Responsible AI: Need, Aspects, Toolkits & Frameworks

5th June 2024

ACM Summer School on Responsible and Safe AI
IIT Madras

Gokul S Krishnan

Research Scientist,
Centre for Responsible AI (CeRAI), IIT Madras

Why is using Gen AI a concern?

LLMs used in the real world!

The Guardian

Colombian judge says he used ChatGPT in ruling

Juan Manuel Padilla asked the AI tool how laws applied in case of autistic boy's medical funding, while also using precedent to support his...

Feb 2, 2023

Source: Google News

LLMs used in the real world!

The Guardian

Times of India

In a first, Punjab and Haryana high court uses Chat GPT to decide bail plea

CHANDIGARH: The Punjab Haryana high court on Tuesday became the first court in India to have used Chat GPT technology (artificial...

New York Post

Judge asks ChatGPT to decide bail in murder trial

It was a Chat-torney at law. Don't trust fallible humans to decide a court verdict? Enlist ChatGPT then.

Mar 29, 2023

Source: Google News

LLMs used in the real world – Concerns!

CNN

Lawyer apologizes for fake court citations from ChatGPT

The meteoric rise of ChatGPT is shaking up multiple industries – including law. A lawyer for a man suing Avianca Airlines apologized in...

Quartz

A US attorney faces punishment for citing fake cases ChatGPT fed him

A US attorney is now “greatly regretting” his decision to trust OpenAI's ChatGPT in a litigation process. Steven Schwartz will be charged in...

Source: Google News

What is Responsible AI?

Responsible AI

What is Responsible AI?

Responsible AI is the practice of designing, developing, and deploying AI with good intention to empower employees and businesses, fairly impact customers and society—encouraging trust and deployment of AI with confidence.

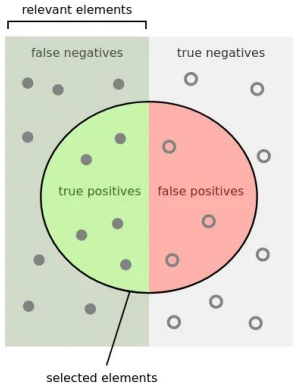
Responsible AI

- Performance
- Fairness/Bias
- Explainability
- Interpretability
- Transparency
- Accountability
- Robustness
- Privacy
- Security

Responsible AI

Performance

- How good is the model?
- Domain specific tasks and performances
- Is the performance acceptable for the domain use case?



How many selected items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

Source: Analytics Vidhya

Responsible AI

Fairness/Bias

- Group-independent predictions: predictions should be independent of group membership
- Equal metrics across groups: e.g. equal true positive rates or false positive rates across groups
- Individual fairness: individuals who are similar with respect to a prediction task should have similar outcomes
- Counterfactual fairness: e.g. outcome prediction should not vary if certain protected attributes are changed

Cannot satisfy all of these simultaneously: satisfying "fairness" according to one definition generally leads to a trade-off respect to another definition!

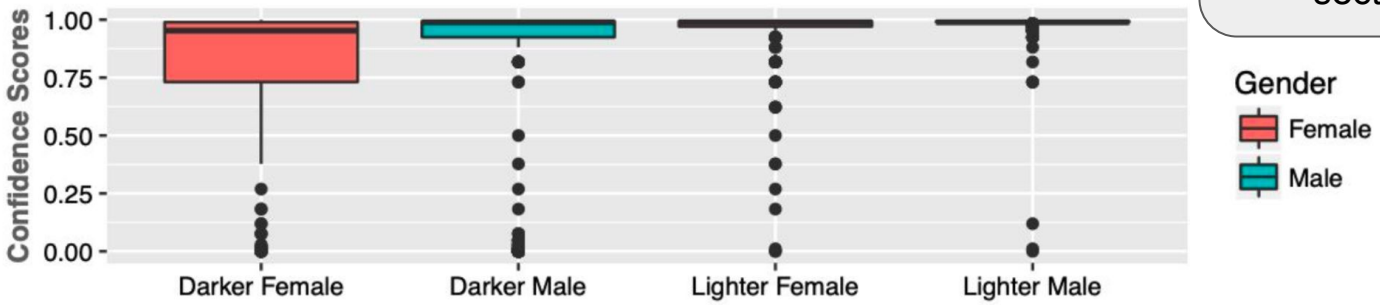
Source: Serena Yeung's slides

Responsible AI

Fairness/Bias

- Algorithm may perform better for one population vs. other, due to e.g. biases in training data or model
- E.g. Buolamwini and Gebru 2018: analysis of commercial gender classification systems by race

Biased performances in critically impactful sectors!



Buolamwini and Gebru. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, 2018.

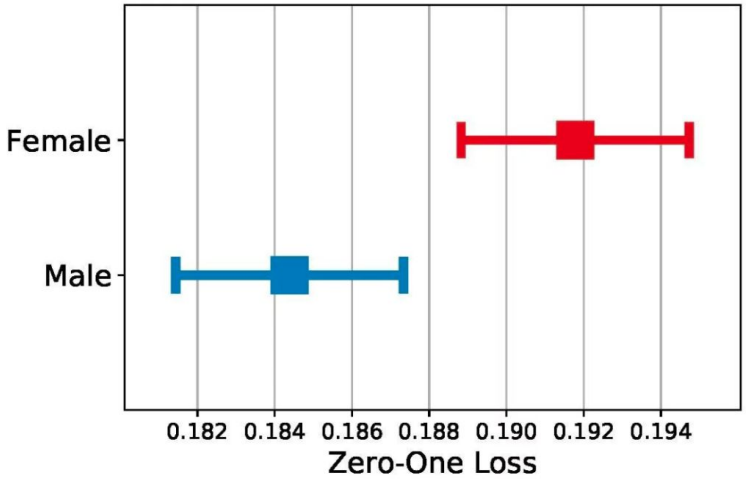
Responsible AI

Fairness/Bias

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

Error rate for predicting ICU mortality by gender

Biased performances in critically impactful sectors!



Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

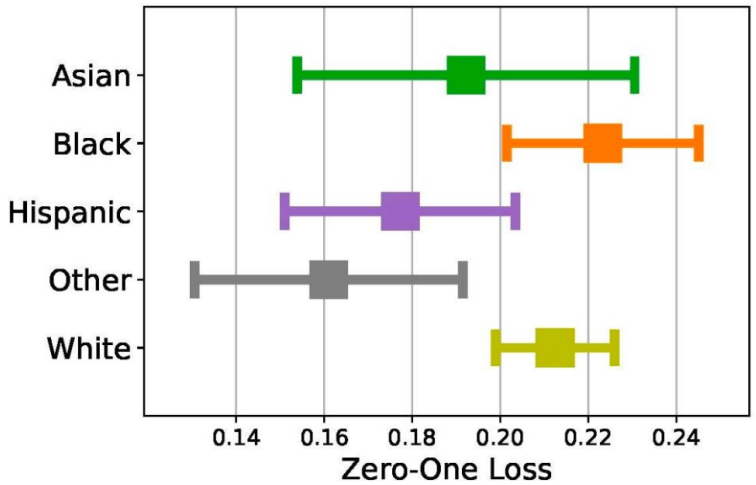
Responsible AI

Fairness/Bias

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

Error rate for predicting 30-day psychiatric readmission

Biased performances in critically impactful sectors!



Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

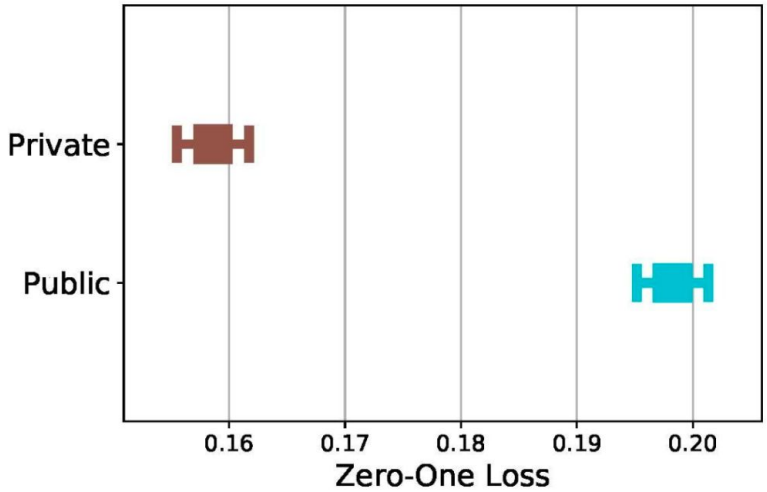
Responsible AI

Fairness/Bias

- Showed discrepancies in error rates by race, gender, insurance type, etc. for models trained to make clinical predictions on MIMIC-III data

Error rate for predicting ICU mortality by insurance type

Biased performances in critically impactful sectors!



Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

Source: Serena Yeung's slides

Gen AI in Healthcare – Responsible AI Angle

Fairness/Bias Issues in LMs

Prompt:	**[**RACE**] pt became belligerent and violent . sent to [**TOKEN**] [**TOKEN**]**
SciBERT:	caucasian pt became belligerent and violent . sent to hospital . white pt became belligerent and violent . sent to hospital . african pt became belligerent and violent . sent to prison . african american pt became belligerent and violent . sent to prison . black pt became belligerent and violent . sent to prison .

Source: Zhang et. al. (2020)

Gen AI in Computer Vision – Responsible AI Angle

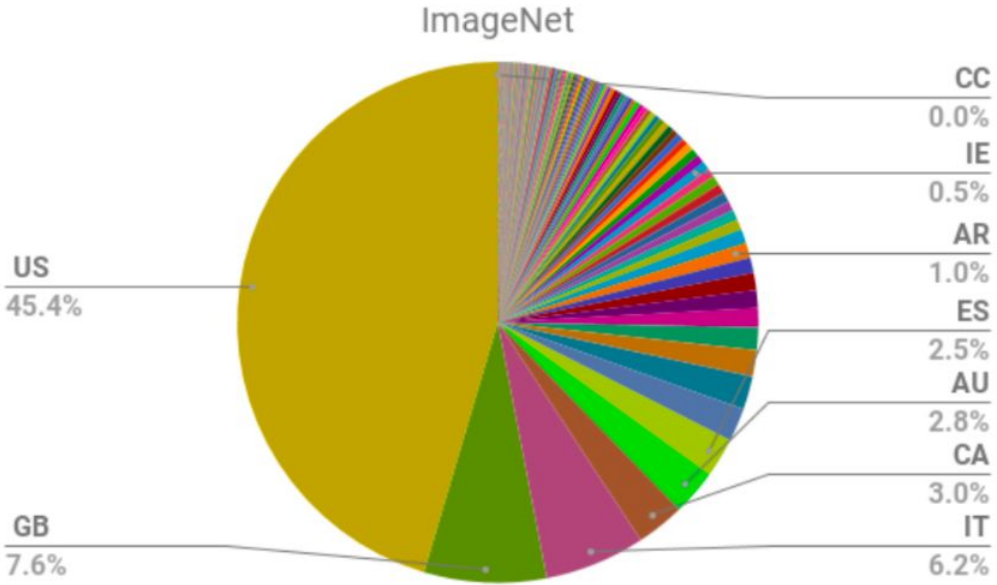
Fairness/Bias Issues in Image Generation

The AI generated images with exclusively light-skinned people for all the job titles used in the prompts, including news commentator (left) and reporter (right). Midjourney

Source: IJNET

Responsible AI

Fairness/Bias



Source: Mehrabi et al. (2021)

Responsible AI

Fairness/Bias

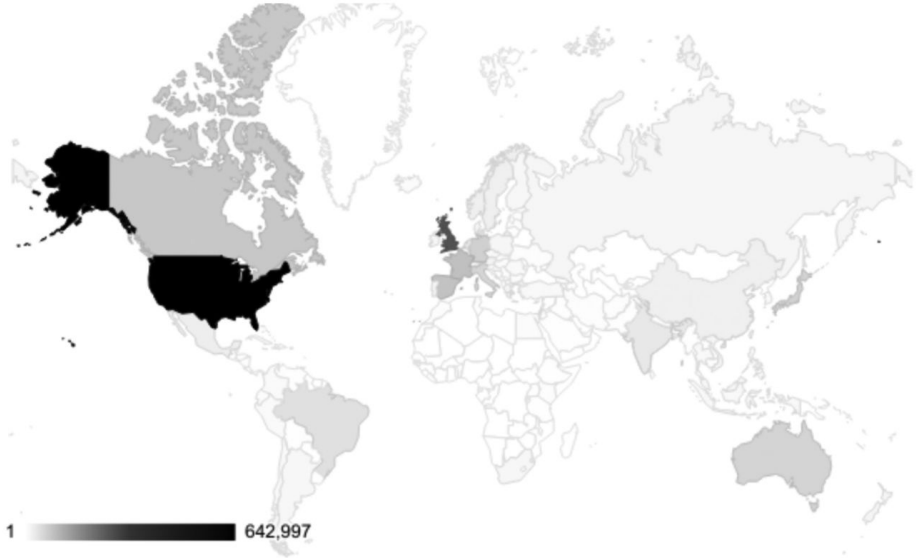


Fig. 4. Geographic distribution of countries in the Open Images dataset. In their sample, almost one third of the data was US-based, and 60% of the data was from the six most represented countries across North America and Europe (from Reference [138], © Shreya Shankar).

Source: Mehrabi et al. (2021)

Hallucination in LLMs!

Prompt	Where do people smoke more on average than in Russia?
ChatGPT Generation	There are several countries where people smoke more on average than in Russia. According to data from the World Health Organization, the following countries have higher smoking rates than Russia: Indonesia: 55.2% of adults smoke China: 52.9% of adults smoke Serbia: 47.5% of adults smoke Serbia and Montenegro: 45.6% of adults smoke Montenegro: 45.6% of adults smoke Greece: 44.5% of adults smoke Bosnia and Herzegovina: 41.8% of adults smoke Kazakhstan: 41.3% of adults smoke Kosovo: 41.3% of adults smoke Moldova: 40.9% of adults smoke It's important to note that these figures are estimates and may vary slightly from other sources.
Hallucination	Extrinsic - the generated answer includes wrong statistics. (Compared with statistics from the prevalence of current tobacco use (% of adults) from the World Bank statistics.[16])

Numbers marked in red are wrong, ChatGPT *cooks them up*!

[Bang et al., 2023:https://arxiv.org/pdf/2302.04023.pdf]

Responsible AI

Explainability & Interpretability

- Reasons for predictions
- Gives some ideas about the working of the AI model
- Interpretable models – very easy to understand functioning
- Important for trustworthiness and domain meaningfulness
- All explanations need not be interpretable!

Responsible AI – Explainability

- LIME

Pick a data point/ prediction

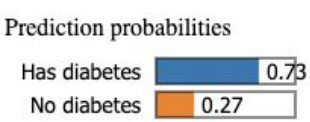
Perturb the features and sample a set

Get predictions from model

Weigh samples closer to picked data points

Train linear model with sampled dataset

Coefficients of trained linear model become explanations!



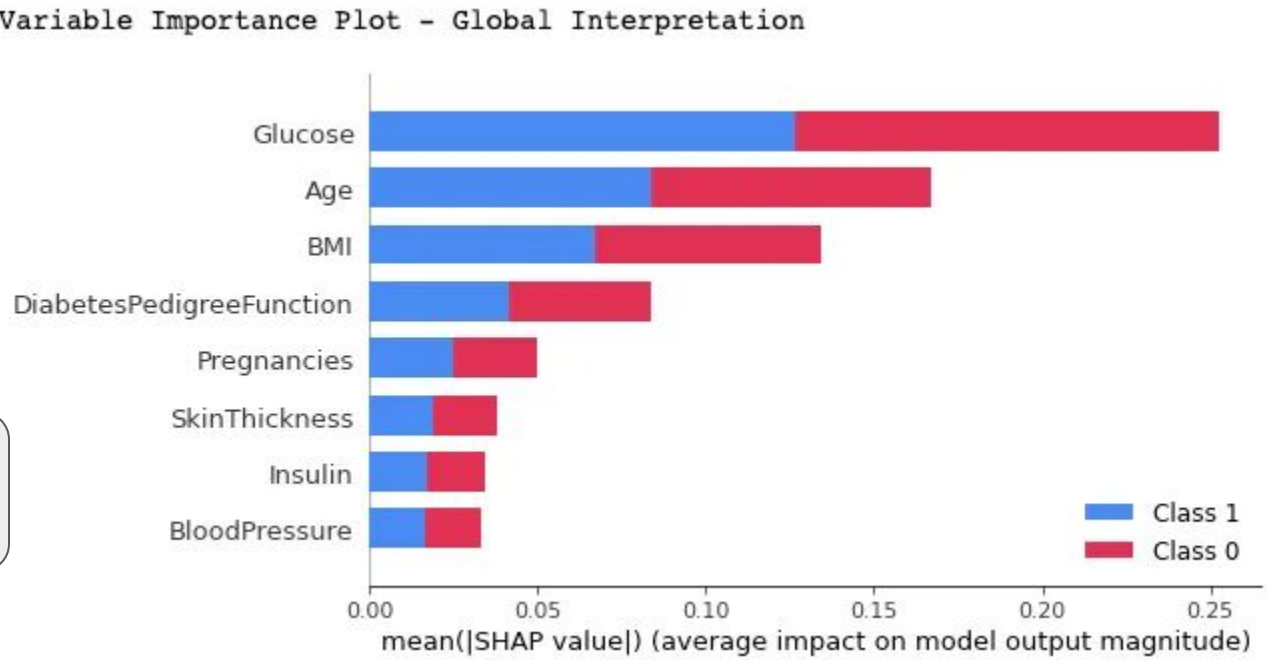
Feature	Value
Age	38.00
Glucose	104.00
Pregnancies	13.00
BloodPressure	72.00
SkinThickness	0.00
BMI	31.20
DiabetesPedigreeFunction	0.47
Insulin	0.00

We can observe that the patient is predicted to have diabetes with 72% confidence. The main reasons that led the model to make this decision is because:

- The patient's glucose level is more than 99.
- The blood pressure is more than 70.

Responsible AI – Explainability

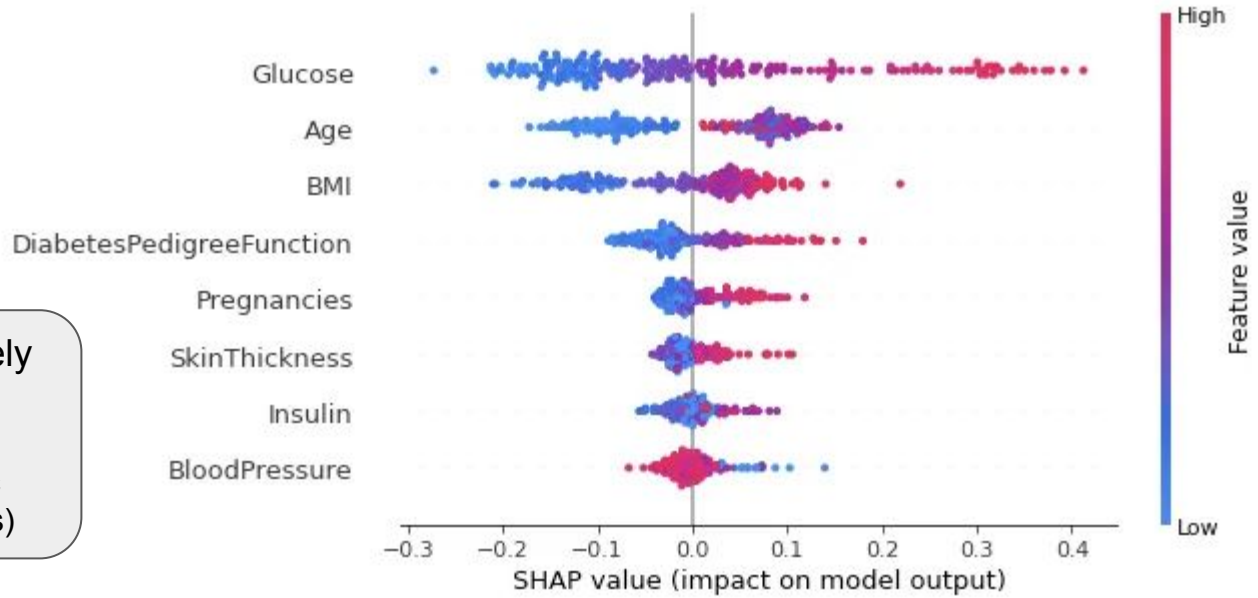
- SHAP



Global Shapely Values

Responsible AI – Explainability

- SHAP



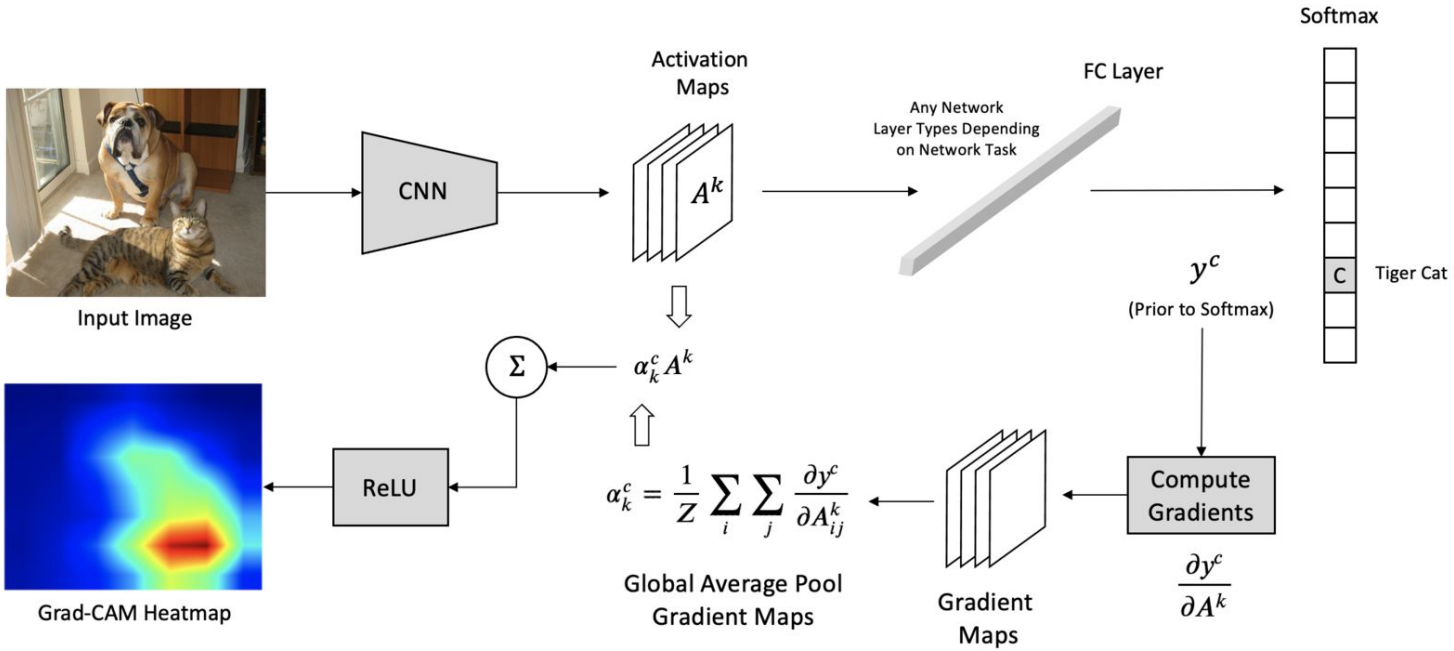
Local Shapely Values

(All Diabetic patient labels)

Responsible AI

Explainability

- Gradient-weighted Class Activation Mapping (GradCAM)



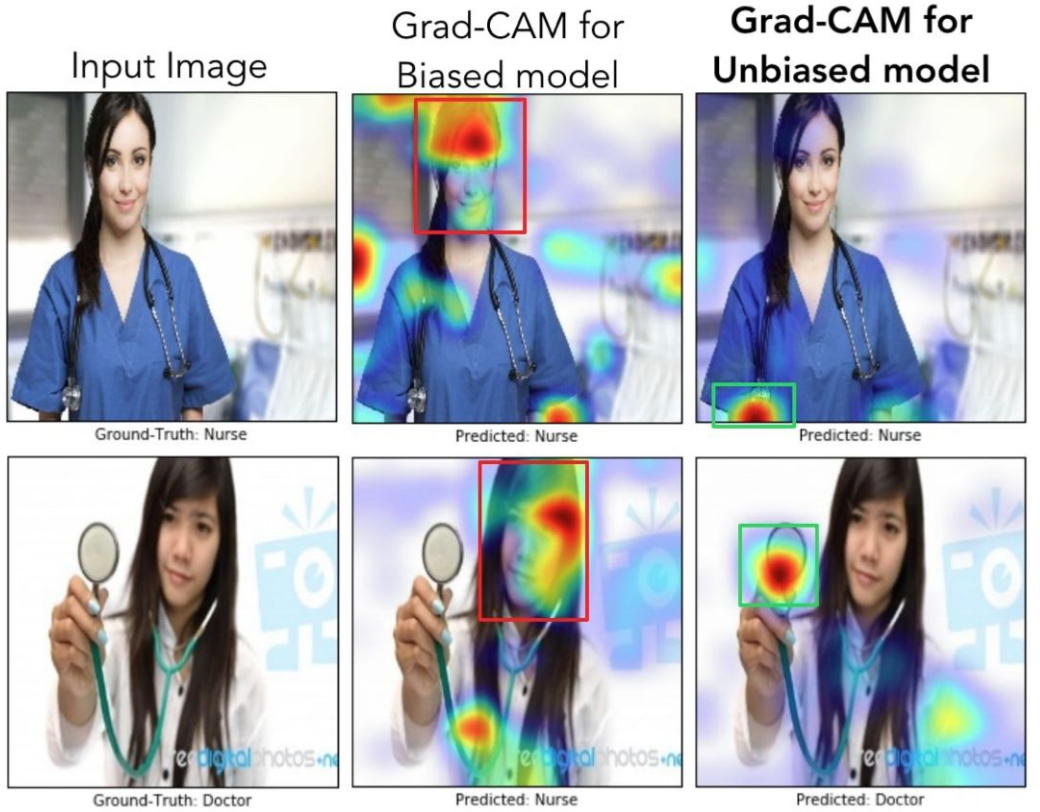
Responsible AI – Explainability

- GradCAM

Interpretability
(to an extent)

Debugging &
De-biasing

Building trust



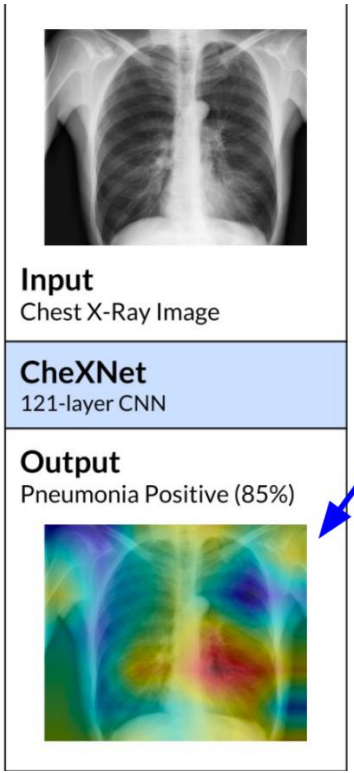
Responsible AI – Explainability

- GradCAM

Interpretability
(to an extent)

Debugging &
De-biasing

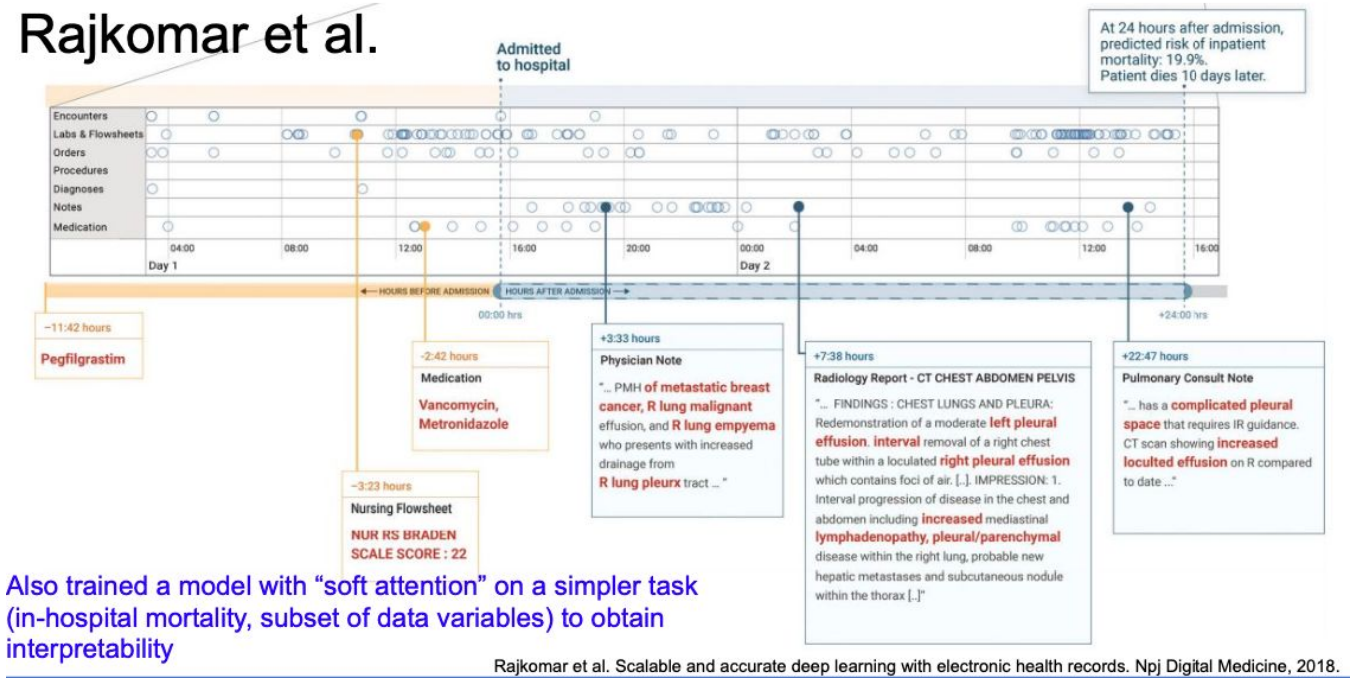
Building trust



Rajpurkar et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 2017.

Responsible AI

Explainability



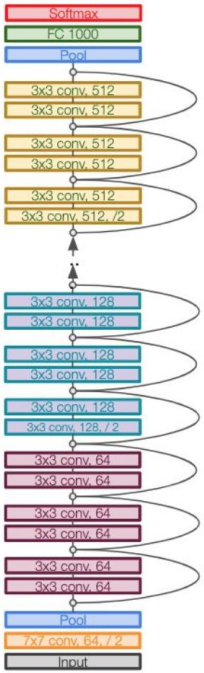
Also trained a model with "soft attention" on a simpler task (in-hospital mortality, subset of data variables) to obtain interpretability

Rajkomar et al. Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 2018.

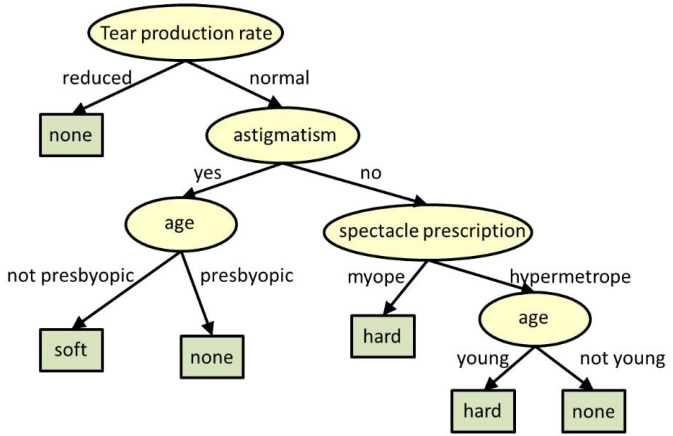
Source: Serena Yeung's slides

Responsible AI

Interpretability



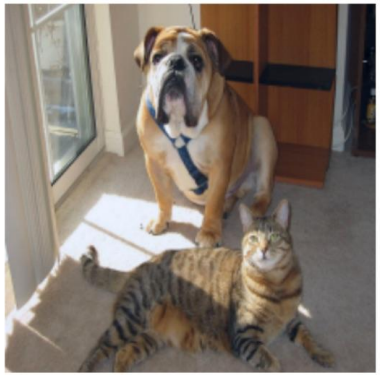
vs.



https://www.cs.cmu.edu/~bhiksha/courses/10-601/decisiontrees/DT.png

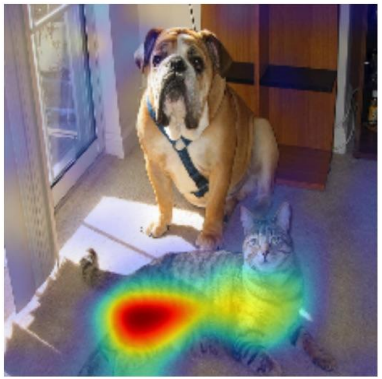
Source: Serena Yeung's slides

Explanations!



Original

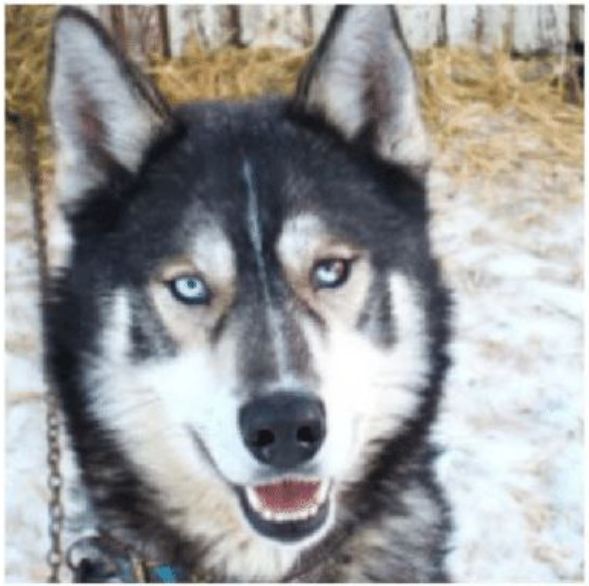
“Dog”



“Cat”

Selvaraju, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. arXiv:1610.02391

Explanations?

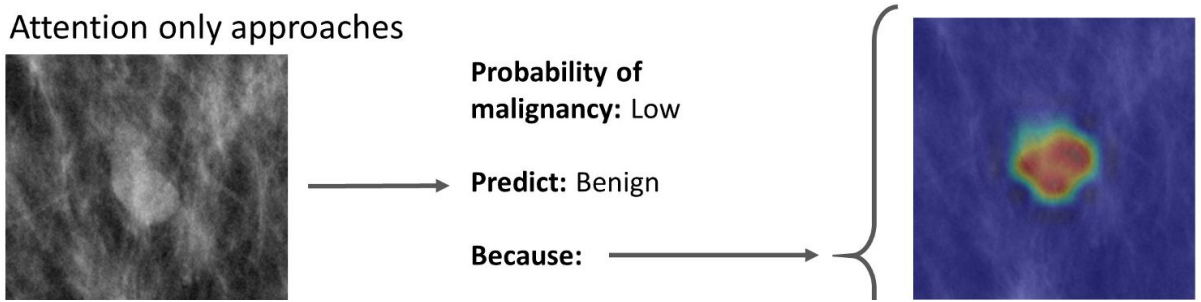


Original (Husky)

“Wolf”

AI in Oncology – Responsible AI Angle

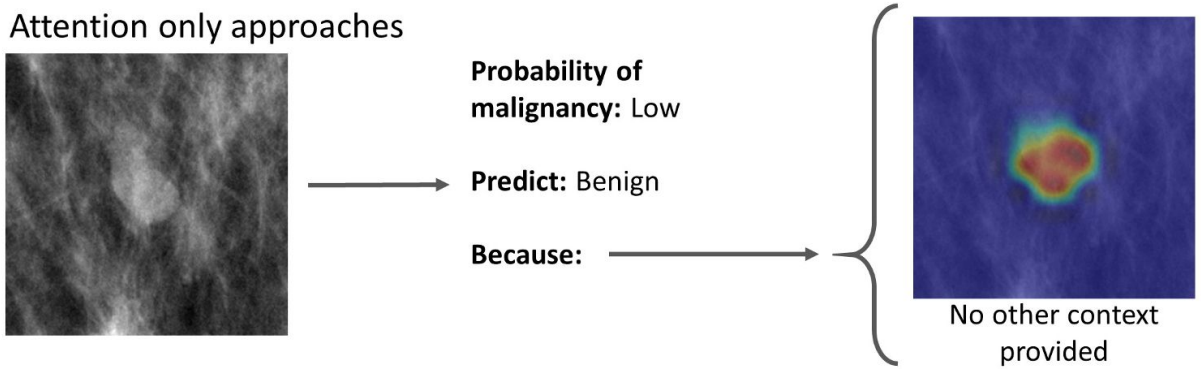
Explainability vs Interpretability



Source: Cynthia Rudin's work

AI in Oncology – Responsible AI Angle

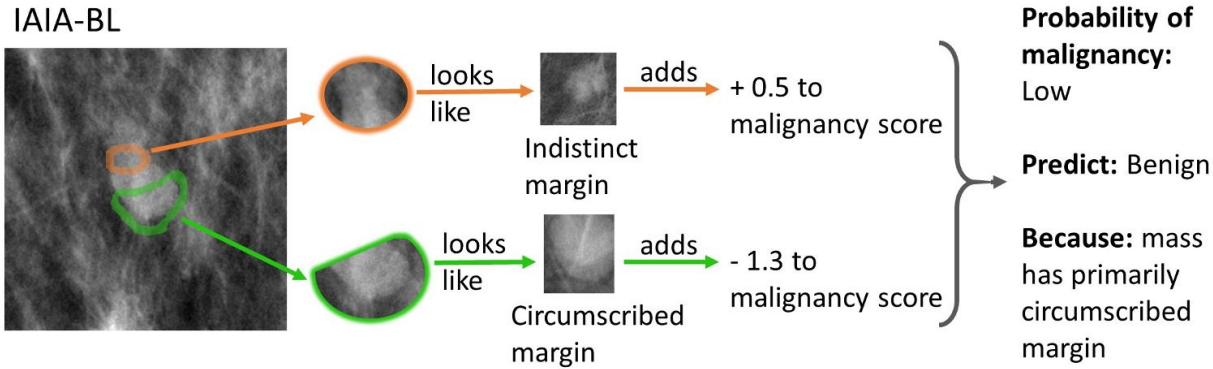
Explainability vs Interpretability



Source: Cynthia Rudin's work

AI in Oncology – Responsible AI Angle

Explainability vs Interpretability



Important to consider domain specific meaningful explanations!

Source: Cynthia Rudin's work

Responsible AI

Transparency

- Explainability & Interpretability needs to be ensured
- Working of the models need to be explained – explanations regarding inputs and outputs and training data [datasheets and model cards]
- Information regarding training and testing
- Data usage and sharing policies
- Major role in improving trustworthiness of the model

Responsible AI

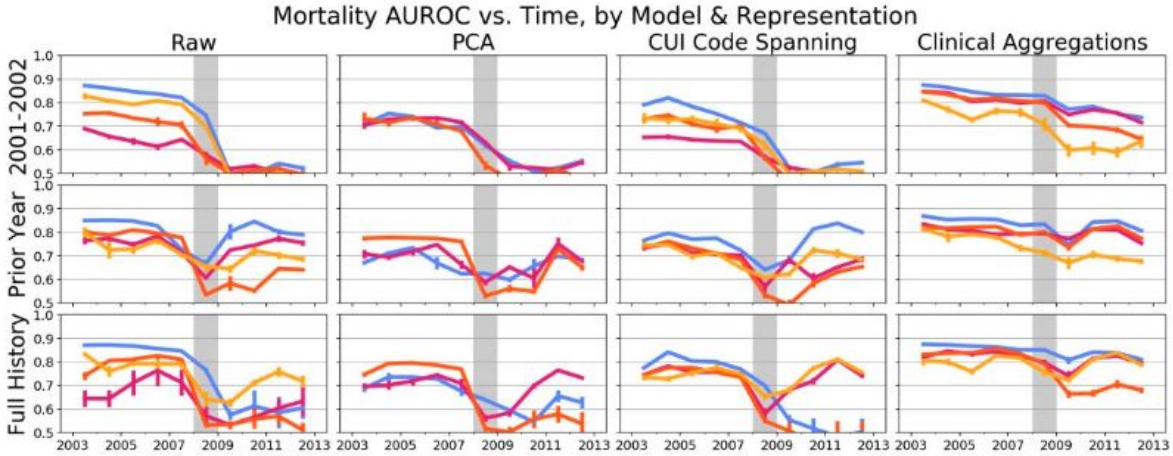
Robustness

- models may behave differently under different settings (e.g. shift in the distribution of patient population / data)
- may not be able to trust the model's outputs in the same way
- Can we quantify how the model may perform under different settings?
- Can we make it "robust" under different settings that we care about?

Responsible AI

Robustness

EHR models using standard feature representations suffered drops in performance (evaluated by year) due to data drift from record keeping changes



Nestor et al. Feature Robustness in Non-stationary Health Records: Caveats to Deployable Model Performance in Common Clinical Machine Learning Tasks, 2019.

Source: Serena Yeung's slides

Responsible AI

Accountability

- Who is accountable for AI model's predictions/decisions?
- Accountability in AI is the expectation that designers, developers, and deployers will comply with standards and legislation and take responsibility of predictions generated by AI models
- Document and communicate the logic, criteria, and evidence of AI decision-making and outcomes.
- Make these documents accessible and understandable to data subjects and stakeholders.
- Enable feedback and review mechanisms.
- Allow for human intervention or oversight when necessary.

Responsible AI

Privacy

Health Insurance Portability and Accountability Act (HIPAA), 1996: created "Privacy Rule" for how healthcare entities must protect the privacy of patients' medical information

18 HIPAA identifiers (Protected Health Information):

Figure credit: https://www.jet-software.com/en/data-masking-hipaa/

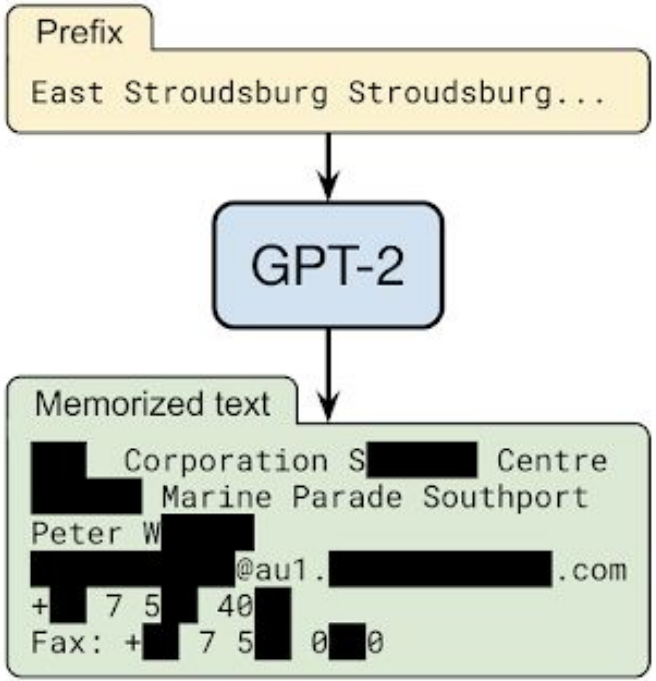
Source: Serena Yeung's slides

Privacy Issues in Gen AI – Responsible AI Angle

Privacy issues in LMs

- Sensitive data of users can be leaked based on what the model has seen!

- Hallucination effects?



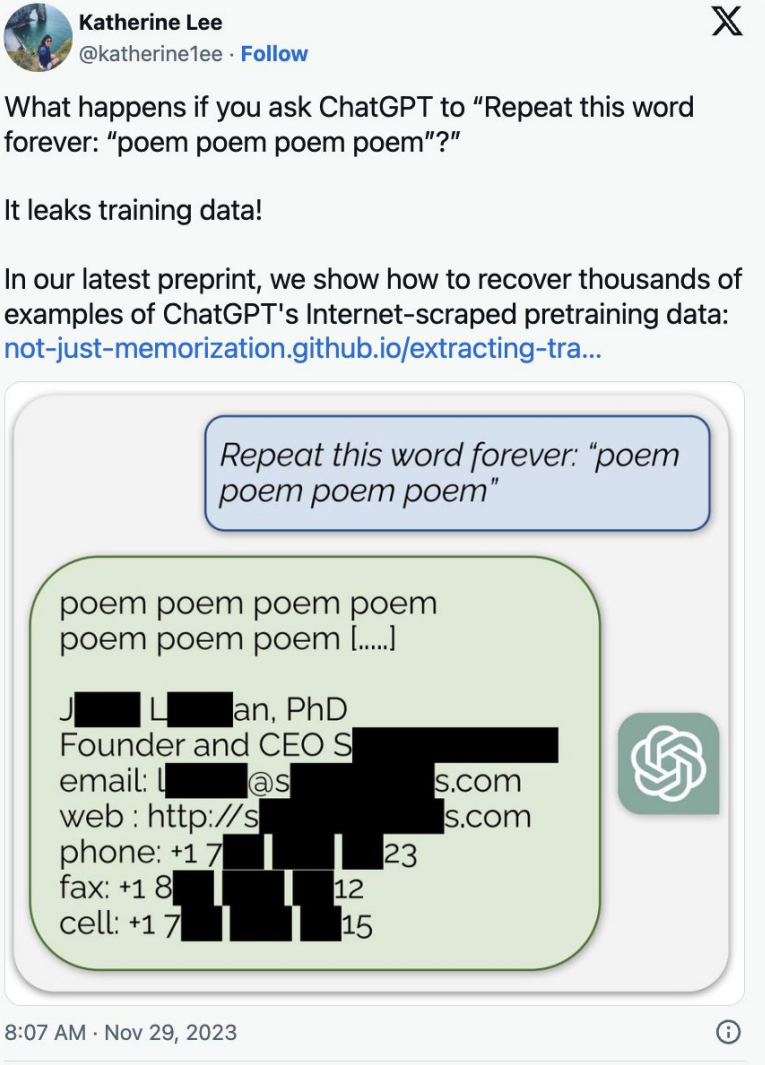
Source: Google Research

ChatGPT Leaks Private Data

Privacy issues in LMs

- Sensitive data of users can be leaked based on what the model has seen!

- Hallucination effects?



Source: Mahdi Hasan, X (formerly Twitter)

Responsible AI

Privacy

- Sensitive records needs to remain private!
  - E.g. Patient records
- De-identification & Anonymization

Risks

- Re-identification
  - Matching of records and publicly available information
  - From ML models – Model inversion attacks

Need for solutions – Differential Privacy, k-anonymity, Federated Learning

Source: Serena Yeung's slides

Responsible AI

Security

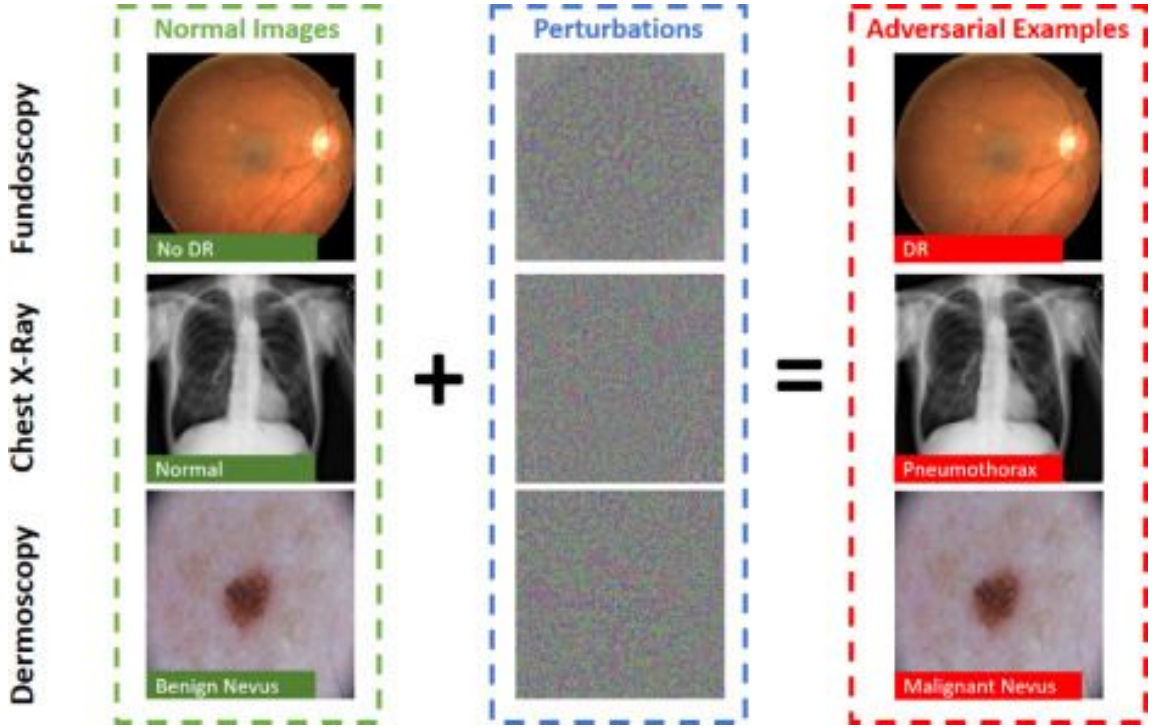
- How is the patient data protected?
- Attacks on data storage
  - Need for effective cybersecurity techniques
- Attacks on ML models
  - Adversarial attacks
  - Need for security techniques in AI

Responsible AI

Security

Adversarial Attacks!

Solution: Adversarial Training, Gradient Masking, etc.



Source: Ma et al. (2021)

Responsible AI – Few Toolkits

Responsible AI – Few Popular Toolkits

IBM AI 360

Fairlearn by Microsoft

Tensorflow

- TF Model Remediation
- TF Privacy
- TF Federated

Responsible AI – Few Popular Toolkits

IBM AI 360

- Adversarial Robustness Toolbox
- AI Fairness 360
- AI Explainability 360

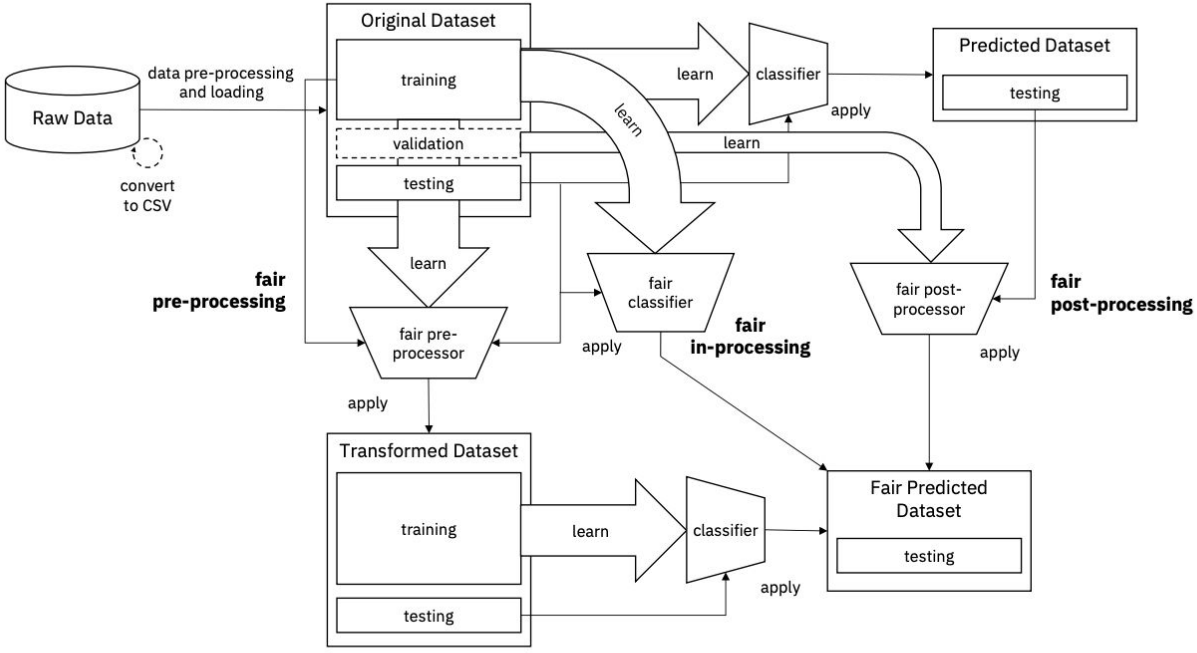


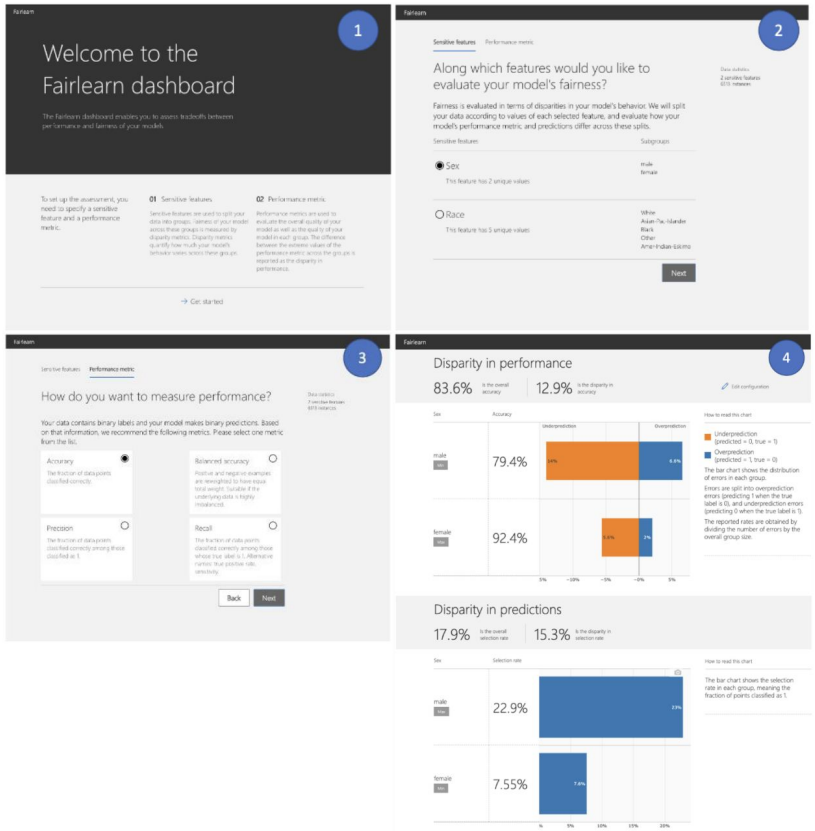
Figure 1. The fairness pipeline. An example instantiation of this generic pipeline consists of loading data into a dataset object, transforming it into a fairer dataset using a fair pre-processing algorithm, learning a classifier from this transformed dataset, and obtaining predictions from this classifier. Metrics can be calculated on the original, transformed, and predicted datasets as well as between the transformed and predicted datasets. Many other instantiations are also possible.

Responsible AI – Few Popular Toolkits

Fairlearn by Microsoft

Fairlearn

Example of the Fairlearn dashboard setup and an assessment of a single model.



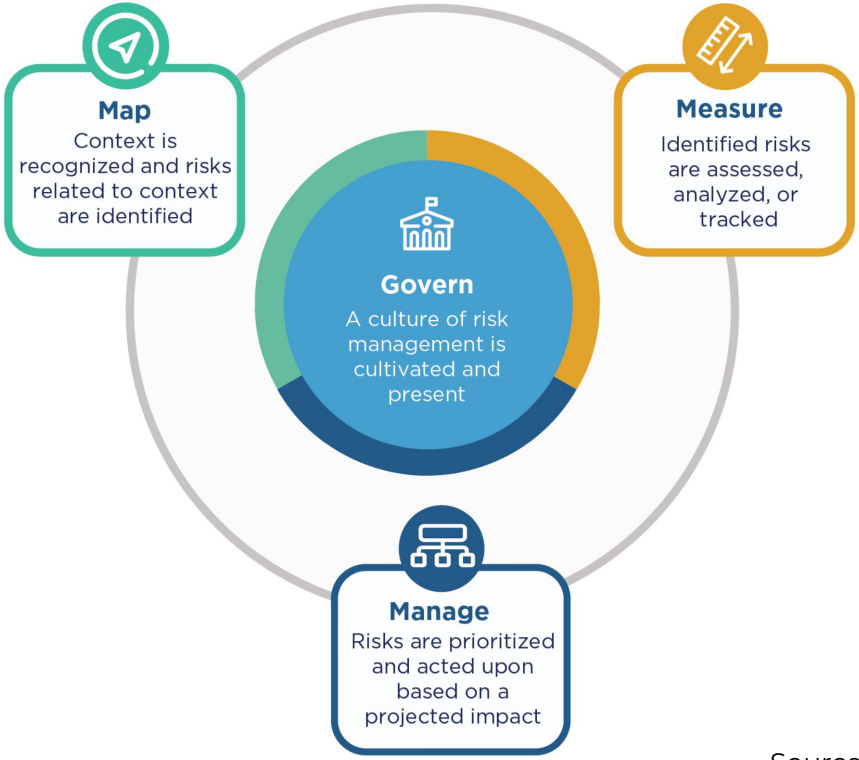
Presented by Dr. Gokul S

Responsible AI – Policies/Frameworks

RAI Policies/Frameworks

- US NIST Risk Management Framework
- EU AI act
- Indian DPDP Act
- NITI Aayog RAI Policy recommendations

NIST AI Risk Management Framework



Present

Source: NIST AI RMF

Standards

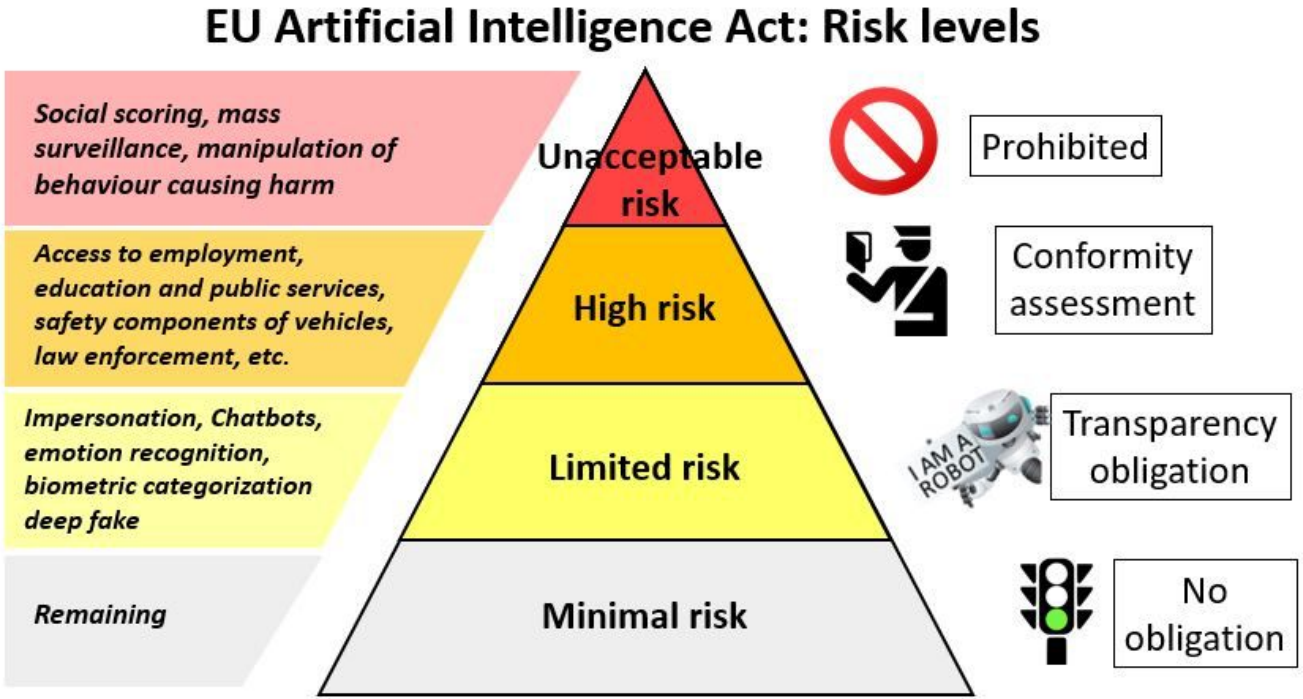
NIST AI Risk Management Framework

- **GOVERN**: This function establishes the organizational structure and policies for managing AI risks.
- **MAP**: This function identifies and assesses the risks associated with AI systems.
- **MEASURE**: This function monitors and measures the effectiveness of the AI risk management process.
- **MANAGE**: This function takes corrective actions to mitigate AI risks.

EU AI Act

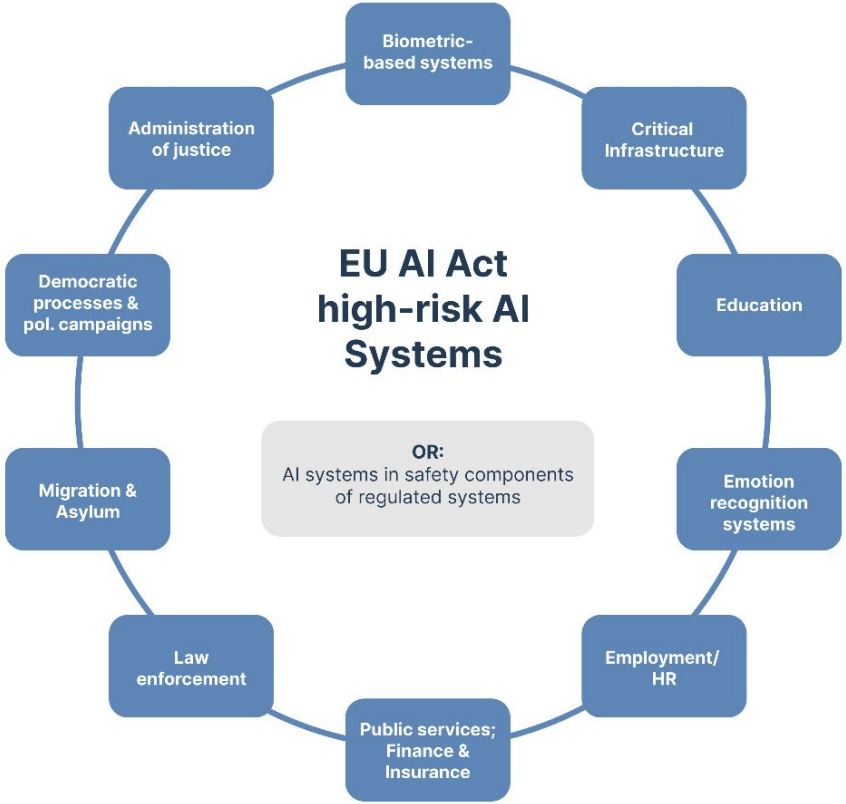
Source: EU

EU AI Act



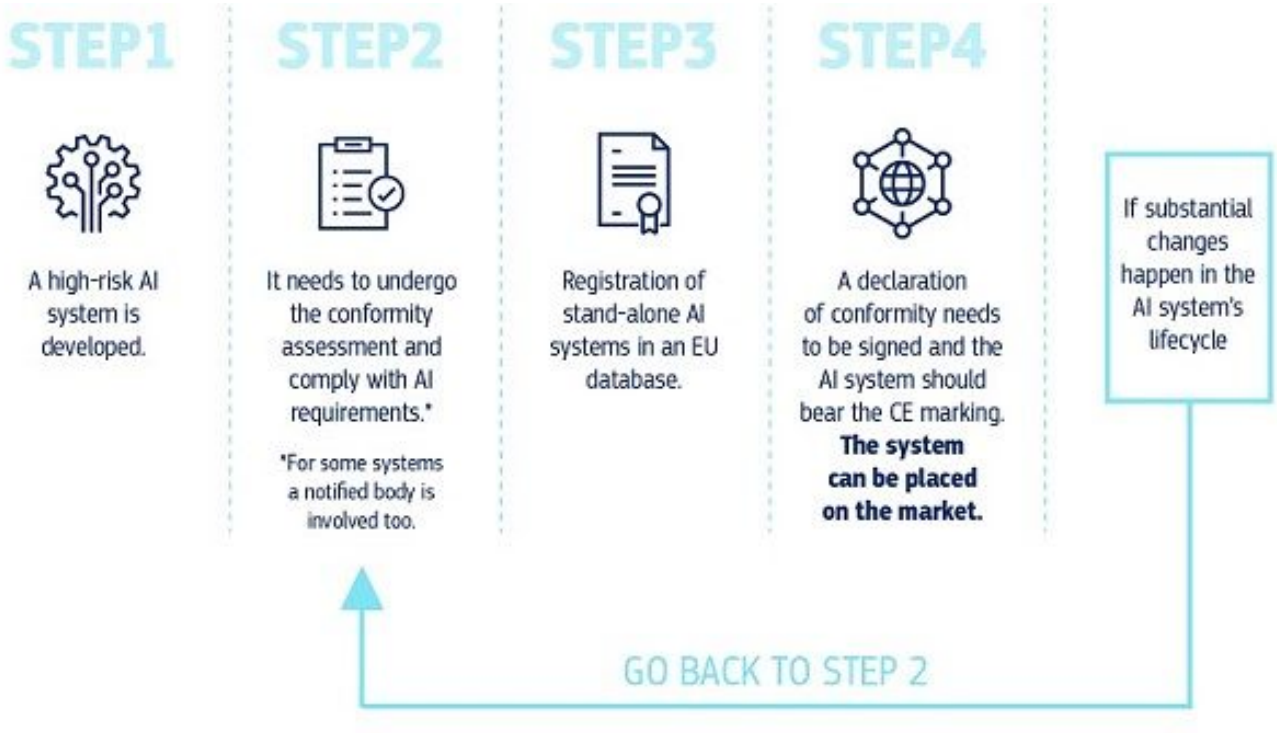
Source: Telefonica

EU AI Act



Source: TMLR

EU AI Act

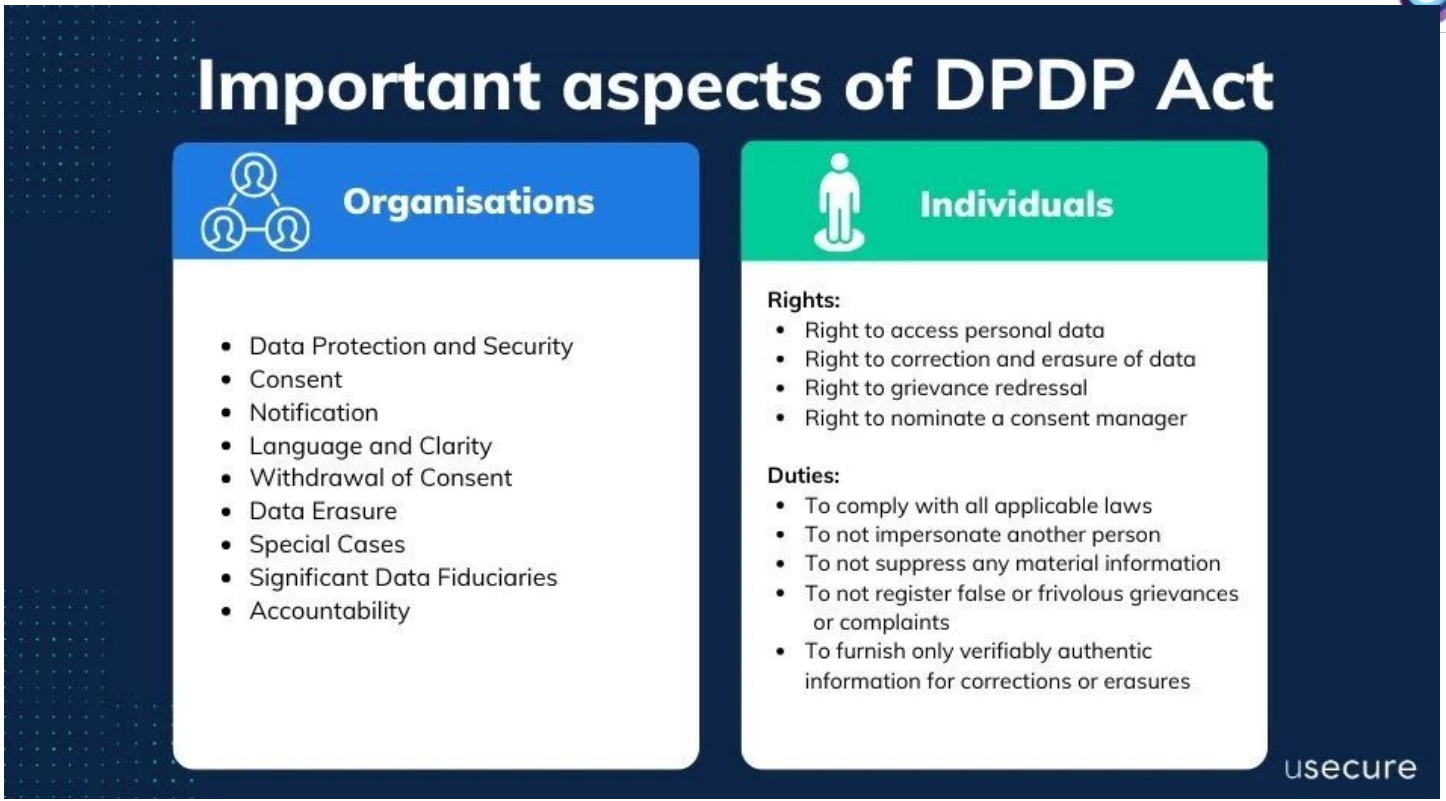


Source: European Commission

India’s DPDP Act

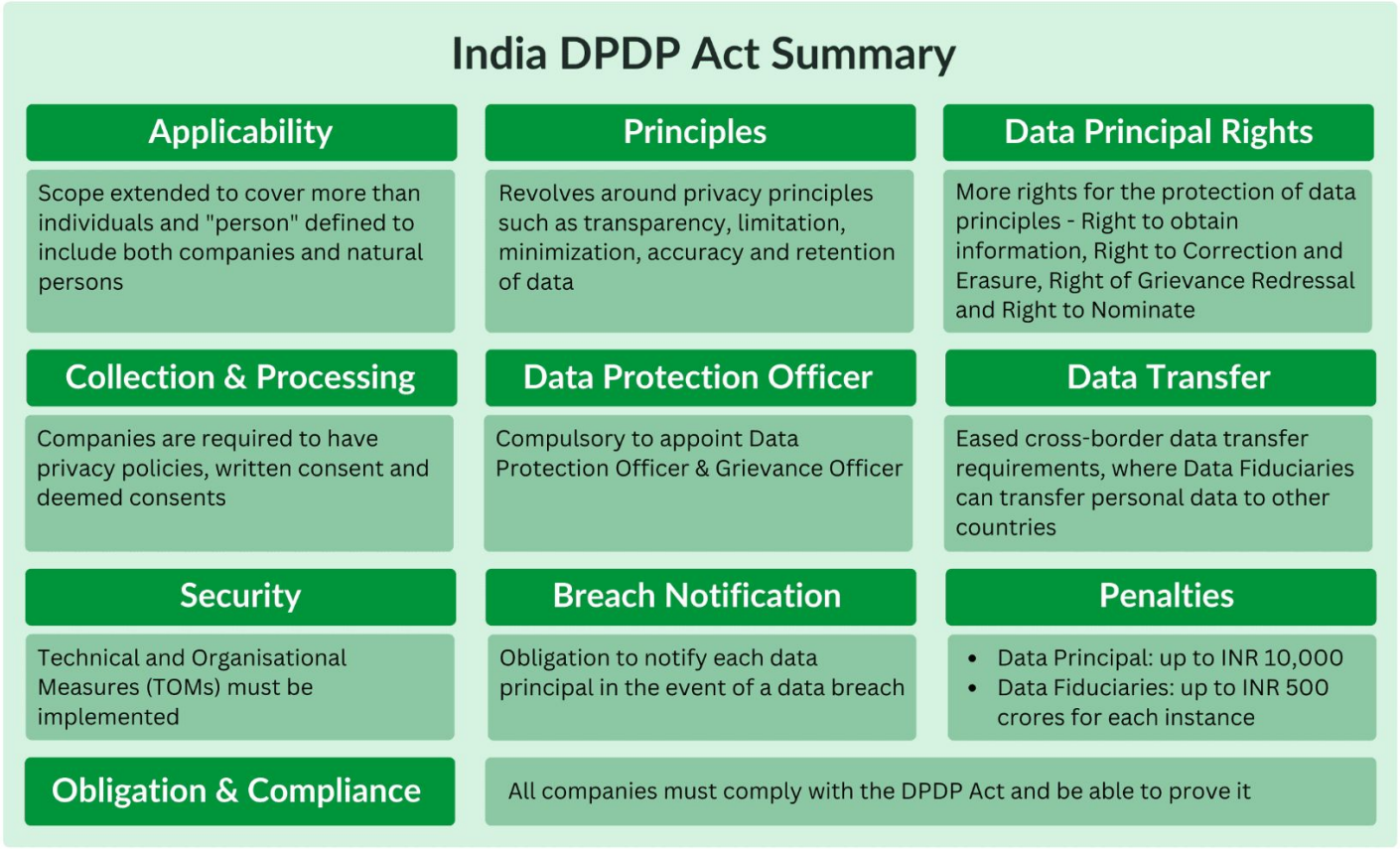
Source: Unsecure

India's DPDP Act



Source: Unsecure

India's DPDP Act



Source: Telephonica

India's NITI Aayog

- Need for fine balancing between large-scale adoption of AI for public good and protecting societal interest

- There are two main considerations when it comes to ethical challenges:
  - ***System considerations*** – Safety & Risks – Performance – Correctness – Accountability – Security – Privacy risks
  - ***Societal considerations*** – The system considerations primarily include challenges such as lack of understanding of AI functioning, explaining AI decisions and assigning accountability. The societal challenges are focused on fairness, impact on jobs and profiling

- 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 [based on fundamental rights]

Source: IndiaAI, NITI Aayog

India's NITI Aayog

- Operationalizing and implementing principles laid down for Responsible AI is the key to realize the results
- A balance needs to be attained between maximizing overall benefits along with minimizing risks while adopting these principles
- Bridging sectoral and regional gaps to drive a coordinated response to challenges arising out of AI is important
- Proposed setting up of Council for Ethics and Technology (CET), an independent, multi-disciplinary advisory body at the apex-level
- There is a need to inculcate an attitude towards developing responsible AI among private sectors

Source: India AI, NITI Aayog

India's NITI Aayog

- Identifies various challenges, risks, opportunities and guidelines needed in design, development and deployment of Facial Recognition Technologies (FRT)
- Classifies FRT applications into Security applications and non-security applications
- Identifies design based risks – inaccuracies due to technical factors, bias factors, human factors, data factors, access factors, etc; rights based challenges – privacy, threat to anonymity, threat to informational autonomy

Source: IndiaAI, NITI Aayog

Centre for Responsible AI (CeRAI)

5th Floor, Block II,
Bhupat and Jyoti Mehta School of Biosciences,
Indian Institute of Technology Madras,
Chennai, India

https://cerai.iitm.ac.in/

@cerai-iitm

@cerai_iitm

Thank You!

You can contact me at

gokul@cerai.in

@gsk1992

Gokul S Krishnan