

# **ENSF 611 – Machine Learning for Software Engineers**

**Week 1 – Introduction to AI Ethics**



# Lecture Goals

- Discuss different ethical concerns with AI
- Watch a video for one example and discuss

# Background

- Previously, we discussed the ethical concerns surrounding generative AI tools
- Is it possible to create a biased AI model?
- A large component of AI models is the dataset
- What do we need to consider so that we don't introduce bias into our model?



# How do we avoid bias in our model ?

# Types of Bias (1/2)

- 1. Reporting bias:** Data does not reflect the real-world frequency of events or outcomes, because people tend to document the unusual or memorable more than the ordinary.
- 2. Automation bias:** Results from automated systems are favored over those from non-automated systems, regardless of their error rates.
- 3. Selection bias:** Data is not representative of the real-world distribution, because of improper or non-random selection methods.

# Types of Bias (2/2)

4. **Group attribution bias:** Assumptions are made about individuals based on the group they belong to, resulting in in-group preference or out-group stereotyping.
5. **Implicit bias:** Assumptions are made based on one's own mental models and personal experiences, which may not apply more generally. A common form of implicit bias is confirmation bias, where data is processed in ways that affirm preexisting beliefs.

# Boston Housing Dataset

- Data collected by US Census and published in 1978
- Dataset was widely used for practicing simple machine learning models (scikit-learn)
  - Was included in multiple textbooks for this course
- The goal for this dataset is to predict the median home price based on 13 different features
- One of the features was related to the number of black people in the neighbourhood
- Dataset was removed from scikit-learn in Fall 2022



# Examples of bias in AI

- **Job search:** A job search platform may discriminate against women by offering them lower positions than men with similar or lower qualifications. This may happen because the platform's algorithm relies on historical data that shows a gender gap in the labor market (IBM Data and AI Team, 2023).
- **Finance:** A loan approval system may unfairly judge applicants based on their gender and marital status, as it assigns lower scores to women and single people. This may happen because the system's algorithm uses historical data that reflects social norms and stereotypes (Pazzanese, 2023).



# Bias in Image Recognition

- Please watch this [YouTube video](#)
  - More information at <https://gendershades.org/>
- What did you think?
- What are the implications of poor image recognition for certain populations?

# Book Recommendations

- ***Weapons of Math Destruction*** by Cathy O'Neil
- ***Algorithms of Oppression*** by Safiya Noble

# Assignment #1

- Reflect on the video we watched
- Research your own example
  - Summarize the example
  - Explain why you picked it
  - Think about how you might resolve the bias
- Remember: if you use generative AI to help you edit, you need to cite it!!!

# What's next?

- We will introduce the machine learning workflow
- Then, we will discuss data processing techniques

# References

- IBM Data and AI Team. Shedding light on ai bias with real world examples. <https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/>, 2023. [Online].
- Christina Pazzanese. Great promise but potential for peril. <https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/>, 2023. [Online].
- Google Developers. Types of bias. <https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias>, 2023. [Online].