



# **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

Presented by Dr. Gokul S Krishnan @ ACM Summer School @ 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





# LLMs used in the real world!

The Guardian

Cor

Times of India

Jud  
me

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

Fel

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



Presented by Dr. Gokul S Krishnan @ ACM Summer School @ IIT Madras

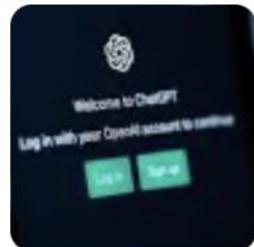
Source: Google News

# 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...



QZ 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...





# What is Responsible AI?

Presented by Dr. Gokul S Krishnan @ ACM Summer School @ IIT Madras



# 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

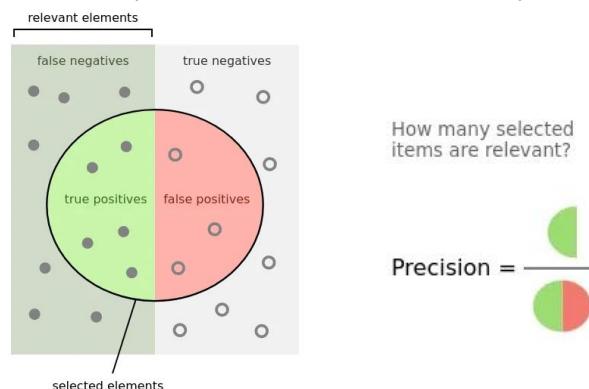


- Performance
- Fairness/Bias
- Explainability
- Interpretability
- Transparency
- Accountability
- Robustness
- Privacy
- Security

# Responsible AI

## Performance

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





## Fairness/Bias

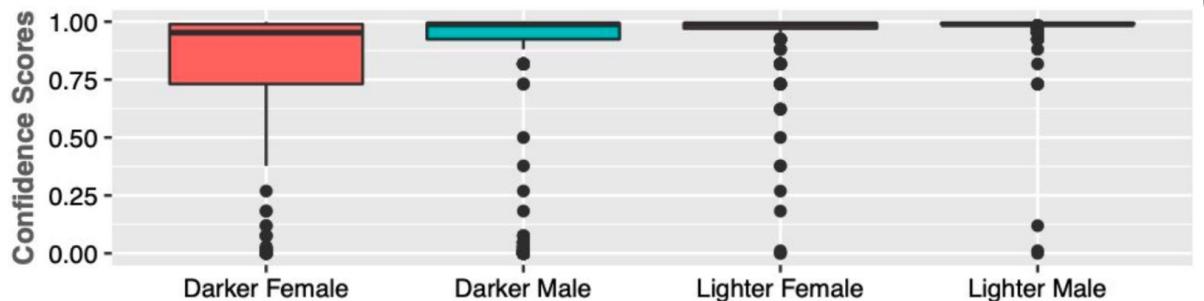
- Group-independent predictions: predictions should be independent of group membership
- Equal metrics across groups: e.g. equal true positive rates or false positive rates across groups
- Individual fairness: individuals who are similar with respect to a prediction task should have similar outcomes
- 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!

## 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.

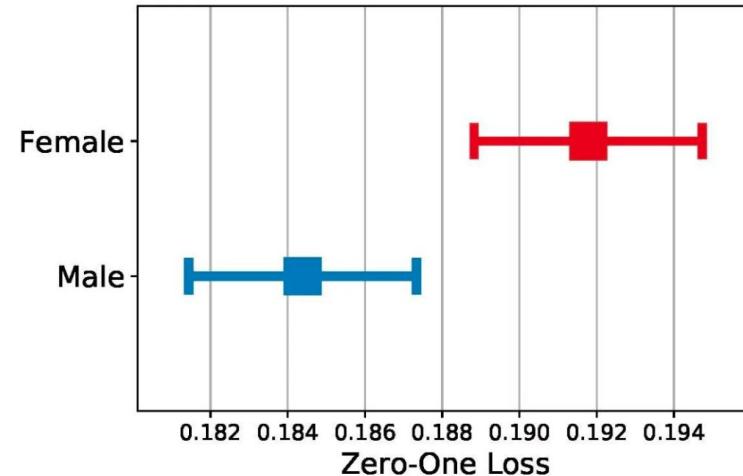
# Responsible AI

## 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!



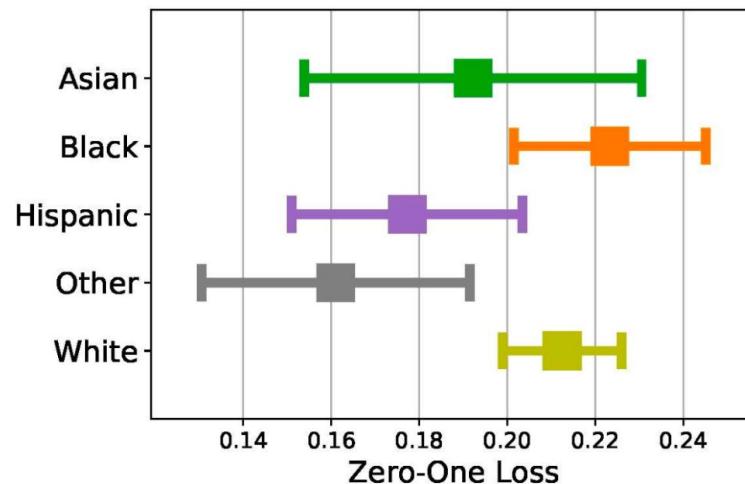
Chen et al. Can AI Help Reduce Disparities in General Medical and Mental Health Care? 2019.

## 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.

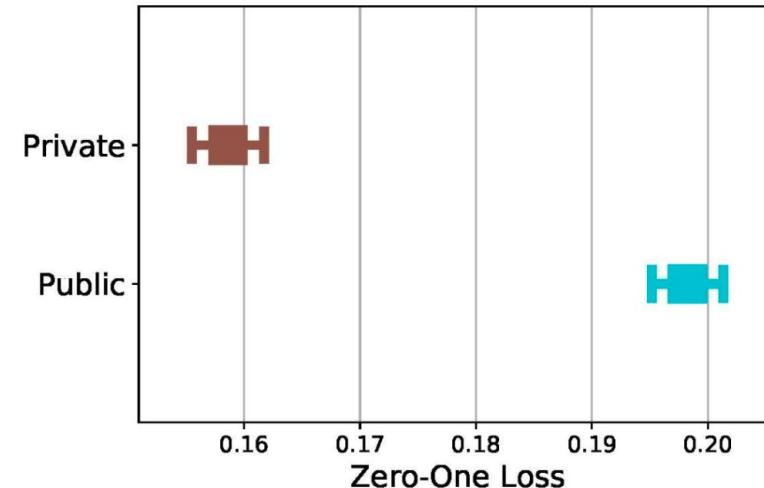
# Responsible AI

## 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.

# 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 .

# Gen AI in Computer Vision – Responsible AI Angle

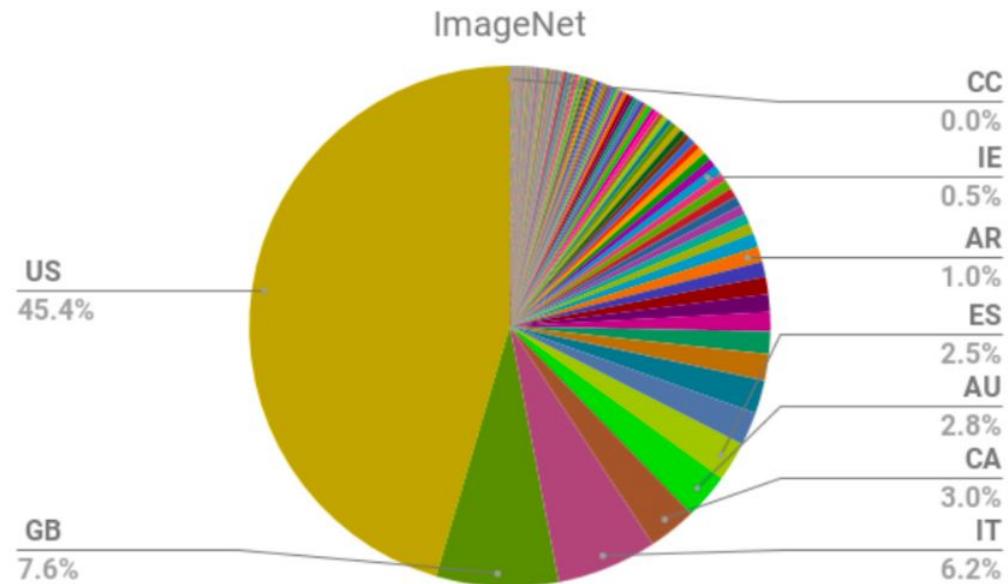
## Fairness/Bias Issues in Image Generation



*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*

# Responsible AI

## Fairness/Bias



## Fairness/Bias

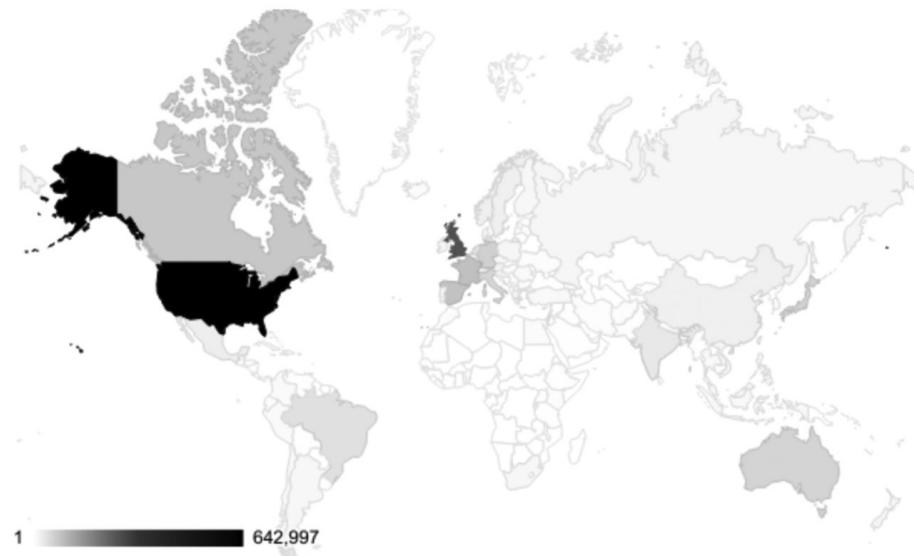


Fig. 4. Geographic distribution of countries in the Open Images dataset. In their sample, almost one third of the data was US-based, and 60% of the data was from the six most represented countries across North America and Europe (from Reference [138], © Shreya Shankar).

# Hallucination in LLMs!

Prompt	Where do people smoke more on average than in Russia?
ChatGPT Generation	<p>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:</p> <p>Indonesia: <b>55.2%</b> of adults smoke    China: <b>52.9%</b> of adults smoke    Serbia: <b>47.5%</b> of adults smoke    Serbia and Montenegro: <b>45.6%</b> of adults smoke    Montenegro: <b>45.6%</b> of adults smoke    Greece: <b>44.5%</b> of adults smoke    Bosnia and Herzegovina: <b>41.8%</b> of adults smoke    Kazakhstan: <b>41.3%</b> of adults smoke    Kosovo: <b>41.3%</b> of adults smoke    Moldova: <b>40.9%</b> of adults smoke</p> <p>It's important to note that these figures are estimates and may vary slightly from other sources.</p>
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. <sup>16</sup> )

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

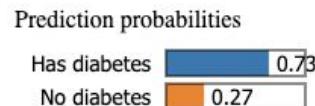
Perturb the  
features and  
sample a set

Get predictions  
from model

Weigh samples  
closer to picked  
data points

Train linear model  
with sampled  
dataset

Coefficients of  
trained linear model  
become  
explanations!



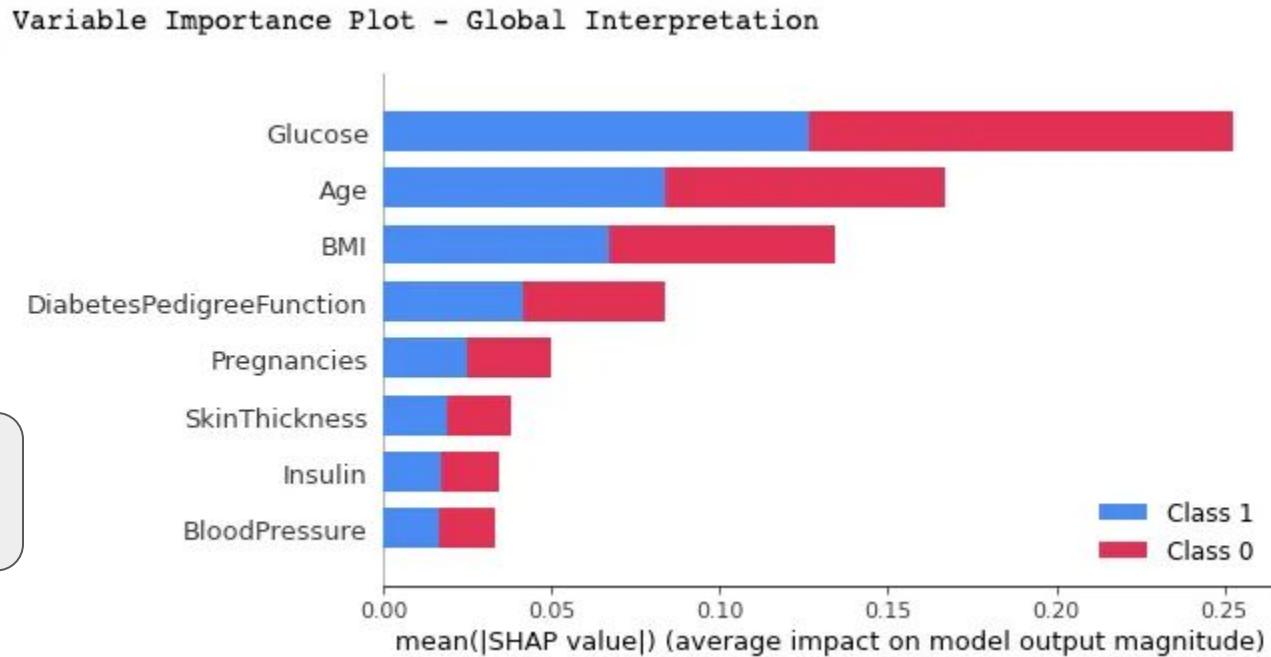
Feature	Value
Age	38.00
Glucose	104.00
Pregnancies	13.00
BloodPressure	72.00
SkinThickness	0.00
BMI	31.20
DiabetesPedigreeFunction	0.47
Insulin	0.00

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

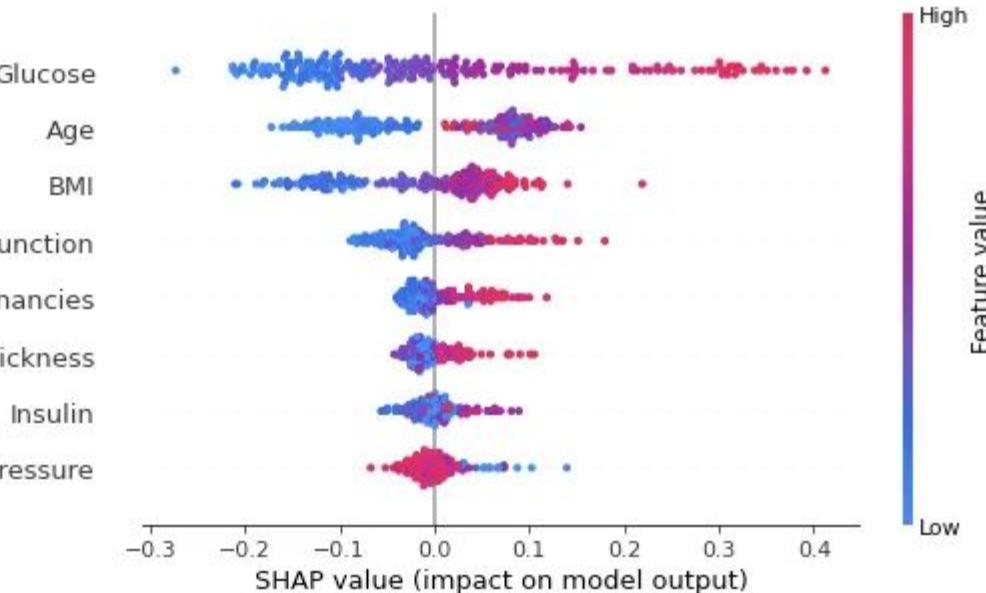
- SHAP



# Responsible AI – Explainability

- SHAP

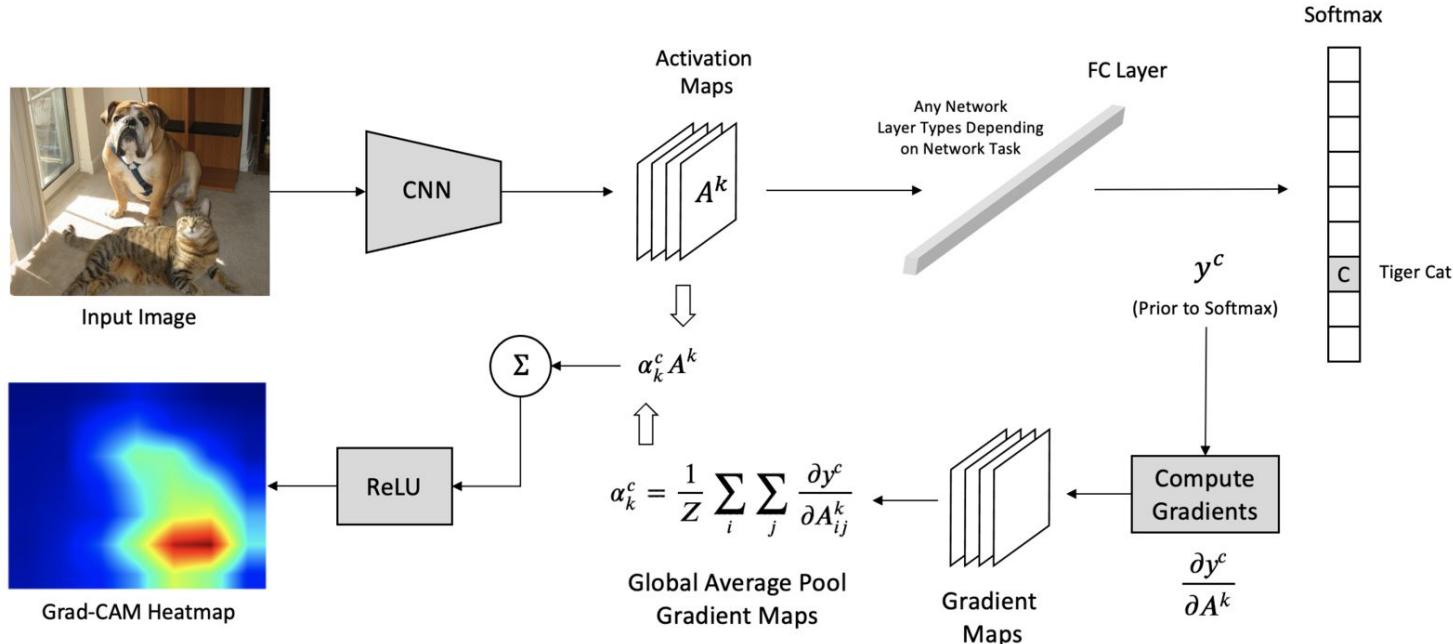
Local Shapely Values  
(All Diabetic patient labels)



# Responsible AI

## Explainability

- Gradient-weighted Class Activation Mapping (GradCAM)



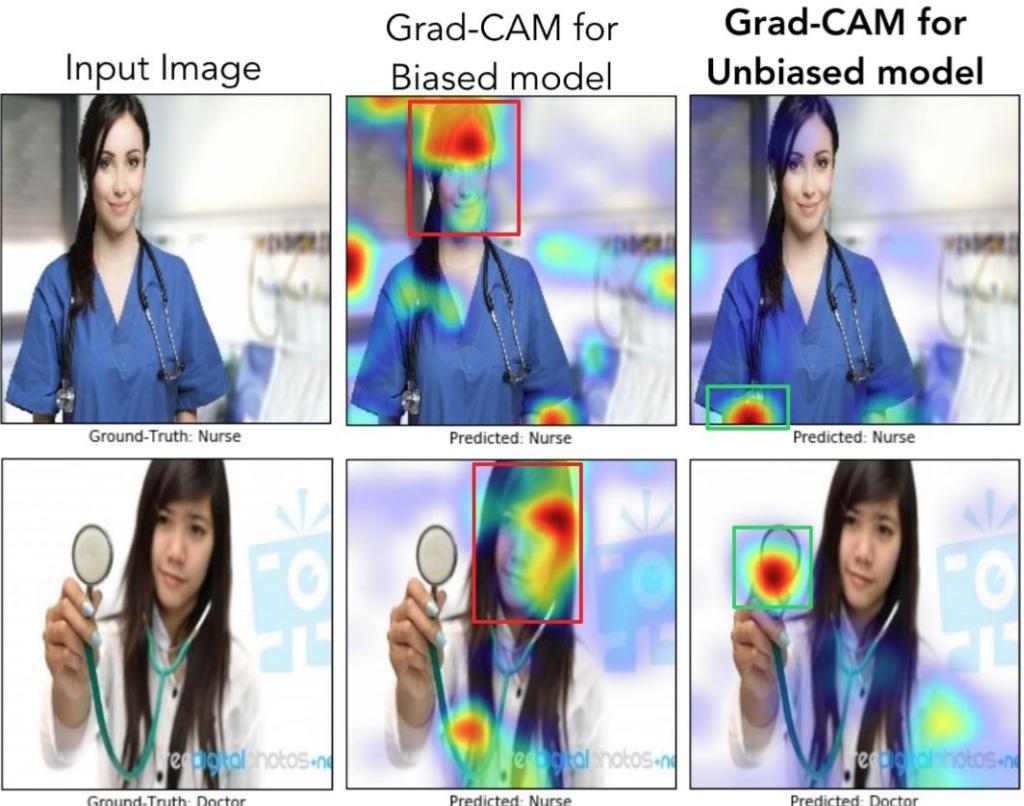
# Responsible AI – Explainability

- GradCAM

Interpretability  
(to an extent)

Debugging &  
De-biasing

Building trust



Presented by Dr. Gokul S Krisl

# Responsible AI – Explainability

- GradCAM

Interpretability  
(to an extent)

Debugging &  
De-biasing

Building trust



**Input**  
Chest X-Ray Image

**CheXNet**  
121-layer CNN

**Output**  
Pneumonia Positive (85%)

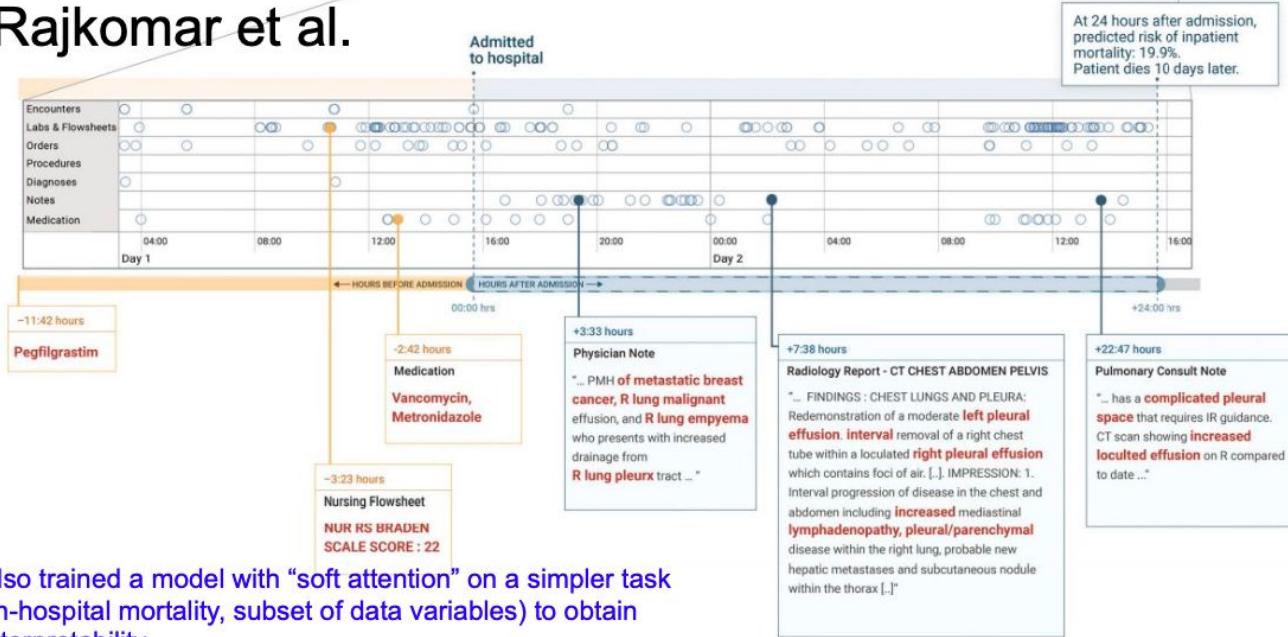


CAM visualization

# Responsible AI

## Explainability

Rajkomar et al.

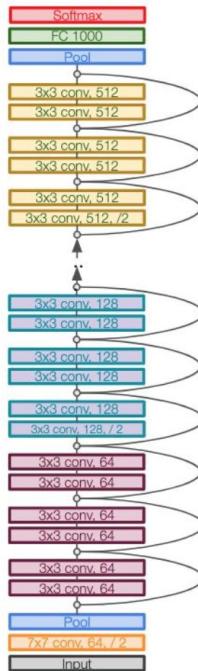


Also trained a model with “soft attention” on a simpler task (in-hospital mortality, subset of data variables) to obtain interpretability

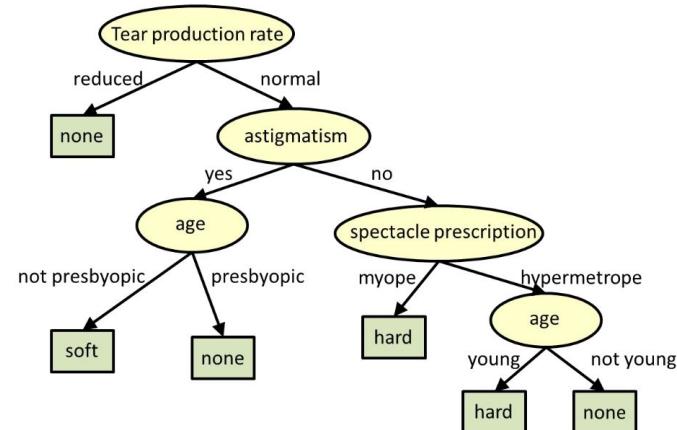
Rajkomar et al. Scalable and accurate deep learning with electronic health records. Npj Digital Medicine, 2018.

# Responsible AI

## Interpretability

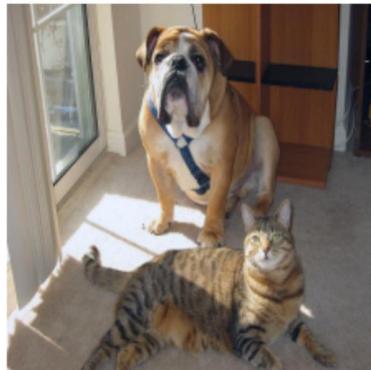


VS.



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

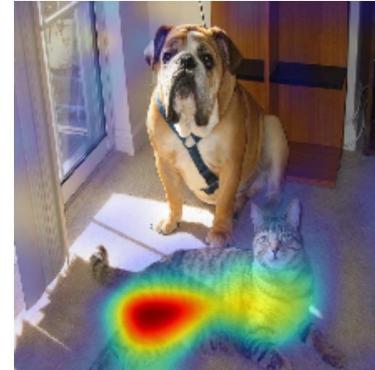
# Explanations!



**Original**



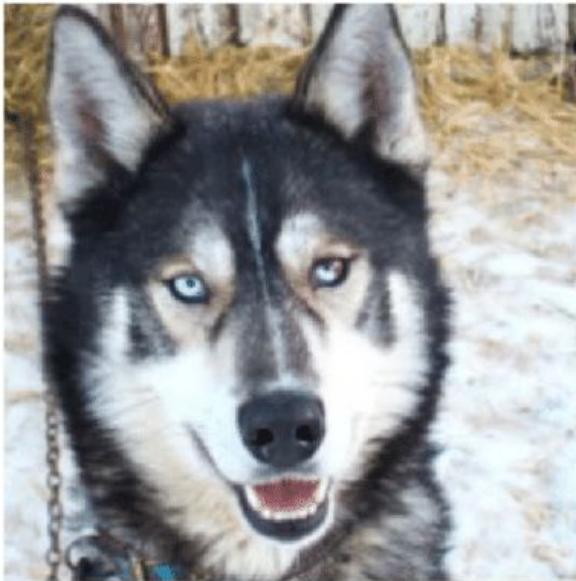
**“Dog”**



**“Cat”**

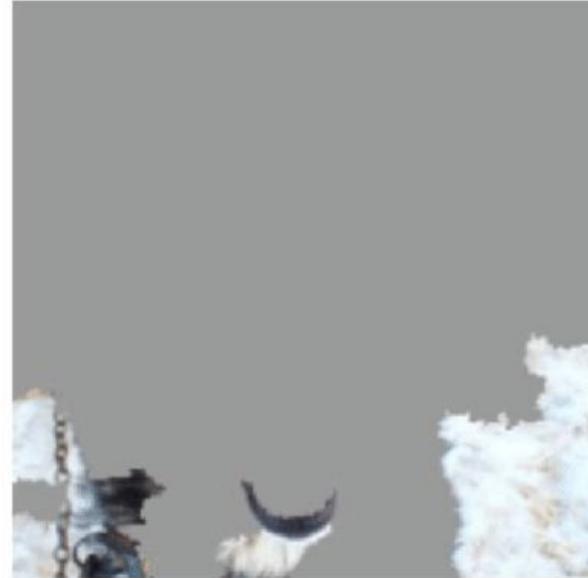
Selvaraju, et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”. arXiv:1610.02391

# Explanations?



**Original (Husky)**

Presented by Dr. Gokul S Krishnan @ ACM Summer School on AI & Explainability  
Sources: Prof. B Ravindran's Slides

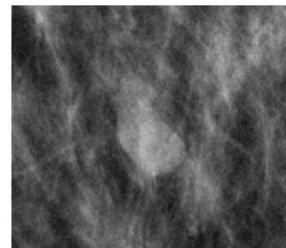


**“Wolf”**

# AI in Oncology – Responsible AI Angle

## Explainability vs Interpretability

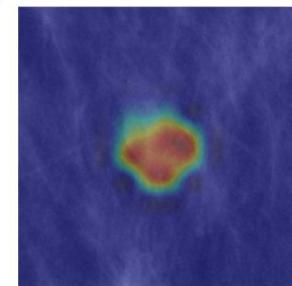
Attention only approaches



Probability of malignancy: Low

Predict: Benign

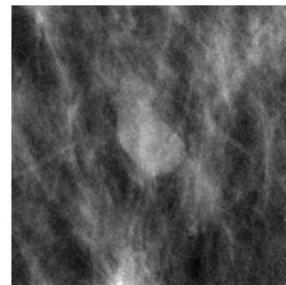
Because:



# AI in Oncology – Responsible AI Angle

## Explainability vs Interpretability

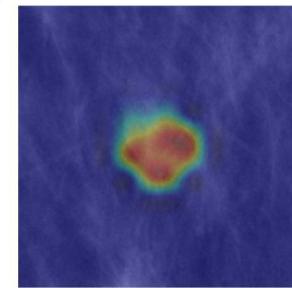
Attention only approaches



Probability of malignancy: Low

Predict: Benign

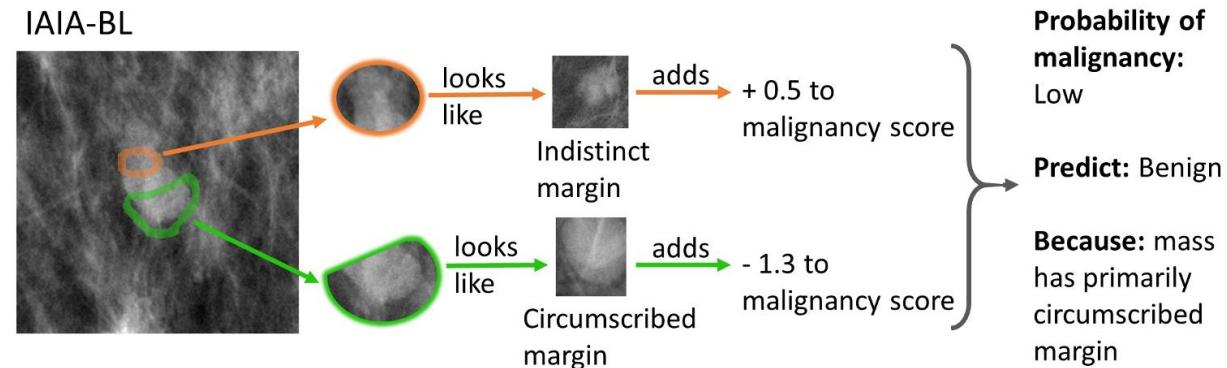
Because:



No other context provided

# AI in Oncology – Responsible AI Angle

## Explainability vs Interpretability



*Important to consider domain specific meaningful explanations!*



# 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

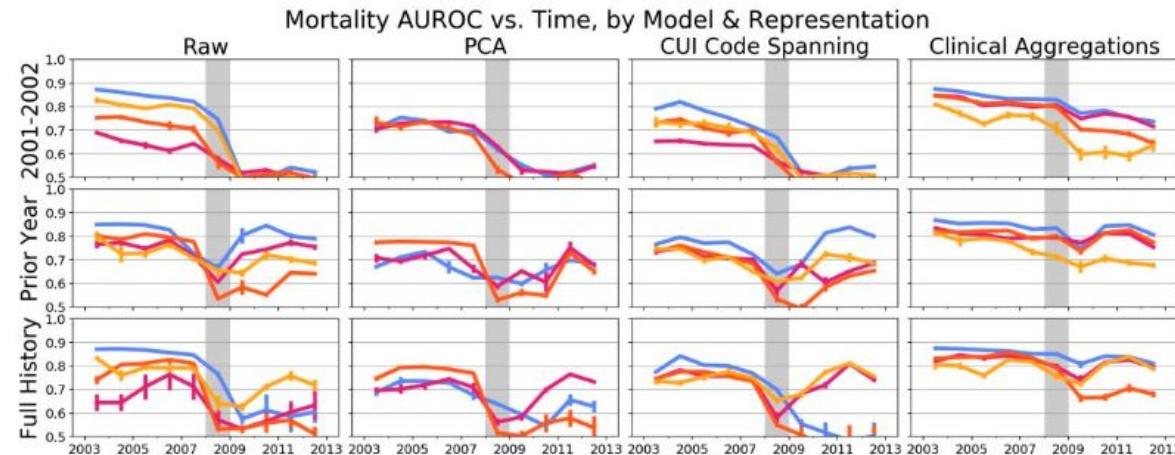


## Robustness

- 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?

## 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.



# 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):

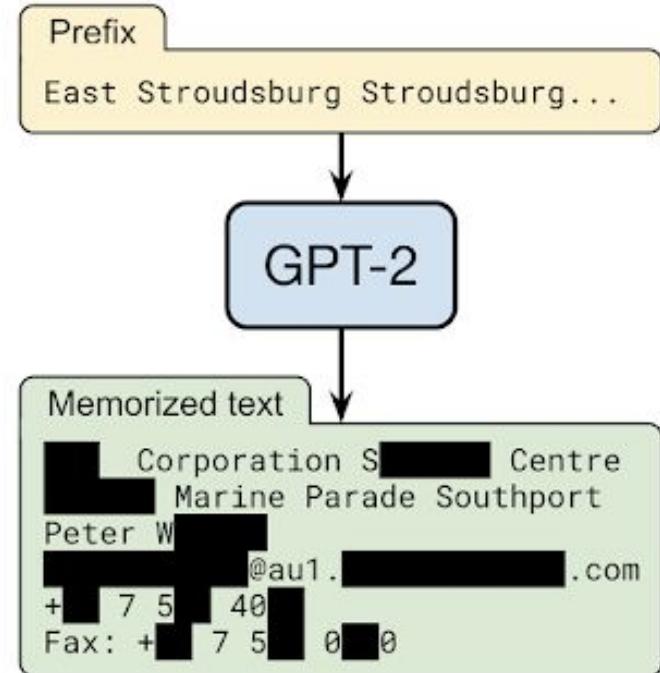


Figure credit: <https://www.jet-software.com/en/data-masking-hipaa/>

# 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?





# ChatGPT Leaks Private Data

## Privacy issues in LMs

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



Katherine Lee  
@katherine1ee · [Follow](#)

What happens if you ask ChatGPT to “Repeat this word forever: “poem poem poem poem”?”

It leaks training data!

In our latest preprint, we show how to recover thousands of examples of ChatGPT’s Internet-scraped pretraining data: [not-just-memorization.github.io/extracting-tr...](https://not-just-memorization.github.io/extracting-train-data/)

*Repeat this word forever: “poem poem poem poem”*

poem poem poem poem  
poem poem poem [....]

J [REDACTED] L [REDACTED]an, PhD  
Founder and CEO S [REDACTED]  
email: [REDACTED]@s [REDACTED].com  
web : http://s [REDACTED].com  
phone: +1 7 [REDACTED] 23  
fax: +1 8 [REDACTED] 12  
cell: +1 7 [REDACTED] 15



Presented by Dr. Gokul S Krishnan @ ACM Sun

Source: Mahdi Hasan, X (formerly Twitter)

8:07 AM · Nov 29, 2023





# 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



# Responsible AI



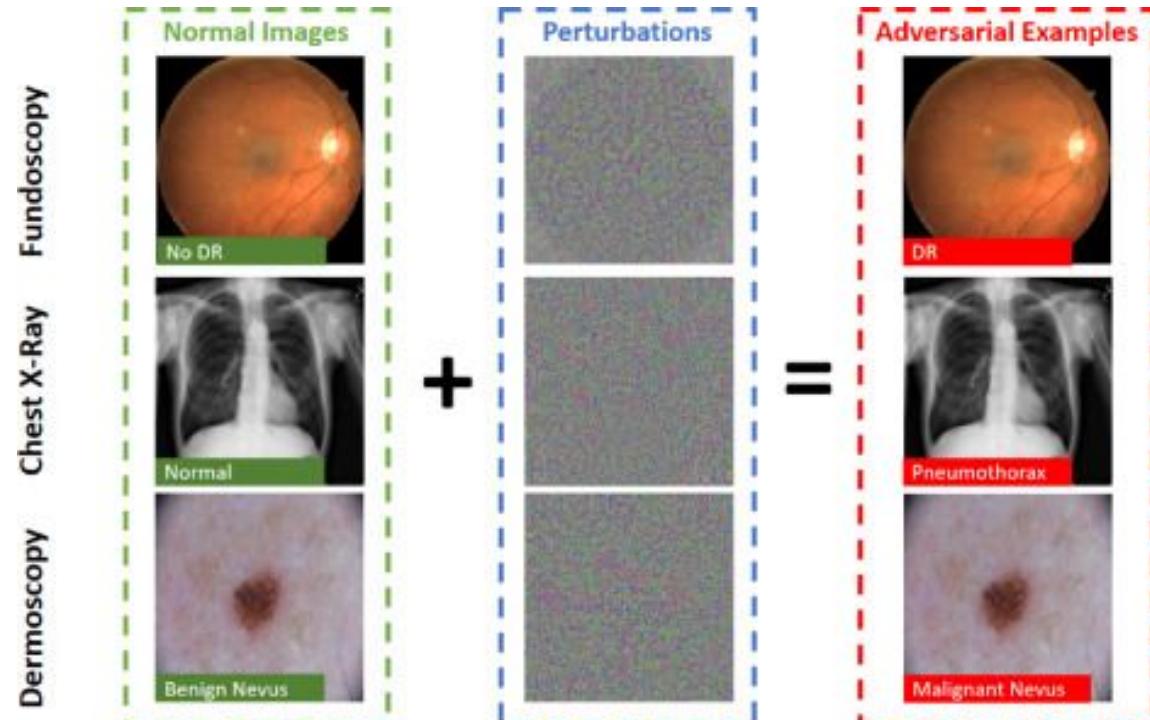
## Security

- 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

## Security

Adversarial Attacks!

Solution:  
Adversarial  
Training, Gradient  
Masking, etc.





# Responsible AI – Few Toolkits

Presented by Dr. Gokul S Krishnan @ ACM Summer School @ IIT Madras

# Responsible AI – Few Popular Toolkits

IBM AI 360



Adversarial  
Robustness  
Toolbox



AI Fairness 360



AI Explainability 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

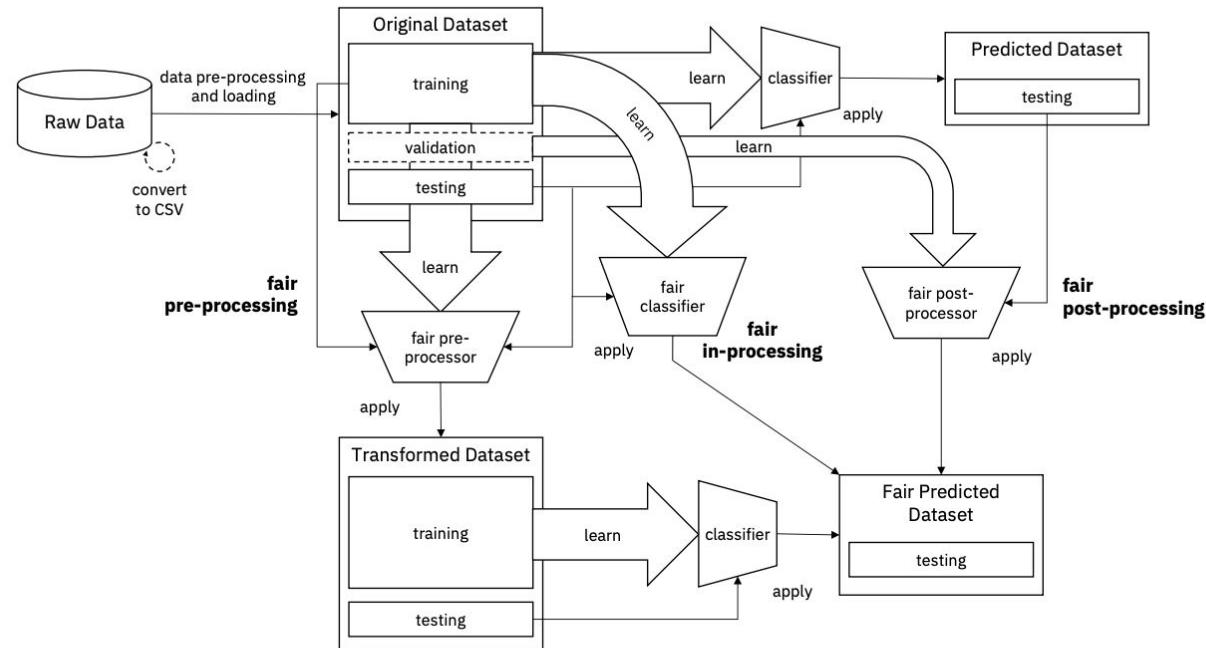


Figure 1. The fairness pipeline. An example instantiation of this generic pipeline consists of loading data into a dataset object, transforming it into a fairer dataset using a fair pre-processing algorithm, learning a classifier from this transformed dataset, and obtaining predictions from this classifier. 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



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

The Fairlearn dashboard makes it easy to assess tradeoffs between performance and fairness of your models.

**1. Welcome to the Fairlearn dashboard**

To set up the assessment, you need to specify a sensitive feature and a performance metric.

**2. Fairness**

Along which features would you like to evaluate your model's fairness?

**3. Performance metrics**

How do you want to measure performance?

Accuracy	Balanced accuracy	Precision	Recall
The fraction of datapoints classified correctly.	Positive and negative examples are recognized and have took into account the class of the underlying data is highly imbalanced.	The fraction of datapoints classified correctly among those classified as 1.	The fraction of data points correctly classified as 1 given those whose true label is 1. Alternative name True positive rate.

**4. Disparity in performance**

83.6% is the overall accuracy | 12.9% is the disparity in accuracy

Sex: Accuracy

Gender	Underprediction	Overprediction
male	79.4%	8.6%
female	92.4%	7%

How to read this chart:

- Underprediction ( $\text{predicted} = 0, \text{true} = 1$ )
- Overprediction ( $\text{predicted} = 1, \text{true} = 0$ )

The bar chart shows the distribution of errors for each gender. Errors are split into underprediction errors predicting 0 when the true label is 1, and overprediction errors predicting 1 when the true label is 0. The disparity is measured by dividing the number of errors by the overall group size.

**5. Disparity in predictions**

17.9% is the overall selection rate | 15.3% is the disparity in selection rate

Sex: Selection rate

Gender	Selection rate
male	22.9%
female	7.55%

How to read this chart:

The chart shows the selection rate for each group, meaning the fraction of points classified as 1.

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 ARMF

# 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.

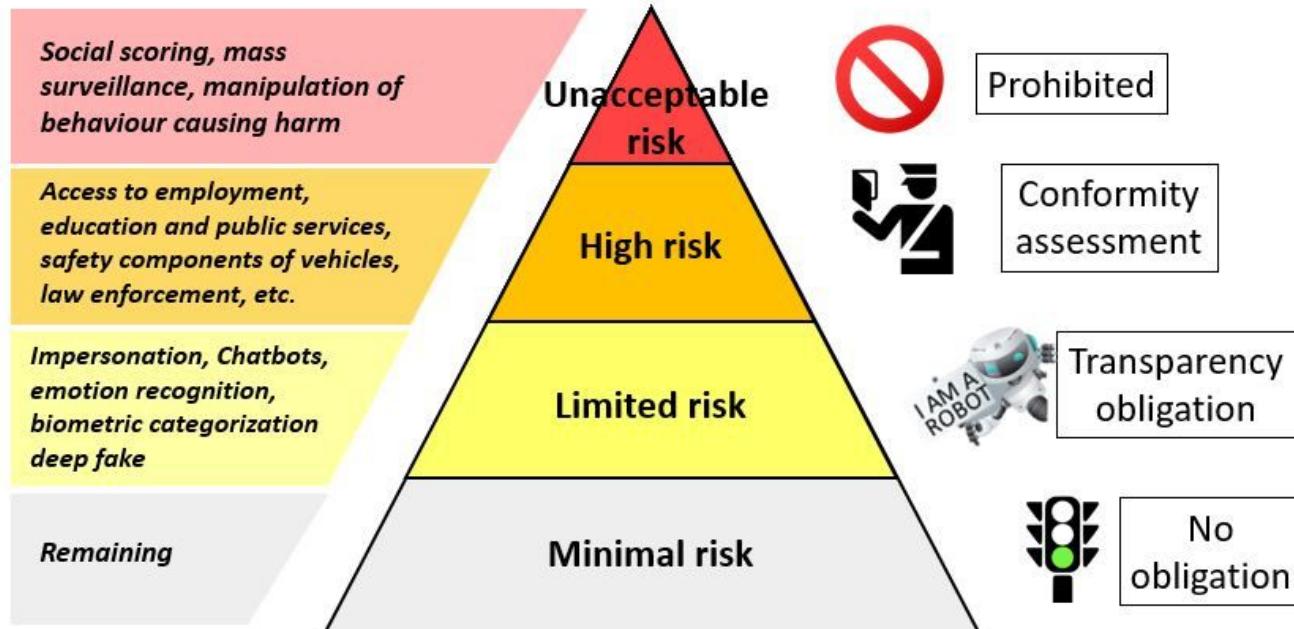


# EU AI Act

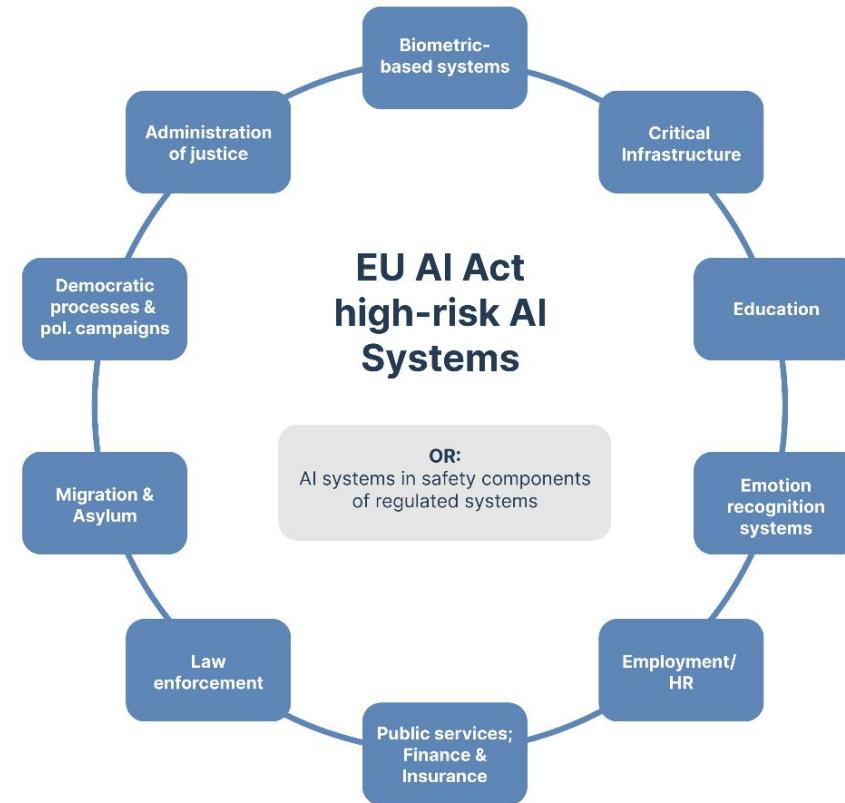


# EU AI Act

## EU Artificial Intelligence Act: Risk levels



# EU AI Act



# EU AI Act

## STEP1



A high-risk AI system is developed.

## STEP2



It needs to undergo the conformity assessment and comply with AI requirements.\*

\*For some systems a notified body is involved too.

## STEP3



Registration of stand-alone AI systems in an EU database.

## STEP4



A declaration of conformity needs to be signed and the AI system should bear the CE marking.

**The system can be placed on the market.**

If substantial changes happen in the AI system's lifecycle



GO BACK TO STEP 2



# India's DPDP Act



## Digital Personal Data Protection Act (DPDP) 2023

# INDIA



## Important aspects of DPDP Act



### Organisations

- Data Protection and Security
- Consent
- Notification
- Language and Clarity
- Withdrawal of Consent
- Data Erasure
- Special Cases
- Significant Data Fiduciaries
- Accountability



### Individuals

#### Rights:

- Right to access personal data
- Right to correction and erasure of data
- Right to grievance redressal
- Right to nominate a consent manager

#### Duties:

- To comply with all applicable laws
- To not impersonate another person
- To not suppress any material information
- To not register false or frivolous grievances or complaints
- To furnish only verifiably authentic information for corrections or erasures

usecure

# India's DPDP Act

## India DPDP Act Summary

Applicability	Principles	Data Principal Rights
Scope extended to cover more than individuals and "person" defined to include both companies and natural persons	Revolves around privacy principles such as transparency, limitation, minimization, accuracy and retention of data	More rights for the protection of data principles - Right to obtain information, Right to Correction and Erasure, Right of Grievance Redressal and Right to Nominate
Collection & Processing	Data Protection Officer	Data Transfer
Companies are required to have privacy policies, written consent and deemed consents	Compulsory to appoint Data Protection Officer & Grievance Officer	Eased cross-border data transfer requirements, where Data Fiduciaries can transfer personal data to other countries
Security	Breach Notification	Penalties
Technical and Organisational Measures (TOMs) must be implemented	Obligation to notify each data principal in the event of a data breach	<ul style="list-style-type: none"> <li>• Data Principal: up to INR 10,000</li> <li>• Data Fiduciaries: up to INR 500 crores for each instance</li> </ul>
Obligation & Compliance	All companies must comply with the DPDP Act and be able to prove it	

# 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]



# 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



# 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





# Centre for Responsible AI (CeRAI)



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



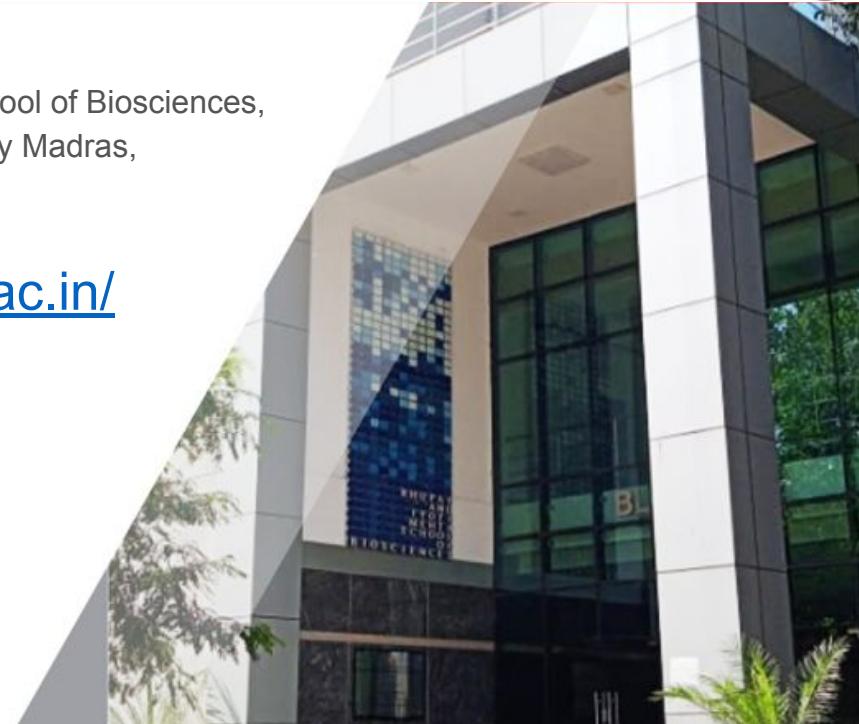
<https://cerai.iitm.ac.in/>



@cerai\_iitm



@cerai\_iitm





# Thank You!



You can contact me at

 gokul@cerai.in

 @gsk1992



Presented by Dr. Gokul Krishnan @ AI and Machine Learning School @ IIT Madras