arc training centre for **information resilience**

Bias in Information Resilience

Dr. Junliang Yu

This Centre is funded by the Australian Government through the Australian Research Council Industrial Transformation Research Program in partnership with the University of Queensland and Swinburne University of Technology.

What is bias?

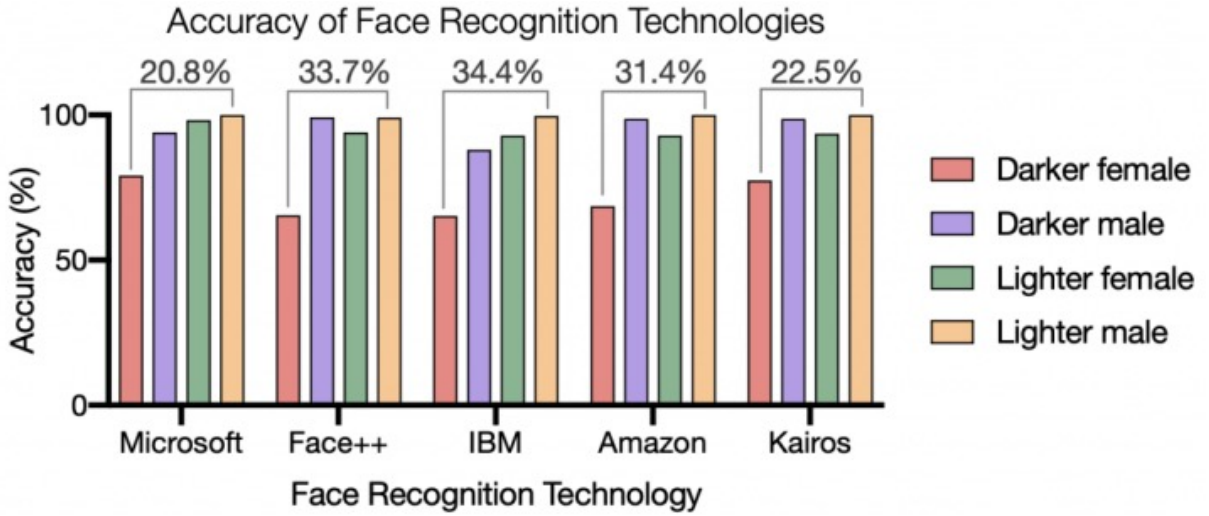
- Psychological perspective:

Bias often refers to a tendency to prefer one thing over another, often based on personal opinions or preconceived notions.

Example: People may be more likely to hire or promote individuals who graduated from the same university as they did.

- When it comes to AI:

Bias is any aspect in which the algorithm, its input, or its output shows systematic and unfair favoritism or discrimination towards certain individuals or groups [1].



Source: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/

Australian Government
Australian Research Council

Where does bias come from?

- Data

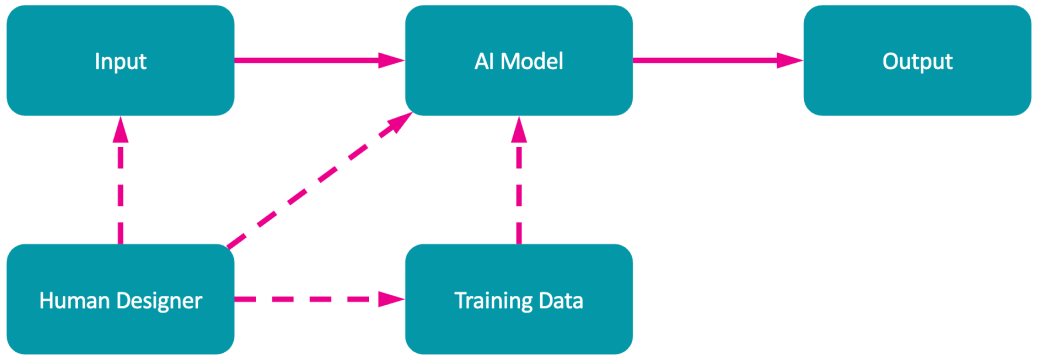
If the data used to train an AI system is biased or incomplete, the system will likely reproduce and amplify those biases in its outputs.

- Algorithm / Model

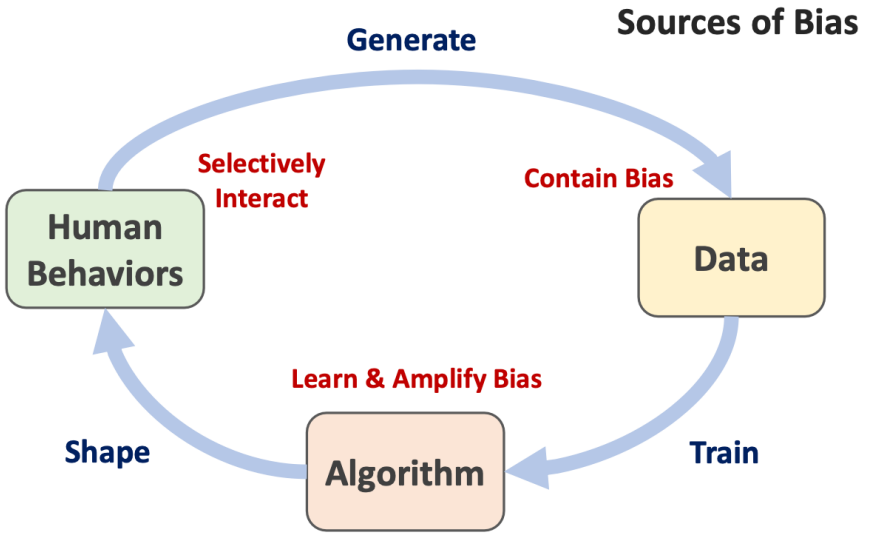
If the algorithms are based on assumptions or values that are not universally shared or if they are optimized for one metric at the expense of others.

- Human Behaviors

Humans may consciously or unconsciously introduce their own biases into the system, either through the data they select, the assumptions they make, or the feedback they provide to the system.

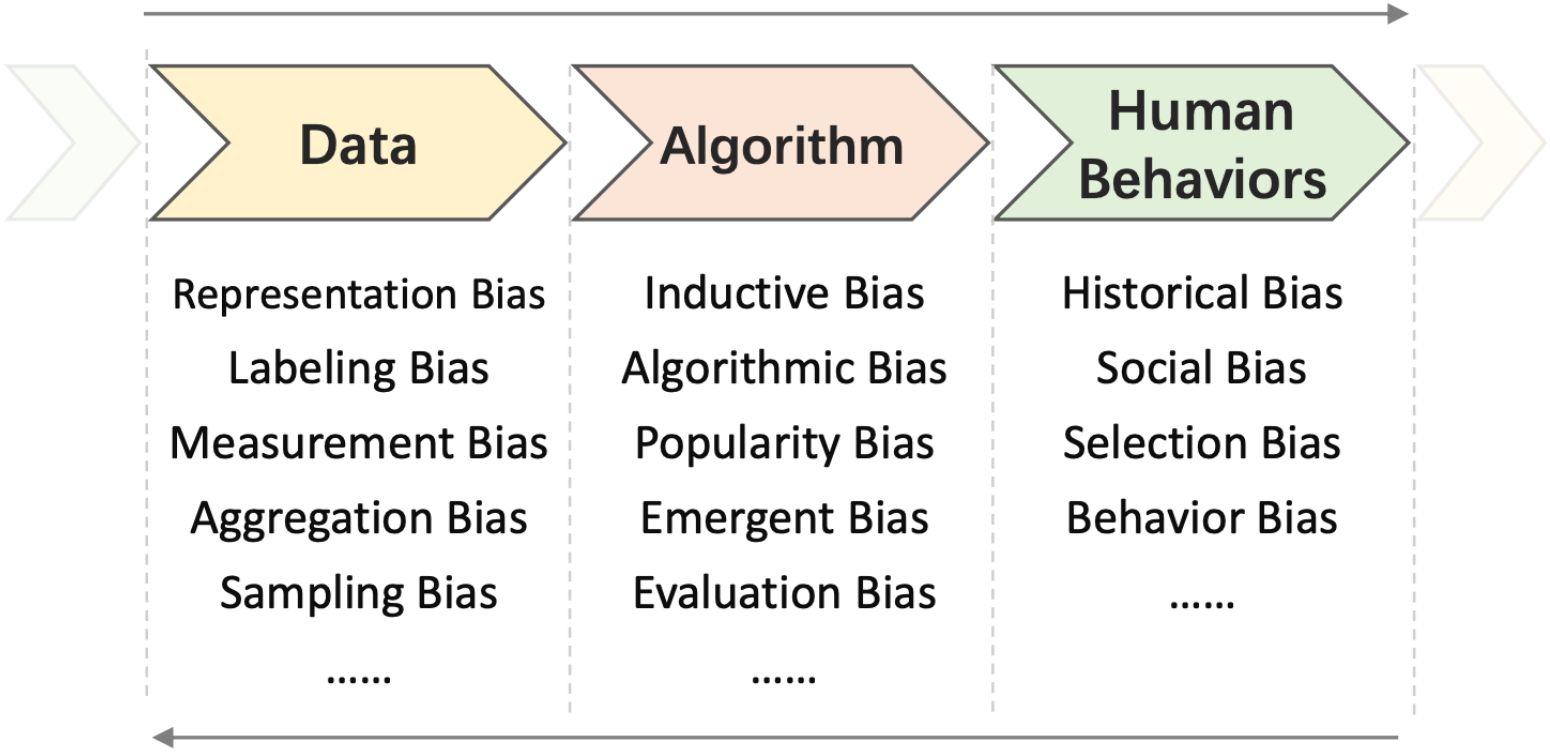


How do AI systems work?



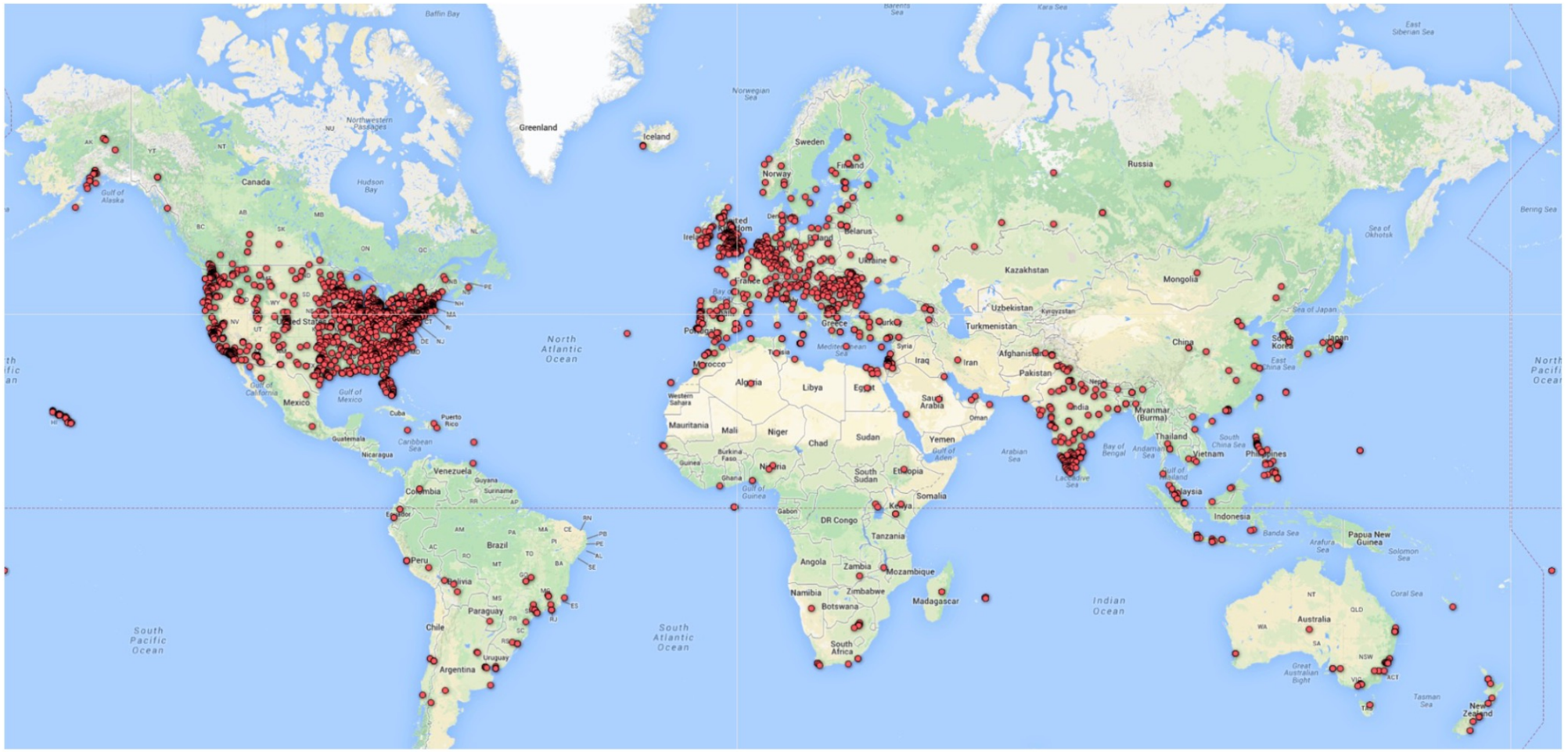
The Human-Data-Algorithm Loop: Source of Bias [2]

What types of biases can exist in AI systems



Australian Government
Australian Research Council

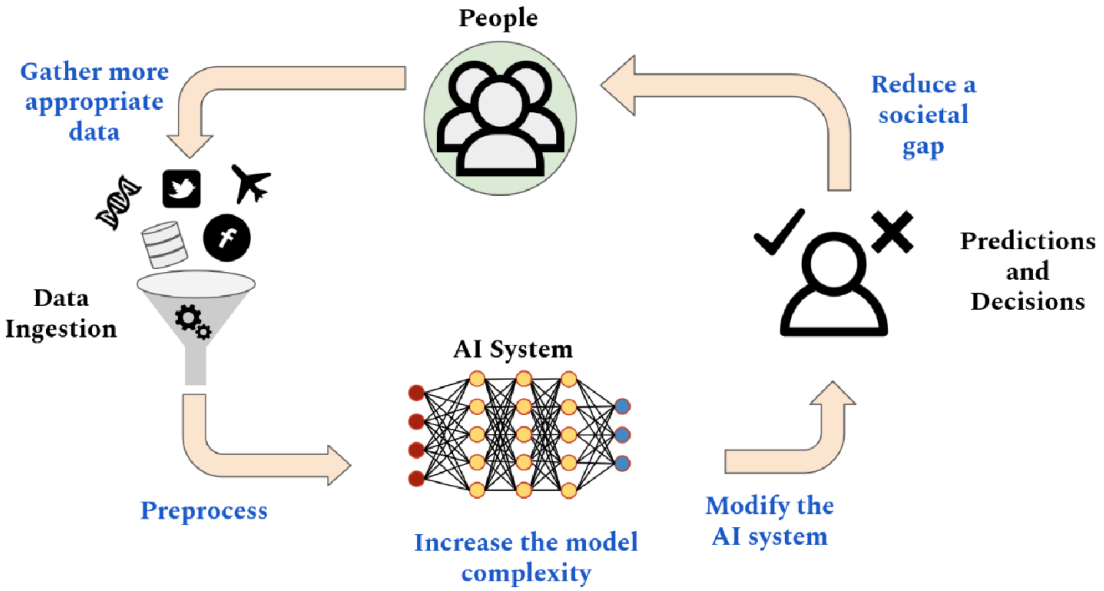
Example: Representation Bias



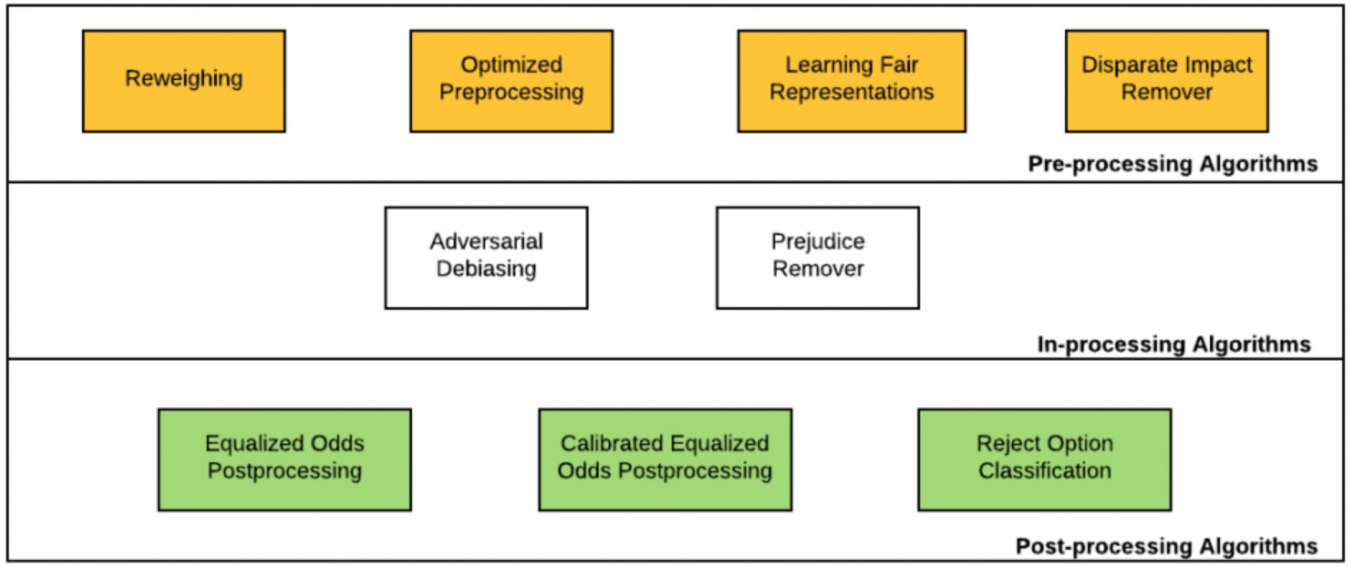
Source: A map of 50,000 Mechanical Turk workers

Australian Government
Australian Research Council

How do we mitigate bias?



Pre-processing	Modify **data** before training
In-processing	Modify **algorithm** that is trained
Post-processing	Modify **predictions** of model



Bias Mitigation Strategies for ML Models [3]

Data level

Algorithm level

Prediction level

Bias in Information Resilience

Impact of Bias on Information Resilience

- **Reduced Accuracy**: Biased data/models lead to inaccurate predictions and decisions, compromising system reliability.
- **Inequity**: Results in unfair treatment of certain groups, eroding trust in the system.
- **Vulnerability**: Biased systems are more susceptible to exploitation, as attackers can predict and manipulate outcomes based on known biases.

Addressing Bias to Enhance Information Resilience

By recognizing and addressing various types of bias, organizations can enhance the robustness, fairness, and effectiveness of their information/AI systems.

Implementing comprehensive bias management strategies is essential towards resilient and trustworthy information systems.

Australian Government
Australian Research Council

Other Concerns to Be Addressed in Information Resilience

Data Quality:
- **Missing Data**: Incomplete datasets can lead to inaccurate models.
- **Noisy Data**: Errors or random variations in data can affect model performance.

Data Labeling and Annotation:
- **Labeling Quality**: Ensuring that labeled data is accurate and representative.
- **Annotation Cost**: The expense and effort involved in manually labeling data, especially for large datasets.

Data Representation:
- **Feature Engineering**: Selecting and creating meaningful features that improve model performance.
- **Dimensionality Reduction**: Reducing the number of features to avoid the curse of dimensionality while retaining important information.

…….

Key Takeaways

Understanding Bias:
• Bias in AI occurs when algorithms or data fail to reflect the real world accurately.

Sources of Bias:
• **Data**: Incomplete or skewed data.
• **Algorithms**: Flawed assumptions or values.
• **Human Behaviors**: Conscious or unconscious biases.

Impact on Information Resilience:
• **Reduced Accuracy**: Leads to inaccurate predictions.
• **Inequity**: Causes unfair treatment of groups.
• **Vulnerability**: Increases susceptibility to exploitation.

Mitigating Bias:
• Address bias at data, algorithm, and prediction levels.
• Promote fairness, robustness, effectiveness and inclusion for better resilience.

Australian Government
Australian Research Council

Reference

[1]. Mehrabi, Ninareh, et al. "A survey on bias and fairness in machine learning." *ACM computing surveys (CSUR)* 54.6 (2021): 1-35.

[2]. Junliang Yu, et al. “Bias in AI: A succinct Introduction.”

[3]. Wang, Zeyu, et al. "Towards fairness in visual recognition: Effective strategies for bias mitigation." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.

Australian Government
Australian Research Council

Thank you !

Australian Government
Australian Research Council