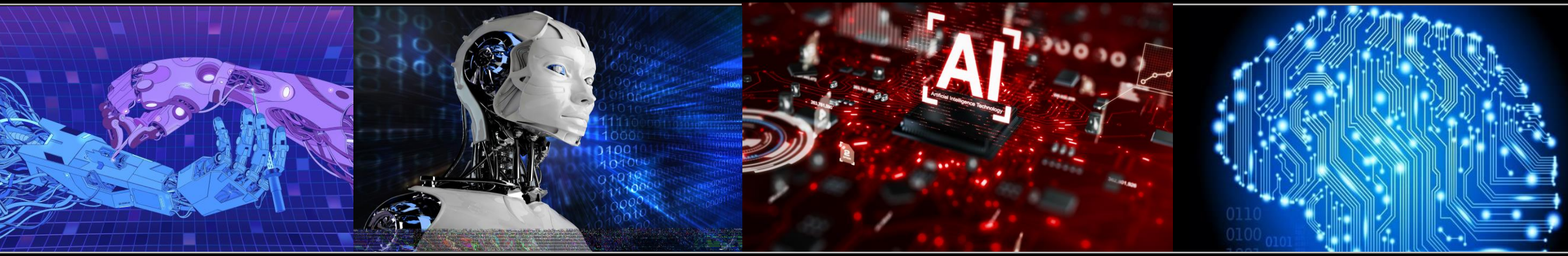


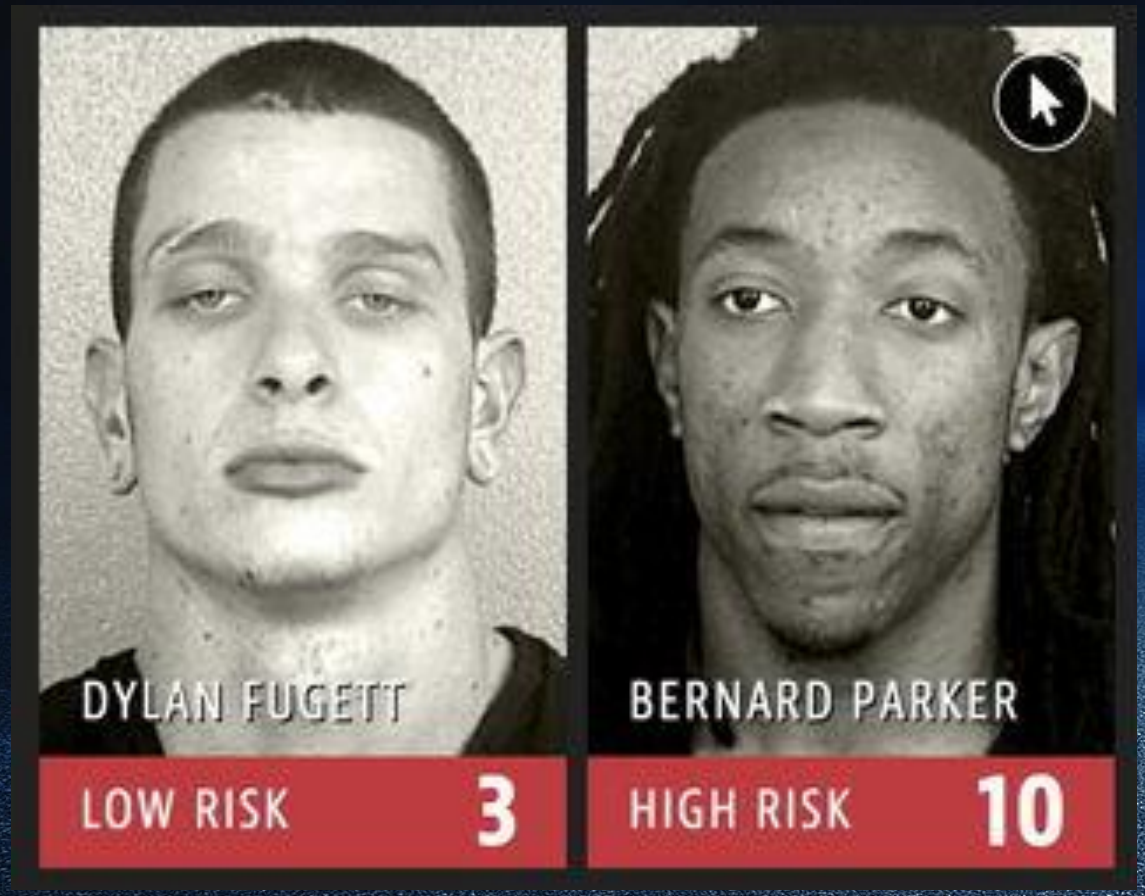
Fairness and bias in AI systems

Prof. dr. Philip Brey
University of Twente

Fairness and Bias

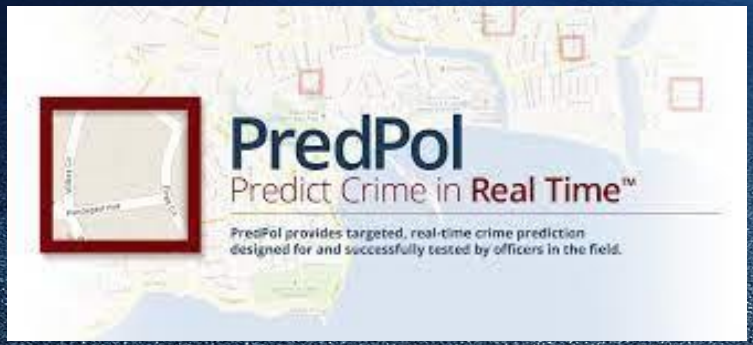
- Algorithms are often biased: they advantage some groups and disadvantage others

- E.g., COMPAS AI system to predict recidivism in sentencing guidelines (proprietary)

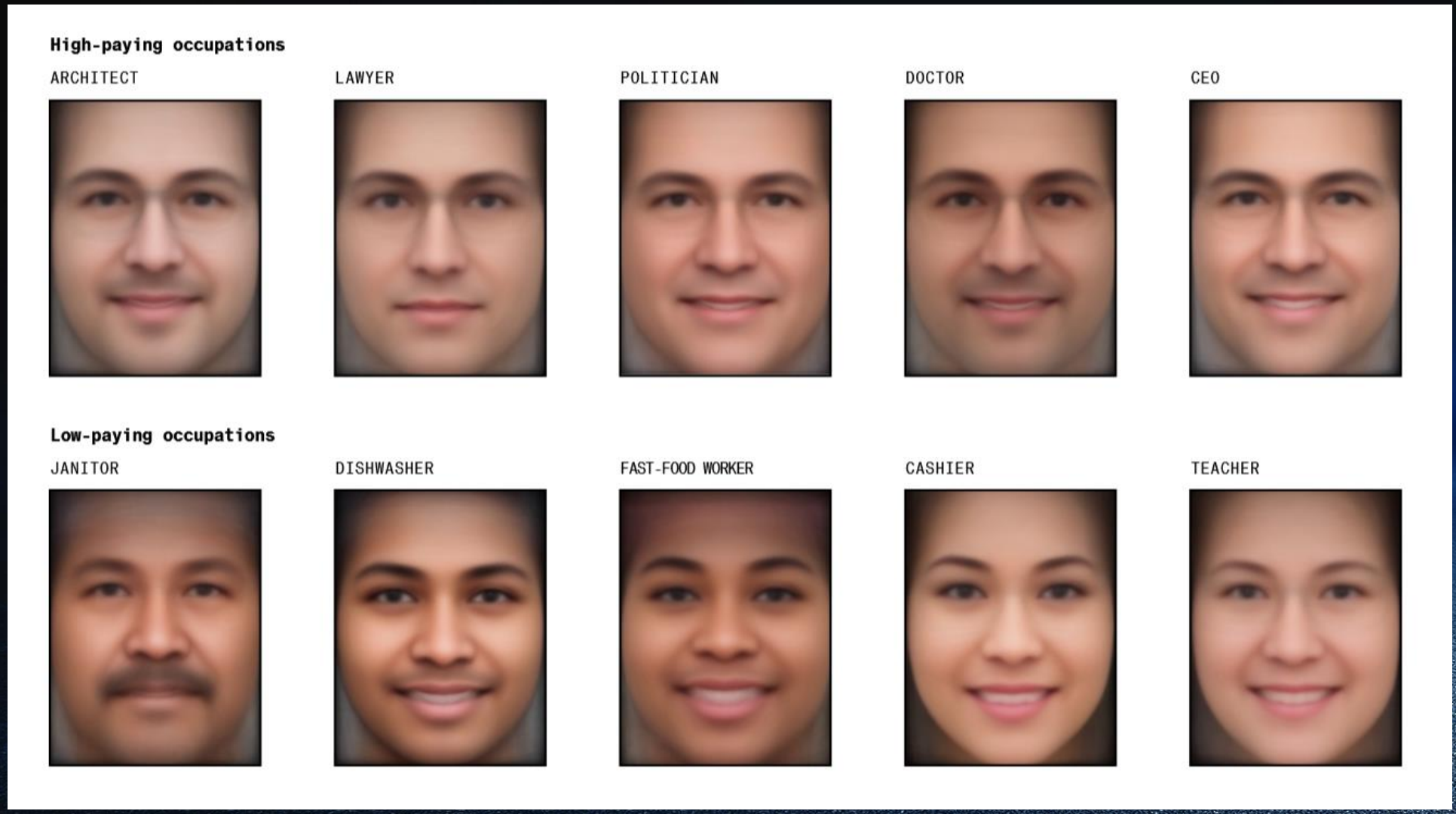


- E.g., Amazon.com job applicant resume review system:
  Downgraded women for technical positions

- E.g., PredPol crime prediction system:
- Targets neigborhoods with many minorities, based on reported incidents rather than actual crime rates



Stable diffusion (generative AI)



Outline of today's class

1. Understanding fairness, equality and nondiscrimination
2. Fairness, equality and AI
3. Access
4. Bias in AI systems
5. Unfair and discriminatory socioeconomic impacts

1. Understanding fairness, equality and nondiscrimination

Discussion of the following moral values:

equality	equity	nondiscrimination
justice	fairness	inclusiveness

Equality

A foundational value because theories and conceptions of justice, fairness, diversity, nondiscrimination, equity and inclusion are all based on the assumption that people are equal.

The most fundamental type of equality is *moral equality*: people have the same dignity and worth and are equally deserving of consideration and respect

Moral equality implies *equal rights*, especially human rights:

It is not the case that some people have more right to life, security, liberty, privacy, etc. than others.

Other forms of equal treatment are also implied:

- *Equality of opportunity:* the idea that any office or position should be open to all applicants, and applications should solely be assessed on one's merits in filling the position. Not qualities that are not relevant, such as gender, race and class.
- *Equal access to public services*
- *Equal pay for equal work*
- *Equality before the law*
- *Equal protection against discrimination*

Equality does not imply equal outcomes (everyone should have the same goods or resources or social status)

Equality does not imply that people are always treated the same. People are different, and these differences justify different treatments. E.g., in hiring or friendship

So when is unequal treatment wrong? When people have a moral claim to be treated equally. Especially unequal treatment based on inalienable parts of one's identity (social identities and physiological features):

Race, color, sex, language, religion, political or other opinion, nationality, national or social origin, property, birth or other status, sexual orientation, gender identity, disability, age, marital status, and genetic characteristics. (See human rights law)

Discrimination

Unequal treatment of individuals or groups based on one of the protected characteristics as well as other irrelevant characteristics such as hair color, weight, or fashion choices.

Groups of people with protected features are called protected groups or protected classes.

Actions that are typically prohibited are:

- discrimination in employment, housing, education and access to goods and services
- insults, hate speech, incitement to hatred, and harassment of members of protected groups

Discrimination can be direct or indirect, and intentional or unintentional

Direct discrimination: someone is treated poorly because of their membership of a protected social group.

E.g., a restaurant owner refuses a customer in a wheelchair because he does not want wheelchair users in his restaurant.

Indirect discrimination: members of a protected social group are treated worse than others because policies or practices that are intended to be neutral systematically work against their interests.

E.g., a restaurant with a single entrance that is not wheelchair accessible.

E.g., a company with a policy requiring employees to work on Saturdays. This policy indirectly discriminates against practicing Jews who want to observe the sabbath.

Indirect discrimination is not always easy to spot, and whether policies or practices count as indirect discrimination is not always clear-cut.

DISCUSSION

What are other examples of indirect discrimination?

Justice and fairness

Modern conceptions of justice are defined in terms of fairness. Justice is the fair treatment of individuals in society. Justice requires respect for equality.

Actions that do not respect equality are by definition unjust and unfair.

But the notions of justice and fairness also apply to issues other than equal treatment. They also apply to issues involving desert– the idea that some people deserve more reward or punishment than others because of their behavior.

E.g. it is fair that an employee who worked twice as many hours as another employee receives twice the salary, and that someone who carried out a crime should be punished more than someone who was only an accessory to the crime.

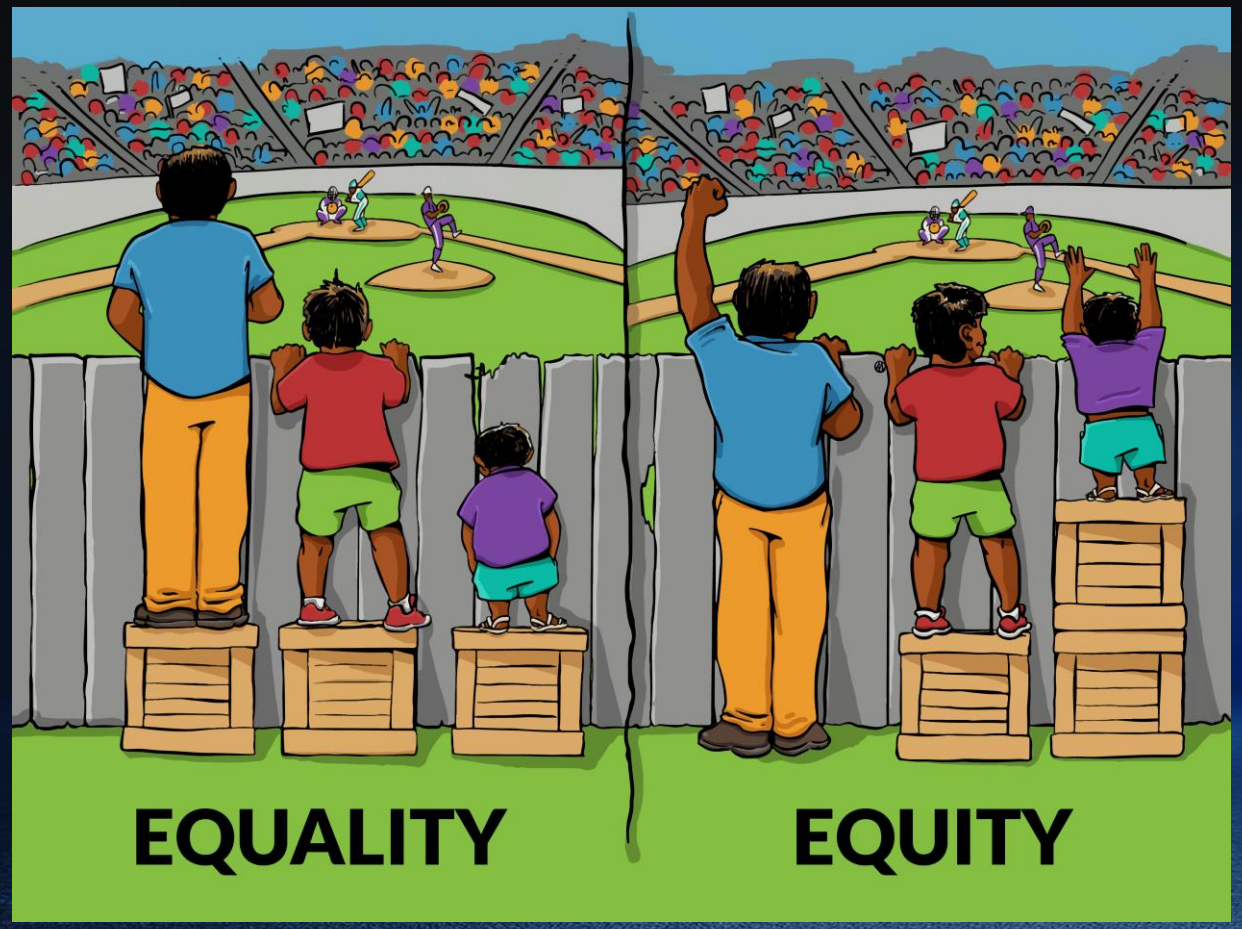
DISCUSSION

What would be a good definition of fairness?

Equity

Fairness sometimes seems to require that equality is not adhered to. When someone is from a disadvantaged background, it may be fair to give them more resources or opportunities, so that they will have a more equal chance to succeed.

This is *equity*: creating a fair playing field by giving disadvantaged people more resources or opportunities, so that there can be genuine equality of opportunity



Equity can be a way of overcoming structural inequalities and injustices in society.

But when are allocations of additional resources of opportunities fair, and when is it unfair? Can it lead to reverse discrimination?

Accepted equity measures:

- Housing subsidies for lower income families
- Financial aid for lower income students
- Extra facilities for employees with disabilities

Controversial equity measures:

- Quotas in hiring (e.g. for women, ethnic minorities)
- Reparations: financial compensation to social groups for historical injustices

DISCUSSION

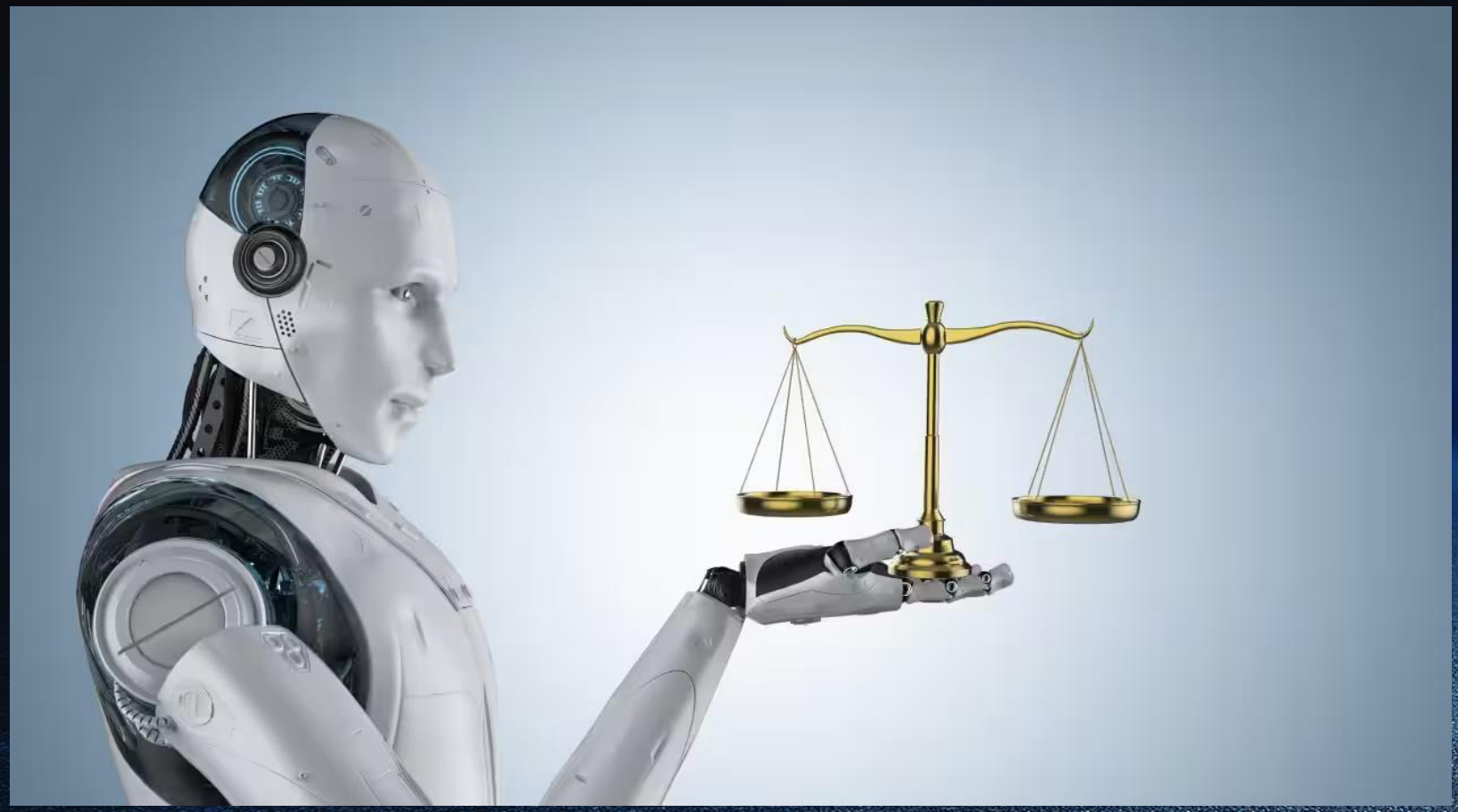
Is there way of assessing either objectively or democratically when equity measures are fair and when they go too far?

Inclusiveness

Inclusiveness or inclusion is support for participation and positive valuation of people with different characteristics and identities.

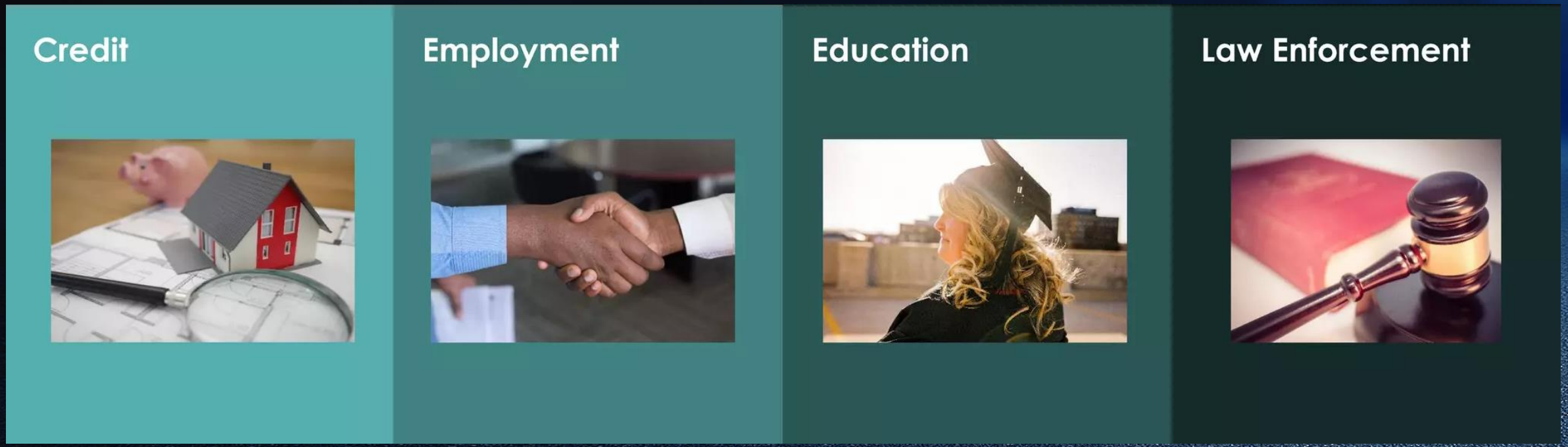
Inclusiveness is a way of supporting equality and equity in organizations and in society. It involves a continued effort to create conditions that diverse people with different identities can fully participate, at all levels, and that differences are being valued and supported.

2. Fairness, equality and AI



AI systems have an increasingly large role in how people are treated.

They can have a major impact on individuals and groups, e.g., in determining access to credit, hiring and promotions, educational opportunities, criminal and civil punishment, and healthcare outcomes.



How can we ensure that people are treated fairly and equally by AI systems?

Users: are they treated equally and fairly, both in their opportunity to access AI systems and in the funcitonality of the system for their interests and purposes?

Data subjects: are they represented in a fair and nondiscriminatory way, and are they treated fairly and equally by the system?

Other stakeholders: does the use of AI systems affect them fairly and equally?

Section 3: Fairness and equality in relation to users: universal access and functional bias

Section 4: Fairness and equality in relation to data subjects: data, algorithmic and user bias

Section 5: Fairness and equality in relation to affected stakeholders: discrimination, unfair socioeconomic impacts

3. Access

Key data point: 54% of US men are using AI in either or both their personal and professional life, but only 35% of women (35%)

Digital divide in AI: Four barriers to access (J. van Dijk):

1. *Lack of motivation.* Possible reasons: perceptions of a lack of usability or benefits, perceptions of risks and dangers, a lack of money, or a lack of skills.

2. *Lack of physical access.* Physical access to computer hardware, internet connectivity, and AI systems. Possible reasons: lack of funds, lack of connectivity, prohibitions, or other reasons. There is also the issue of quality of access, as determined by the quality of the hardware, bandwidth, internet data caps, and other factors.

Generative AI can often be used for free, but advanced generative and other AI systems can be expensive.

3. *Lack of skills:* Lack of skills in using AI systems, including basic operational skills and content-related and strategic skills like using AI systems for personal and professional goals like applying for jobs and publishing videos.

4. Lack of usage. Is the user successful in using AI technology for an extended amount of time, and for a diversity of advanced usage applications? Does the user succeed in using the technology for serious usage that benefits their human capital and resources (e.g., for work, career, study, and societal participation)?

Lack of usage could also be the result of *functional bias*: the system is designed for particular types of users with particular types of goals. These users are supported, while other users are marginalized.

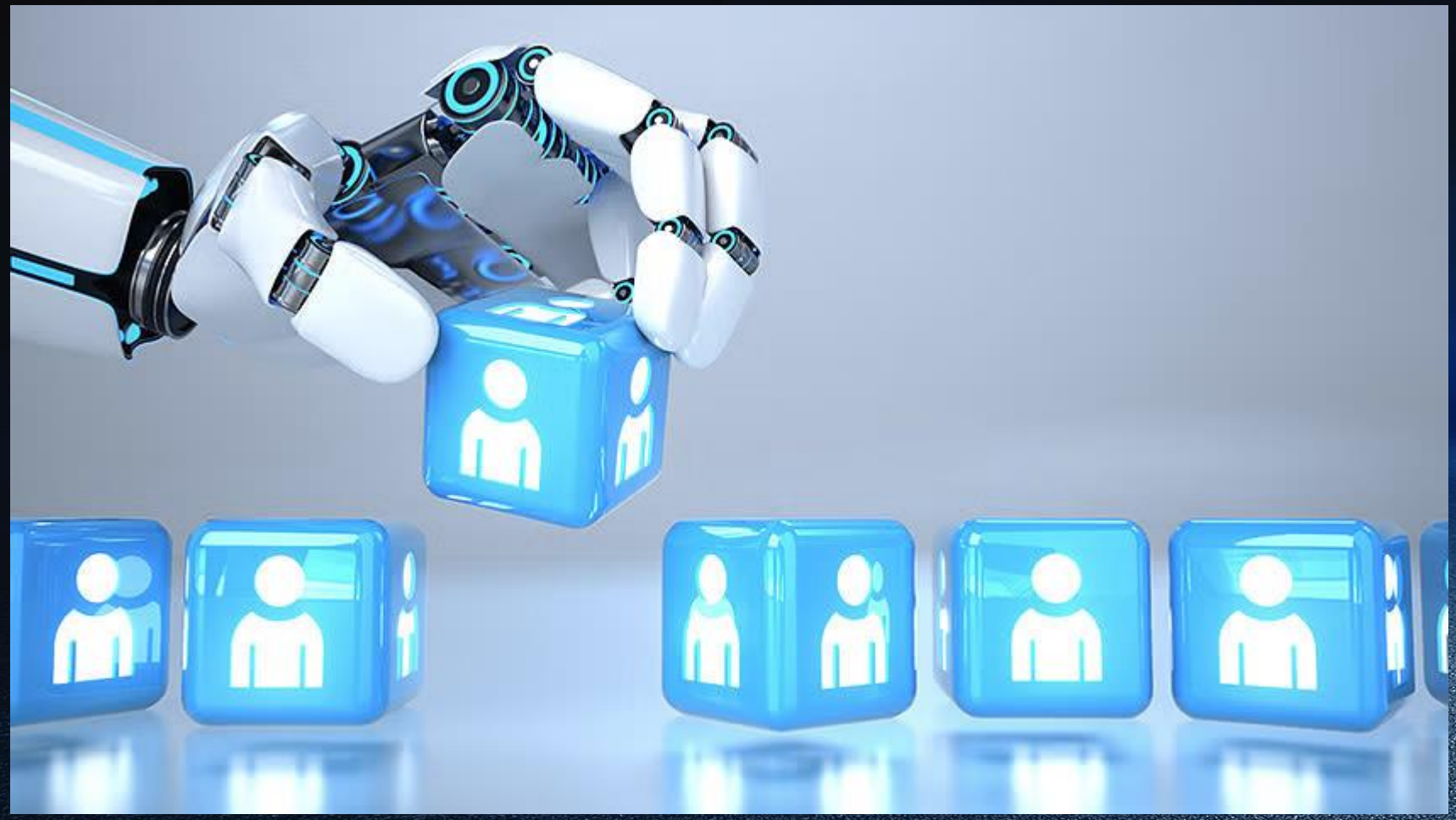
E.g., surveillance systems strengthen governments and companies but may weaken citizens.

E.g., a business support system may be designed to work well for for-profit businesses but not for non-profit organizations.

The challenge: ensuring universal access

- *Lack of motivation:* educating people about the benefits and risks of AI
- *Lack of physical access:* helping users overcome physical and economic barriers
- *Lack of skills:* training users in AI skills, making AI easier to use
- *Lack of usage:* ensuring a diversity of use cases, stimulating low educated users to engage in serious use cases

4. Bias in AI systems



A *bias* in an information system is any systematic error or distortion in data, algorithms, or decision-making processes that results in inaccurate, unfair, or discriminatory outcomes.

Social bias: bias that relates to individuals or groups represented by the system and that involves unfair or discriminatory outcomes.

Nonsocial bias: bias that leads to inaccurate outcomes, e.g., by use of skewed scientific data in AI system for astronomy or mechanical engineering.

In this class, we are interested in *social bias* in AI systems.

Types of social bias (by origin):

Data bias (or input bias): the data used to train the machine learning model is unrepresentative or incomplete, leading to biased outputs.

Algorithmic bias: inherent biases in algorithms used in machine learning models due to biased assumptions or decision-making criteria.

User bias: bias that emergers when users introduce their own biases or prejudices into the system, consciously or unconsciously. E.g., when users provide biased training data or interact with the system in a biased way.

Data bias: three types

Historical bias:

Due to biases in society, data sources are biased and result in biased data sets. E.g., due to marginalization of minority groups, they may be underrepresented in, e.g., health data or census data. E.g., minorities may have poorer health due to socioeconomic disadvantages, leading to biased health data in which poor health correlates with minority status.

Representation bias:

bias due to a dataset not accurately representing the population it is meant to model, for example due to over- or undersampling of certain social groups

Measurement bias:

Bias arising from how particular features in a data set are selected, utilized, and measured, leading to inaccuracies or inconsistencies.

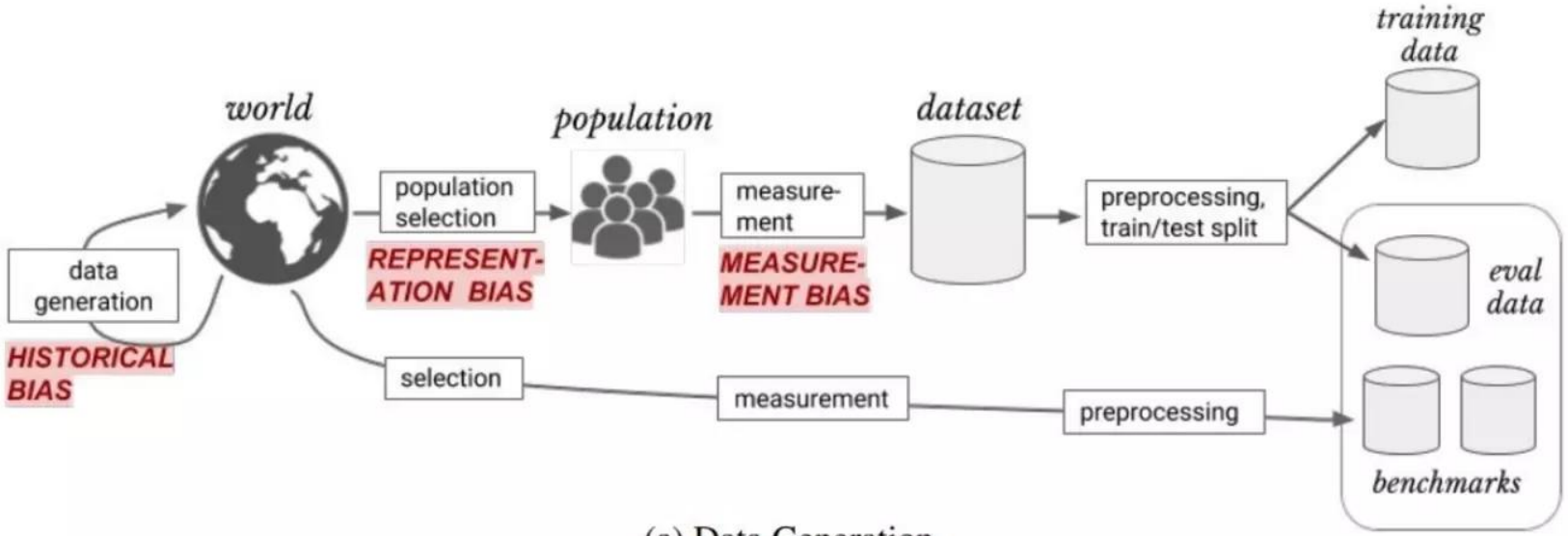
Types of algorithmic bias

Evaluation bias:

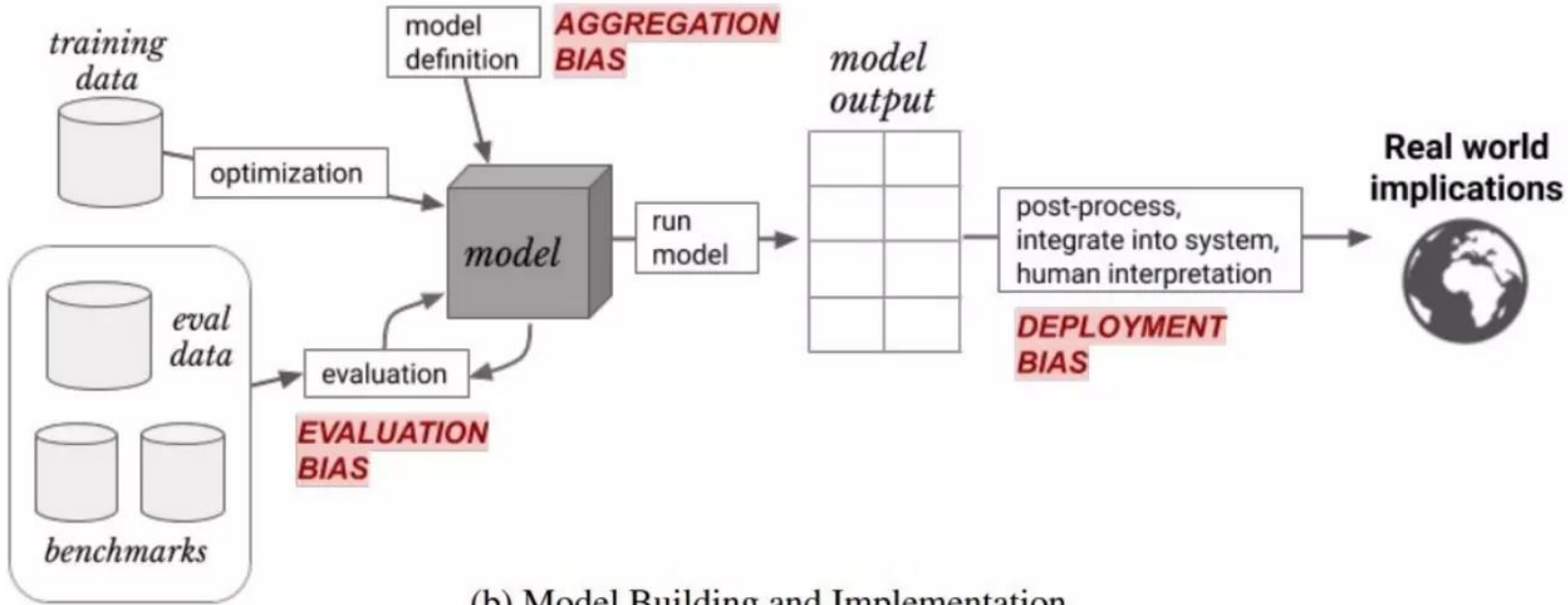
Occurs when the criteria used to evaluate the performance of an algorithm are biased or flawed. E.g., an algorithm is evaluated based on metrics that do not account for fairness or equity.

Aggregation bias: the combining or aggregating of data in a way that distorts and leads to false or misleading conclusions. E.g., aggregating data on housing affordability or availability at the city or regional level may obscure disparities within neighborhoods or communities.

Other: Algorithms may use biased categories for categorizing data, biased assumption in the application of regression models to analyze the relationship between one or more independent variables and a dependent variable, may apply biased optimizing functions, etc. etc.



(a) Data Generation



(b) Model Building and Implementation

Source: Suresh, Guttag 2020

Mitigating bias in AI systems

Sometimes called: *Algorithmic fairness*

1. Pre-processing of data: Using pre-processing techniques on biased data sets before the model is trained to try to transform the data so that the underlying bias is removed. This can involve oversampling, undersampling, synthetic data and adversarial debiasing (training the model to be resilient to specific types of bias).

2. Model selection: This refers to techniques that focus on ensuring that fairness properties are incorporated directly into the process of choosing or designing a machine learning model. They include *in-processing* techniques (techniques to change state-of-the-art learning algorithms in order to remove bias during the model training process)

Some techniques:

- *Fair representation learning:* learning representations of the data that disentangle sensitive attributes such as race or gender from other features, so as to avoid that resulting representations are biased

- *Fair Loss Functions:* functions that explicitly incorporate fairness considerations into the optimization process by penalizing models for making decisions that result in unfair outcomes.

3. Post-processing: adjusting the output of AI models to remove bias and ensure fairness.

Some techniques:

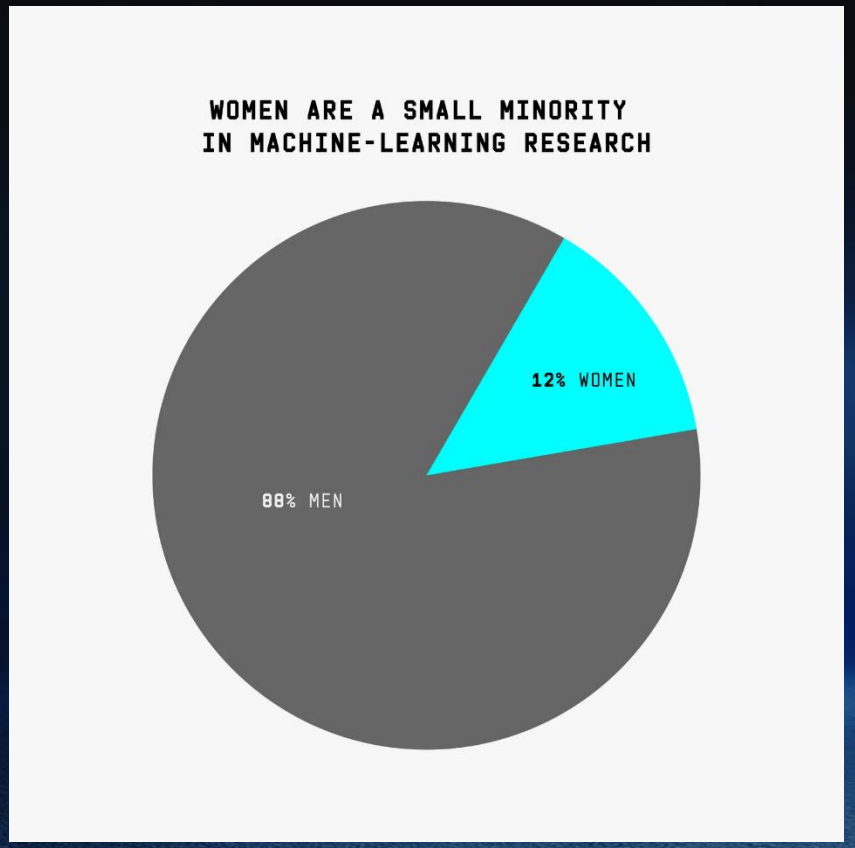
- *Calibration:* adjustment of predicted probabilities outputted by a model to match the observed frequencies of outcomes in different demographic groups.

- *Equalized odds post-processing:* adjusting the predicted probabilities of a classifier to ensure equalized odds or equal false positive and negative rates across different demographic groups.

Apart from these techniques, the composition and training of development teams is considered to be key:

4. *Diversification of development teams.* Diverse teams are more sensitive to fairness, equality and inclusion and can think up more strategies.

5. *Training of developers.* Training developers will make them more sensitive and will equip them with techniques and strategies



Source: WIRED

Algorithmic fairness gone wrong

Challenges for fairness research (Mehrabi et al., 2022)

Synthesizing a definition of fairness, with different dimensions

Incorporating equity measures in addition to equality measures

Searching for bias in data sets (since there are many types of bias)

5. Unfair and discriminatory socioeconomic impacts

How are the benefits and risks of AI distributed?

Even if AI was not biased in data, modeling or user interaction, it could still have unfair consequences.

- Use or misuse that is unfair or discriminatory (intentionally or unintentionally)

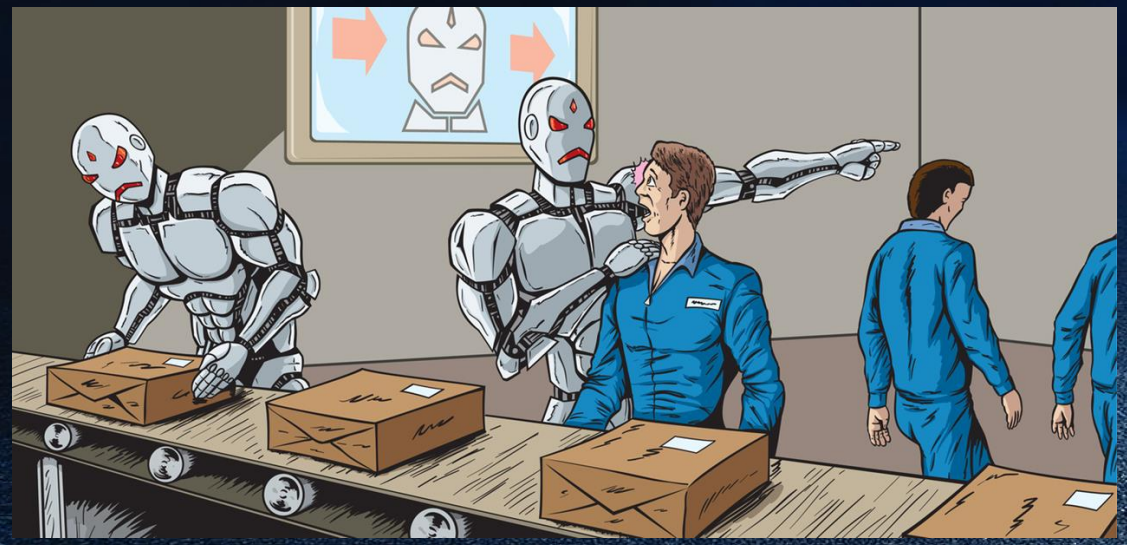
E.g., a facial recognition system deployed for security purposes is used for racial profiling

E.g., the use of online censorship to suppress minorities or political dissidents in dictatorial states

- Complex interactions and systemic effects: complex interactions with social, economic, and political systems that lead to systemic effects that perpetuate unfairness.

  E.g., a predictive policing system produces biased outcomes because it is influenced by systemic biases in law enforcement practices

  E.g., job losses due to generative AI disproportionately affect women

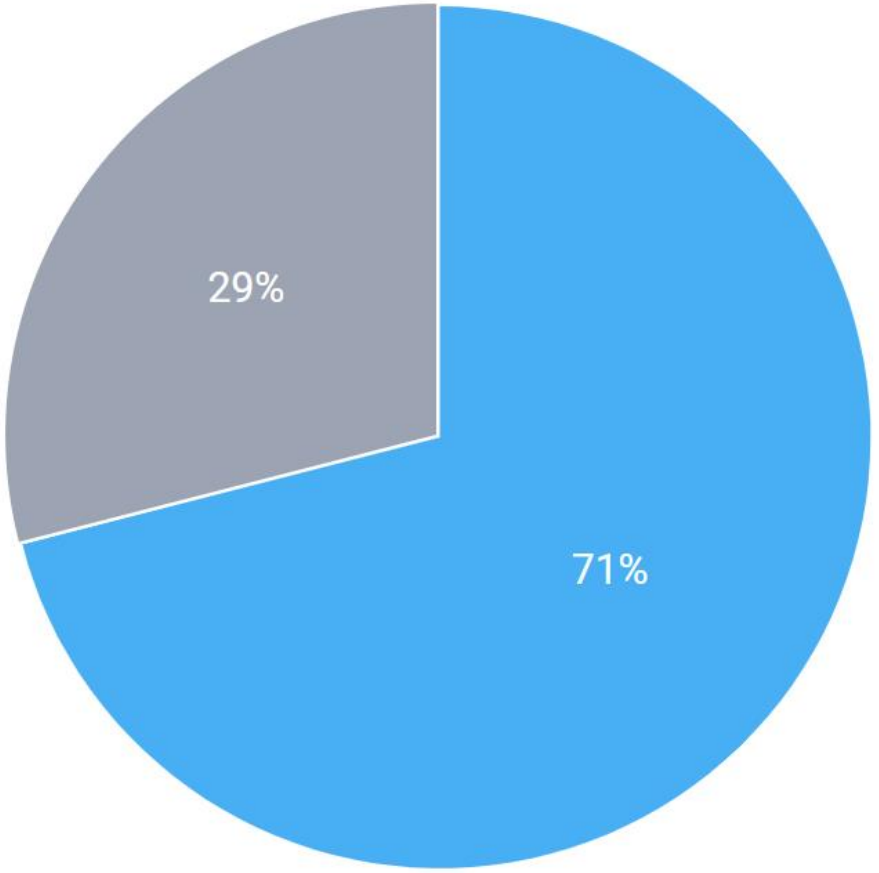


- Feedback loops and reinforcement of biases and inequalities

  E.g., recommendation systems on social media platforms recommend content that aligns with users' preexisting beliefs or preferences, thereby amplifying their biases

71% of employees in AI-exposed jobs are women

Gender breakdown of AI-exposed jobs



revelio labs

Using AI to promote fairness, equality and inclusion in society

- *Data-driven policy making:* AI can analyze large datasets to identify patterns of inequality and discrimination in sectors such as education, healthcare, and criminal justice

- *Replacement:* AI can replace human practices by ones that are less biased

- *Support of marginalized groups:* AI system can be used by marginalized groups and communities by supporting better information, communication, education and influence on policy-making

- *AI for Social Good Initiatives:* AI systems can be deployed in social good initiatives to address social challenges and promote equity and inclusion. E.g., AI-driven healthcare solutions for underserved communities.

THANK YOU