Working Document: Towards Responsible #AIforAll

DRAFT FOR DISCUSSION

सत्यमेव जयते

NITI Aayog

Draft Document for discussion

The content of this draft document is solely for the purposes of discussion with stakeholders on the proposed subject.

The document was prepared for an expert consultation held on 21 July 2020. The information contained herein is neither exhaustive nor final and is subject to change.

All stakeholders are requested to review the documents and provide comments on or before 28th August 2020, preferably on email at annaroy@nic.in

AI Strategy pushes for rapid adoption of AI

The National Strategy for Artificial Intelligence highlights the potential of AI and lays down recommendations to accelerate adoption

Economic and sectoral potential

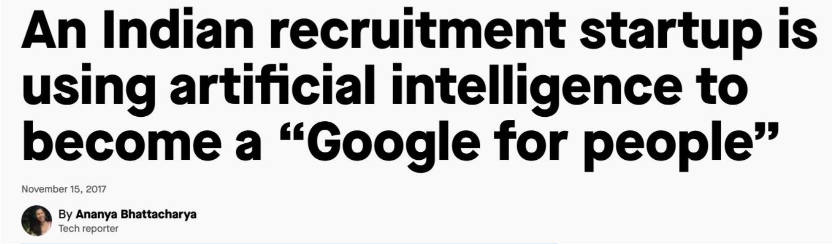
- AI expected to **boost India's annual growth rate by 1.3%** by 2035;
- Emphasized **potential of large scale adoption of AI in a variety of social sectors,** towards 'AI for All';
- **AI Garage for 40% of the world,** or the export of relevant social sector products to other emerging economies;

Increasing adoption

- **In the government,** as Ministries and Departments are looking to deploy AI solutions (IPO, Telangana Gov, etc);
- **In private sector and startups,** with ML powered solutions in health, agriculture, etc (NIRAMAI, Satsure, etc);
- **In academia,** where research institutions are pushing the frontiers of AI through Basic and Applied research (Safe and Stable RL, reasoning, comprehension, etc)

Adoption has seen challenges globally

Recent examples of instances raising concerns of ethical use of AI

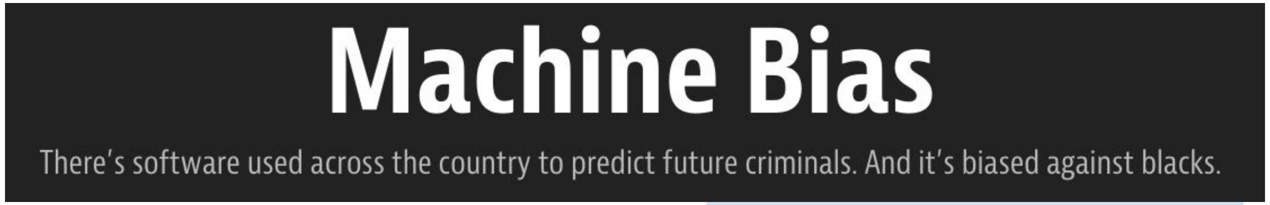


An Indian recruitment startup is using artificial intelligence to become a "Google for people"

November 15, 2017

By Ananya Bhattacharya
Tech reporter

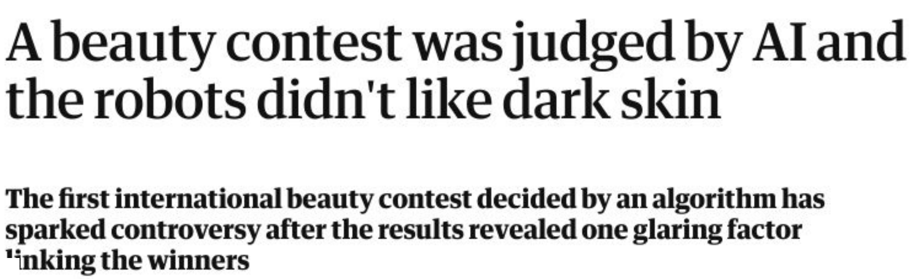
- QZ India, Nov 2017



Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

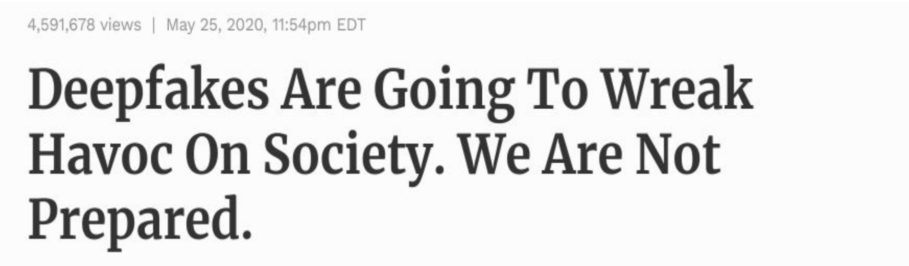
- ProPublica, May 2016



A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

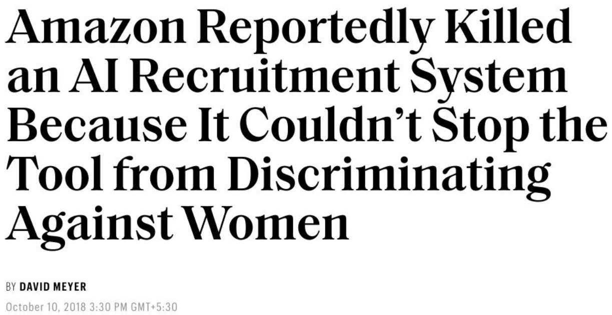
- Guardian, Sep 2016



4,591,678 views | May 25, 2020, 11:54pm EDT

Deepfakes Are Going To Wreak Havoc On Society. We Are Not Prepared.

- Forbes, May 2020

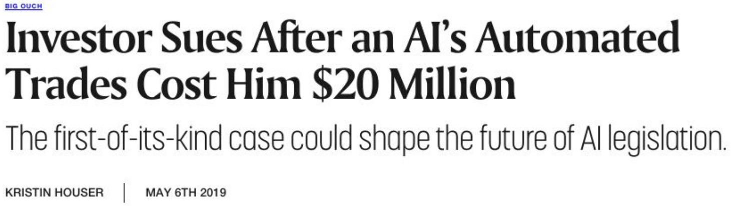


Amazon Reportedly Killed an AI Recruitment System Because It Couldn't Stop the Tool from Discriminating Against Women

BY DAVID MEYER

October 10, 2018 3:30 PM GMT+5:30

- Fortune, Oct 2018



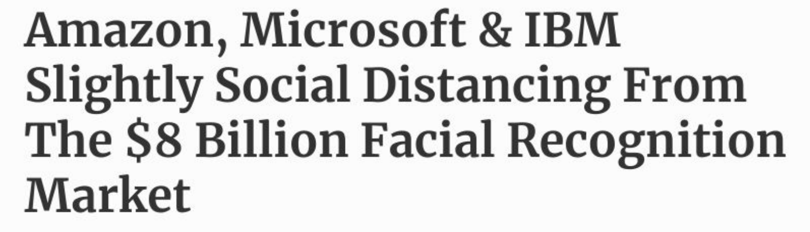
BIG OUCH

Investor Sues After an AI's Automated Trades Cost Him $20 Million

The first-of-its-kind case could shape the future of AI legislation.

KRISTIN HOUSER | MAY 6TH 2019

- MIT Tech Review, Jan 2020

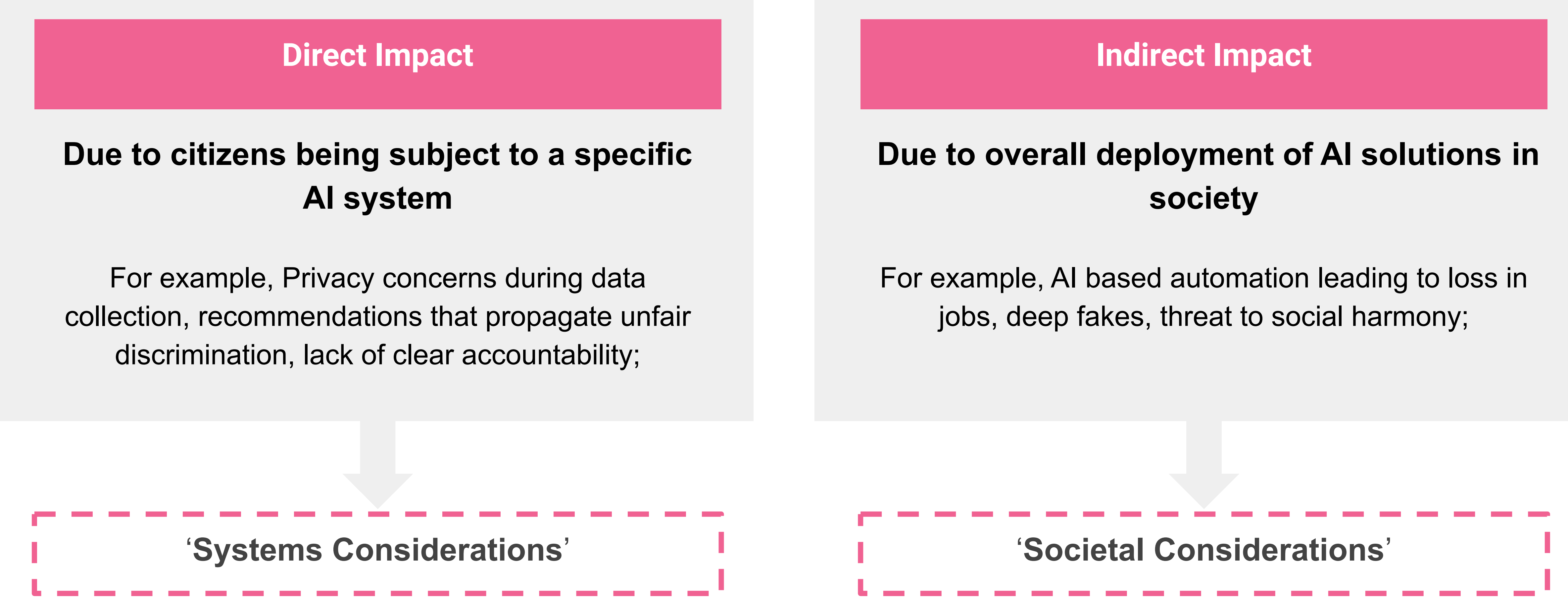


Amazon, Microsoft & IBM Slightly Social Distancing From The $8 Billion Facial Recognition Market

- Forbes, June 2020

Studying the Challenges- Approach

Challenges are studied under 2 broad areas depending on nature of the impact



Direct Impact

Due to citizens being subject to a specific AI system

For example, Privacy concerns during data collection, recommendations that propagate unfair discrimination, lack of clear accountability;

‘Systems Considerations’

Indirect Impact

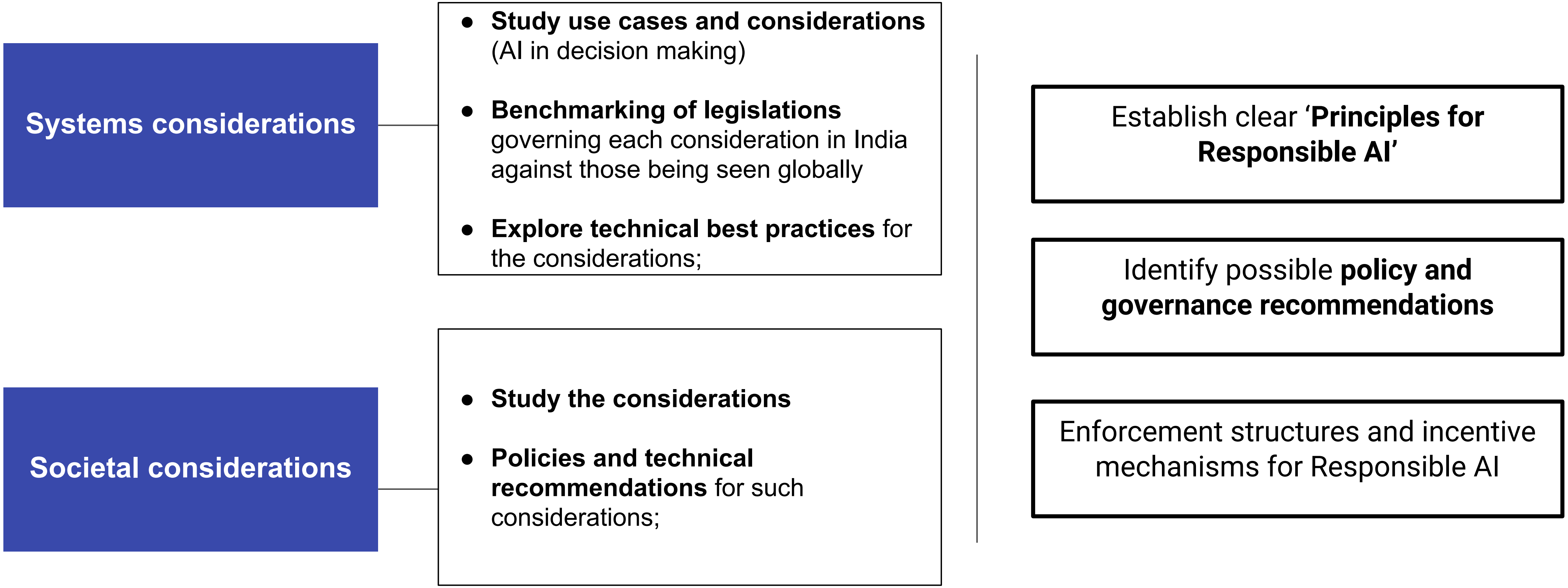
Due to overall deployment of AI solutions in society

For example, AI based automation leading to loss in jobs, deep fakes, threat to social harmony;

‘Societal Considerations’

Methodology and Objectives

Scope of paper limited to ‘**Artificial Narrow Intelligence**’



Systems considerations

- **Study use cases and considerations** (AI in decision making)
- **Benchmarking of legislations** governing each consideration in India against those being seen globally
- **Explore technical best practices** for the considerations;

Societal considerations

- **Study the considerations**
- **Policies and technical recommendations** for such considerations;

Establish clear ‘**Principles for Responsible AI**’

Identify possible **policy and governance recommendations**

Enforcement structures and incentive mechanisms for Responsible AI

The paper aims to create a foundation for an ecosystem of Responsible AI in India

Study of system considerations

Note:
These considerations 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 today.

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

The issue

- While accuracy gives a reasonable view into how a system performs, understanding decision making process is important to ensure safe and reliable deployment

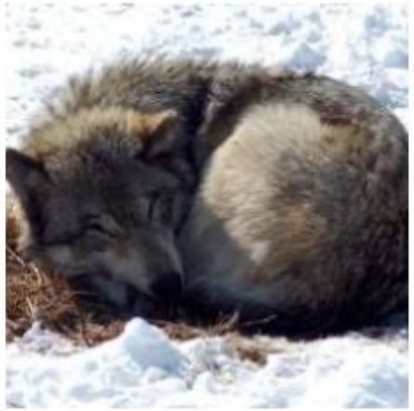
Its implications

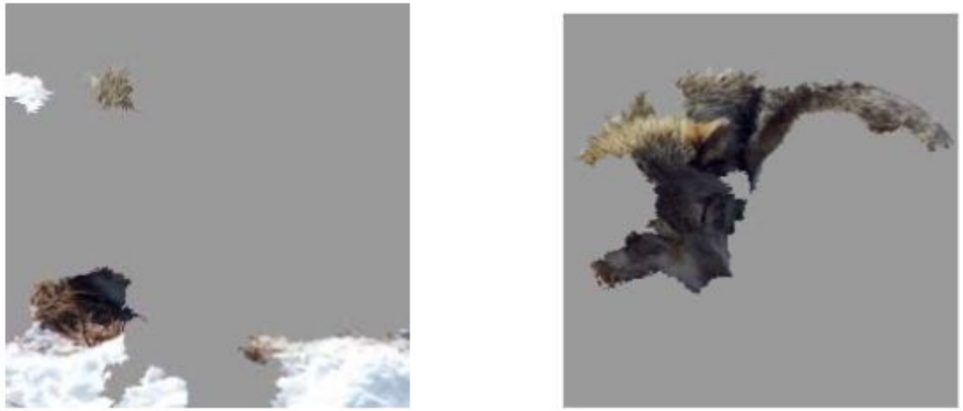
- The system could pick spurious correlations, in the underlying data, leading to good accuracy in test datasets but significant errors in deployment

Example:

2 separate classifiers are used to distinguish between wolf and husky

The classifiers have similar accuracy





Classifier 1

Classifier 2

Parts of the image determining classification

Classifier 1 detects wolf because of environment (snow)

Classifier 2 detects wolf because of its body features

Source: https://www.youtube.com/watch?v=TBJqgvXYhfo

Systems Consideration 2:
Post deployment, can users of the AI system understand why a specific decision was made?

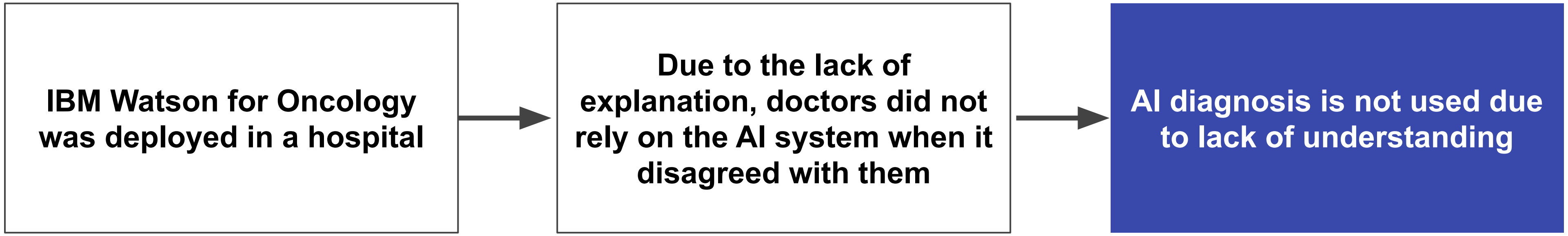
The issue

- 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;

Its implications

- Leads to:
  - 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;

Example: Deployment for Cancer Diagnosis



IBM Watson for Oncology was deployed in a hospital → **Due to the lack of explanation, doctors did not rely on the AI system when it disagreed with them** → **AI diagnosis is not used due to lack of understanding**

Source: https://spectrum.ieee.org/biomedical/diagnostics/how-ibm-watson-overpromised-and-underdelivered-on-ai-health-care

Systems Consideration 3:
Consistency across stakeholders

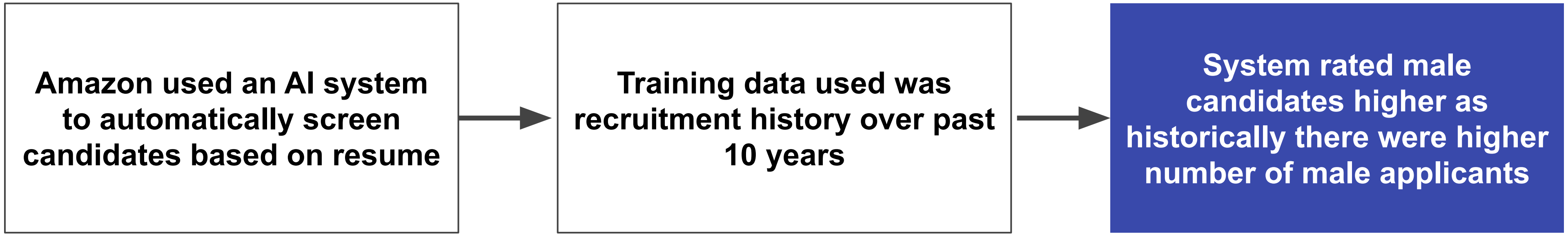
The issue

- Different types of cognitive biases have been identified and tend to be 'unfair' for certain groups (across religion, race, caste, gender);
- 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;

Its implications

- 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 for social order**

Example: Amazon's Resume screening application



Amazon used an AI system to automatically screen candidates based on resume

Training data used was recruitment history over past 10 years

System rated male candidates higher as historically there were higher number of male applicants

Source: https://in.reuters.com/article/amazon-com-jobs-automation/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idINKCN1MK0AH

Systems Consideration 4:
Incorrect decisions leading to exclusion of service/ benefits

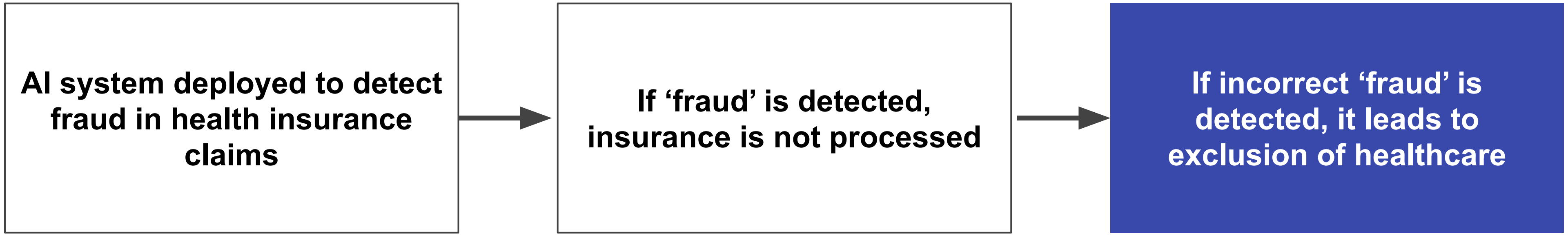
The issue

- 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;

Its implications

- May lead to exclusion of citizens from services guaranteed by the state;

Example:



AI system deployed to detect fraud in health insurance claims

If ‘fraud’ is detected, insurance is not processed

If incorrect ‘fraud’ is detected, it leads to exclusion of healthcare

Systems Consideration 5: Accountability of AI decisions

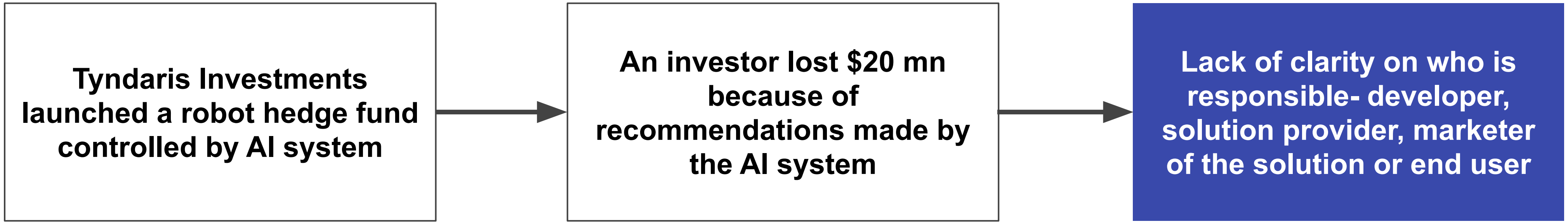
The issue

- 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

Its implications

- Lack of consequences reduces incentive for responsible action
- Difficulty in grievance redressal

Example:



Tyndaris Investments launched a robot hedge fund controlled by AI system

An investor lost $20 mn because of recommendations made by the AI system

Lack of clarity on who is responsible- developer, solution provider, marketer of the solution or end user

Source: https://futurism.com/investing-lawsuit-ai-trades-cost-millions

Systems Consideration 6:
Privacy risks

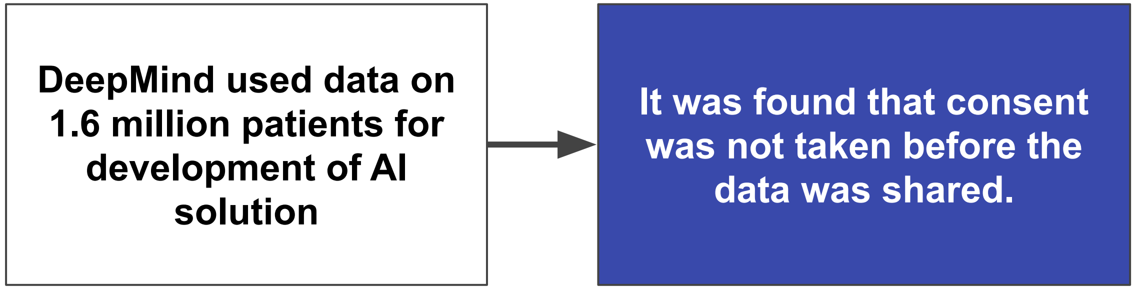
The issue

- 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;

Its implications

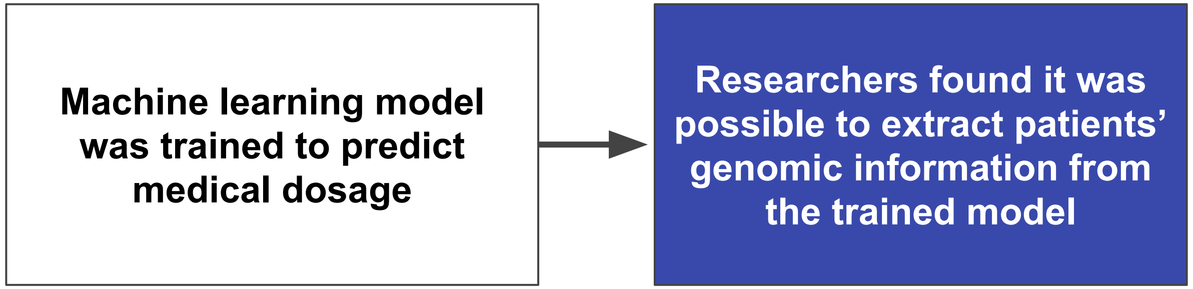
- Infringement of Right to Privacy;

Example:



DeepMind used data on 1.6 million patients for development of AI solution → It was found that consent was not taken before the data was shared.

Source:
https://venturebeat.com/2019/12/21/ai-has-a-privacy-problem-but-these-techniques-could-fix-it/



Machine learning model was trained to predict medical dosage → Researchers found it was possible to extract patients' genomic information from the trained model

Source:
https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf

Systems Consideration 7:
Security risks

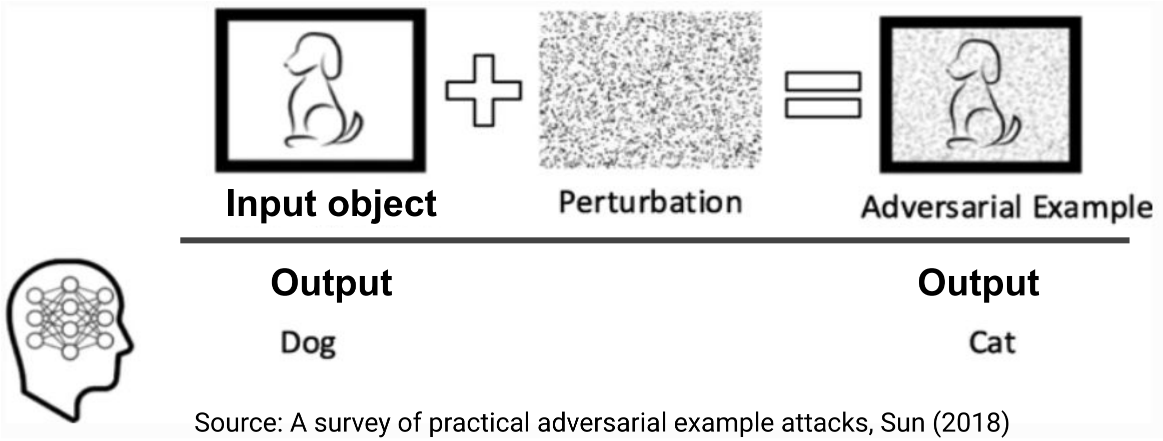
The issue

- 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;

Its implications

- In real world deployments, may lead to malfunctioning of system;
- Risk to IP protection due to potential of ‘model steal’ attacks;

Example:



Source: A survey of practical adversarial example attacks, Sun (2018)

- ‘Perturbation’ included by attacker has the potential to alter the output of the system;
- In usages such as ‘**AI systems for traffic management’, this may have serious real world repercussions;**

Impact of system considerations on citizen rights

- System considerations may affect citizens in a variety of ways and present tangible challenges;
- We explore practices in Legislation and Product development for ensuring “Responsible AI”

Legislation - Summary

Relevant legal protection for AI-related concerns largely exists, but needs to adapt

Protection for citizens across sectors largely exists

- Many of the harms caused by AI is not new
- Existing legislations cover most of the consequences raised under ‘Systems Consideration’

The protection, however, needs to adapt for AI

- The nuances associated with AI requires a relook at the existing regulations
- Ex: While accountability laws exist (‘Consumer Protection Act’), a framework is required to assign accountability for AI systems

Sector specific regulations/ guidelines may be required in certain cases

- Some sectors have unique considerations that may require sector-specific laws for AI
- Ex: Use of AI in administrative decisions by the State would be required to explain the decision making process

Legal benchmarking: Singapore

Countries are using a combination of sectoral regulations and broader AI guidelines

What exists today?

Guidelines or regulations established specifically for AI

- **‘Model AI Governance Framework’** released by Infocomm Media Development Authority (IMDA) to **serve as only a guide** to implement ‘explainable, fair, transparent, and human centric AI;

Sector specific regulations that may be applied to AI

- **‘FEAT Principles’ for AI in financial services**, released by Monetary Authority of Singapore (MAS) meant to serve as **non-prescriptive guidance document** to encourage adoption of fair, explainable, ethical, and accountable AI;

Sector agnostic laws that are relevant to AI

- **Personal Data Protection Act (PDPA) 2012** released by the Personal Data Protection Committee (PDPC) **establishes a data protection law** that comprises various rules governing the collection, use, disclosure and care of personal data;

For data protection, specific laws exist - but other regulations are in the form of ‘guides’

Legal benchmarking: EU

Countries are using a combination of sectoral regulations and broader AI guidelines

What exists today?

Guidelines or regulations established specifically for AI

- **EU Ethics Guidelines for Trustworthy AI released by High Level Expert Group on AI,** a **non-binding document** that put forward a set of 7 key requirements that AI systems should meet in order to be deemed ‘trustworthy’;

Sector specific regulations that may be applied to AI

- Certain use cases under **few sectors are termed ‘high-risk’** and have specific requirements. Such use-cases have an accompanying **‘oversight’ mechanism**

Sector agnostic laws that are relevant to AI

- **General Data Protection Rules (GDPR) 2016, a regulatory framework** for protection of personal data and relevant to AI, establishes need for ‘**privacy by design**’ when developing automated solutions;

GDPR is very exhaustive and ethics guidelines released - but no overarching legislation yet

Legal benchmarking: USA

Countries are using a combination of sectoral regulations and broader AI guidelines

What exists today?

Guidelines or regulations established specifically for AI	● **10 “Principles for the Stewardship of AI Applications”** released by the US White house **establishes priorities** for US federal agencies drafting and implementing regulations on AI, including fairness and non-discrimination;
Sector specific regulations that may be applied to AI	● **Fair Credit Reporting Act (FCRA) and the Equal Credit Opportunity Act (ECOA),** which mandates contain provisions for outcome based explanations for adverse action and mandates for non-discrimination; **HIPAA Privacy Rule (2000) and Graham Leech Bliley Act (1999) for governance of data in healthcare and finance respectively;**
Sector agnostic laws that are relevant to AI	● (Proposed Bill) **Algorithmic Accountability Act, 2019**, which would establish a law to reduce biased decisions and outcomes; **California Consumer Privacy Act, 2018** established in California for **data protection containing provisions relevant to the use of AI**;

Legal and regulatory scenario in India

In India, there are gaps in legal protections for impacts of systems considerations

What exists today?	Example(s)	Gap Analysis
Guidelines or regulations established specifically for AI	● Not yet defined;	● Overarching principles would help to guide formation of standards and regulations;
Sector specific regulations that may be applied to AI	● For example, **Medical Device Rules, 2017** laying out standards and regulations for medical devices; ● **SEBI's Circular on AI/ML applications** offered by market intermediaries;	● Sectors with risk for implication to citizens have already defined some form of ethical framework;
Sector agnostic laws that are relevant to AI	● Draft **PDP Bill** for Data Privacy, **Consumer Protection Act**, SPDI Rules (2011) and IT Act (2000), Right to Information Act;	● For areas such as privacy, inclusiveness and accountability, regulations already exist but need to adapt for AI specific challenges;

Technical Approach- Summary

NSAI recommended using technology to manage AI risks; It is an evolving field

Technical mechanisms of managing AI specific challenges is growing

- Growth of AI is relatively recent;
- However, there is a growing interest in both private sector and Government agencies in developing tools to manage the risks;

Open sourcing of these tools has been vital for its development

- Open sourcing of such tools has increased both usage and development;

Ethics in AI is a growing field of research and must be encouraged

- Popular conference in AI has seen a spike in research papers on Ethical AI;
- However, the applications are increasing at a rapid rate, both in scale and performance, and such research must be encouraged

Technical best practices

Technical means are being explored globally to mitigate risks

Technology can help by:	Example:
Interpreting decision of AI solutions to instil trust	● **‘Pre hoc’ techniques** such as Exploratory Data Analysis (**EDA**), concept extraction, dataset summarization, distillation techniques; ● **‘Post hoc’** techniques for model explanation through input attribution (**LIME, SHAP, DeepLift**) and example influence matching (**MMD critic, influence function, etc**);
Allowing processing of data in a manner that is ‘privacy preserving’	● Usage of methods such as **federated learning, differential privacy, Zero Knowledge Protocols or Homomorphic Encryption;**
Assessing data sets for representation or “fairness”	● Tools such as I**BM ‘AI Fairness 360’, Google ‘What-If’ Tool, Fairlearn** and open source frameworks such as **FairML**;

Urgent need for countries to enable international and multi-disciplinary research in the field

Study of societal considerations

Note:
Societal considerations look to highlight broader ethical ramifications of the usage of AI such as impact on jobs, or in specific uses such as psychological profiling for malicious use

Societal Consideration 1:
Impact on Jobs

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

- A report by the Oxford Economic Forecast indicates a **high potential for automation of tasks performed by the Indian workforce**

- Manufacturing and IT services, sectors that account for 10 million and 3 million jobs a year are particularly impacted

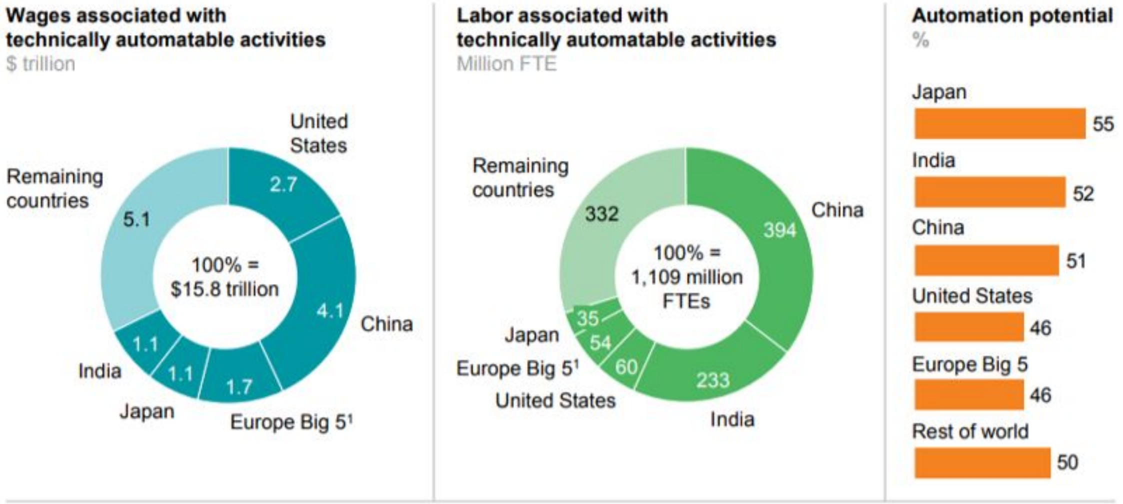
- Impact of technology and innovations in the job landscape is not new. Major technology introductions in the past have resulted in enhanced productivity and redefined job profiles

- In the short term, tracking changes in job profiles, both locally and internationally, would enable data driven policies

- **Skilling, adapting legislations and regulations** to the change have historically allowed countries to **leverage benefits and harness new job opportunities.**

Technical automation potential is concentrated in countries with the largest populations and/or high wages

Potential impact due to automation, adapting currently demonstrated technology (46 countries)



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

Societal Consideration 2:
Malicious use - Psychological Profiling

- Psychological profiling enabled by AI and the ease of spreading propaganda through online platforms has potential to cause social disharmony and disrupt democratic process
  - Cambridge Analytica scandal involved using data of millions of users, without their consent on matters of National and Political interest around the world
  - In Myanmar, online platforms were used to spread hate speech and fake news was targeted against a particular community, leading to ethnic violence
- **Legislation**: The PDP Bill identifies obligations for social media intermediaries with regards to **actions that may impact democracy, public order or sovereignty and security of the State**
- **Technology**: Proactive identification and flagging of propaganda and hate speech is less advanced when dealing with posts in local languages. Research efforts must be dedicated to improve technology advancements in these areas

Principles

Why Principles

Different Groups shape the future of AI

The Government

Develop policies that may impact AI
Procure AI systems

Citizens

Influenced by AI systems, either directly or indirectly

Regulators

Oversee sector specific/ sector agnostic rules and regulations

Private sector

Develop AI products and solutions
Use AI products and solutions

Research Community

Research across social sector, regulation, technologies and innovation in AI

Standard Setting Bodies

Set standards for research or technology that may use AI. Ex: BIS, ICMR

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

How are the Principles developed

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

Principles of Responsible AI

The following principles are based on the core principle of ensuring AI does not cause harm

- Principle of Safety and Reliability
- Principle of Equality
- Principle of Inclusivity and Non-discrimination
- Principle of Privacy and security
- Principle of Transparency
- Principle of Accountability
- Principle of protection and reinforcement of positive human values

The changing nature of technology necessitates regular update of the principles.
An institutional mechanism for this is proposed in a later section.

Enforcement Mechanisms

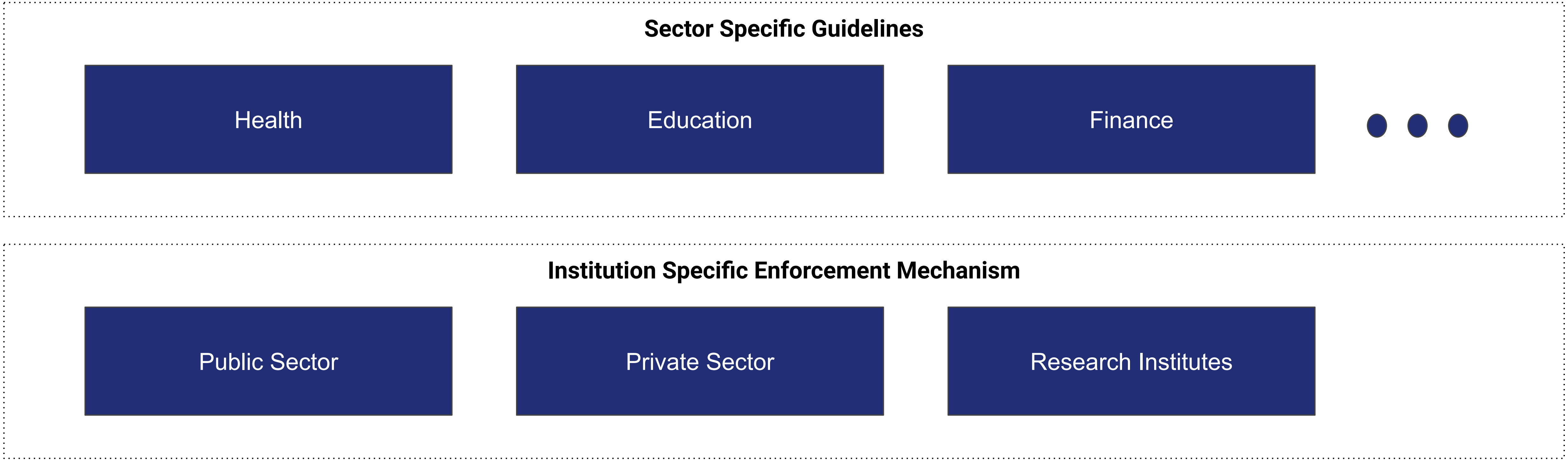
Structures for implementation of Principles

Management and Update of Principles for Responsible AI

- Update principles as per emerging use cases and examples of arising challenges;
- Guide various bodies involved in setting standards and regulations for AI

Entity managing the Principles must include experts from technology, sector, and legal/policy fields;

It is recommended that a research institution with the necessary expertise manage the Principles;



Sector Specific Guidelines

Health

Education

Finance

● ● ●

Institution Specific Enforcement Mechanism

Public Sector

Private Sector

Research Institutes

Appendix

Encourage Research into Responsible AI

- Low cost, results oriented financing models for startups looking to develop tools to detect bias, explainable AI models, privacy preserving techniques;

- The Government may fund specific research projects in responsible AI;

- Host high quality international conference on ‘Responsible AI’, with a focus on recognizing quality efforts on responsible AI;

- Introduce ethics of AI into the university curriculum

Self assessment guide for Responsible AI (abridged*)

1

Problem scoping:

- Assess the potential ‘degree of harm’ by engaging with social scientists, humanists, development sector researchers and other relevant experts;
- Develop a plan of action for unintended consequences on an ongoing basis.
- Establish a grievance redressal mechanism
- Identify mechanisms to handle errors in decision by the AI system
- Ensure provision for public auditing without opening up the system for unwarranted manipulation

2

- Identify and document goals for equality, non-discrimination and inclusion
- Identify Explain-ability goals and requirements of the system

Data collection:

- Identify all relevant rules and regulations for handling data
- Document known sources of data and steps taken to ensure privacy and safety
- Assess the representativeness of the dataset and how its use over time will impact different datasets

3

Data labeling

- Assess and account for human variability and bias in annotation

Data processing:

- Ensure only relevant data is being used and personal and sensitive data is being adequately masked

Training

- Assess explainability of the model used
- Ensure fairness goals are reflected in training of the system
- Ensure training model is not memorizing sensitive data

*Longer detailed version will be presented in the paper

Self assessment guide for Responsible AI (abridged)

4

Evaluation:

- Assess working of the system by engaging with sector and data experts for safe and reliable deployment
- Evaluate if the system meets the fairness goals across anticipated deployment scenarios
- Evaluate the system against adversarial inputs
- Evaluate error rates across sub population groups and assess potential social impact

5

Deployment:

- Ensure easy accessibility of grievance redressal mechanisms
- Assess impact of real world bias and feedback loops it may create

Ongoing

- 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

6

- 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

Terms of Reference of Ethical Committees (1/2)

Ethical Committees are accountable for enforcement of principles

- EC should assess the “potential of harm” and potential benefits, evaluate plan for mitigating risks and provide recommendations on whether the AI solution should be approved.
- Ethical Committees (EC) must ensure the AI system is developed, deployed, operated and maintained in accordance with the Principles
- EC should determine the extent of review needed for an AI system depending on inherent risks and benefits.
- EC should ensure accessible and affordable grievance redressal mechanisms for decisions made by the AI system.

Terms of Reference of Ethical Committees (2/2)

- EC should ensure creation of structures within the entity for protection of ‘whistleblowers’ reporting unethical practices
- Every EC should have a documented Standard Operating Protocol (SOP) on functioning. The SOP may be reviewed and updated periodically to reflect changing requirements
- Every EC review must be documented, including the risks identified, mitigation strategy, and comments from the committee members

Composition of Ethical Committee (1/2)

Ethical Committees should have multi-disciplinary composition without Conflict of Interest

Member	Definition
Chairperson	Nodal point of contact, accountable for independent and efficient functioning of the committee Must be able to ensure active participation of all members in discussions and deliberations Ratify minutes of EC meetings
Member Secretary	Must be a member of the organization or institute and should be able to dedicate time for EC reviews Ensure effective procedures and protocols for EC review
Data Science and/or AI expert (one or more depending on requirement)	Must be a qualified data scientist Must identify procedural or technical risks during development and deployment including, data collection, annotation, management, storage, processing, training, maintenance, and monitoring.

Composition of Ethical Committee (2/2)

Member	Definition
Sector expert	Must have expertise in the sector and wide ranging deployment scenarios Must evaluate safety, reliability, access and affordability of grievance redressal mechanism
Legal expert	Must have expertise in relevant rules and regulations relevant to the AI system Must evaluate legal considerations for the AI system
Social scientist/ ethicist (one or more depending on requirement)	Must have background in social or behavioural science or relevant expertise. Must be sensitive to local cultural and moral values. Must assess impact on community, socio-cultural, religious, philosophical context
Representative of Stakeholder community (one or more, depending on requirement)	Must be a stakeholder of the AI solution. Serve as a representative of the user community

Thank you