Overview

Realizing the benefits that AI-led technologies can bring to the economy, has been working towards creating an AI-friendly technological ecosystem in India. Some of the noteworthy steps that have been taken are:

- The Ministry of Commerce and Industry has set up an AI Taskforce in 2017 which focused on the various sectors of importance in AI and the challenges faced in its adoption
- In 2018, the government think-tank, NITI Aayog was directed to initiate programs on AI and its applications. The MeitY also formed four committees to analyse issues related to leveraging AI and key policies required for its adoption. It also looked into the legal and ethical issues to AI
- The Personal Data Protection Bill, 2019 which seeks to provide for protection of personal data of individuals, and establishes a Data Protection Authority (DPA) for the same
- In 2020, NITI Aayog recommended setting up of AIRAWAT, an AI-explicit computer framework, to assist in the processing need of AI startups and research and innovation

Areas for government intervention

Creating principles and guidelines for responsible AI is just one half of the task with the other half mostly focusing on implementation and incorporation of the principles into the system. In a country as diverse as India with limited digital literacy, successful implementation requires government intervention in some form. NITI Aayog's document on Responsible AI outlines the following areas of intervention to create a strong structure of the responsible AI ecosystem.

- Regulatory interventions towards creating a trusted AI ecosystem
- Policy interventions to enable a responsible AI adoption
- Awareness and capacity building on responsible AI in the public sector
- Facilitate alignment of procurement mechanisms with responsible AI principles

In addition to this, the government must mandate responsible AI practices in all public sector procurement of AI services. Such a mandate will create a demand for good practices and boost the adoption of ethics-by-design practices in the country.

The future of AI regulation

In the wake of concern about the way AI algorithms are being used by corporates and governments, rules and regulations are being drafted to help safeguard the use of data and its privacy. As high-quality data sets are used with or without the knowledge of the customers, it becomes a moral responsibility to let the public know how the data is being used and the outcome. Hence, in the coming years, both private companies and government need to introduce regulations in sync with the various use of AI. For example, as autonomous vehicles become popular, countries using such vehicles need to update their traffic laws and regulations. Similarly, as companies increase their use of chatbots and virtual assistants, laws need to be in place to ensure that personal and sensitive data are not being shared. Overall, AI laws are in the offing.

APPENDIX

I. The hand of Law: Legislation and regulatory scenario

Any new technology wave is accompanied by its own sets of pros and cons. It is hard to predict how a new technology will be misused. It takes man-years to understand such misuse and put in place laws to regulate such technological misuse. In this regard, most governments are adopting a wait and see approach to laws and regulations on AI.

A comparative analysis of legal benchmarking guidelines: India vis-à-vis US, EU and Singapore

The EU is the most active in proposing new rules and regulations. Countries in the EU are using a combination of sectoral regulations and broader AI guidelines. There already exists sector-specific regulations for AI and sector agnostic laws that are relevant to AI. For example, the General Data Protection Rules (GDPR) 2016, which is a regulatory framework for the protection of personal data and relevant to AI, already exists. Similarly, in the US, there are 10 "Principles for the Stewardship of AI Applications" which are used by US federal agencies for drafting and implementing regulations on AI. In addition to this, there are also proposed bills to reduce biased decisions and outcomes.

However, the case in India is slightly different. While there are some defined areas of ethical framework, AI-specific laws have still not been devised. For example, in areas such as privacy, inclusiveness and accountability, regulations exist but need to be streamlined for AI-specific purpose.

Legal Benchmarking: A Comparative Analysis

	EU	USA	Singapore	India
Guidelines or regulations established specifically for AI	Yes	Yes	Yes	Not defined
Sector-specific regulations that may be applied to AI	Yes	Yes	Yes	Yes, but needs to be updated
Sector agnostic laws that are relevant to AI	Yes	Yes (Proposed)	Yes	Yes (draft version)

Source: Towards Responsible #AIforAll, NITI Aayog

Furthermore, the EU, having realized the need to design the future course of AI regulatory framework, have put in place certain guidelines and requirements identified by the High-Level Expert Group (HLEG) in 2019. The key requirements were in sync with global standards and included human agency and oversight, robustness and safety, privacy and data governance, transparency, diversity and fairness, societal and environmental well-being and accountability. Subsequently, to support the creators of AI systems and assess the compliance of the AI systems with European values, the European Commission published the Checklist for Trustworthy Artificial Intelligence in 2020 that guides developers and users of AI when implementing the above key EU requirements for building trustworthy AI.

The EU's approach to build a trustworthy AI follows a risk-based approach. It classifies risk into 4 categories:

Minimal risk

A proposal that allows free use of applications such as AI-backed video games. Minimum interference from regulators in this area as the system proposes no risk towards users/ consumers.

Limited risk

A proposal that allows free use of applications such as AI-backed video games. Minimum interference from regulators in this area as the system proposes no risk towards users/ consumers.

High risk

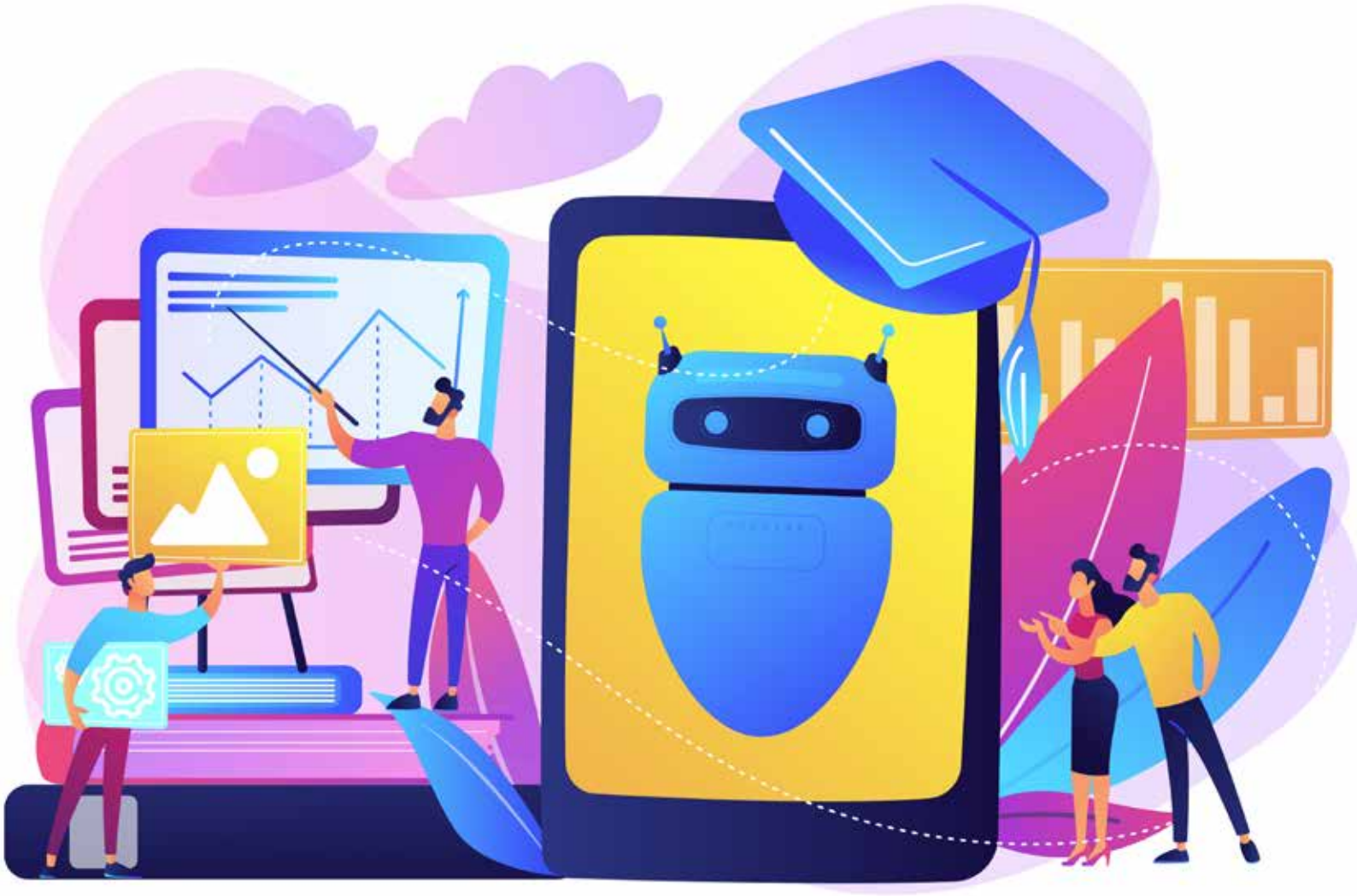
A proposal that allows free use of applications such as AI-backed video games. Minimum interference from regulators in this area as the system proposes no risk towards users/ consumers.

Unacceptable risk

A proposal that allows free use of applications such as AI-backed video games. Minimum interference from regulators in this area as the system proposes no risk towards users/ consumers.

Other legal aspects gaining regulatory attention

Laws that concern data are relevant for AI as those laws can impact the use and growth of AI. Hence, certain countries have strict laws in place when it comes to sharing and exchange of data without prior consent. The introduction of the General Data Protection Regulation (GDPR) in 2018 by the EU has forced many member countries to follow a prohibitive regulatory approach towards data and its usage. Nevertheless, there are still no specific laws around ethical AI and only time will say whether the government will regulate laws or private companies will self-monitor to ensure that they are on track.




INDIAai
A MEITY, NEGD & NASSCOM INITIATIVE