Main Sources of Biases in AI

Data

Measurement Bias

Happens during **feature selection or collection**, such as predicting age based on height **without considering sex or ethnicity differences**, causing inaccuracies

Representation Bias

Arises when **datasets don't represent all groups well**, leading to poor generalization, as seen in under-served populations

Algorithmic Selection

Aggregation Bias

Occurs with "one-size-fits-all" models that **overlook data diversity**, such as binary gender models excluding non-binary identities

Learning Bias

Arises when **model choices amplify disparities**, like an AI favoring male resumes over female resumes during hiring **due to data validity assumptions**

Deployment

Deployment Bias

Occurs when **AI is used in different contexts than it was developed for**, leading to inappropriate outcomes, like improper associations in language models

Post-Deployment Feedback Bias

Happens when **models are adjusted based on user feedback** without considering demographic diversity, introducing new biases in systems like recommenders or search engines

Recreated based on *Challenging systematic prejudices: an investigation into bias against women and girls in large language models*, International Research Centre on Artificial Intelligence, UNESCO, 2024. https://unesdoc.unesco.org/ark:/48223/pf0000388971