



NITI Aayog



RESPONSIBLE AI

#AIFORALL

Approach Document for India
Part 1 – Principles for Responsible AI

FEBRUARY 2021

RESPONSIBLE AI

#AIFORALL

Approach Document for India
Part 1 – Principles for Responsible AI

February 2021

Acknowledgements

In writing this report; Towards Responsible AI for All, Rohit Satish and Tanay Mahindru from NITI Aayog have made valuable contributions.

We are pleased to have collaborated with the World Economic Forum Centre for the Fourth Industrial Revolution as the Knowledge partner in developing the Responsible AI for All approach document. The valuable contributions of Ms. Kay Firth Butterfield and Ms. Arunima Sarkar from World Economic Forum is acknowledged. Legal inputs enabled by Microsoft is also acknowledged.

We are also grateful for the support and contributions of several experts from India and globally including Prof Amit Sethi, Prof Balaraman Ravindran, Prof Gary Marchant, Google, Mr John Havens and Srichandra (IEEE), Prof Mayank Vatsa, Dr Shefalika Goenka and her team at PHFI, Dr P Anandan and Dr Rahul Panicker from Wadhwani Institute for Artificial Intelligence, Dr Rohini Srivatsa, and Vidhi Center for Legal Policy. Valuable inputs were also provided by various Ministries/ Departments of the Government of India and regulatory institutions, namely MeitY, DST, DBT, PSA's Office, RBI and NHA.



Anna Roy
Advisor,
NITI Aayog

Dr. Rajiv Kumar

Vice Chairman

National Institution for Transforming India

Government of India

New Delhi, India



FOREWORD

The economic potential of deployment of Artificial Intelligence has been widely highlighted by policy makers, technologists, academics and civil society around the world. In India, the National Strategy on Artificial Intelligence (NSAI) released by NITI Aayog in 2018 highlights the potential of AI to solve social challenges faced by its citizens in areas such agriculture, health and education, in addition to the pure economic returns that are brought by this technology.

Since 2018, the deployment of AI in India has only grown, through the support of enthusiastic state governments, research institutions, leading applications from the private sector and a vibrant evolving AI start-up ecosystem. Though AI is often deployed with intentions of improving access and quality and higher efficiency and solving pressing problems, risks and challenges of leveraging AI have also emerged across a number of different areas.

AI is a technology that continues to advance rapidly and the discourse on AI ethics and governance is also evolving. Globally, a number of different sets of 'AI ethics principles' have been put forward by multilateral organizations, private sector entities and various nation states. For India, these principles are grounded in the fundamental rights afforded to citizens by the Constitution. Apart from establishment of principles however, it is also necessary for India to frame means of implementing the principles across the public sector, private sector and academia in a manner that balances innovation and governance of potential risks.

Building further on the National Strategy on AI, this approach paper, the first part of the strategy titled "Towards Responsible AI for All", aims to establish broad ethics principles for design, development and deployment of AI in India – drawing on similar global initiatives but grounded in the Indian legal and regulatory context. The second part of the strategy which will be released shortly explores means of operationalization of principles across the public sector, private sector and academia. Within this framework, it is hoped that AI can flourish, benefitting humanity while mitigating the risks and is inclusive bringing the benefits of AI to all.

The paper incorporates insights, feedback and experiences consolidated through inter-ministerial consultations, large-scale global multi-stakeholder consultations and a series of 1-1 consultations with AI ethics experts in India and globally, as well as wider public consultations, conducted over the last 15 months. This paper is meant to serve as an essential roadmap for the AI ecosystem, encouraging adoption of AI in a responsible manner in India and building public trust in the use of this technology, placing the idea of 'AI for All' at its very core.

February, 2021
New Delhi,
India

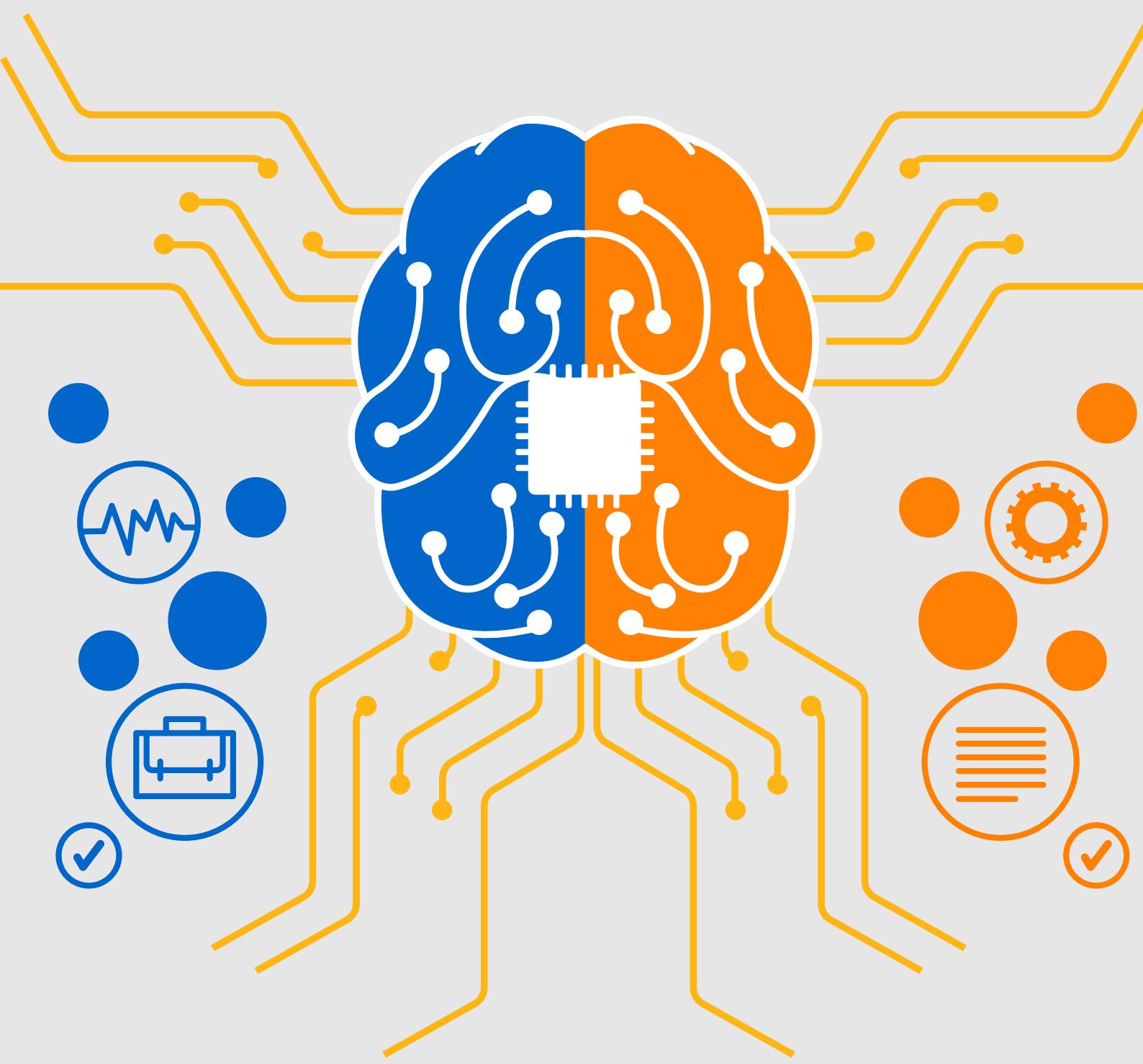
A handwritten signature in blue ink, appearing to read "Rajiv Kumar".
Dr. Rajiv Kumar

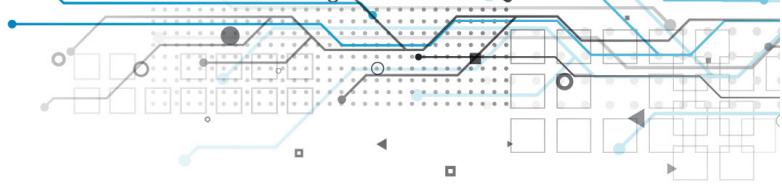


Contents

| | |
|---|-----------|
| Introduction | 01 |
| 1. The Need for Responsible AI | 05 |
| Exploring the Ethical Considerations | 10 |
| Systems Considerations | 11 |
| <i>Systems Consideration 1: Understanding the AI system's functioning for safe and reliable deployment</i> | 12 |
| <i>Systems Consideration 2: Post-deployment—can the relevant stakeholders of the AI system understand why a specific decision was made?</i> | 13 |
| <i>Systems Consideration 3: Consistency across stakeholders</i> | 15 |
| <i>Systems Consideration 4: Incorrect decisions leading to exclusion from access to services or benefits</i> | 16 |
| <i>Systems Consideration 5: Accountability of AI decisions</i> | 18 |
| <i>Systems Consideration 6: Privacy risks</i> | 21 |
| <i>Systems Consideration 7: Security risks</i> | 23 |
| Societal Considerations | 26 |
| 2. Legal and Regulatory Approaches for Managing AI Systems | 28 |
| 3. Technology Based Approach for Managing AI Systems | 33 |
| 4. Principles for Responsible Management of AI Systems | 37 |
| Appendix | 43 |
| 1. Self-Assessment Guide for AI Usage | 44 |
| 2. Review of Global Regulatory Landscape | 50 |
| 3. Model Transparency Mechanisms | 53 |

Introduction





Introduction

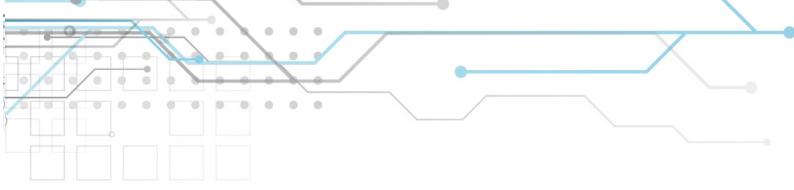
NDI Aayog released the National Strategy for Artificial Intelligence (NSAI) discussion paper in June 2018, in pursuance of the mandate entrusted to it by the Hon'ble Finance Minister in the Budget Speech of 2018 – 2019. NSAI while highlighting the potential of Artificial Intelligence (AI) for accelerating growth also emphasised the social potential of large scale adoption of AI with a focus on themes of inclusivity, adopting the theme of 'AI for All'. Towards promoting development as well as adoption of AI, the NSAI made broad recommendations for supporting and nurturing an AI ecosystem in India under four heads, (a) promotion of research; (b) skilling and reskilling of the workforce; (c) facilitating the adoption of AI solutions; and (d) the development of guidelines for 'responsible AI'. While underlining the role of private sector and collaboration, NSAI identified key focus sectors where the Government was expected to play the lead, viz. health, education, agriculture, smart cities and mobility.



NSAI recommended establishment of clear mechanisms to ensure that the technology is used in a responsible manner by instilling trust in their functioning as a critical enabling factor for large scale adoption in a manner that harnesses the best that the technology has to offer while protecting citizens. Need for a fine balance between protecting society (individuals and communities) without stifling research and innovation in the field was underlined.

The future of AI is determined by a diverse group of stakeholders, including researchers, private organisations, Government, standard-setting bodies, regulators and general citizens. Around the world, various countries and organisations have defined principles to guide responsible management of AI for various stakeholders.

'Towards the Development of Responsible 'AI for All'', proposes principles for the responsible management of AI systems that may be leveraged by relevant stakeholders in India. Case studies of AI systems in India and around the



world are studied and the principles for responsible AI are derived from the Constitution of India and various laws enacted thereunder.

The case studies and considerations in this paper are limited in context to 'Narrow AI' solutions. They have been grouped into two broad buckets: 'Systems considerations' arising as a result of the system design choices and deployment processes, and have the potential to impact stakeholders interacting with a specific AI system; and 'Societal' considerations, that are broader ethical challenges pertaining to risks arising out of the very usage of AI solutions for specific functions, and have potential repercussions on the society beyond the stakeholder interacting directly with specific systems.

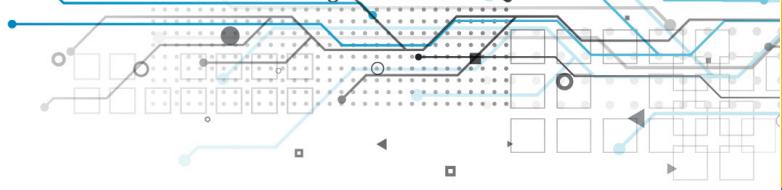
The Paper examines following system considerations:

- Lack of understanding an AI system's functioning makes it difficult to reliably and safely deploy AI systems
- Challenges in explaining specific decisions of AI systems makes it difficult to trust them
- Inherent bias could make the decisions prejudiced against segments of population
- Potential for exclusion of citizens in AI systems used for delivering important services and benefits
- Difficulty in assigning accountability
- Privacy risks
- Security risks;

and following Societal Considerations:

- Impact on Jobs
- Malicious psychological profiling

The Supreme Court of India, in various cases such as *Naz Foundation* and *Navtej Johar* has defined the prevailing morality of our country to be based on the principle of Constitutional morality. The Supreme Court has stressed time and again on adherence to constitutional morality over social morality, with the former's reach extending beyond the mere text of the Constitution to encompassing the values of a diverse and inclusive society while remaining faithful to other constitutional principles. The Paper studies the various



considerations under the lens of the Constitutions and identifies 'Principles for Responsible Management of Artificial Intelligence in India'.

On the basis of Systems and Societal considerations, the Paper identifies the following **broad principles for responsible management of AI**:

1. Principle of Safety and Reliability
2. Principle of Equality
3. Principle of Inclusivity and Non-discrimination
4. Principle of Privacy and security
5. Principle of Transparency
6. Principle of Accountability
7. Principle of protection and reinforcement of positive human values

The manner and degree of implementation of principles must provide an enabling environment for promoting a responsible AI ecosystem in India. The measures may be suitably calibrated according to the specific risk associated with different AI applications in a manner that keeps pace with technology advances.

This is an evolving area of work. NITI Aayog is already working on Part-2 of the Paper that would provide the approach towards ongoing update of Principles and enforcement mechanisms of the responsible AI in the public sector, private sector and academia. This Paper is expected to be released shortly.

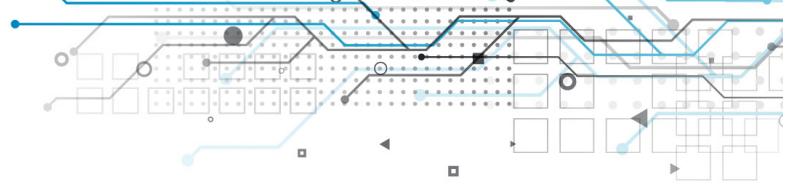
I consider this document to be a critical step towards #AIforAll and hope it starts a dialogue on ensuring that the significant and transformative potential of Artificial Intelligence is used for the benefit of Indian citizens and humanity overall



Amitabh Kant
CEO, NITI Aayog

The Need for Responsible AI





The Need for Responsible AI

Artificial Intelligence (AI) systems have gained prominence over the last decade due to their vast potential to unlock economic value and help mitigate social challenges. Thus not only the development but also adoption of AI has seen a global surge in recent years. It is estimated that AI has the potential to add USD 957 billion, or 15 percent of current gross value added to India's economy in 2035¹. It is projected, that the AI software market will reach USD 126 billion by 2025, up from USD 10.1 billion in 2018². The rapid increase in adoption can also be attributed to the strong value proposition of the technology.

The National Strategy for Artificial Intelligence (NSAI) has successfully brought AI in the centre-stage of the reform agenda of the Government by underlining its potential to improve outcomes in sectors such as healthcare, agriculture, or education. Role that AI plays in facilitating improved scale of delivery of specialized services (remote diagnosis or precision agriculture advisory) and improved inclusive access to government welfare services (regional language chatbots or voice interfaces) implies a whole new path for government interventions in these sectors. Further the NSAI underlines the need for a robust ecosystem that facilitates cutting edge research to not only solve for these societal problems and serve as the test bed of AI innovations but at the same time enable India to take a strategic global leadership by scaling these solutions globally.

As these factors continue to favour the increased application of AI to a variety of private and public use cases, it is expected that AI usage will become ingrained and integrated with society. In India, large scale applications of AI are being trialled everyday across sectors³. In Uttar Pradesh, for example, 1,100 CCTV

1. Rewire for Growth: Accelerating India's Economic Growth with AI, Accenture (2018)
2. Artificial Intelligence Market Forecasts | Omdia; DC FutureScape: Worldwide IT Industry 2018 Predictions
3. <https://indiaai.gov.in/case-studies>

Places of Worship

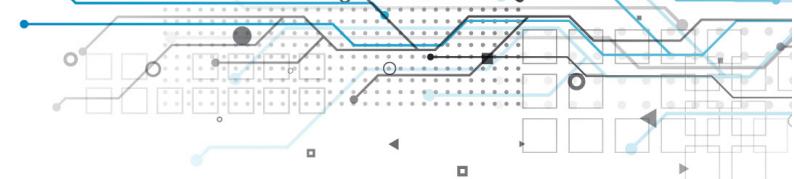
cameras were installed for the 'Prayagraj Kumbha Mela' in 2019. The cameras would raise an alert when the crowd density exceeded a threshold, and the connected Integrated Command and Control Centres provided the security authorities with relevant information⁴. Wadhwani AI is testing an AI-powered smartphone-based anthropometry tool that will empower health workers to screen low-birth-weight babies without any specialised equipment⁵. NIRAMAI, a startup, has developed an early-stage breast cancer detection system using a portable, non-invasive, non-contact AI-based device.⁶ Researchers from IIT Madras are looking to use AI to predict the risk of expectant mothers dropping out of healthcare programmes, to improve targeted interventions and increase positive healthcare outcomes for mothers and infants⁷.

Box 1: Artificial Intelligence

In this document, the scope and definition of AI is similar to the one mentioned in the National Strategy for AI, 2018 (NSAI, 2018)- a constellation of technologies that enable machines to act with higher levels of intelligence and emulate the human capabilities of sense, comprehend and act. Computer vision and audio processing can actively perceive the world around them by acquiring and processing images, sound and speech. The natural language processing and inference engines can enable AI systems to analyse and understand the information collected. An AI system can also take decisions through inference engines or undertake actions in the physical world. These capabilities are augmented by the ability to learn from experience and keep adapting over time.

This paper studies the ethical implications of 'Narrow AI', which is a broad term given to AI systems that are designed to solve specific challenges that would ordinarily require domain experts. Both systems and societal considerations are explored from the perspective of narrow AI only. Broader ethical implications of 'Artificial General Intelligence' (AGI) or 'Artificial Super Intelligence' (ASI) are not considered in this paper. Further, systems considerations considered in this document mainly arise from decisions taken by algorithms.

4. Artificial Intelligence real showstopper of Kumbh Mela 2019
5. <https://www.wadhwaniai.org/work/maternal-newborn-child-health/>
6. <https://www.niramai.com/>
7. <https://www.livemint.com/technology/tech-news/google-funds-six-ai-based-research-projects-in-india-11582016278056.html>



While the potential of these solutions to improve productivity, efficiency and outcome is well established, the NSAI (2018) also advocated for managing the AI systems responsibly. Around the world, instances of harm caused by deployment of AI systems have been realised. AI systems appear to have prejudices in certain decisions and this gets amplified when used in large scale, such as when the system to allocate healthcare in the USA was found to discriminate against black people⁸. The blackbox nature of AI and its 'self-learning' ability make it difficult to justify its decisions and in apportioning liability for errors. AI systems often lack transparency and the user is unaware that they are dealing with a chatbot or an automated decision-making system, this awareness being key to build trust with the user. Safety and robustness of AI systems can pose serious challenges especially in high risk prone applications; unequal access to AI powered applications for marginalized populations can further accentuate digital divide.

According to a Capgemini report, 85% of the surveyed organisations in India have encountered ethical concerns from the use of AI¹¹. There are also concerns of AI systems leading to job loss due to automation. The usage of AI for malicious intent for e.g. deep fakes to create misinformation have shown to have serious repercussions on society with instances of AI system enabled targeted propaganda, leading to social discord.

The risks of not managing AI systems responsibly also has a significant economic impact. Multiple firms placed a moratorium on facial recognition technology after issues around bias against specific population groups emerged⁹. A survey by Capgemini shows that ethical AI interactions drive customer trust and satisfaction- with AI systems that are seen as ethical have a 44 point Net-Promoter-Score (NPS) advantage over the ones that are not. Over 50% executives agreed that it is important to ensure that AI systems are ethical and 41% of senior executives report to have abandoned an AI system due to ethical concerns.¹⁰

This paper aims to study the risks from the use of AI systems in India and around the world. In this regard, the impact of AI systems may broadly be divided into following two groups:

8. <https://www.nature.com/articles/d41586-019-03228-6>

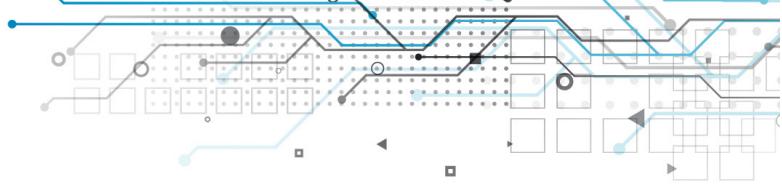
9. <https://gcn.com/articles/2020/06/10/ibm-quits-facial-recognition.aspx>

10. https://www.capgemini.com/wp-content/uploads/2019/08/AI-in-Ethics_Web.pdf

- 
- a. **Direct impacts**—defined as the implications that are caused due to citizens (or primary ‘affected stakeholders’) being subject to decisions of a specific AI system. These typically result from system design choices, development and deployment practices and are studied under **Systems considerations**. For example, **AI for cancer screening** needs consideration for the patient’s privacy in its design
 - b. **Indirect impacts**—defined as implications caused due to the overall deployment of AI solutions in society. This has potential repercussions on society **beyond the stakeholder** directly interacting with the system and are studied under **Societal considerations**. Such considerations may require policy initiatives by the Government.

This document examines the potential risks, followed by a study of legislative practices and technology approaches of managing them and goes on to recommend Principles for responsible management of AI systems. The Principles are expected to safeguard public interest and also promote innovation through increased trust and increased adoption.

Besides establishment of Principles there is a need to formulate enforcement mechanisms that would ensure the Principles are adopted across board, including the public sector, private sector and academia in a manner that balances innovation and potential risks. Part-II of the series on *Responsible AI for All* will explore the enforcement mechanisms to translate Principles to practice.



Exploring the Ethical Considerations

The considerations in this section were chosen on the basis of expert consultations, desk review of examples of AI deployment globally, and interviews with agencies deploying AI solutions in India. The causes for considerations may be deeply interconnected and, in some cases, partially overlapping. Considerations have thus been divided in a manner that identifies *distinct risks* they pose to various stakeholders.

Systems Considerations

~~Box~~ 2: The Blackbox problem

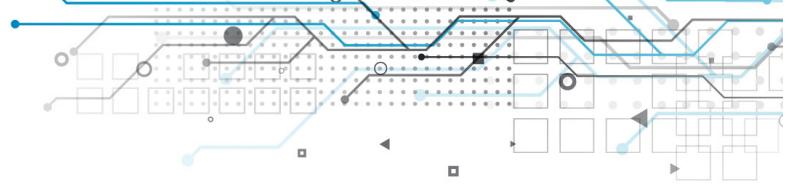
While the advances in machine learning algorithms and techniques have greatly contributed to higher accuracy, the underlying AI systems have also become increasingly opaque. Such systems have been successful in using a large number of features to make complex and sometimes consequential decisions but without exposing the underlying rationale.

In traditional statistical approaches, human programmers influence the choice of parameters and the mechanism to influence a prediction. In AI systems, input to the model (called features) are provided along with the 'correct' output through annotated labels during the training. The AI system then identifies the relationship between input features and the labels. Understanding this relationship becomes harder as the models become increasingly complex. This manifests itself as the inability to fully understand an AI's decision-making process and the inability to predict the AI's decisions or outputs—also known as the "black box problem".¹¹

The blackbox problem does not exist for all forms of machine learning solutions, and there are means of performing similar functions using more "rule-based" techniques, although the accuracy may be significantly lower. The accuracy vs interpretability trade-off has limited the applicability of AI systems in several high-stakes decision making.¹² One of the major global research efforts is around identifying models that are highly accurate and explainable.¹³

One left
network
will

-
11. Bathaei, Yavar. *THE ARTIFICIAL INTELLIGENCE BLACK BOX AND THE FAILURE OF INTENT AND CAUSATION*. *Harvard Journal of Law & Technology* Volume 31, Number 2 Spring 2018
 12. <https://arxiv.org/pdf/1811.10154.pdf>
 13. <https://www.darpa.mil/program/explainable-artificial-intelligence>



The considerations emerging from this property of deep learning technology manifest themselves in several ways. For example, it is difficult to establish if an AI system can be deployed safely and reliably without understanding and controlling how it works; trusting an AI system's decision becomes a challenge if there is a disagreement due to lack of explanation; improving a model performance is difficult when the root cause is unknown.¹⁴

Systems Consideration 1: Understanding the AI system's functioning for safe and reliable deployment

As mentioned in Box 2, machine learning models learn by identifying relationships between input features and output labels. Model evaluation is typically done by holding out a portion of the datasets as a *test dataset*. This may not necessarily reflect the various real world deployment scenarios and when the relationship between the input features and output is not understood, it becomes difficult to predict its performance in a new environment. This makes it difficult to reliably deploy and scale such AI systems.

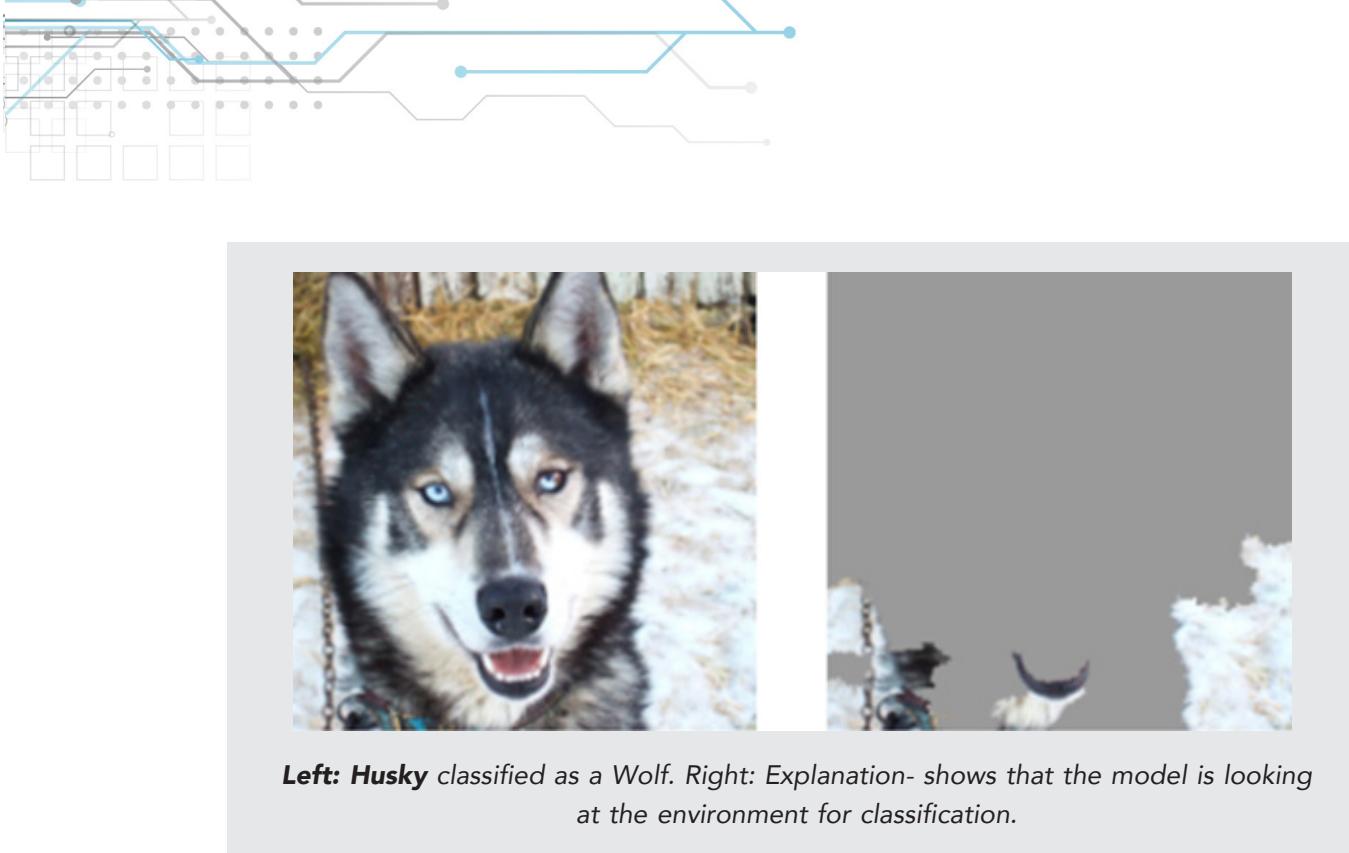
| The Issue | Its Implications |
|--|---|
| While accuracy gives a reasonable view into how a system performs, understanding decision making process is important to ensure safe and reliable deployment | The system could pick spurious correlations, in the underlying data, leading to good accuracy in test datasets but significant errors in deployment |

Box 3: Wolf or Husky?

In an example referenced in Ribeiro et al., (2016)¹⁵, an image classification algorithm performed reasonably well in its prescribed task—classify the image of an animal as either a wolf or a husky. When the model was analysed, it was found that the system was classifying images based on the background and not the animal itself. While the model performed reasonably well on the data used to test- it would clearly not do as well in the real world.

14 <https://storage.googleapis.com/cloud-ai-whitepapers/AI%20Explainability%20Whitepaper.pdf>

15. Ribeiro, M., Singh, S., & Guestrin, C. (2016, August 09). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Retrieved August 10, 2020, from <https://arxiv.org/abs/1602.04938>



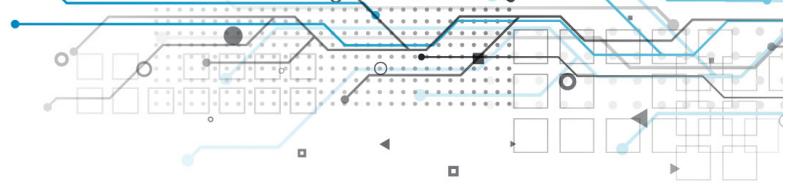
Left: Husky classified as a Wolf. Right: Explanation- shows that the model is looking at the environment for classification.

Systems Consideration 2: Post-deployment—can the relevant stakeholders of the AI system understand why a specific decision was made?

While the previous consideration was on understanding the overall principles behind decision making, in certain cases, individual decisions may have significant impact and may require an explanation. There are various examples for such decisions—credit scoring, fraud detection, loan eligibility, insurance qualification, access to Government services, etc. As algorithms make these decisions, very often, an end user has an expectation of factual assessment. Particularly in blackbox systems, the user is sometimes neither aware of the inputs considered nor of the exact contours of the decision made by the algorithm. Such explanations also satisfy a strong imperative for *reason giving*, a key component of procedural fairness in law.¹⁶

In addition to providing explanations, the deployment environment and stakeholders interacting with the AI system should also be considered. The stakeholders may come from diverse backgrounds and the explanation offered by the system must be in a manner that can be understood by them. It is also important to note that the stakeholders are not only limited to the users, but also audit agencies, regulatory bodies, standard-setting entities, people affected by the decisions, etc.

16. C. Coglianese & D. Lehr, 'Transparency and Algorithmic Governance' (2019) 71 ADMIN. L. REV. 1



| The Issue | Its Implications |
|---|---|
| With 'Deep Learning' systems have become opaque, leading to the 'black box' phenomenon; Simple linear models, offer interpretable solutions but their accuracy is usually lower than deep learning models; | Leads to: <ul style="list-style-type: none"> • A lack of trust by users, discouraging adoption; • Difficulty in audit for compliance and liability; • Difficult to debug/maintain/verify and improve performance; • Inability to comply with specific sectoral regulations; |

Box 4: IBM Watson for Oncology

The absence of explanation of output or decision may affect the adoption of the technology depending on the severity of implications. In a specific deployment of IBM Watson, for example – in particular, Watson for Oncology–when Watson's results agreed with physicians, it provided confirmation but didn't help reach a diagnosis. When Watson didn't agree, then physicians simply thought it was wrong. This resulted in the system being abandoned in many hospitals around the world.^{17,18}

Lack of **understanding of specifics** in individual decisions has several consequences which discourages adoption, especially for consequential decisions. Individual decisions are difficult to audit by regulatory, standards and compliance agencies **besides the lack of redressal available to an aggrieved recipient given the difficulty in determining the grounds for challenging it in a court of law.**

For the developers of the system, identifying specific errors and making improvements to its performance is a challenge as the **inability to track the source of the error makes targeting interventions difficult**. In cases where the law requires an explanation of individual decisions, it becomes prohibitive to use, even if the models are highly accurate.

17. Forbes: Time To Open The AI 'Black Box'

18. <https://theconversation.com/people-dont-trust-ai-heres-how-we-can-change-that-87129>

Systems Consideration 3: Consistency across stakeholders

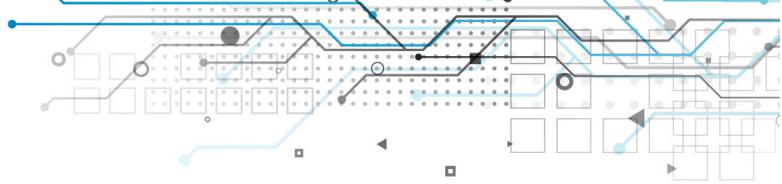
Though automated solutions are often expected to introduce objectivity to decision making, recent cases globally have shown that AI solutions have the potential to be 'biased' against specific sections of society. This can lead to inconsistent output across a diverse demography who are otherwise similarly placed. Real life manifestations of such bias tie into historically discriminatory behaviour, where members of a certain caste, class, sex or sexual orientation, among others, are denied opportunities on the basis of an identifying characteristic even though they are completely similar in all ways relevant to the decision being made.¹⁹

Interesting !

The emergence of bias in AI solutions is attributed to a number of factors arising from various decisions taken across different stages of the lifecycle and the environment in which the system learns. In more rule-based machine learning techniques, the performance of the AI solution is largely dictated by the rules defined by its developers. In deep learning methods, the performance of the solution is defined by the data used, models chosen, parameters used, goals defined, etc, and the inherent complexity of such models makes it difficult to identify specific sources of bias. While individual human decisions are not without bias, AI systems are of particular interest due to their potential to amplify its bias across a larger population due to large-scale deployment.

| The Issue | Its Implications |
|--|--|
| <ul style="list-style-type: none">Different types of cognitive biases have been identified and tend to be 'unfair' for certain groups (across religion, race, caste, gender, genetic diversity);Since AI systems are designed and trained by humans, based on examples from real-world data, human bias could be introduced into the decision-making process; | <ul style="list-style-type: none">Large scale deployment of AI, leads to a large number of high frequency decisions, amplifying the impact of unfair bias.Leads to lack of trust and disruption of social order |

¹⁹ <https://www.brookings.edu/blog/techtank/2019/11/18/highlights-addressing-fairness-in-the-context-of-artificial-intelligence/>



Box 5: Bias in the real world

Bias has already led to instances of discrimination in the real world. In 2015 Amazon experimented a machine learning-based solution to evaluate applicants by observing

patterns in resumes submitted to the company a previous 10-year period²⁰. The system rated male applicants higher than females because historically, there was a higher number of male applications and trends in the data showed a historical preference for male candidates, as well. **In effect, Amazon's system taught itself that male candidates were preferable.**

Is the
in rm
right??

Another example that has made headlines recently was when a passport photo checker used AI to check if a person has blinked¹⁶. This model, however, had issues when checking people of Asian descent- which was mostly attributed **to the lack of Asian faces in the training dataset.**

Systems Consideration 4: Incorrect decisions leading to exclusion from access to services or benefits

AI systems are inherently probabilistic systems and it is uncommon to find systems that are 100 percent accurate in their predictions. For consequential decisions, like the beneficiary identification system, criminal identification system, the *social cost* of an incorrect decision is very high and typical performance indicators may not be sufficient. In a beneficiary identification system, an incorrect decision could lead to exclusion of services and benefits guaranteed by the State and in criminal identification systems, it could lead to loss of fundamental rights. When the AI systems are used, particularly for critical services by the Government, it is important to have processes and systems in place for raising an objection.

The 'systematic' exclusion from access to services and benefits could undermine trust in the system. General lack of awareness could also lead to over-dependence due to false or exaggerated belief in such technologies (*automation bias*) and may further aggravate the problem.²¹ A typical approach towards this is to introduce a human intervention whenever such consequential decisions are made.

20. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

21. <https://doi.org/10.2514/6.2004-6313>

| The Issue | Its Implications |
|---|--|
| <ul style="list-style-type: none"> There are a variety of means of assessing or evaluating the performance of an AI system (Accuracy, precision, recall, sensitivity, etc); In some cases, despite a high accuracy a system may fail in other measures; | <ul style="list-style-type: none"> May lead to exclusion of citizens from services guaranteed by the state; |

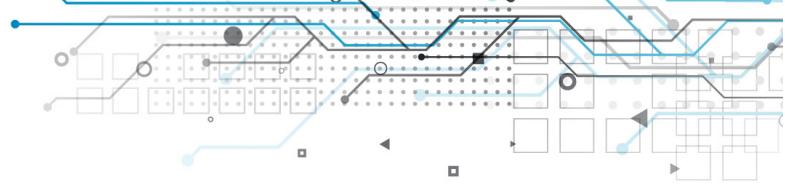
Box 6: Ensuring no one is left behind- Fraud Detection in Health Insurance

National Health Authority (NHA) is the apex body responsible for implementing India's public health insurance scheme under **Pradhan Mantri Jan Arogya Yojana (PM-JAY)**. It is established at the national level and implementation is carried out by the State Health Agencies. Beneficiaries can avail cashless treatment at any of the empanelled hospitals across India and PM-JAY makes the payment directly to the hospital.

The NHA includes a fraud detection cell at the national level called National Anti-Fraud Unit (NAFU) and at the state level it is called State Anti-Fraud Unit (SAFU). A large number of transactions are processed for insurance claims on a daily basis. To detect and flag fraudulent transactions, AI systems are employed.

Considering the high social cost of a potential incorrect decision, no treatment is stopped because of flagging by the AI system. When the AI system flags a case, the reasons for flagging is forwarded to the SAFU and investigated. While the patient always receives treatment without delay, the payment is disbursed to the hospital only after all the queries related to the case are adequately resolved.

The AI system has been developed by a vendor hired through a public RFP. The RFP document emphasizes the need to **minimize false positives in the system**. For evaluation of the bidders, 8% is reserved for "Adaptability of the solution to incorporate feedback to reduce false positives and handle errors". In addition, the payment structure is outcome based and has a variable component for the ratio of true positive cases to the



total amount of cases. On the other hand, in order to ensure genuine fraudulent cases do not pass undetected, a minimum criterion is defined for the total number and the total value of fraudulent cases identified by the AI system²².

Systems Consideration 5: Accountability of AI decisions

This consideration emerges mainly in more opaque forms of AI in which a specific decision, action or inaction of the system is influenced by a variety of parameters, such as data used for training, algorithms, processes, training parameters, deployment environment etc. Different entities may be involved in each step of the development and deployment process. In self-learning systems, the deployment environment itself could influence the decision-making process. The '*many hands problem*', associated with complex computer systems, complicates the issue of assigning responsibility under extant regimes of accountability and legal recourse. Establishing cause of action is the first step of a civil suit and an opaque AI system coupled with a large number of interconnected factors behind individual decisions makes it difficult for attribution of errors and assigning liabilities.²³ Examples of real-world instances of such issues is presented in *Box 7* and *Box 8*.

| The Issue | Its Implications |
|--|--|
| <ul style="list-style-type: none"> Decisions by AI systems are influenced by a complex network of decisions at different stages of its lifecycle. Deployment environment also influences self-learning AI Assigning accountability for harm from a specific decision is a challenge | <ul style="list-style-type: none"> Lack of consequences reduces incentive for responsible action Difficulty in grievance redressal |

²² RFP, "Selection of an agency to design, develop, implement, operate and maintain Fraud Control Analytics Platform for National Health Authority"

²³. <https://stanford.library.sydney.edu.au/archives/sum2010/entries/computing-responsibility/#2.2.1>

Box 7: The case of Elaine Herzberg

In 2018, Elaine Herzberg was hit by a test vehicle operating in a self-driving mode. The collision led to the first recorded case of fatality involving a self-driving car. The Advanced Technology Group at Uber Technologies had modified the vehicle with a proprietary automated driving system. A human-backup safety driver was sitting in the car during the collision but was looking at a cellphone during the crash. The road was dry and illuminated by the street light.

Following the collision, the National Transport Safety Board (NTSB) launched an investigation and identified the following:

The Uber ATG automated driving system detected the pedestrian 5.6 seconds before impact. Although the system continued to track the pedestrian until the crash, it never accurately identified the object crossing the road as a pedestrian — or predicted its path.

- Had the vehicle operator been attentive, the operator would likely have had enough time to detect and react to the crossing pedestrian to avoid the crash or mitigate the impact.
- While Uber ATG managers had the ability to retroactively monitor the behaviour of vehicle operators, they rarely did so. The company's ineffective oversight was exacerbated by its decision to remove a second operator from the vehicle during testing of the automated driving system.

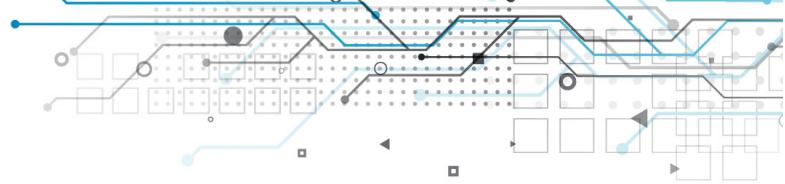
Uber ATG made several changes to address the deficiencies identified, including the implementation of a safety management system²⁴

In this situation, it was difficult to ascertain liability (safety driver or ATG group or the technology itself). After the incident, Uber stopped testing its self-driving vehicles across all cities. The incident also caused other companies to cease road testing of their self-driving vehicles²⁵.

In Nov 2019, the NTSB released a report³¹ with the following recommendations,

24. <https://www.ntsb.gov/news/press-releases/Pages/NR20191119c.aspx>

25. <https://spectrum.ieee.org/view-from-the-valley/transportation/self-driving/jensen-huang-on-the-uber-tragedy-and-why-nvidia-suspended-testing>



To the National Highway Traffic Safety Administration:

1. Require entities who are testing or who intend to test a developmental automated driving system on public roads to submit a safety self-assessment report to your agency.
2. Establish a process for the ongoing evaluation of the safety self-assessment reports as required in Safety Recommendation 1 and determine whether the plans include appropriate safeguards for testing a developmental automated driving system on public roads, including adequate monitoring of vehicle operator engagement, if applicable.

To the state of Arizona:

3. Require developers to submit an application for testing automated driving system (ADS)-equipped vehicles that, at a minimum, details a plan to manage the risk associated with crashes and operator inattentiveness and establishes countermeasures to prevent crashes or mitigate crash severity within the ADS testing parameters.
4. Establish a task group of experts to evaluate applications for testing vehicles equipped with automated driving systems, as described in Safety Recommendation 3, before granting a testing permit.

To the American Association of Motor Vehicle Administrators:

5. Inform the states about the circumstances of the Tempe crash and encourage them to (1) require developers to submit an application for testing automated driving system (ADS)-equipped vehicles that, at a minimum, details a plan to manage the risk associated with crashes and operator inattentiveness and establishes countermeasures to prevent crashes or mitigate crash severity within the ADS testing parameters, and (2) establish a task group of experts to evaluate the application before granting a testing permit.

To the Uber Technologies, Inc., Advanced Technologies Group:

6. Complete the implementation of a safety management system for automated driving system testing that, at a minimum, includes safety policy, safety risk management, safety assurance, and safety promotion.²⁶

26. <https://www.ntsb.gov/news/events/Documents/2019-HWY18MH010-BMG-abstract.pdf>

Box 8: Investment management

In 2017, Hong Kong-based Li Kin-kan let an AI-led system manage \$250 mn of his own cash and additional leverage from Citigroup Inc, totalling up to \$2.5 billion. The AI system was managed by London-based Tyndaris Investments. The system was developed by an Austria-based company. It works by scanning through online sources like real-time news and social media and makes predictions on US stocks.

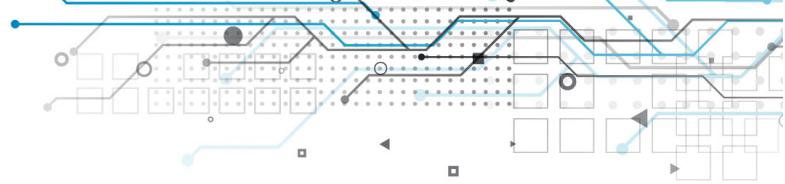
By 2018, the system was regularly losing money, including over \$20 mn in a single day. The investor decided to sue Tyndaris Investments for allegedly exaggerating the AI's capabilities.²⁷ In this case, assigning liability- across developers, marketers and users of the system- is not straight-forward.

The existing legal systems allocate responsibilities for action and consequences assuming a human agent. While some legislations and protocols dealing with the regulation of technology and data are focused more on ensuring that accountability is built into the system, providing accountability of a remedial nature for AI systems is not easy.

Systems Consideration 6: Privacy risks

AI systems rely on large amounts of training data and when an individual's personal data is used there are bound to be considerable privacy concerns. Lack of adequate privacy safeguards may permit technology to wholly record and analyse an individual's personal life without their consent or knowledge, significantly harming an individual interest by disregarding their preferences on the use of data. Such harm may be economic – stealing an individual's credit card information; or emotional – where an individual's personal details is the subject of public discussion. There are also examples of the impact on democratic institutions which have been examined under the 'Societal considerations' section.

27. <https://www.bloombergquint.com/technology/who-to-sue-when-a-robot-loses-your-fortune>



| The Issue | Its Implications |
|---|--|
| <ul style="list-style-type: none"> AI is highly reliant on data for training, including information that may be personal and/or sensitive (PII), giving rise to: Risk that entities may use personal data without the explicit consent of concerned persons; Possible to discern potentially sensitive information from the outputs of the system; | <ul style="list-style-type: none"> Infringement of Right to Privacy |

Box 9: Privacy and AI

Using facial recognition technology for surveillance: Clearview AI, a start-up based in the USA, gained attention around the world following a *New York Times* article. The company trained AI models to recognise people from over 3 billion images scraped from social media platforms and provided the technology to law enforcement agencies. The company website states that their mission is to **provide assistance to law enforcement agencies** for the **identification of criminals**. In many cases, law enforcement agencies have **cited success** in using the software **to identify criminals—despite having limited knowledge of how it works**²⁸. Several civil groups in the US have raised objections particularly around the known vulnerabilities in facial recognition technology and possibility of mass surveillance for malicious purposes.²⁹ Social media sites from which data was scraped also issued '**cease-and-desist**' notice against the company for violating the terms and policies for data use.^{30,31}

Ri -
to
Privacy

Model inversion: Machine learning models may sometimes use sensitive and personal information for training and the model itself may be available for general use. However, research has shown ways in which the training data can be inferred from the trained model. This is particularly relevant for 'ML-as-a-service' models that are trained on a large number

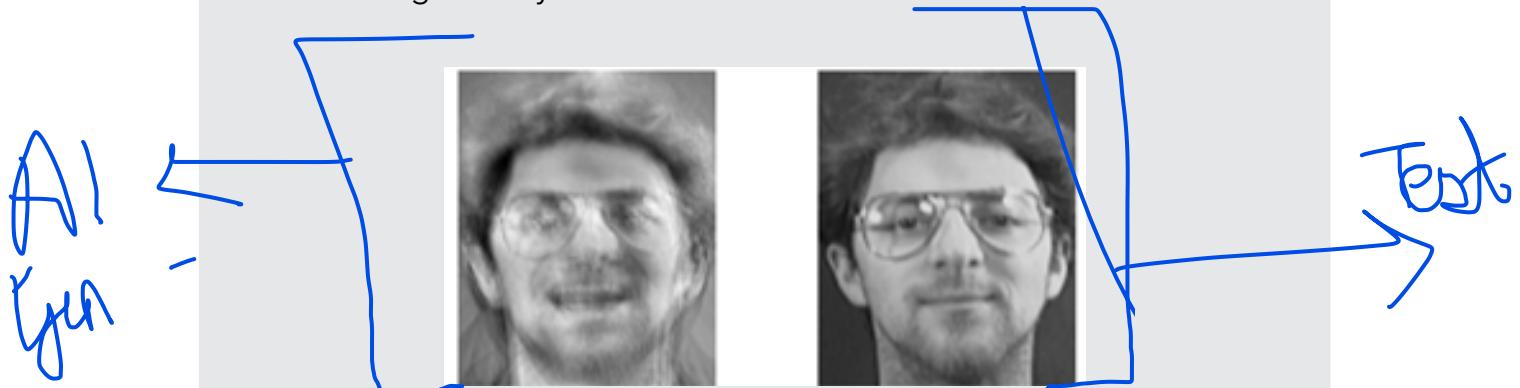
28. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>

29. <https://aboutblaw.com/Oqa>

30. <https://www.bbc.com/news/technology-51220654>

31. <https://www.cnet.com/news/clearview-ai-hit-with-cease-and-desist-from-google-over-facial-recognition-collection/>

of potentially personal and sensitive datasets. This technique has been demonstrated to be able to infer personal and sensitive information from non-personal data- sensitive genomic information was identified from patient demographic and dosage data. Fredrikson et al. (2015) also demonstrated the ability to extract dataset images that were used to train a facial recognition system.^{32,33}



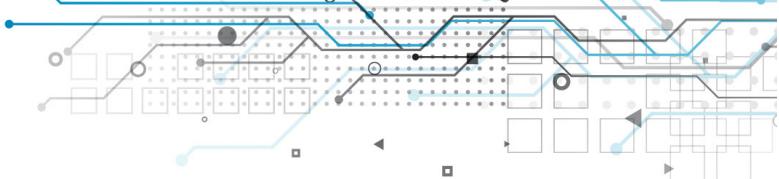
Left: Generated image from the AI model. **Right:** Image used for training the model

Membership inference attack: Shokri et al (2017) proposed *membership inference attack* through which it is possible to know if a particular dataset was used for training a model. This is possible even when the model architecture and parameters are not known. This can lead to a privacy breach as, for example, knowing that a person's clinical record was used for training a diagnostic model could mean that the person had the disease.³⁴

Systems Consideration 7: Security risks

Security risks in AI systems arise from its reliance on data and from its design and deployment environment. Some of these attacks are unique to machine learning systems and affect different parts of the machine learning development cycle. Adversarial machine learning attacks are designed to take advantage of vulnerabilities in the machine learning model with potentially harmful real-world consequences.

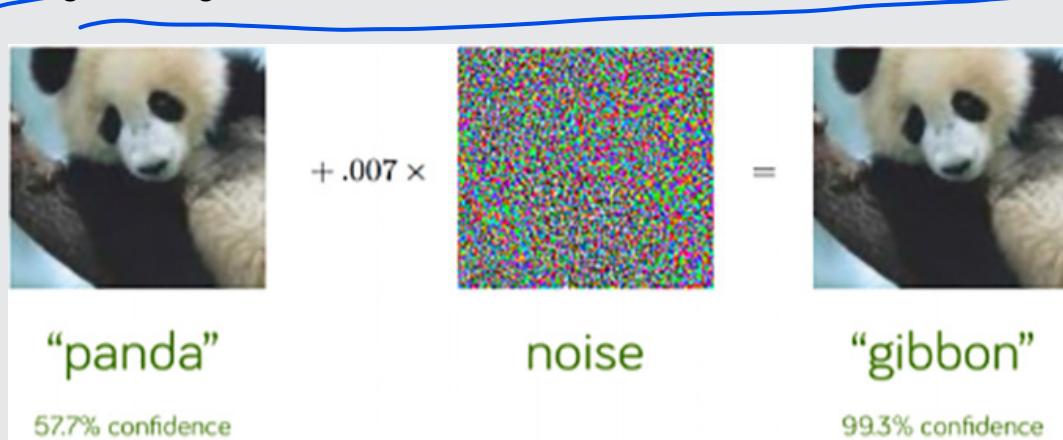
32. Fredrikson, Matt, et al. "Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures." *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, doi:10.1145/2810103.2813677.
33. M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In USENIX Security Symposium, pages 17–32, 2014
34. R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership Inference Attacks Against Machine Learning Models," 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, 2017, pp. 3-18, doi: 10.1109/SP.2017.41



| The Issue | Its Implications |
|--|---|
| <ul style="list-style-type: none"> AI systems are susceptible to attack such as manipulation of data being used to train the AI, manipulation of system to respond incorrectly to specific inputs, etc; Given some AI systems are 'black boxes', the issue is made worse | <ul style="list-style-type: none"> In real world deployments, may lead to malfunctioning of system; Risk to IP protection due to potential of 'model steal' attacks |

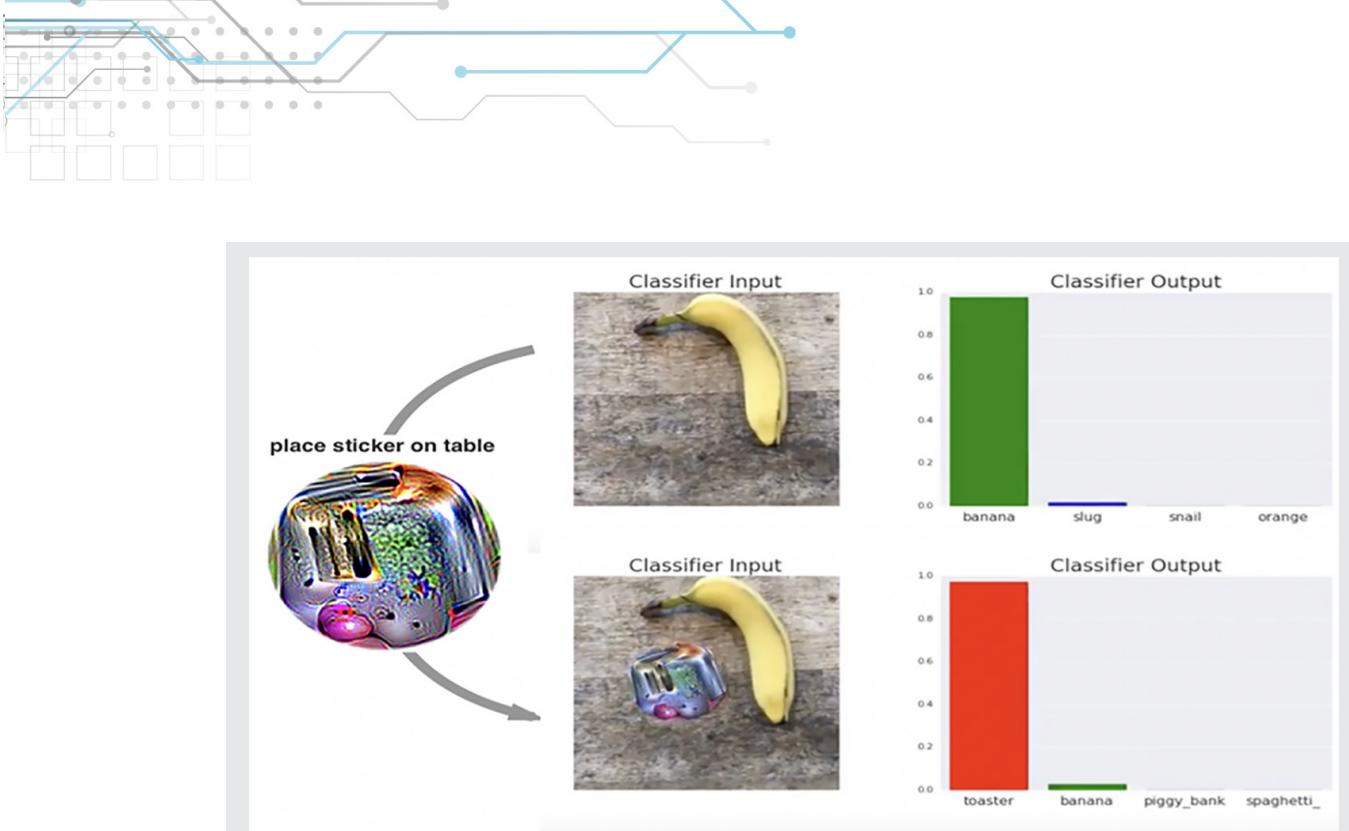
Box 10: Adversarial attack

Adversarial attacks affect the output of an AI system by introducing 'perturbations' to the input data. Researchers have found, for example, that by carefully manipulating even less than one percent of an image, it was possible to cause the model to make mistakes in classification²⁶. In another example, researchers have demonstrated how a patch, known as 'adversarial patch' may be generated and placed anywhere in the input image leading to misclassification²⁷.



The image of a Panda is correctly classified with 57.7% confidence. By adding a small noise to the image, the image is classified as a Gibbon with 99.3% confidence³⁵

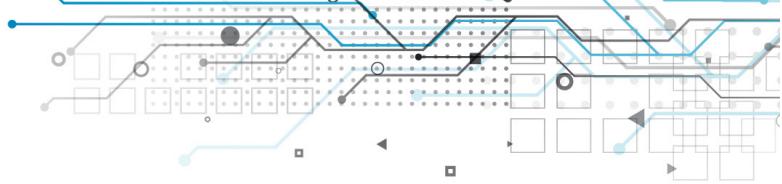
35. <https://arxiv.org/abs/1412.6572>



When the top image is presented to the classifier, it reports 'banana' with 97% confidence. When the image is accompanied by the 'adversarial patch', the classifier reports 'toaster' with 99% confidence³⁶

The attacks on the AI system may render the purpose of the system redundant and, in some cases, have the potential to be dangerous. Such attacks on autonomous vehicles may lead to accidents. In reinforcement learning systems, such attacks can lead to reduced performance or make it behave in an unintended manner.³⁷

36. <https://arxiv.org/pdf/1712.09665.pdf>
 37. <https://openai.com/blog/adversarial-example-research/>



Societal Considerations

Malicious Use of AI

The Cambridge Analytica scandal that broke out in 2018 is a quintessential example of the real-world consequence of privacy breach and the impact of psychological profiling. The data from millions of users was used without their consent, to sway public opinion on matters of national and political interest around the world. This was facilitated through a Facebook app called '*This is your Digital Life*' that paid users to take a psychological survey. Users logged in through Facebook and the survey responses were captured along with the user's likes and profile information. In addition to this, the app also pulled information on the user's Facebook friends. The allegation was that the data was used to create psychological profiles of users by corresponding answers to the survey with Facebook profile information. This profiling was used to target political campaign messages. While this gained media attention for its role in the US Presidential elections, subsequently, its involvement in other countries was revealed. This included its role in the Brexit campaign, elections in Kenya, Thailand, Indonesia and its role in Indian elections.

This episode was a watershed moment for data protection around the world. Facebook confirmed that, though only 2,70,000 users had provided consent and downloaded the app, by tapping into the user's friends network, information of up to 87 million users was used. Of these, 5,62,455 users were from India³⁸.

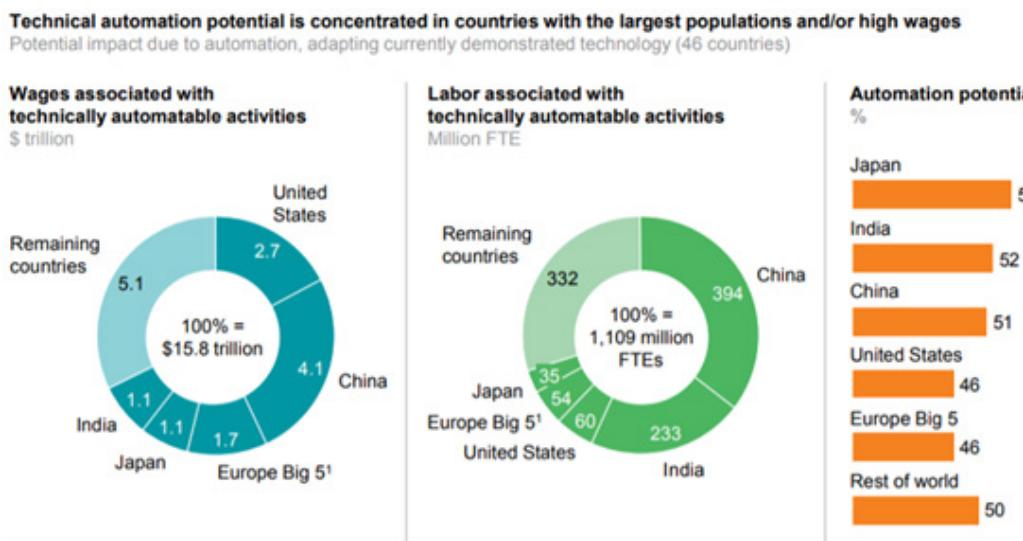
It also highlighted the role of AI in enabling profiling and ease of spreading targeted messages. It was alleged that the psychological profiling further helped in transmitting fake news to the susceptible population and was used as a 'propaganda machine'. This was seen as a violation of the fundamental user choice and democratic process around the world.

Impact on Jobs

The rapid rise of AI has led to the automation of a number of routine jobs.

38. <https://about.fb.com/news/2018/04/restricting-data-access/>

A report by the Oxford Economic Forecast³⁹ indicates a high potential for automation of tasks performed by the Indian workforce.



¹ Pakistan, Bangladesh, Vietnam, and Iran are largest countries by population not included.

² France, Germany, Italy, Spain, and the United Kingdom.

NOTE: Numbers may not sum due to rounding.

SOURCE: Oxford Economic Forecasts; Emsi database; US Bureau of Labor Statistics; McKinsey Global Institute analysis

A report by NASSCOM notes that automation has been heavily tested and implemented during the pandemic.⁴⁰ Frequent newspaper reports stress the snowballing adoption of robotic devices in manufacturing processes.

This is an evolving area requiring more research, for the immediate next steps it is proposed to:

- study the on-ground impact on job automation more rigorously, track the changes in job landscape and develop targeted policies;
- build human capacity to adapt to the changing landscape through the introduction of incentives and programs for lifelong learning and relevant reforms to education and skilling;
- with the changing job landscape recognise and safeguard the interests of citizens under new job roles, such as gig workers;
- have a long-term strategy to harvest the economic potential of AI. The National Strategy for AI (2018) identifies the need to invest in research, adapting skilling programs for the AI age, and accelerating adoption.

39. <https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx>

40. <https://nasscom.in/knowledge-center/publications/covid-19-tipping-point-automation>

Legal and Regulatory Approaches for Managing AI Systems





Legal and Regulatory Approaches for Managing AI Systems

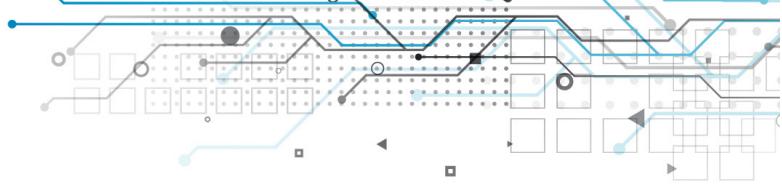
Malicious Use of AI

The previous sections highlighted a number of ways in which AI systems may impact the rights of citizens. It must be noted, however, that not all of these concerns are new relating to an emerging technology but already exist in various forms. In certain high-risk sectors such as health and finance, various sector specific legal protections and guidelines for products and services exist. However, a simple applicability of these laws to AI-based decision-making processes may not be appropriate. For specific aspects of algorithmic decisions, new legal protections may be needed. For example, while no anti-discrimination law directly regulates decision making by AI, the extant laws are equally silent about the means of decision-making that they do govern⁴¹. Therefore, it will fall within the jurisdiction of anti-discrimination legislation to regulate decisions arrived at through the use of AI as well, particularly when the decision-making AI is being used by an entity having constitutional or legal obligations to be unbiased.

In the study of regulatory and legal approaches, it is important to identify the specific role legislation may play. The greatest risk of adopting this approach to manage AI systems is that regulations have historically not kept pace with technology. AI is still an evolving field and the risks are not well understood, making it difficult to design concrete long term regulatory approaches. Regulating AI is a complex topic and there are diverse views regarding what degree and what forms of regulation will be effective for its varied applications. AI is a rapidly advancing technology, and a one size fits all approach may not be the most suitable approach. There is a need to balance soft governance measures with regulation depending on the use case and risks involved.⁴² While

41. "Responsible AI: A Global Policy Framework". ITechLaw. Available at <https://www.itechlaw.org/ResponsibleAI/> access.

42. <https://www.weforum.org/whitepapers/ai-governance-a-holistic-approach-to-implement-ethics-into-ai>



overarching AI ethics principles will guide the overall design, development and deployment of AI in the country, a graded risk-based approach to varying use cases across different sectors need to be adopted.

At the same time, the AI ecosystem has multiple stakeholders- private sector, research, government, legal bodies, regulators, standard setting bodies, etc. It is important to bring in a common understanding on acceptable behaviour among different stakeholders and clarify applicability of existing policies and regulations through creation of Principles and guidance framework. Principles offer a technology agnostic framework for communicating expectations from responsible AI systems and identifying governance mechanisms.

Some relevant legislation for protection from AI related concerns exist in certain cases but would need to be adapted to cater to challenges posed by AI. Some sectors have unique considerations that may require sector specific laws for AI. Moreover, the review of AI ethics principles and guidelines will need to be done on an ongoing basis given the rapid pace of development of this emerging technology.

Global Approaches

Around the world, countries have identified a broad framework through Principles and other guidance documents to guide the design, development and use of AI systems. *Ethics Guidelines for Trustworthy AI* released by the High Level Expert Group in the European Union is a non-binding document that proposes a set of 7 key requirements that AI systems should meet in order to be deemed 'trustworthy'.⁴³ Along similar lines, Singapore has a *Model AI Governance Framework*⁴⁴ and the United States of America has *Principles for the Stewardship of AI Applications*.⁴⁵

In addition to the overall guidance framework, specific actions have been identified in high risk sectors to guide their development and adoption. These are also typically non-binding and ensure that sector-specific issues are considered. The 'FEAT Principles' for AI in financial services, released by Monetary Authority of Singapore (MAS) serves as a non-prescriptive guidance document to encourage adoption of fair, explainable, ethical, and accountable

43. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>

44. <https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework>

45. <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>

AI.⁴⁶ European Union has identified certain sectors as high risk and suggests an oversight mechanism.

Binding regulations and acts of the Parliament are generally reserved for aspects that have been well understood. Globally, such instruments mostly cover data protection and are not restricted to just AI systems. The Personal Data Protection Act (PDPA) 2012 released by the Personal Data Protection Committee (PDPC) in Singapore establishes a data protection law that comprises various rules governing the collection, use, disclosure and care of personal data. General Data Protection Rules (GDPR) 2016, in the EU is a regulatory framework for protection of personal data and establishes the need for 'privacy by design' when developing automated solutions. In the USA, the Algorithmic Accountability Act of 2019 is a proposed bill that *requires specified commercial entities to conduct assessments of high-risk systems that involve personal information or make automated decisions, such as systems that use artificial intelligence or machine learning.*⁴⁷ The USA also has the HIPAA Privacy Rule (2000) and Graham Leech Bliley Act (1999) for the governance of data in healthcare and finance respectively.

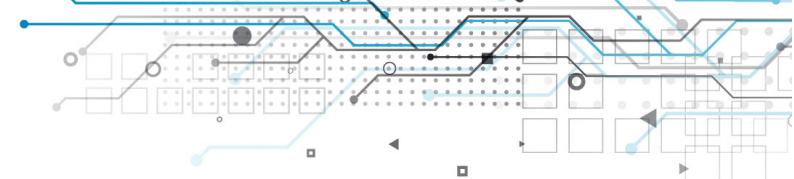
Status in India

Currently, India does not have an overarching guidance framework for the use of AI systems. Establishing such a framework would be crucial for providing guidance to various stakeholders in responsible management of Artificial Intelligence in India.

There are certain sector specific frameworks that have been identified for development and use of AI. In finance, SEBI issued a circular in Jan 2019 to Stock Brokers, Depository Participants, Recognized Stock Exchanges and Depositories and in May 2019 to All Mutual Funds (MFs)/ Asset Management companies (AMCs)/ Trustee Companies/ Board of Trustees of Mutual Funds/ Association of Mutual Funds in India (AMFI) on reporting requirements for *Artificial Intelligence (AI) and Machine Learning (ML) applications and systems offered and used.* The reporting is towards creating an inventory of AI systems in

46. <https://www.mas.gov.sg/~media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf>

47. <https://www.congress.gov/bill/116th-congress/house-bill/2231/all-info#:~:text=Official%20Title%20as%20Introduced,and%20data%20protection%20impact%20assessments>.



the market and guide future policies.^{48,49} The strategy for National Digital Health Mission (NDHM) identifies the need for creation of guidance and standards to ensure reliability of AI systems in health.⁵⁰ The Data Empowerment and Protection Architecture (DEPA) by NITI Aayog presents a technical framework for people to retain control of their personal data, and the means to leverage it to avail services and benefits.⁵¹

India currently does not have overarching legislation specific to AI. The closest to this is the draft Personal Data Protection Bill (2019) (**PDP**) designed as comprehensive legislation outlining various facets of privacy protections that AI solutions need to comply with. It covers limitations on data processing, security safeguards to protect against data breaches and the provision of special provisions relating to vulnerable users such as children. Additionally, the PDP Bill provides for a vibrant data protection legislation where the law shall be supplemented with regulations and codes of practice, thereby making it easier for privacy to evolve with evolving technologies. For example, if a certain aspect of privacy regarding AI requires clarity, the Authority may simply issue a code of practice to provide the same. As of writing of this paper, the PDP bill is yet to be passed.

The **Information Technology Act, 2000 (IT Act)** is the backbone of data protection legislation in India. The provisions of the IT Act, combined with the Information Technology (Reasonable security practices and procedures and sensitive personal data or information) Rules, 2011 (SPDI Rules) establish a technology-agnostic regime for the protection of sensitive personal information for all bodies corporate.

48. https://www.sebi.gov.in/legal/circulars/jan-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-market-intermediaries_41546.html

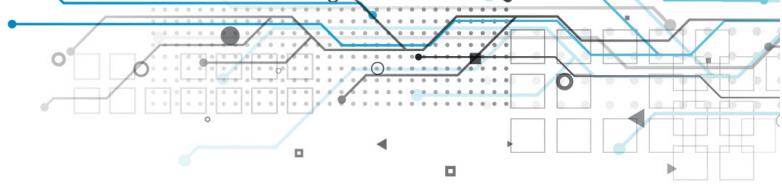
49. https://www.sebi.gov.in/legal/circulars/may-2019/reporting-for-artificial-intelligence-ai-and-machine-learning-ml-applications-and-systems-offered-and-used-by-mutual-funds_42932.html

50. https://ndhm.gov.in/assets/uploads/NDHM_Strategy_Overview.pdf

51. https://niti.gov.in/sites/default/files/2020-09/DEPA-Book_0.pdf

Technology Based Approach for Managing AI Systems

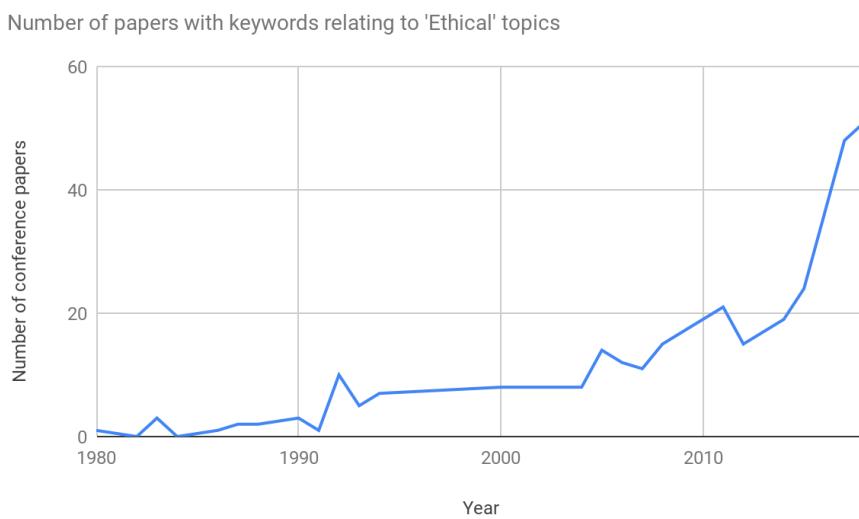




Technology Based Approach for Managing AI Systems

As discussed in the previous section, guidance framework or Principles may serve to set expectations on responsible management of AI systems. It is also important for technology to respond to these expectations. The National Strategy for Artificial Intelligence (2018) advocates for leveraging technology to manage AI systems responsibly. Technology has the potential to be agile and respond to evolving requirements.

There has been a growing interest in using technology and statistical methods to address AI-related considerations—increasing not only the body of research in the field but also promoting a sense of responsibility amongst solution developers in academia and industry. Figure 1 shows the increase in the number of papers on ‘ethical’ topics is on the rise in AI, robotics and Computer Science related conferences in the last decade. The advances in this field are nascent, and must be promoted to keep pace with the general growth in some of the classical and trending topics in AI. This is the area where countries can collaborate to fund research.



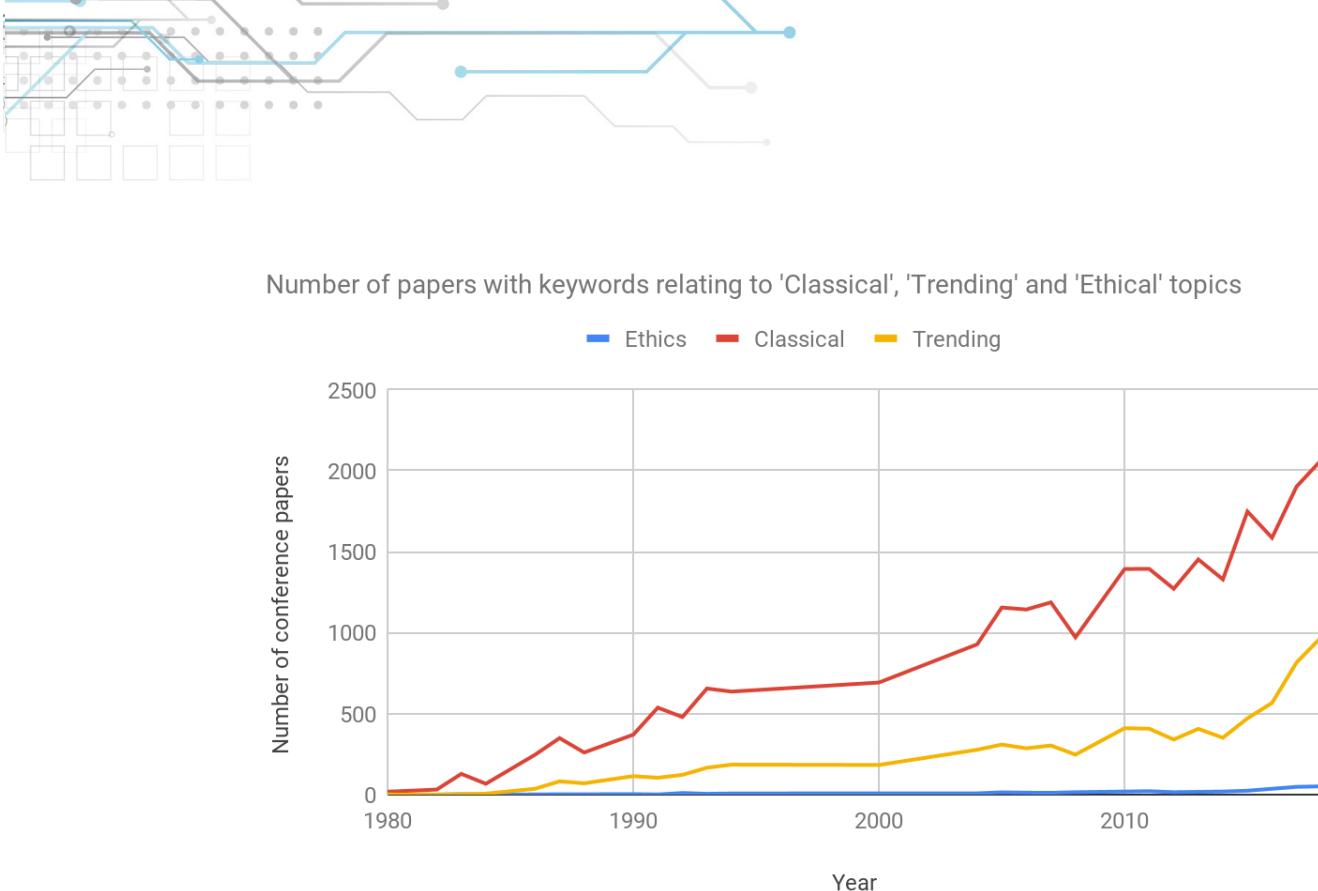


Figure 1: Left: Number of conference papers on ethical topics has steadily seen an increase over the past decade. Right: The increase has not kept pace with the developments in other areas of AI. Source: Prates (2018)⁵² and Stanford AI Index 2019⁵³

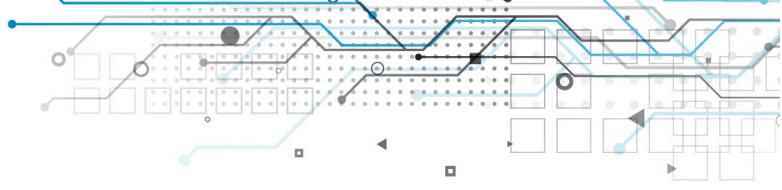
Private sector, Academic Institutes, Government organisations and International bodies around the world have contributed to research and development of technology tools to manage AI systems responsibly. Defence Advanced Research Projects Agency (DARPA) has dedicated programs on Explainable AI (XAI), Guaranteeing AI Robustness against Deception (GARD), Understanding Group Biases (UGB), and Machine Common Sense (MCS). Global Partnership on Artificial Intelligence (GPAI) has a working group on responsible AI. The World Economic Forum has launched the [Global AI Action Alliance to accelerate the adoption of trusted, transparent and inclusive AI globally and across sectors](#)⁵⁴. Google, Microsoft and IBM have also released open-source toolkits to understand [bias in datasets](#) and the ML model. LIME and SHAP, developed at research institutions and used to explain individual decisions through input attribution, are also available as open source libraries. In general, open sourcing of tools and techniques has increased both the development and adoption.

Most of the techniques mentioned above have evolved over the last decade. Technologies to manage privacy, such as differential privacy and zero knowledge

52. <https://arxiv.org/pdf/1809.08328.pdf>

53. https://hai.stanford.edu/sites/default/files/ai_index_2019_report.pdf

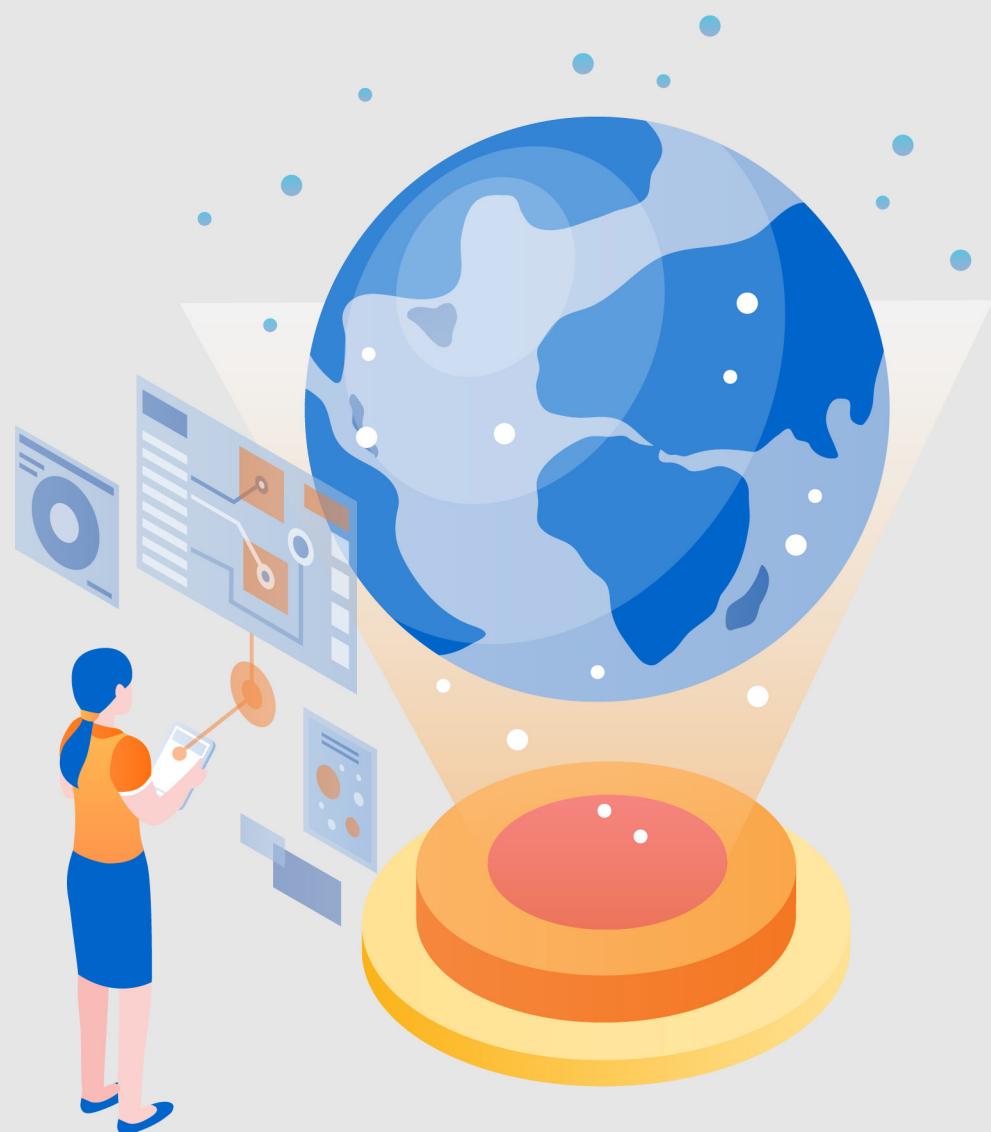
54. <https://www.weforum.org/projects/global-ai-action-alliance>

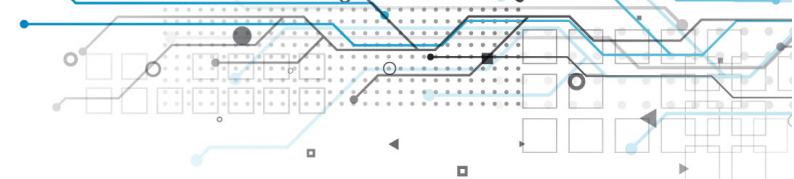


proofs, have a longer history and are being adapted to machine learning. The rise in computing power and storage capacity, coupled with lowering cost has enabled novel techniques such as federated learning.

There has been significant progress in technology approaches to managing AI responsibly and this must be encouraged. The National Strategy for Artificial Intelligence (2018) highlighted the need for collaborative research in Responsible AI. The Government may consider identifying relevant areas for research in responsible AI tools and techniques and incentivise creation and adoption.

Principles for Responsible Management of AI Systems





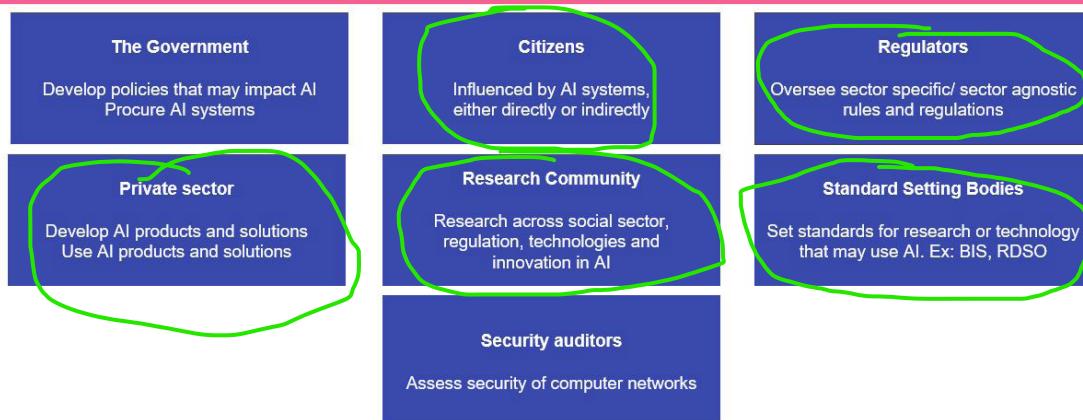
Principles for Responsible Management of AI Systems

The previous section identifies the need for guiding principles for responsible management of AI in India. The principles are expected to serve as a guide for various stakeholders in the AI ecosystem. These principles, to be effective, must be grounded on the nation's accepted value system and compatible with International standards.

Different stakeholders of the AI ecosystem shape the future of AI, and it is essential to have a common set of principles which can guide all stakeholder groups towards responsible AI.

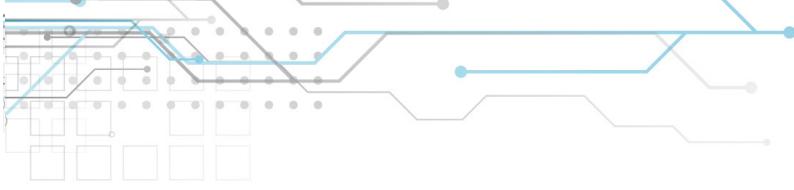
Box 11: AI ecosystem stakeholder groups

Different Groups shape the future of AI



A common set of principles across these entities helps ensure AI is used beneficially

The Supreme Court of India (**Supreme Court**) in cases such as *Naz Foundation* and *Navtej Johar*, has defined the prevailing morality of our country to be based on the principle of Constitutional morality. The Supreme Court has stressed time and again on adherence to constitutional morality over social morality,



with the former's reach extending beyond the mere text of the Constitution to encompassing the values of a diverse and inclusive society while remaining faithful to other constitutional principles. Constitutional morality has been described as the basis on which the rights of minorities can be upheld in the face of majoritarianism, and is to be followed over societal morality, especially when the latter infringes the basic rights guaranteed by the Constitution of India (**Constitution**).

Box 12: Considerations in the context of Constitution of India

The considerations mentioned in the previous section also find expression in the Constitution under Fundamental Rights. The relevant articles are summarized below,

Article 14: Right to Equality

The Constitution guarantees equal treatment of equally placed persons and groups before the law, and equal protection of the law to all.

Articles 15 & 16: Right against Discrimination

The Constitution prohibits discrimination on the basis of religion, race, caste, sex, descent, place of birth or residence in matters of education, employment, access to public spaces, etc.

While the Constitution prohibits discrimination based on certain markers, it also provides for positive discrimination in the form of affirmative action.

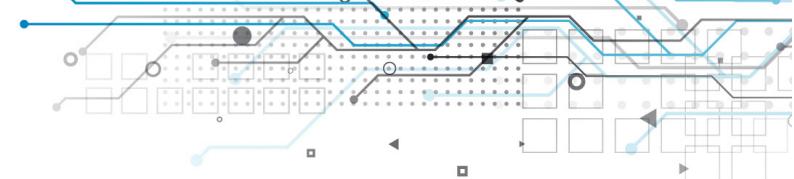
Article 15, while prohibiting discrimination, empowers the Government to make special provision for the advancement of any socially and educationally backward classes of citizens or for the Scheduled Castes and the Scheduled Tribes, and to make provisions for their admission to educational institutions, whether private, aided or unaided.

Article 21: Right to Life and Healthcare

The Constitution guarantees the right to life to all persons. Various High Courts have read the right to healthcare, including the right to avail health insurance, to be part of the right to life.

Article 21: Right to Privacy

The Supreme Court has held that the right to privacy is an intrinsic part of



the right to life and liberty guaranteed under Article 21 of the Constitution and as part of the freedoms enshrined in Part III thereof.

Article 38: State Directive for Economic Equality

The Constitution directs the State to ensure economic welfare of the people and minimise inequalities in income, status, facilities and opportunities, both between individuals and between groups of people. The State is also directed to ensure a living wage for all workers, including agricultural workers.

Transparency and accountability

The Supreme Court, in its interpretation of the Constitution, has held that transparency in decision making is critical even for private institutions. The Constitution guarantees accountability of all State action to individuals and groups.

Box 13: Creation of Principles

Principles were developed after consultation with diverse set of stakeholders

AI case studies in India and around the world



Instances of harm caused by AI systems around the world were studied to identify relevant considerations in Indian context

Rights according to the Indian Constitution



Supreme court, in various instances, has defined the prevailing morality of India to be based on the principle of Constitutional morality. Principles thus flow from the constitution and all laws enacted thereunder

International standards for AI

Various International bodies such as GPAI, UNESCO, IEEE have developed standards for AI. For effective global collaboration on AI, it is important for India's principles to be **compatible with relevant international standards**

Ethics is an emerging field and should be an ongoing research

The considerations were identified following several expert interviews and multi-stakeholder workshops with experts from India and globally across public and private sectors, start-ups, academia and civil society. The Principles may be derived from the Constitution and all laws enacted thereunder. The following Principles are recommended for the responsible management of artificial intelligence in India and are based on the underlying principle of ensuring AI systems are designed in a manner that enables fundamental rights:

- 
- 1. Principle of Safety and Reliability:** AI should be deployed reliably as intended and sufficient safeguards must be placed to ensure the safety of relevant stakeholders. Risks to all stakeholders should be minimized and appropriate grievance redressal, care and compensation structures should be in place, in case of any unintended or unexpected harm. The AI system needs to be monitored through its lifecycle so it performs in an acceptable manner, reliably, according to the desired goals.
 - 2. Principle of Equality:** AI systems must treat individuals under same circumstances relevant to the decision equally
 - 3. Principle of Inclusivity and Non-discrimination:** AI systems should not deny opportunity to a qualified person on the basis of their identity. It should not deepen the harmful historic and social divisions based on religion, race, caste, sex, descent, place of birth or residence in matters of education, employment, access to public spaces, etc. It should also strive to ensure that unfair exclusion of services or benefits does not happen. In case of an adverse decision, appropriate grievance redressal mechanism should be designed in a manner affordable and accessible to everyone irrespective of their background.
 - 4. Principle of Privacy and Security:** AI should maintain privacy and security of data of individuals or entities that is used for training the system. Access should be provided only to those authorized with sufficient safeguards.
 - 5. Principle of Transparency:** The design and functioning of the AI system should be recorded and made available for external scrutiny and audit to the extent possible to ensure the deployment is fair, honest, impartial and guarantees accountability.
 - 6. Principle of Accountability:** All stakeholders involved in the design, development and deployment of the AI system must be responsible for their actions. Stakeholders should conduct risk and impact assessments to evaluate direct and indirect potential impact of AI systems on end-users, set up an auditing process (internal and if required external) to oversee adherence to principles and create mechanisms for grievance redressal in case of any adverse impact.
- Actions of AI Model*

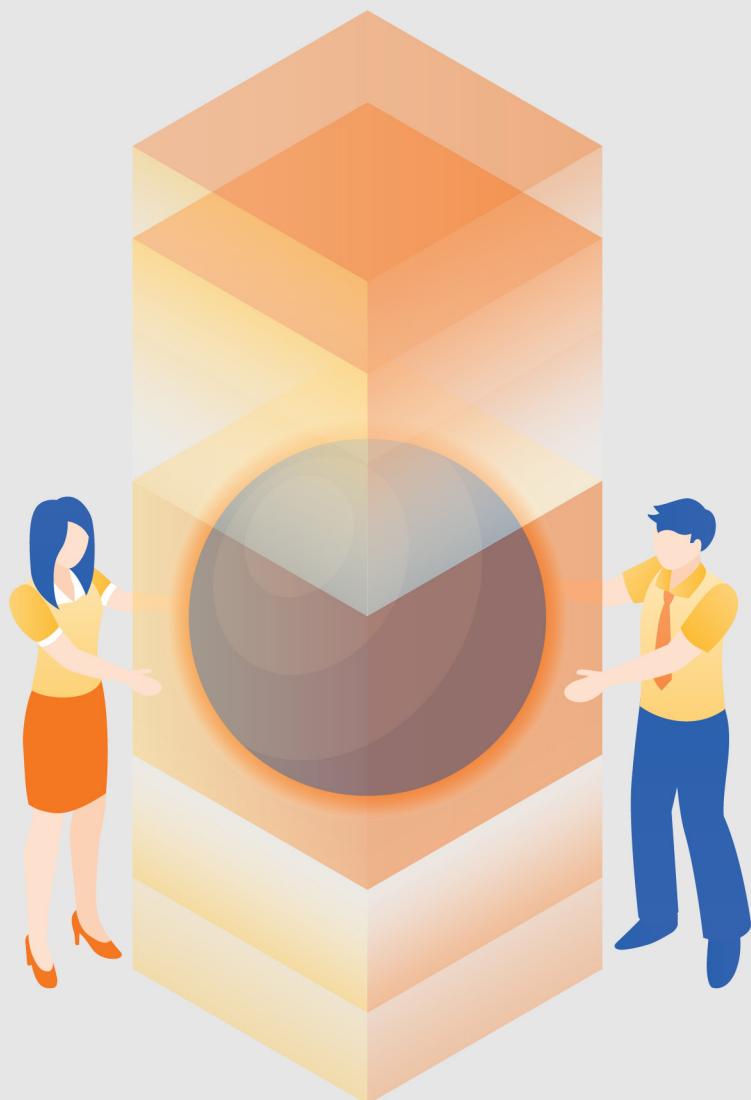


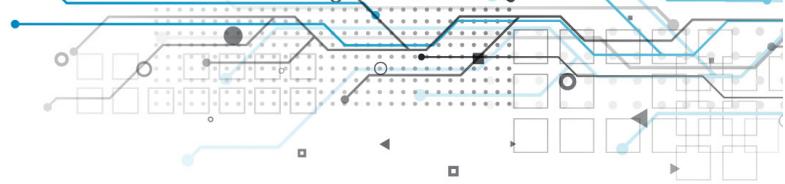
7. Principle of protection and reinforcement of positive human values:

AI should promote positive human values and not disturb in any way social harmony in community relationships

It is important to ensure that these Principles are updated in the future to reflect the latest knowledge, innovation and technology advances. A mechanism for the same and a framework for the enforcement of these Principles will be explored in Part 2 of the paper. Subsequent versions will also explore specific policy interventions for Responsible AI.

Appendix





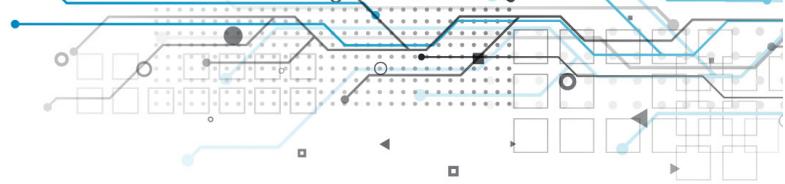
Appendix 1

Self-Assessment Guide for AI Usage

Stakeholders who design, develop, procure, deploy, operate, maintain AI systems may use this guide to assess the various considerations and potential mitigation approaches across the system's life cycle. This is not an exhaustive list and requires an ongoing update with latest advances. It is only intended to serve as a guide to help assess the AI governance readiness of stakeholders as per the Responsible AI principles in this document.

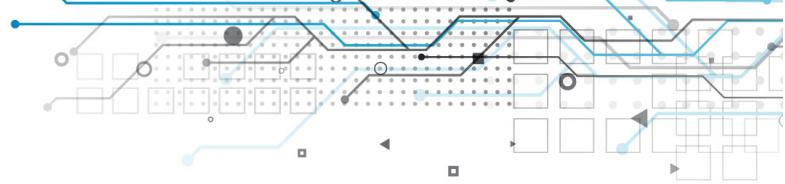
| Problem Definition and Scoping | |
|---|---|
| Consideration | Mitigation Strategy |
| Have you assessed the potential 'degree of harm' from the AI system being deployed in the short term and the long term? | Constitute an ethical committee consisting of sector experts, social scientists, data and <u>other relevant experts</u> to assess the potential degree of harm due to development and deployment of the AI system |
| | The group may recommend guidelines to follow to ensure social risks are appropriately managed |
| | Document the concerns identified and plan appropriate measures and incentive mechanisms to mitigate them |
| Is there an appropriate grievance redressal mechanism for stakeholders who may be impacted by the AI system? | Establish a grievance mechanism for anyone impacted by the AI system |
| | Document the measures taken to make stakeholders aware of the grievance redressal mechanism through appropriate channels |
| | The support mechanism should be easily accessible, ideally at no additional cost |

| | |
|---|--|
| Have you engaged with the stakeholders to understand the degree of explainability that may be required for individual decisions? | <p>Identify stakeholders for the AI system and specify their profiles and needs in the procurement document</p> |
| | <p>Engage with the stakeholders to understand the purpose and the degree of explanation that may be required</p> |
| | <p>In the procurement document, specify the role and responsibility of the vendor</p> |
| | <p>Specify individual roles and responsibilities and, if a part of the work is subcontracted, identify roles and responsibilities of each agency</p> |
| Have you identified mechanism to handle errors in decision by the AI system? | <p>Specify who has the right to make changes to the system during development, launch and post launch stage to manage any social risks</p> |
| | <p>If the potential degree of harm for a decision is expected to be high, have appropriate mechanisms in place so stakeholders can contest and humans can get involved in the decision making process</p> |
| | <p>Ensure the terms of service allows for audit by research institutions, audit agencies to probe the system, identify bias, risks and review behaviour</p> |
| Can the terms of use allow for public auditing to understand behaviour and identify risks without causing unintended consequences | <p>Consult legal counsel and data science experts to identify means to allow for audit without exposing sensitive and personal information or causing any other unintended harm to the system and its stakeholders</p> |
| | <p>If data must be made available for audit, ensure steps are taken to expose the data in a manner that preserves privacy, security and protects legal rights of associated people or businesses</p> |



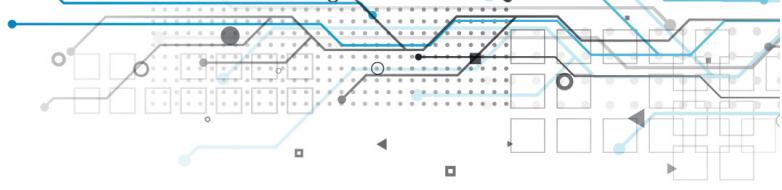
| Has goals with respect to Equality, Non-Discrimination and Inclusion been defined? | Define the fairness goals with respect to the use case, in terms of the social cost of inclusion/exclusion, in the procurement document. These goals may be defined by the ethical committee |
|--|--|
| Data Collection | |
| Consideration | Mitigation Strategy |
| Ensuring various laws regarding data collection are adhered | Identify laws, regulations and any other guidelines that may apply to the use case and specify it in the procurement document |
| | Document all known data points that are used for training the system. Ensure only the parameters required for training are used |
| | Identify and document data parameters that are personal and/or sensitive |
| | Create and document a process to continually scan for and identify new sources of personal and/or sensitive data |
| Have appropriate measures been taken to protect privacy | Document who has access to personal and sensitive data and have a SOP for when employees leave |
| | Consult with experts to identify risks where personal and/or sensitive data can be inferred, for eg: by combining either internal or external datasets |
| | If personal and/or sensitive data must be used, identify ways to mask the data using best encryption practices or 'coarsen the data resolution' so individuals cannot be identified |

| Have measures been taken to ensure the dataset is a fair reflection of real-world use cases and frequencies? | <p>Assess your datasets to ensure representativeness, skews, and correlations in features and labels</p> <p>If a continuous training process is used, track and analyse ongoing representativeness of the data and document strategies and techniques to mitigate negative experiences and discriminatory outcomes</p> |
|---|--|
| Data Labelling | |
| Consideration | Mitigation Strategy |
| Have various bias-inducing factors during labelling, such as human variability, accessibility, memory and inherent bias been accounted for? | <p>Document the data annotation process and any bias mitigation strategy employed</p> <p>Understand the variability among annotators through a set of standardized tests</p> <p>Design clear tasks, incentive structures and feedback mechanisms to ensure accurate and unbiased annotation</p> <p>Ensure diversity within annotators and the kappa score for rate-reliability may be recorded</p> |
| Model selection | |
| Consideration | Mitigation Strategy |
| Does the model selection satisfy explainability requirements of the system? | <p>If explanation is crucial to the system and accuracy requirements can be met, use explainable AI models</p> <p>If explainable AI models cannot be used, document a strategy for both decision-process-summarization and individual decision explainability</p> <p>In this regard, techniques such as input attribution, example influence matching, concept extraction, distillation techniques may be considered</p> |



| | |
|--|---|
| Does the model parameters reflect non-discrimination and inclusion goals set for the use case? | Identify metrics to ensure non-discrimination and inclusion goals are tracked |
| | Document the measures taken to ensure the algorithm, objective function and thresholds reflect the non-discrimination and inclusion goals of the system |
| Consideration | Mitigation Strategy |
| Have security considerations been taken into account during training? | <p>Ensure the model does not 'memorize' sensitive/personal data during training</p> <p>Techniques such as Zero Knowledge protocol, edge computing, federated learning may be considered for additional protection</p> |
| Evaluation | |
| Consideration | Mitigation Strategy |
| Is the system adequately evaluated for bias? | Organize a diverse focus user group of testers from diverse background for adversarial testing of the system |
| | Calculate and document error rates for different sub-population groups and evaluate if the performance is in line with fairness goals set for the system |
| | Stress test the system for particularly difficult cases and ensure the performance for each sub-population groups is documented |
| | Identify situations where the AI system may be error prone and develop mechanisms- such as alert and human intervention- to ensure stakeholders in such situations are not harmed |

| Deployment | |
|---|--|
| Consideration | Mitigation Strategy |
| Has the performance been reviewed by the ethical committee and the system considered safe for deployment? | Organize a review with the ethical committee to assess performance, functioning, various risk mitigation strategies to ensure safe and reliable deployment |
| Ongoing Monitoring | |
| Consideration | Mitigation Strategy |
| Is the system being evaluated on ongoing basis and tested for performance, accuracy, unintended consequences, fairness? | Ensure risk mitigation strategy for changing development environment |
| | Ensure documentation of policies, processes and technologies used |
| | Monitor Fairness goals over time and ensure mechanisms to constantly improve |
| | Track performance of the system and changes over time |
| | Ensure policies and mechanisms to ensure third party agencies can probe, understand and review behaviour of the system |
| | Ensure engagement with open source, academic and research community for auditing the algorithm |



Appendix 2

Review of Global Regulatory Landscape

European Union

The EU's Ethics Guidelines for Trustworthy AI is based on an approach founded on fundamental rights. It offers sector-agonistic guidelines that require AI practitioners to respect the proportionality between means and ends, and carefully create a balance between competing interests and objectives. It also states that the development, deployment and use of AI systems must account for both substantive and procedural fairness.

Under General Data Protection Regulation, 2016 (GDPR), entities processing personal data or determining the means for the processing of personal data are required to implement technical and organisational measures that ensure a level of security appropriate to the risk involved in processing such personal data. These measures include – the pseudonymization and encryption of data; preserving the confidentiality, integrity and resilience of processing systems and services; and restoring the availability of, and access to, personal data in a timely manner.

The EU Cybersecurity Act, 2019 (Cybersecurity Act) entrusts the European Union Agency for Cybersecurity (ENISA) with the responsibility of developing certification frameworks for cybersecurity, including the development of sectoral frameworks for cybersecurity with regards to products.

In a white paper on Artificial Intelligence, the European Commission highlighted the importance of suitably amending the Product Liability Directive to enhance security related aspects for AI. The paper noted that the existing product safety legislation already protects against all kinds of risks arising from the product according to its use, including AI products. However, it noted that certain amendments may be introduced in the Product Liability Directive to address risks arising out of new technologies. These amendments include



risk assessment, human oversight at the time of design, specific requirements addressing the risk of faulty data at the design stage and provisions discussing and requesting cooperation between economic operators in the supply chain to ensure that safety standards are adequately preserved.

Singapore

The privacy and security regime in Singapore is consolidated under a single law – the Personal Data Protection Act, 2013 (**PDPA**). Regulating privacy issues in AI is rooted in two core data protection principles, *consent obligation* and *purpose limitation*.⁵⁵ The Personal Data Protection Commission (**PDPC**) is Singapore's data protection regulator. PDPC published a revised Model Artificial Intelligence Governance Framework (Framework) in 2019, containing a roadmap for data protection compliance for organizations deploying AI.

The PDPA also requires an organization to protect personal data in its possession or under its control by employing such '*reasonable security arrangements*' to prevent unauthorized access to data and mitigate similar risks. These arrangements also intend to cover an assessment of the adequacy of existing safeguards, with the PDPC stressing the importance of covering all foreseeable scenarios that can potentially lead to a data breach/security risk to the personal data.⁵⁶ Towards this, it proposed a framework that stressed on the importance of having a 'human in the loop' or a 'human over the loop' based on the degree of severity and harm occasioned by the particular processing of personal data.⁵⁷

The equivalent data protection standard for cross-border transfers is also adopted by the PDPA. Businesses deploying AI systems in a manner that moves personal data from one jurisdiction to another must ensure that the entity receiving such data is bound by a series of safeguards that provide the same standard of data protection as the PDPA itself.

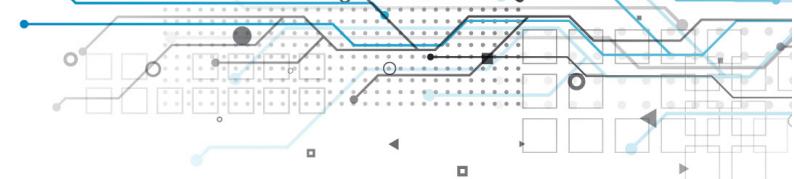
The Model AI Governance Framework (**Model Framework**), was drafted based on a discussion paper issued by the Personal Data Protection Commission and part of Singapore's National AI Strategy.⁵⁸ It provides a means for entities

55. Benjamin Wong YongQuan, 'Data privacy law in Singapore: the Personal Data Protection Act 2012' [2017] 7(4) International Data Privacy Law 287

56. In the matter of an investigation under Section 50(1) of the Personal Data Protection Act 2012 and L'Oreal Singapore Pte. Ltd, Case No. DP-1812-B3091. [Singapore]

57. Personal Data Protection Commission, 'Discussion Paper On Artificial Intelligence (AI) And Personal Data – Fostering Responsible Development And Adoption Of AI' PDPC Discussion Paper (05 June 2018) [Singapore]

58. https://www.smartnation.gov.sg/docs/default-source/default-document-library/national-ai-strategy.pdf?sfvrsn=2c3bd8e9_4



employing AI to demonstrate their implementation of the accountability-based practices in data management and protection contained therein.⁵⁹ It focuses on prospective accountability by making internal governance processes robust and demonstrating to customers and regulators that the entity has employed practices to foster accountability among the designers and operators of AI, and to ensure that the AI systems, applications and algorithms are transparent and fair in their operation, while providing information and explanation to consumer about where and how AI is being used with respect to their data or services and products made available to them.

In 2018, the Monetary Authority of Singapore (**MAS**) published the Principles to Promote Fairness, Ethics, Accountability and Transparency in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector (**FEAT Principles**) specifically aimed at the use of AI for financial services. The FEAT Principles have identified the proactive disclosure of the use of Artificial Intelligence and Data Analytics (**AIDA**) to data subjects as a key principle to ensure transparency in the context of data protection for AI

USA

The US government, through Executive Order 13859 in February 2019, issued a series of directions to various federal stakeholders to develop policies and principles that promote advancement of AI based technology while also protecting civil liberties.⁶⁰ Pursuant to this Executive Order, the White House in January 2020, issued a set of 10 "Principles for the Stewardship of AI Applications," which called for, among others, fairness and non-discrimination to be top priorities for agencies drafting and implementing regulations on AI.⁶¹

Unlike Singapore and the EU, the US lacks an overarching federal legislation on privacy. However, various sector specific laws regulate aspects of privacy. The state of California, which has recognized privacy as a constitutional right, enacted a comprehensive legislation, the California Consumer Privacy Act, 2018 (CCPA) providing for certain safeguards that directly affect AI systems. Like Singapore and EU, businesses deploying AI systems would be obligated to notify users of the purposes of such processing with users being able to withdraw their consent from such processing.

59. <https://ai.bsa.org/wp-content/uploads/2019/09/Model-AI-Framework-First-Edition.pdf>

60. <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/>

61. <https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf>

Appendix 3

Model Transparency Mechanisms

Model Cards

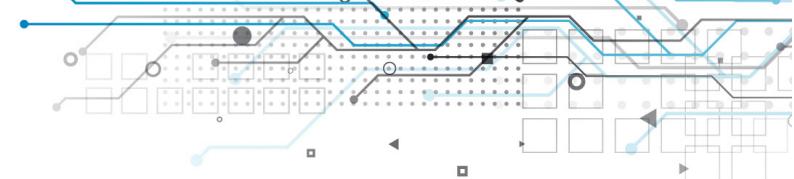
Google released the Model Card Toolkit, a toolset designed to facilitate AI model transparency reporting for developers, regulators, and downstream users. It's based on Google's Model Cards framework, which provide a structured framework for reporting on ML model provenance, usage, and ethics-informed evaluation and give a detailed overview of a model's suggested uses and limitations that can benefit developers, regulators, and downstream users alike.⁶² Model cards are aimed at both experts and non-experts. Developers can use them to design applications that emphasize a model's strengths while avoiding or informing end users of its weaknesses. For journalists and industry analysts, they might provide insights that make it easier to explain complex technology to a general audience. And they might even help advocacy groups better understand the impact of AI on their communities. Google has designed examples for two features of its Cloud Vision API, Face Detection and Object Detection. They provide overviews of both models' ideal forms of input, visualize some of their key limitations, and present basic performance metrics. Both are early proofs of concept, to advance the conversation around the value of transparency in AI.⁶³

Datasheets for Datasets

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, Microsoft proposed datasheets for datasets. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, it is

62. <https://arxiv.org/pdf/1810.03993.pdf>

63. <https://modelcards.withgoogle.com/about>



proposed that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.⁶⁴

Fact Sheet

The goal of IBM's Fact Sheet project is to foster trust in AI by increasing transparency and enabling governance. Increased transparency provides information for AI consumers to better understand how the AI model was created. This allows a consumer of the model to determine if it is appropriate for their situation. AI Governance enables an enterprise to specify and enforce policies describing how an AI model or service should be constructed and deployed. This can prevent undesirable situations, such as a model training with unapproved datasets, models having biases, or models having unexpected performance variations. A Fact Sheet is a collection of relevant information (facts) about the creation and deployment of an AI model or service. Facts could range from information about the purpose and criticality of the model, measured characteristics of the dataset, model, or service, or actions taken during the creation and deployment process of the model or service.⁶⁵

64. <https://arxiv.org/pdf/1803.09010.pdf>
65. <https://aifs360.mybluemix.net/introduction>



सत्यमेव जयते
NITI Aayog