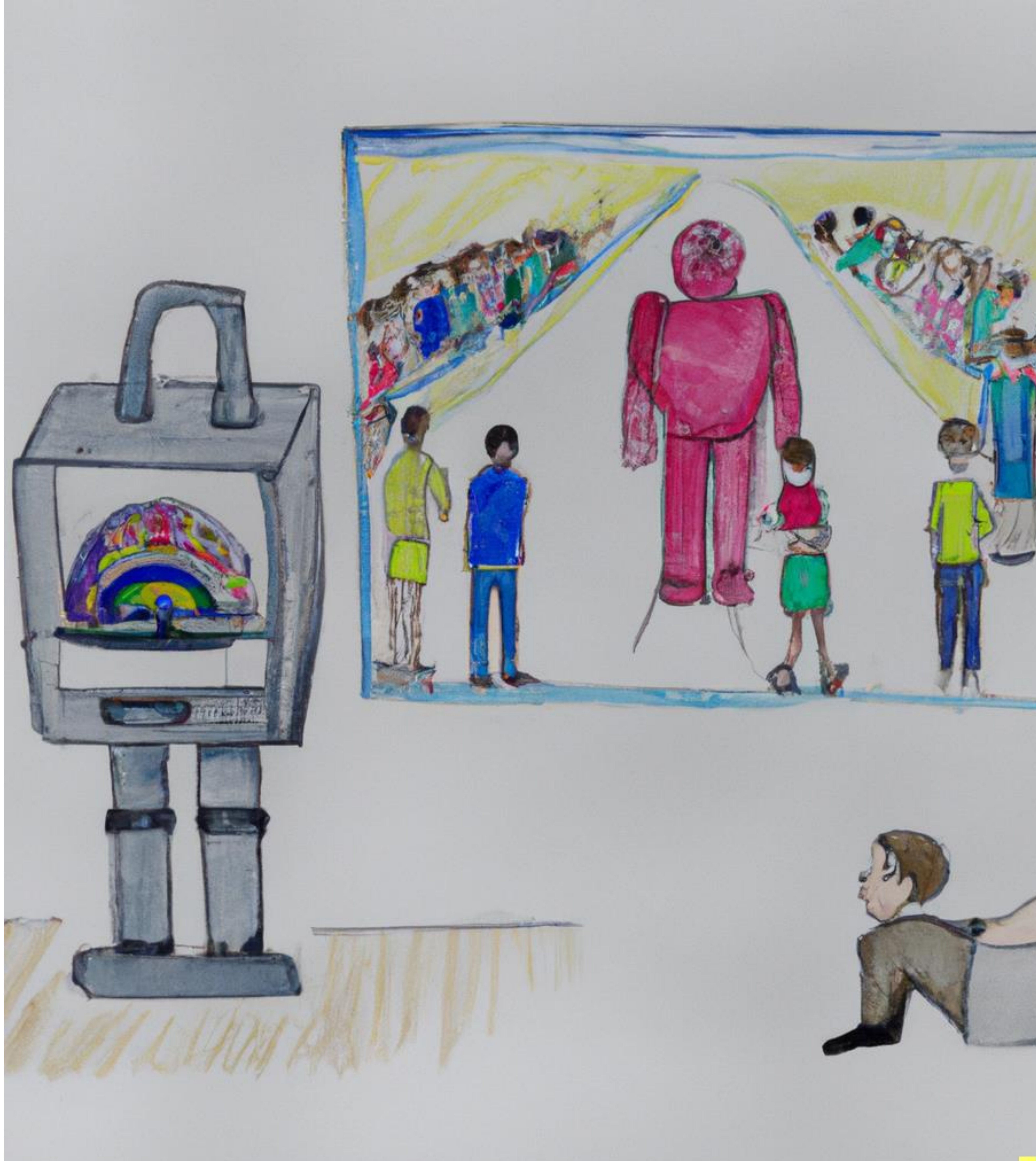
The Long March Towards AI Fairness

Rui Guo
RUC
2022



- Overview of today’s talk
  - To protect people from unfair treatment or discrimination, conventional wisdom from the legal academia points to certain protected factors or social group categories to identify and prevent prohibited behaviors or biases. This has caused problems in the context of Artificial Intelligence (AI).
  - An example of disability discrimination demonstrates the role of stereotypes and the difficulty to achieve AI fairness.
  - Dealing with stereotypes requires our deeper reflection on the problem of moral agency in AI.

人大未来法治研究丛书
Renmin Law and Technology Institute
中国人民大学未来法治研究院
人工智能的伦理和治理
郭 锐 著
The Ethics and Governance of Artificial Intelligence
法律出版社 LAW PRESS · CHINA

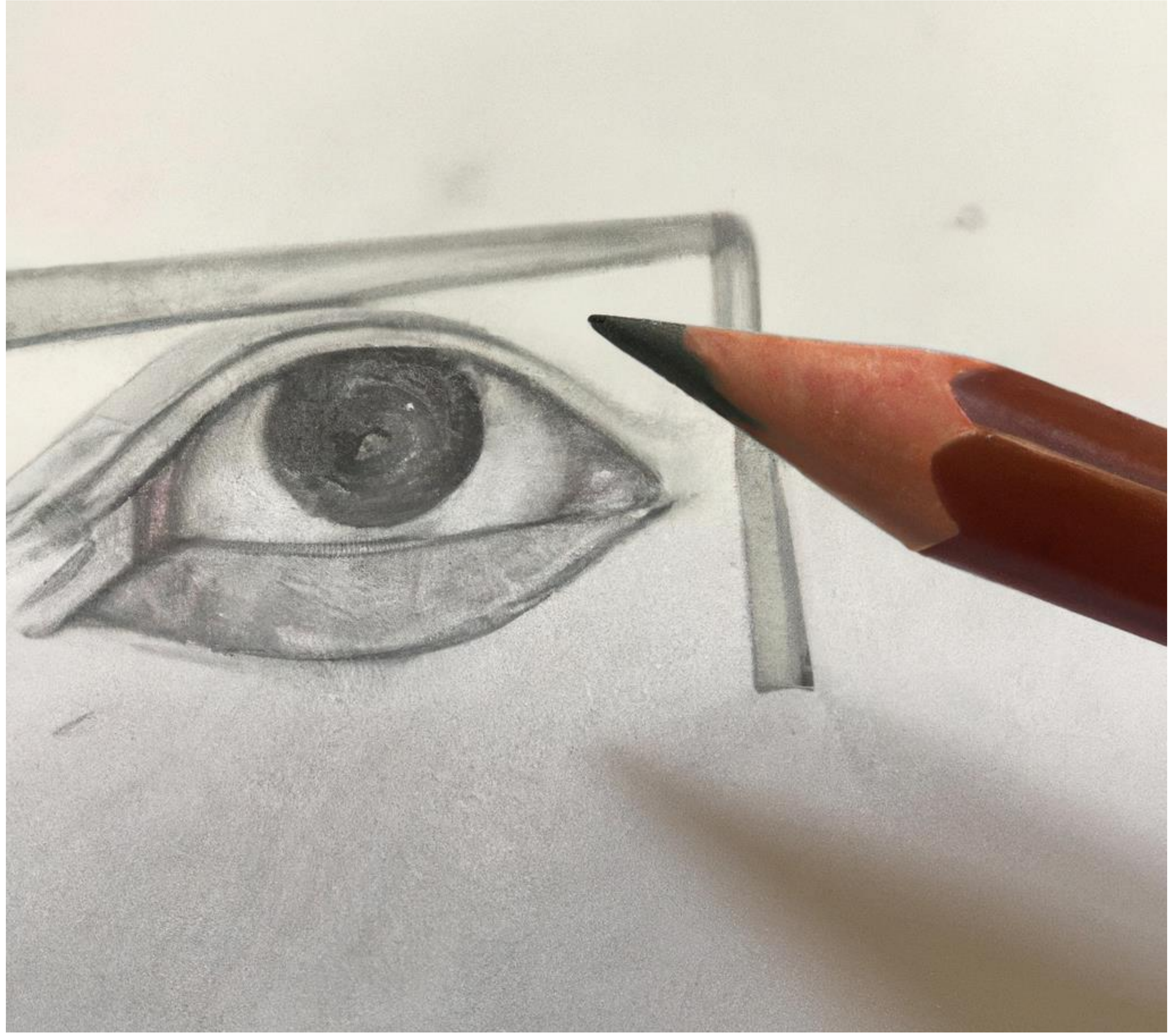
What is discrimination, according to lawyers?

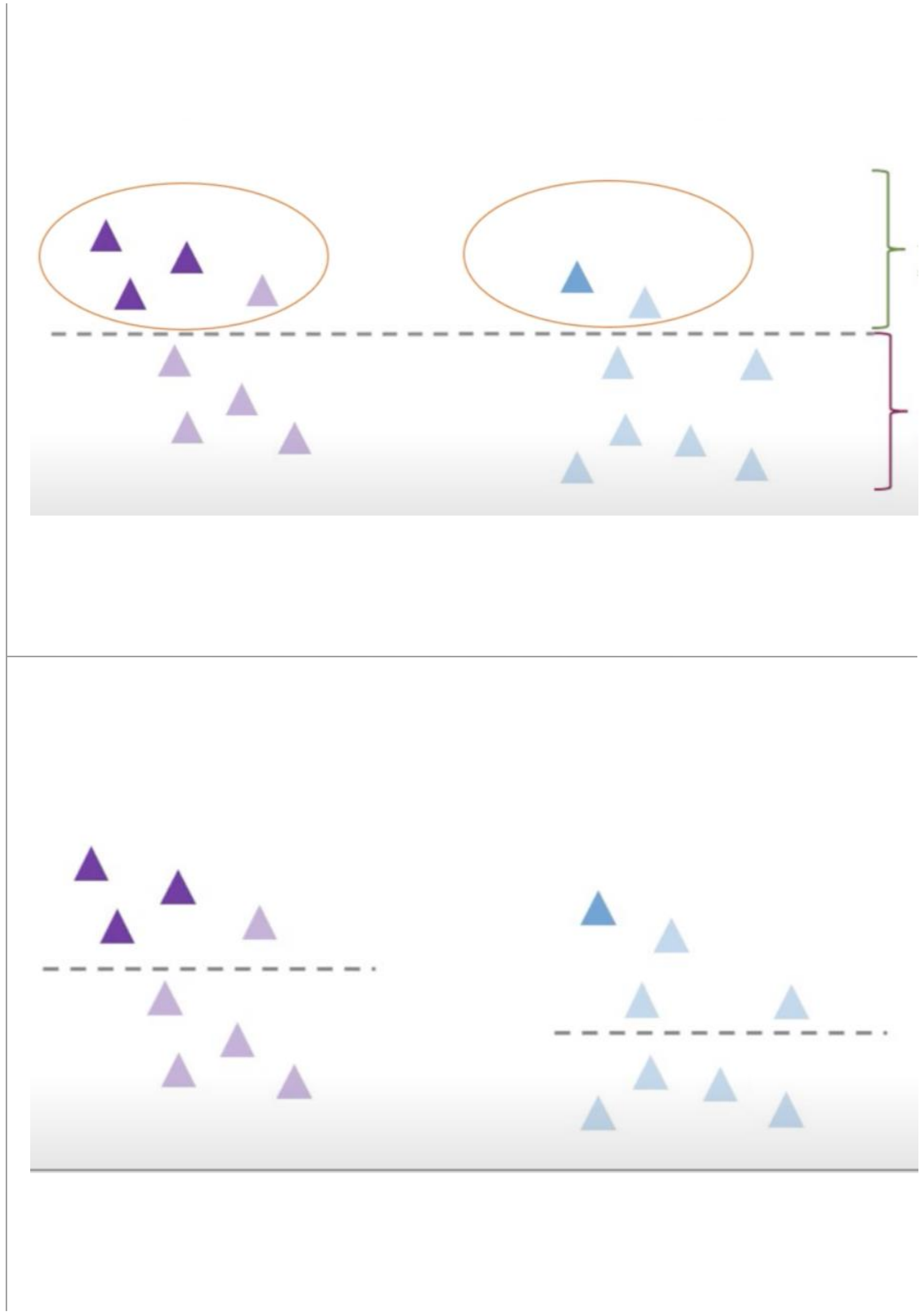
- (Wiki)…the act of making unjustified distinctions between people based on the groups, classes, or other categories to which they belong or are perceived to belong. People may be discriminated on the basis of race, gender, age, religion, disability, or sexual orientation, as well as other categories. Discrimination especially occurs when individuals or groups are unfairly treated in a way which is worse than other people are treated, on the basis of their actual or perceived membership in certain groups or social categories.

- BUT people very rarely admit their discriminatory intentions. It has to be PROVED in court.

    - Direct evidence of discrimination—subjective (“smoking gun” evidence, and those related to subjective conditions e.g. in employment context unequal discipline; Catch-22; Changing the articulated reason during the litigation; etc.

    - Indirect evidence: statistical proof





Getting rid of problematic data? —problems, problems, problems

An hypothetical example of AI disability discrimination

- A person with a physical disability submits a resume to Company X for a management position in an open recruitment at Company X. Company X uses a deep learning model Y to to screen the resume. Y was trained using historical data from Company X for the past 10 years. X have recruited and hired 50,000 employees over the past 10 years, including 25 with disabilities.

- Y has taken into the following factors:

- coming to work on time;

- amount of time away from office seat;

- previous employer evaluation.

Willingness and ability to work? Unbiased data?

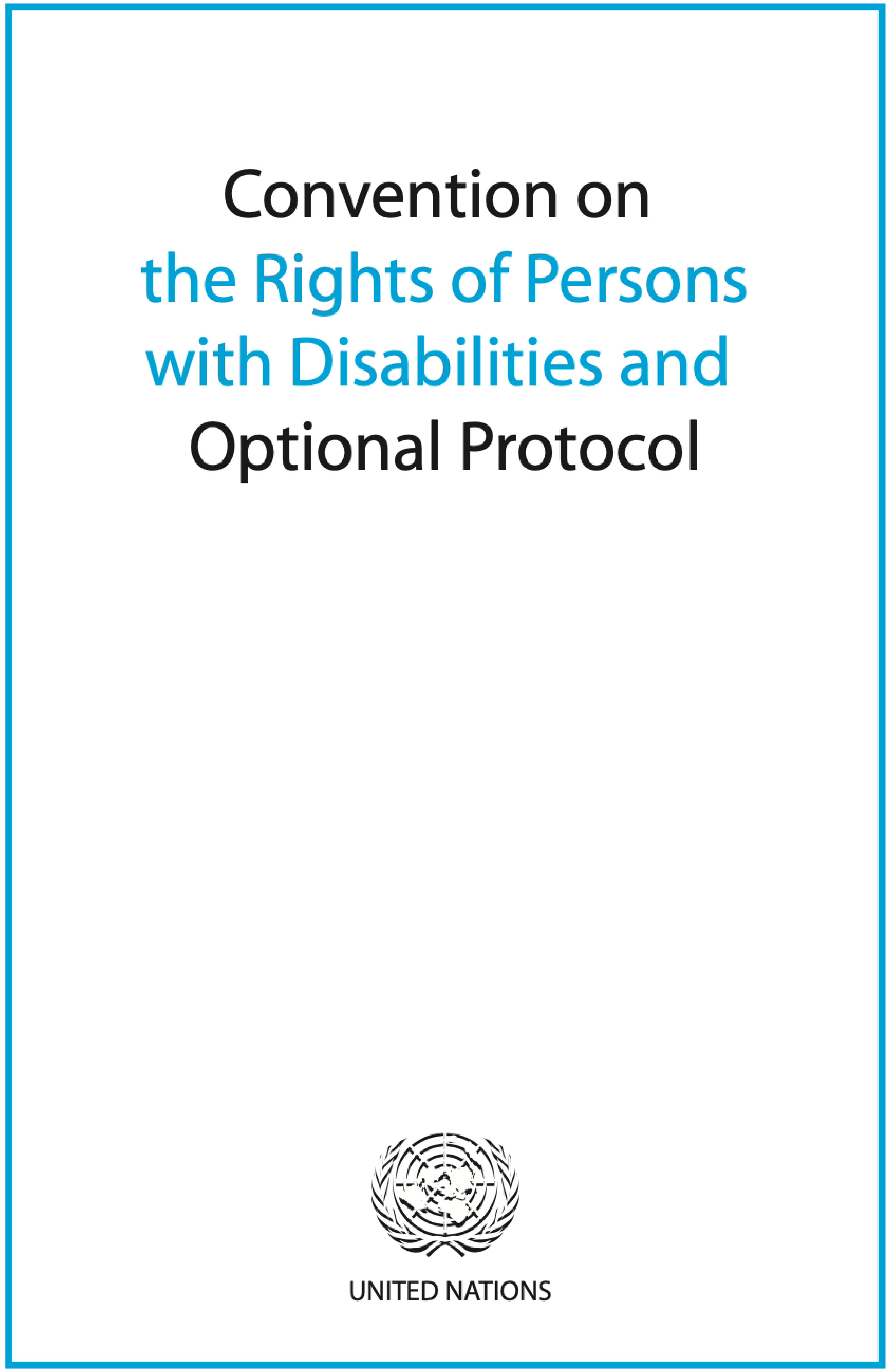
The training data set;

—coming to work on time; amount of time away from office seat;

—previous employer evaluation.

Insights from disability studies

- On Data
  - indicators and proxies
- Stereotypes on disability and their impact



Convention on
the Rights of Persons
with Disabilities and
Optional Protocol

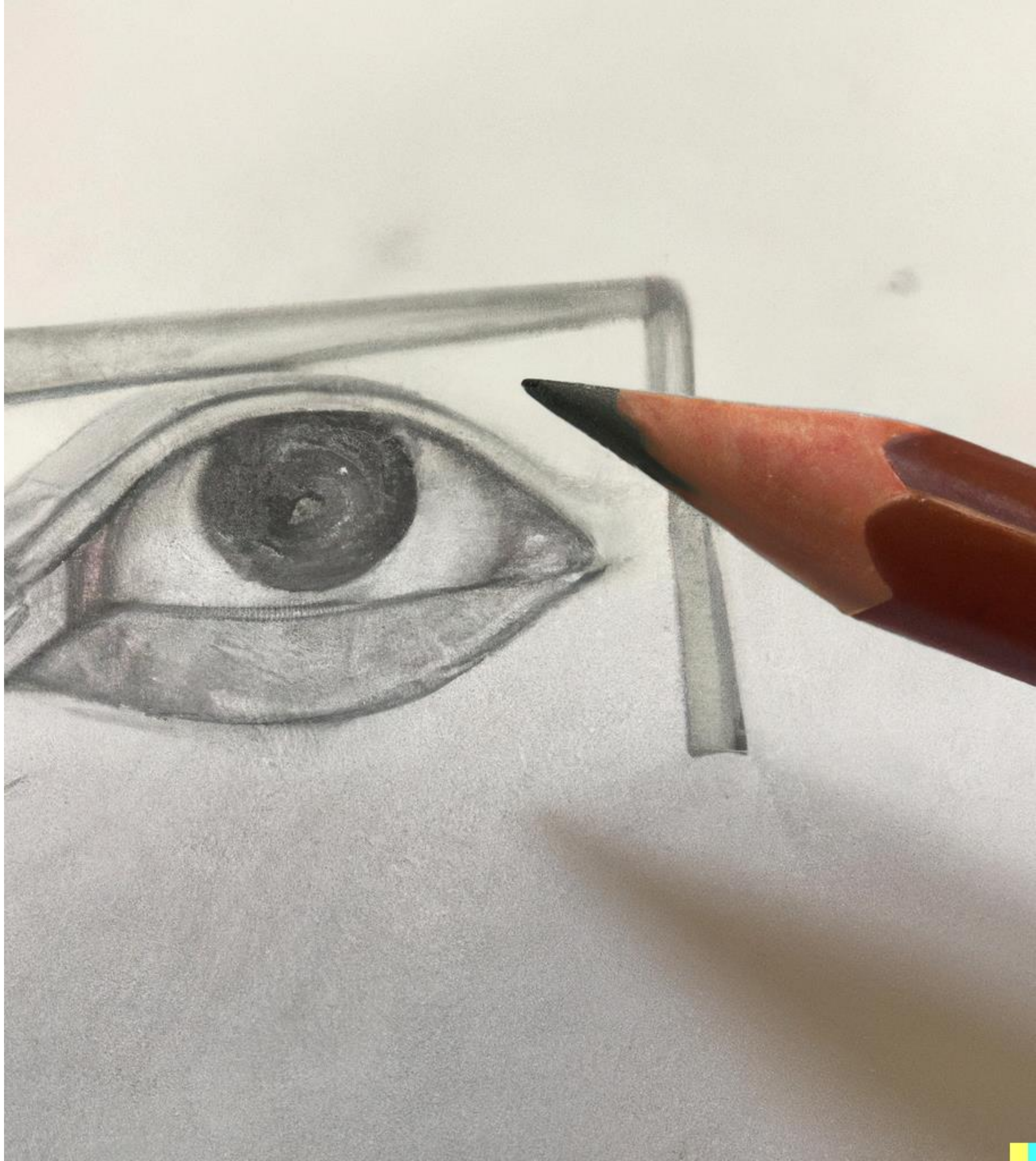
UNITED NATIONS

Reflecting on Moral Agency

- AI Discrimination takes place when
  - AI malfunctions; or
  - AI executes a human will that intended to discriminate;
  - AI executes a human will that did not intend to discriminate yet the consequences misaligned with the original intention.

A Bigger Picture

- Should AI be treated as a moral agent?
    - Consciousness;
    - Freedom; and
    - Will
- Moral agency issues may not be the full story.



Crisis of the Order of Creation

Two Problems

Crisis of the Order of Creation

- AI is not simply extending human reach or amplifying human powers; AI has made human encounter the Crisis of the Order of Creation, which originates from two problems:
  - the problem of ultimate moral principle;
  - the problem of causation.
- Any given AI system is inbuilt with these two unsolvable problems, given the conflicting human desires and limited human rationality.

What does the Crisis of the Order of Creation implies

- It takes societal changes to uproot a stereotype.
  - Awareness
  - Social movement

Thank you!

Rui Guo
rguo@ruc.edu.cn
rguo@law.harvard.edu