STEREOTYPES

AND BIAS IN AI

1. Prototype = Most typical representative of a concept/category (e.g. the typical “doctor” or the typical “bird”)
2. Stereotype = Perceived characteristics of an entire group of people (e.g. „men are not good at socializing“, „people of color are often criminal“, „Asians are good at math“)
3. Bias = Disproportionate weight in favor of or against an idea, thing or group of people

Note: A method from Social Psychology to detect unconscious biases is the **Implicit Association Test** (IAT, Greenwald, McGhee, & Schwartz, 1998).

1. Unconscious (implicit) human biases

* Biases may exist toward any social group
* may still influence decisions and actions we take
* Certain factors can activate unconscious beliefs

1. Widespread assumptions about human vs algorithmic decision-making

* People judge subjectively and have numerous bias effects in their decisions. Decisions can be influenced e.g. by momentary personal moods or even the weather. **TRUE**
* Computers / algorithms / AI systems, on the other hand, are objective and fair and able to judge neutrally **FALSE**

1. Biased AI

* Biased AI refers to AI systems that produce results that systematically and unfairly favor certain groups due to the presence of inherent biases in the data, thus reinforcing existing societal inequalities
* Fairness is “the absence of any prejudice or favoritism toward an individual or group based on their inherent or acquired characteristics”

1. Incomplete data: Selection bias

* Coverage bias: The population represented in the data set does not match the population that the machine learning model is making predictions about
* A model is trained to predict future sales of a new product based on phone surveys conducted with a sample of consumers who bought the product. Consumers who instead bought a competing product were not surveyed, and as a result, this group of people was not represented in the training data.

5. Sampling bias: Data is not collected randomly from the target group.

* Instead of randomly targeting consumers, the surveyer chose the first 200 consumers that responded to an email invitation (who might have been more enthusiastic about the product than average purchasers).
* Participation bias: Members of certain groups opt-out of surveys (or labeling tasks) at different rates than members of other groups
* Consumers who bought the competing product were 80% more likely to refuse to complete the survey, and their data was underrepresented in the sample.

1. How to keep human bias out of AI?

* Integration of bias detection strategies
* More diversity in teams developing AI would be very beneficial
* Awareness and understanding of the various causes of algorithmic biases is the first step
* Ethical governance standards

1. How to identify potential algorithmic bias?

* **Missing Feature Values** : indicator that certain key characteristics of your data set are under-represented
* **Unexpected Feature Values** : Look for examples that contain feature values that stand out as especially uncharacteristic or unusual
* **Data Skew** : certain groups or characteristics may be under or over-represented relative to their real-world prevalence