INTRODUCTION ON AI FAIRNESS

Lili Jiang, associate professor

Department of Computing Science

UMEÅ UNIVERSITY

WHAT IS FAIRNESS

- Fairness: Impartial and just treatment or behaviour without favouritism or discrimination.
- AI fairness: AI systems should treat all people fairly.

- The other side of the coin:

  Unfairness

  Bias

  Discrimination

SOME INFAMOUS EXAMPLES

- COMPAS system used by US courts predicts higher values to the black defendents than their actual risks.
- Image management system labels images of black people with much lower accuracy.
- Amazon's automated recruiting tools was found to be biased agains women.
- Google's Ads tools for targeted advertising was found to serve significantly fewer ads. for high-paid jobs to women than men.
- .....

WHY CHALLENGING

- Bias is as old as human civilization.
- Intentional and indirect discrimination both exist.
- Bias exists in data/model/validation etc. in the whole AI life-cycle.
- Different contexts/applications/stakeholders
- General lacking of regulations.
- No consensus on definition, state-of-the-art methodologies, metrics etc.
- Requires interdisciplinary efforts.
- ……

TRUSTWORTHY AI [1]

AI system should be:

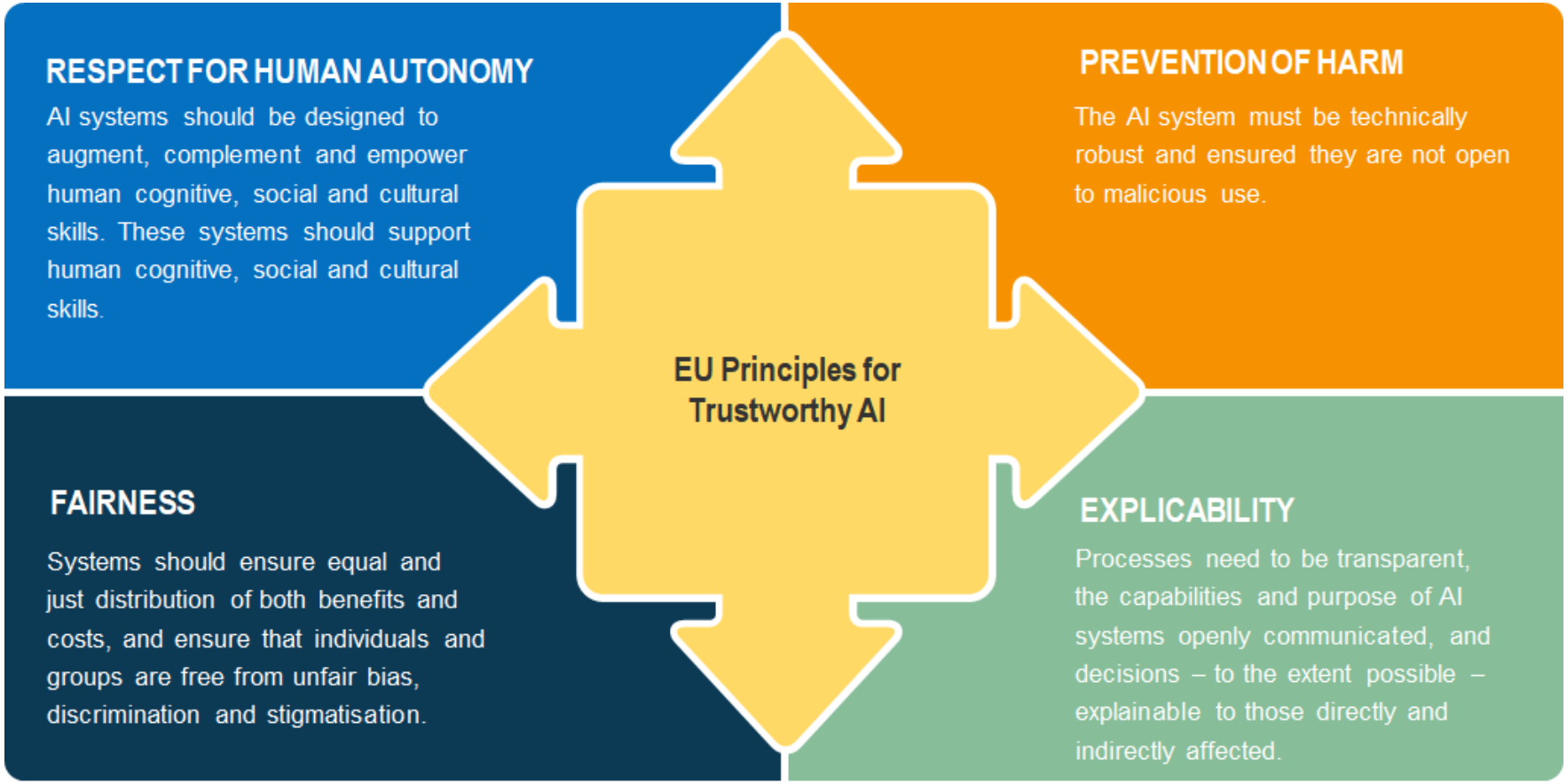
- Lawful
- Ethical
- Robust

GDPR (FAIRNESS AND EXPLAINABILITY)

Article 5(1) requires that personal data shall be: “(a) processed lawfully, fairly and in a transparent manner in relation to individuals ('lawfulness, fairness and transparency’);

Articles 21 and 22 suggest that the right to understand “meaningful information” about and the “significance” of, automated processing is related to an individual’s ability to opt out of such processing.

EU PRINCIPLES FOR TRUSTWORTHY AI [1]



Images from Intrepid Tech Ventures

REQUIREMENTS OF TRUSTWORTHY AI [1]

Ensures that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI:

(1) human agency and oversight,

(2) technical robustness and safety,

(3) privacy and data governance,

(4) transparency,

(5) diversity, non-discrimination and fairness,

(6) environmental and societal well-being,

(7) accountability.

CONTENT

- AI Fairness
  - Definition and notion
  - Fairness-aware machine learning
  - Metrics

FAIRNESS - DEFINITIONS

NO one defintion of fairness applicable in all contexts

AI FAIRNESS – NOTION [2]

Fairness: AI systems should treat all people fairly.

Fairness is most often conceptualized as equality of opportunity.

Narrow view:	Ensure that people who are similarly qualified for an opportunity have similar chances of obtaining it.
Broad view:	Ensure people of equal ability and ambition are able to realize their potential equally well.
Middle view:	Discount differences due to past injustice that accounts for current differences in qualifications.

CONTENT

- AI Fairness
  - Definition and notion
  - Fairness-aware machine learning
  - Metrics

FAIRNESS - METHOD AND TECHNOLOGIES

- Discover bias
- Mitigate bias

FAIRNESS - DISCOVECR BIAS [3]

- Bias in Data
  - Historical bias
  - Representation bias
  - Measurement bias
- Bias in Modeling
  - Aggregation bias
  - Evaluation bias

BIAS IN DATA

- Historical bias
  - E.g., “Man is to Computer Programmer as Woman is to Homemaker,”
- Representation bias
  - Underrepresentation bias (e.g., less data from lower-income background).
  - Overrepresentation bias (e.g., more white faces in image dataset)
- Measurement bias
  - Data that’s easily available is often a noisy proxy for the actual features or labels of interest
  - E.g., black defendants getting harsher sentences than white defendants for the same crime.

BIAS IN MODELING

- Aggregation bias
  - Distinct populations are inappropriately combined, but a single model is unlikely to suit all groups.
  - E..g., in health care, some models used the disesese levels differ in complicated ways across ethnicities.
- Evaluation bias
  - A model is optimized using training data, but its quality is often measured against certain benchmarks, which do not represent the general population.

FAIRNESS – MITIGATION METHODS [4]

- Pre-processing
- In-processing
- Post-processing

See [3] for more details on these approaches

PRE-PROCESSING METHODS

Produce a “balanced” dataset.

- Modifying the original data distribution by altering class labels of carefully selected instances close to the decision boundary.
- Assigning different weights to instances based on their group membership.
- Carefully sampling from each group.
- ….

IN-PROCESSING METHODS

Reformulates the classification problem by explicitly incorporating the model's discrimination behavior in the objective function via regularization or constraints, or by training on latent target labels.

- Modifying the splitting criterion of decision trees to consider the impact of the protected attributes.
- Integrating a regularizer to reduce the effect of “indirect prejudice”.
- Redefining the classification problem by minimizing an arbitrary loss function subject to the individual fairness-constraint.
- Incorporating disparate mistreatment into logistic-regression and SVMs.
- .....

IN-PROCESSING METHODS-2

Assuming the existence of latent fair classes:

- Altering the in-training weights of the instances of those assumed latent classes.
- Extending AdaBoost by considering the cumulative fairness of the learner up to the current boosting round.
- ……

IN-PROCESSING METHODS-3

For unsupervised learning:

- Fair-PCA approach forces equal reconstruction errors for both protected and unprotected groups.
- Fair clustering as having approximately equal representation for each protected group in every cluster and define fair-variants of classical k-means and k-medoids algorithms.

POST-PROCESSING METHODS

Postprocesses the classification model once it has been learned from data.

- Differentiating the decision boundary itself over groups to keep proportionality of decisions among protected versus unprotected groups.
- Wrapping a fair classifier on top of a black-box base classifier.
- ……

CONTENT

- AI Fairness
  - Definition and notion
  - Fairness-aware machine learning
  - Metrics

FAIRNESS – METRICS [5]

- Equalized odds and equality of opportunity
  - A predictor satisfies equalized odds if both the true positive rate (TPR) and (separately) the false positive rate (FPR) are the same across groups.
- Group fairness metrics
  - Disparate impact
  - Statistical parity difference
  - Equal opportunity difference
  - Demogreaphic parity
- Predictive parity: is satisfied when the positive predictive value(PPV) is the same for both groups.
- Calibration: an algorithm is calibrated if for all scores, the individuals who have the same score have the same probability of belonging to the positive class, regardless of group membership.

See [5] for more on these metrics.

SOME FAIRNESS – APPLICATION AND OPEN SOURCE

- **Fairlearn**: Fairlearn is a Python package that empowers developers of artificial intelligence (AI) systems to assess their system's fairness and mitigate any observed unfairness issues.
- **AI Fairness 360**: an. open-source toolkit of metrics to check for unwanted bias in datasets and machine learning models, and state-of-the-art algorithms to mitigate such bias.
- **FairML**: an end-to-end toolbox for auditing predictive models by quantifying the relative significance of the model's inputs.
- ……

XAI – APPLICATION AND OPEN SOURCE

- **SHAP (SHapley Additive exPlanations)** is a model agnostic and works by breaking down the contribution of each feature and attributing a score to each feature.
- **LIME (Local Interpretable Model-agnostic Explanations)** is another model agnostic method that works by approximating the behavior of the model locally around a specific prediction.
- **DALEX, T**he moDel Agnostic Language for Exploration and eXplanation, package Xrays any model and helps to explore and explain its behavior, while helping to understand how complex models are working.
- **TFDV**: TensorFlow Data Validation is a library for exploring and validating machine learning data. It is designed to be highly scalable and to work well with TensorFlow and TensorFlow Extended (TFX).
- **XAI** is a machine learning library that is designed with AI explainability at its core. XAI contains various tools that enable for analysis and evaluation of data and models.

REFERENCES

[1] European Commission, Directorate-General for Communications Networks, Content and Technology, *Ethics guidelines for trustworthy AI*, Publications Office, 2019, https://data.europa.eu/doi/10.2759/346720

[2] Solon Barocas and Moritz Hardt and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities,* fairmlbook.org, 2019.

[3] Suresh, H., & Guttag, J. (2021). *Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle*. MIT Case Studies in Social and Ethical Responsibilities of Computing, (Summer 2021). https://doi.org/10.21428/2c646de5.c16a07bb

[4] Ntoutsi E, Fafalios P, Gadiraju U, et al. (2020). *Bias in data-driven artificial intelligence systems—An introductory survey*. WIREs Data Mining Knowl Discov. 2020;10:e1356.

[5] P. Garg, J. Villasenor and V. Foggo, "Fairness Metrics: A Comparative Analysis," *2020 IEEE International Conference on Big Data (Big Data)*, Atlanta, GA, USA, 2020, pp. 3662-3666, doi: 10.1109/BigData50022.2020.9378025.

TAKE-AWAY

- Bias in, bias out.
- Bias is along with human civilization and the whole life cycle of AI system.
- Bias cannot permanently removed.
- Fairnss is not only about equality of opportunities.
- It is an interdiscipliary topic (norm, law, technology.)