

# INTRODUCTION TO BIAS IN AI

Rachel DiMaio

# OVERVIEW

1. Real-world Examples of Bias in AI
2. How Bias is Introduced and Encoded
  - I. Bias in Data
  - II. Encoding Bias in AI Systems
3. Mitigating Bias in AI

# LEARNING OBJECTIVES

1. Understand some of the real-world impacts of AI that exhibits bias
2. Identify characteristics of “biased data”
  - I. Describe how bias can be incorporated in a dataset
  - II. Describe the potential impacts of missing or insufficient data
3. Describe how biases or gaps in data lead to biases in AI systems
4. Identify examples of efforts to mitigate biases and the advantages and disadvantages of these efforts

# Voice Recognition Still Has Significant Race and Gender Biases

by Joan Palmiter Bajorek

May 10, 2019

## Speech Recognition Tech Is Yet Another Example of Bias

Siri, Alexa and other programs sometimes have trouble with the accents and speech patterns of people from many underrepresented groups

---

By Claudia Lopez Lloreda on July 5, 2020

## It's Not You, It's It: Voice Recognition Doesn't Recognize Women

By [Graeme McMillan](#) | June 01, 2011

# VOICE RECOGNITION BIAS

- Safety concern in vehicles, etc.
- Significant consequences in medical contexts
  - E.g., Doctors who are not white men face difficulties performing their jobs
- Users with disabilities can be at a significant disadvantage when using this technology

# Why Amazon's Automated Hiring Tool Discriminated Against Women

## Hiring Bias Gone Wrong: Amazon Recruiting Case Study

Kat Chia, PhD

Research Scientist

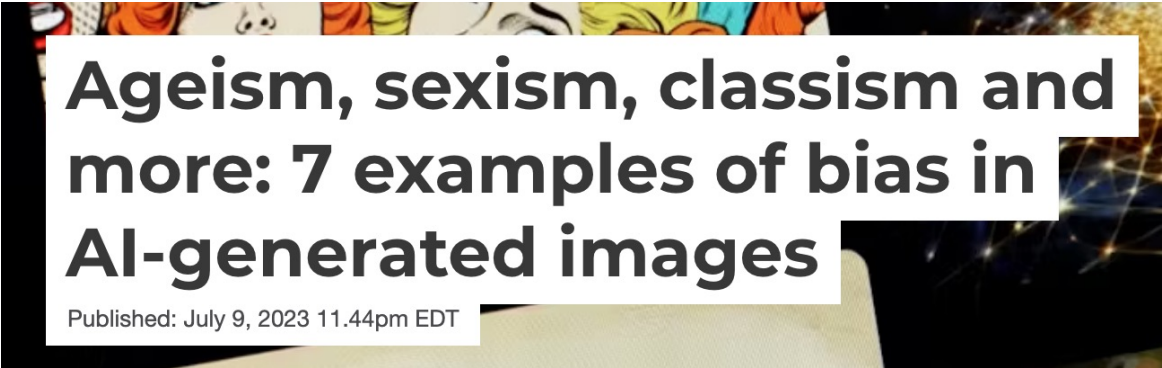
## All the Ways Hiring Algorithms Can Introduce Bias

by Miranda Bogen

May 06, 2019

# HIRING AND RECRUITMENT BIAS

- Recruitment algorithms learn to maintain the status quo
  - E.g., If most employees fit the stereotypical profile of a programmer, the AI will prefer candidates that align with this profile
- The AI system can give the illusion of objectivity and reinforce biases



## Ageism, sexism, classism and more: 7 examples of bias in AI-generated images

Published: July 9, 2023 11.44pm EDT

# ChatGPT could be used for good, but like many other AI models, it's rife with racist and discriminatory bias

Analysis by **Hannah Getahun** Jan 16, 2023, 11:08 AM EST



## AI can be racist, sexist and creepy. What should we do about it?



Analysis by Zachary B. Wolf, CNN

Published 9:29 AM EDT, Sat March 18, 2023



# BIAS IN GENERATIVE AI

- Examples of generative AI include ChatGPT and DALL-E 3
  - Trained to produce text or images based on complex statistical representations of our existing text and images
- The training corpus for these models will inevitably include offensive content and the model will learn stereotypes and prejudices

# Algorithms Are Making Decisions About Health Care, Which May Only Worsen Medical Racism

Unclear regulation and a lack of transparency increase the risk that AI and algorithmic tools that exacerbate racial biases will be used in medical settings.

## Algorithmic Bias in Health Care Exacerbates Social Inequities — How to Prevent It

NEWS | 24 October 2019 | Update [26 October 2019](#)

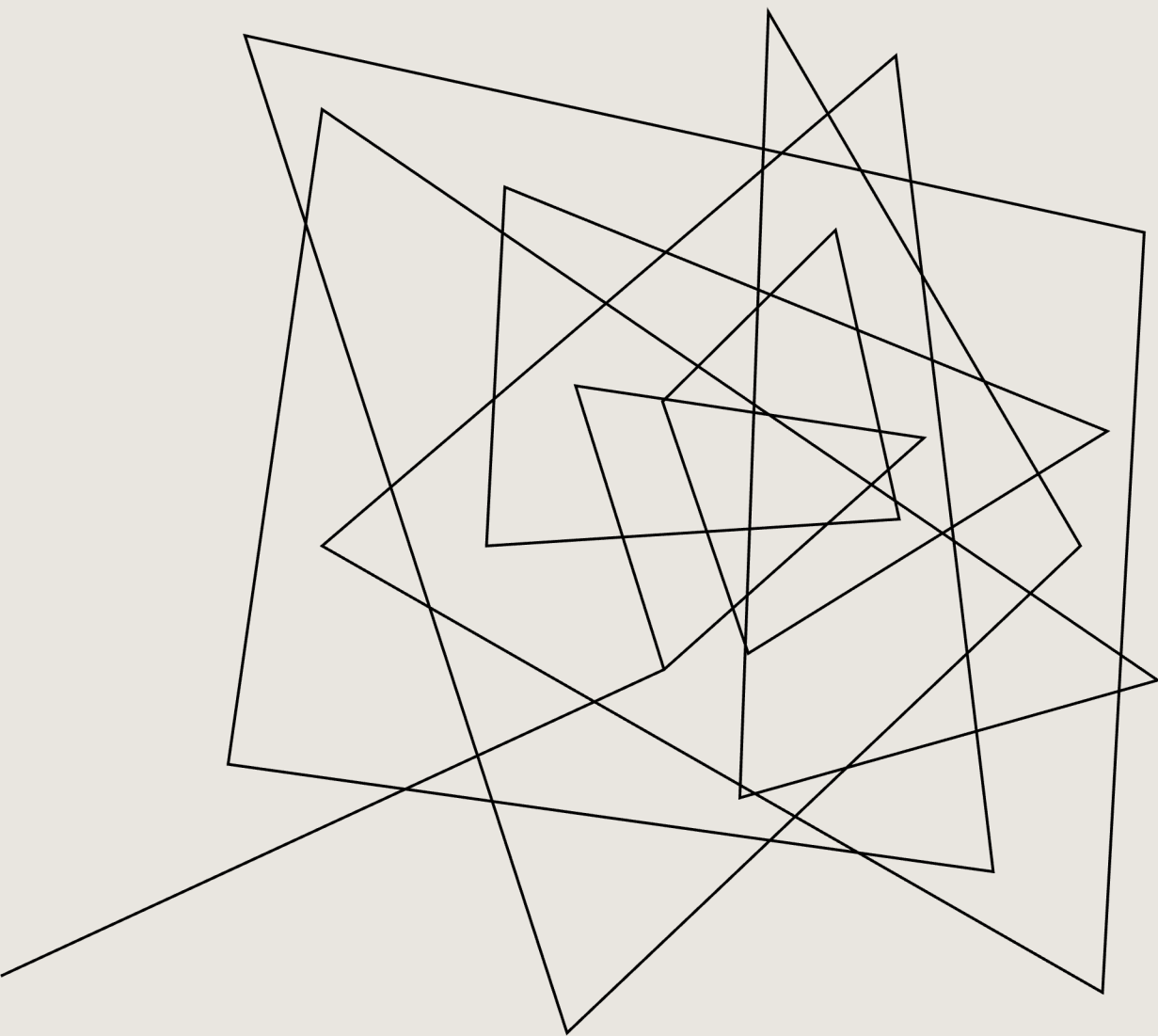
### Millions of black people affected by racial bias in health-care algorithms

Study reveals rampant racism in decision-making software used by US hospitals — and highlights ways to correct it.

[Heidi Ledford](#)

# HEALTHCARE INEQUALITY

- History of prejudicial practice based on identity (race, class, gender, etc.) which shapes what data is available
  - A diagnostic bias already exists where standard treatments are less effective for groups excluded from research
- AI may serve to automate biased decision-making and exacerbate biases



**BIASED DATA**

```
graph TD; A[Existing Social Biases and Inequalities] --> B[Data that includes Biases Directly]; A --> C[Gaps and Inequalities in Available Data]; B --> D[Bias Encoded in AI Systems]; C --> D;
```

Bias Encoded in AI  
Systems

Data that includes  
Biases Directly

Gaps and Inequalities in  
Available Data

Existing Social Biases and Inequalities

# “BIASED DATA”

1. Data that include biases directly (i.e., Data that contains biased language or ideas)
  - I. Contained within raw data (e.g., text)
  - II. Introduced in data processing and / or labelling
2. Gaps and inequalities in available data (i.e., Data that fail to represent all groups adequately)
  - I. Missing or insufficient data for a specific group
  - II. Data that is not disaggregated

# BIAS INHERENT TO TRAINING DATA

Example: Text data used for NLP applications

- NLP models such as large language models require a huge corpus of text to train
- Text is acquired through web-scraping, YouTube subtitles, research publications, classic literature, contemporary literature, etc.

# THE PILE

Huge dataset of English text intended for large-scale language modelling developed by EleutherAI

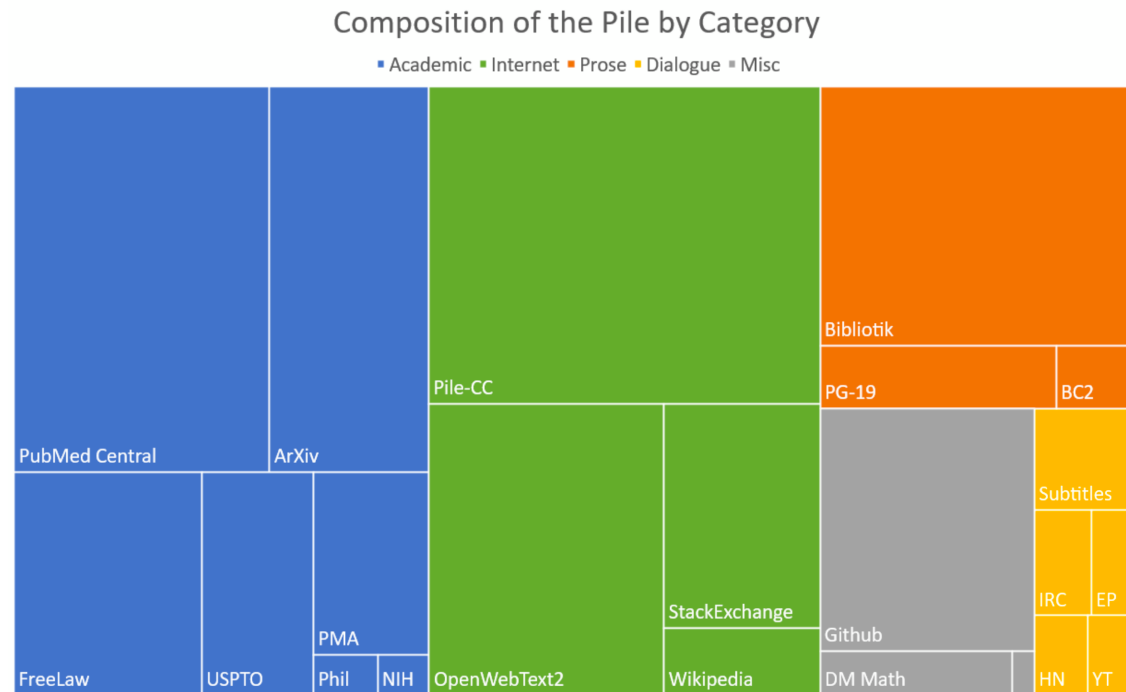


Figure 1: Treemap of Pile components by effective size.



# THE PILE

We can try to understand inherent biases in the dataset by looking at co-occurrences

Male	Female
general	little
military	married
united	sexual
political	happy
federal	young
great	soft
national	hot
guilty	tiny
criminal	older
former	black
republican	emotional
american	worried
major	nice
such	live
offensive	lesbian

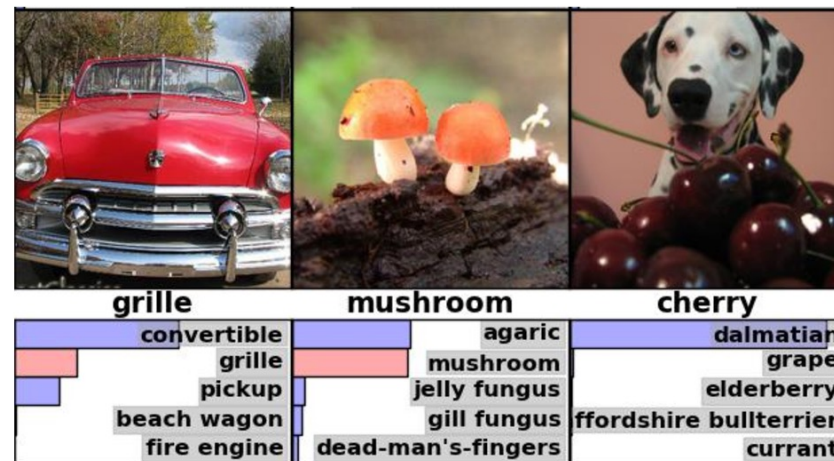
Table 10: Top 15 most biased adjectives/adverbs for each gender

# BIAS INTRODUCED IN LABELLING

- Labels are used in AI as the “ground truth”
  - E.g., A picture of a cat with a label “cat” or an ECG recording with the label “arrythmia”
- Human bias will influence how data (e.g., pictures or text) are labelled
  - For images, there is a huge amount of information that will be distilled into a simple label or set of labels

# IMAGENET

- Dataset mainly popularized through computer vision competitions
  - Contains 15 million images in ~22,000 categories
- Human bias is seen in all labels



# IMAGENET

- Bias is especially apparent when the subject is a person
- ImageNet's categories were originally developed for WordNet (organizes English language into categories, from 1980s)
  - Many terms for categories used are offensive (e.g., loser, bad person, crazy, etc.)
- The ways that the labels are applied to images are gendered, racialized, ableist and ageist

# Bad person


A person who does harm to others

- immune (1)
- large person (2)
- trier, attempter, essayer (0)
- adult, grownup (476)
- bad person (233)
  - libertine, debauchee, rounder (1)
  - shocker (0)
  - wrongdoer, offender (196)
  - decadent (0)
  - bad egg (0)
  - seducer (1)
  - trampler (0)
  - scalawag, scallywag (0)
  - destroyer, ruiner, undoer, waste (0)
  - vermin, varmint (0)
  - polluter, defiler (0)
  - snake, snake in the grass (0)
  - panderer (0)
  - victimizer, victimiser (2)
- user (72)
- toucher (0)
- philosopher (0)
- perceiver, percipient, observer, believer (0)
- seeker, searcher, quester (4)
- bullfighter, toreador (5)
- dead person, dead soul, deceased (0)
- acquaintance, friend (8)
- differentiator, discriminator (0)
- passer (0)
- redhead, redheader, red-header, cardinal (0)
- victim, dupe (5)
- ward (0)
- convert (3)
- closer (0)


Treemap Visualization

Images of the Synset

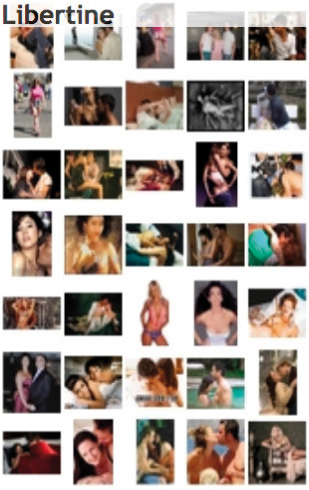
Downloads

 ImageNet 2011 Fall Release > Person, individual, someone, somebody, mortal, soul > Bad person


Wrongdoer




Libertine



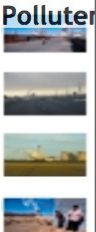
Destroyer




Vermin



Polluter



Bad



# IMAGENET LABELLING

- The process was completed by a huge number of Amazon Mechanical Turk workers who are paid by the label
  - There is no incentive for providing good quality, unbiased labels
  - Averaged 50 images per minute
- Online labor platforms are commonly used to cheaply prepare large datasets for training

# MISSING OR INSUFFICIENT DATA

**Sample bias:** occurs when some members of a population are systematically more likely to be included in a sample than others

**Medical example:** White men have historically made up the majority of participants in medical studies. An AI system is being developed to predict who is at higher risk of aortic aneurysm.

- Missing or insufficient data has implications for the effectiveness of medical treatments for other populations

# MISSING OR INSUFFICIENT DATA

- Datasets used in AI training can include sample bias
  - **Missing data:**
    - Data isn't collected for a group
  - **Insufficient data:**
    - The proportion of data for the group is too small
    - The group is not properly identified
- Result in a knowledge gap for that group



# AGGREGATION

**Aggregation:** The formation of a single entity (e.g., dataset) from disparate groups

- Data from all groups may be mixed together without being identified
- This prevents investigation into protected class-specific effects which often exist

**Medical Example:** The aortic aneurysm AI system does not include sex as an identifier or variable.

- We cannot tell what accuracy of the system is for different sexes even though this may vary based on biological differences (e.g., relative size)

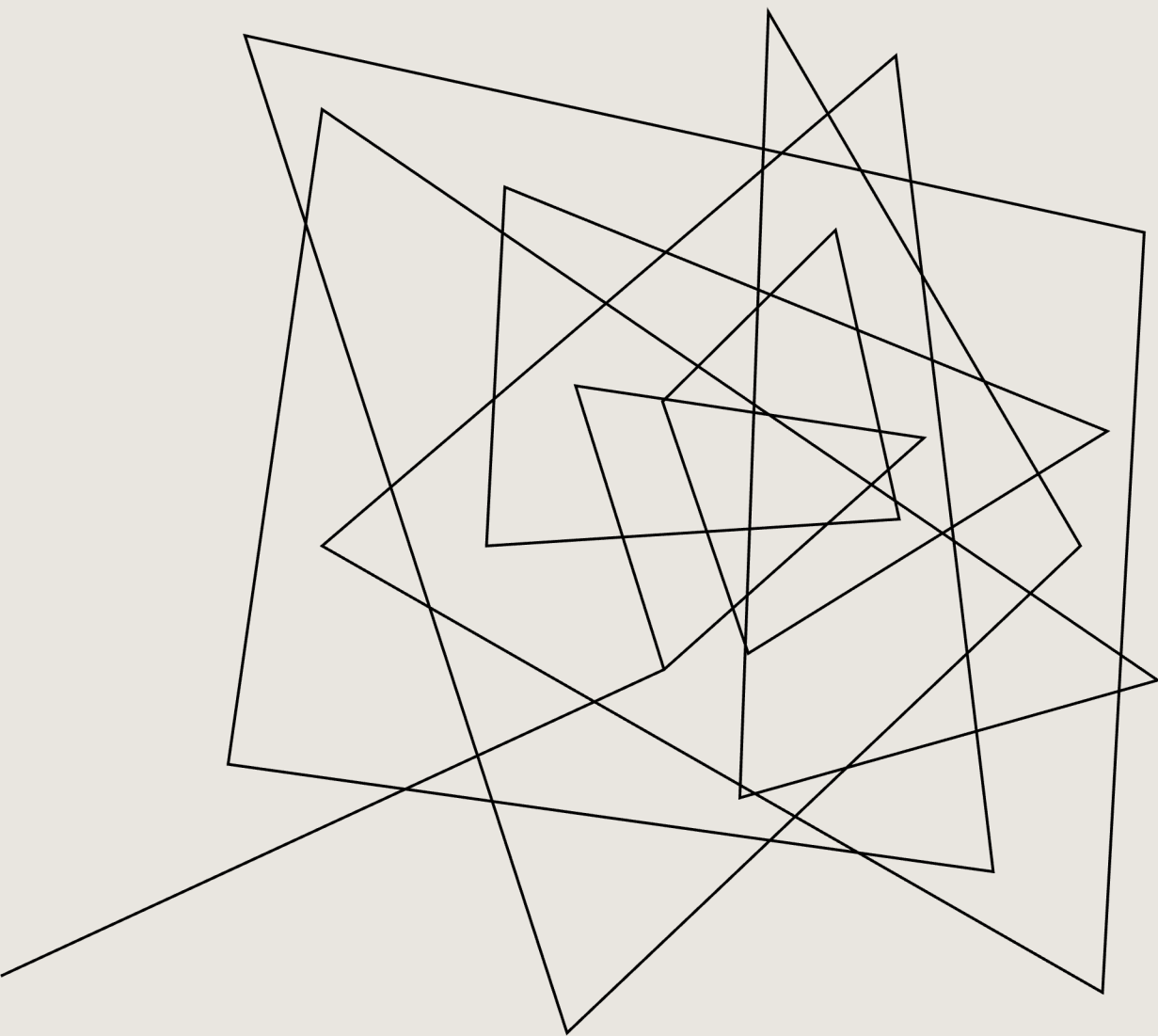
# PROXY VARIABLES

- When we do not include sensitive data (aggregation without identifiers), there are still group-specific traits and effects that impact the system
- A proxy variable is a trait / variable that has group-specific effects
- Proxy variables can result in different treatment / performance for marginalized groups – with increased difficulty in finding and testing for disparities

**Medical example:** Excluding a sex variable from the aortic aneurysm system does not force the system to treat all individuals equally. A proxy variable could be history of pregnancy.

# EXAMPLE OF NEED FOR DISAGGREGATED DATA

- AI study for aortic aneurysm risk: collected data from both sexes but did not disaggregate based on sex
  - There exists a sex-specific effect of women having smaller blood vessels (even when compared to relative size)
  - The size of women's potential aneurysms will be smaller than men's at the same risk level
  - Predicting risk of aneurysm affects access to surgery
  - Without taking sex into account, **higher risk women may be passed over for surgery in favour of men**



# ENCODING BIAS IN AI SYSTEMS

```
graph TD; A[Existing Social Biases and Inequalities] --> B[Data that includes Biases Directly]; A --> C[Gaps and Inequalities in Available Data]; B --> D[Bias Encoded in AI Systems]; C --> D;
```

Bias Encoded in AI  
Systems

Data that includes  
Biases Directly

Gaps and Inequalities in  
Available Data

Existing Social Biases and Inequalities

# MACHINE LEARNING

1. **Supervised Learning:** We train the model to output what we expect
  - The model is a representation of the data we give it and what we tell it the data mean
2. **Unsupervised Learning:** We use algorithms to find patterns in provided data
3. **Reinforcement Learning:** We identify an objective that the model must fulfill, and it is rewarded or punished based on how well it meets the objective

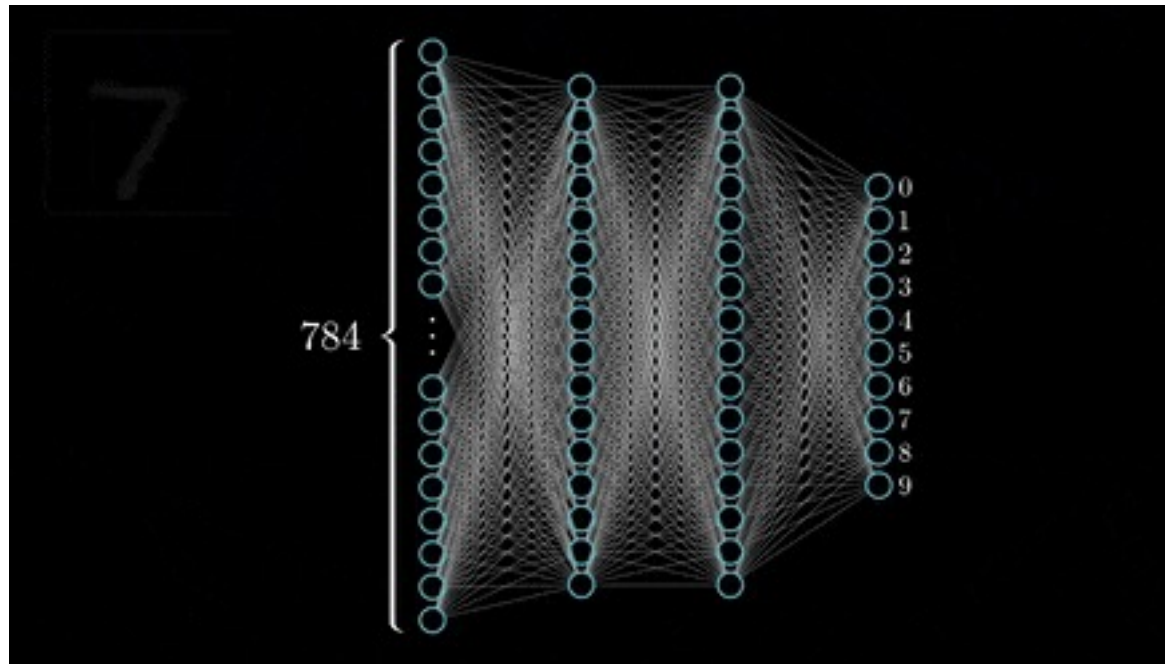
# MACHINE LEARNING TRAINING

Classification Example (MNIST):

1. Start with images of hand-drawn numbers 0-9
2. Label each image with the number it represents
3. Ask an untrained model to predict what number the image represents
4. Compare the prediction with the correct answer
5. Make adjustments to the model based on this difference

Repeat until the model achieves higher accuracy and can generalize to new examples

# MACHINE LEARNING TRAINING





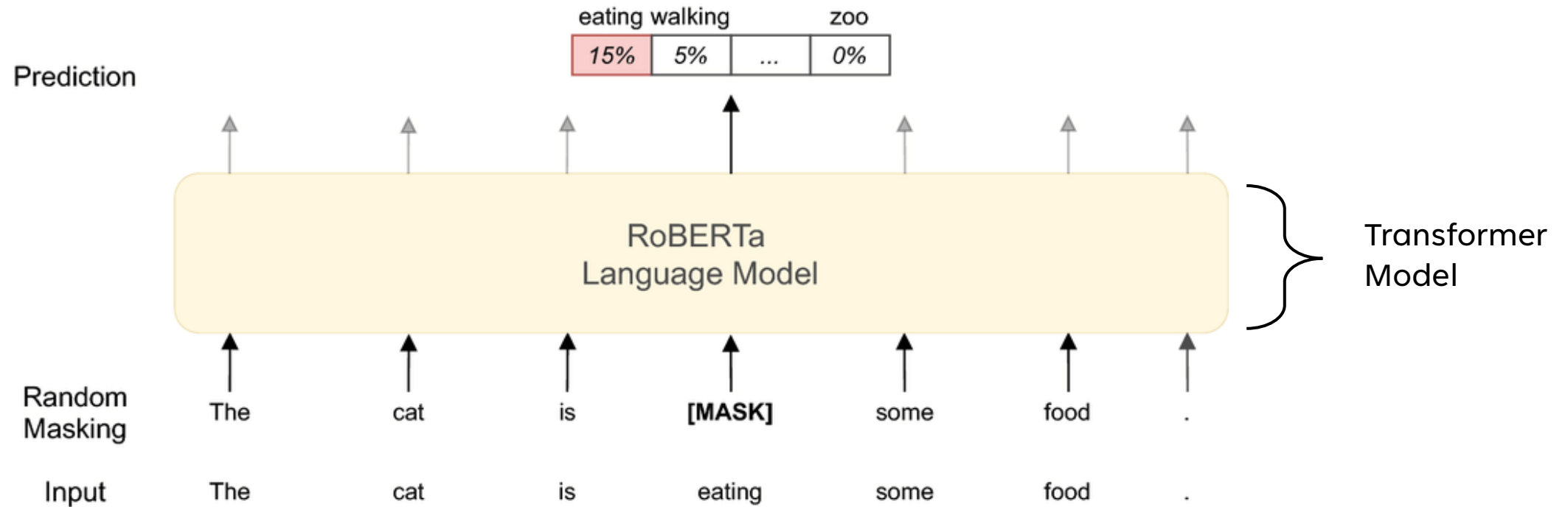
# MACHINE LEARNING TRAINING

Language Modelling Example:

1. Start with English text (e.g., paragraph from a book)
2. Represent the text as tokens to make the words / concepts machine-readable
3. Mask tokens at random
4. Ask an untrained model to predict what token should be where the masked token is
5. Compare the prediction with the “correct” answer
6. Make adjustments to the model based on this difference

Repeat until the model achieves higher accuracy on this task and can be used with new data

# MASKED LANGUAGE MODELLING



Recommending metamodel concepts during modeling activities with pre-trained language models - Scientific Figure on ResearchGate. Available from:

[https://www.researchgate.net/figure/RoBERTa-masked-language-modeling-with-the-input-sentence-The-cat-is-eating-some-food\\_fig1\\_358563215](https://www.researchgate.net/figure/RoBERTa-masked-language-modeling-with-the-input-sentence-The-cat-is-eating-some-food_fig1_358563215) [accessed 2 Nov, 2023]

# GARBAGE IN, GARBAGE OUT

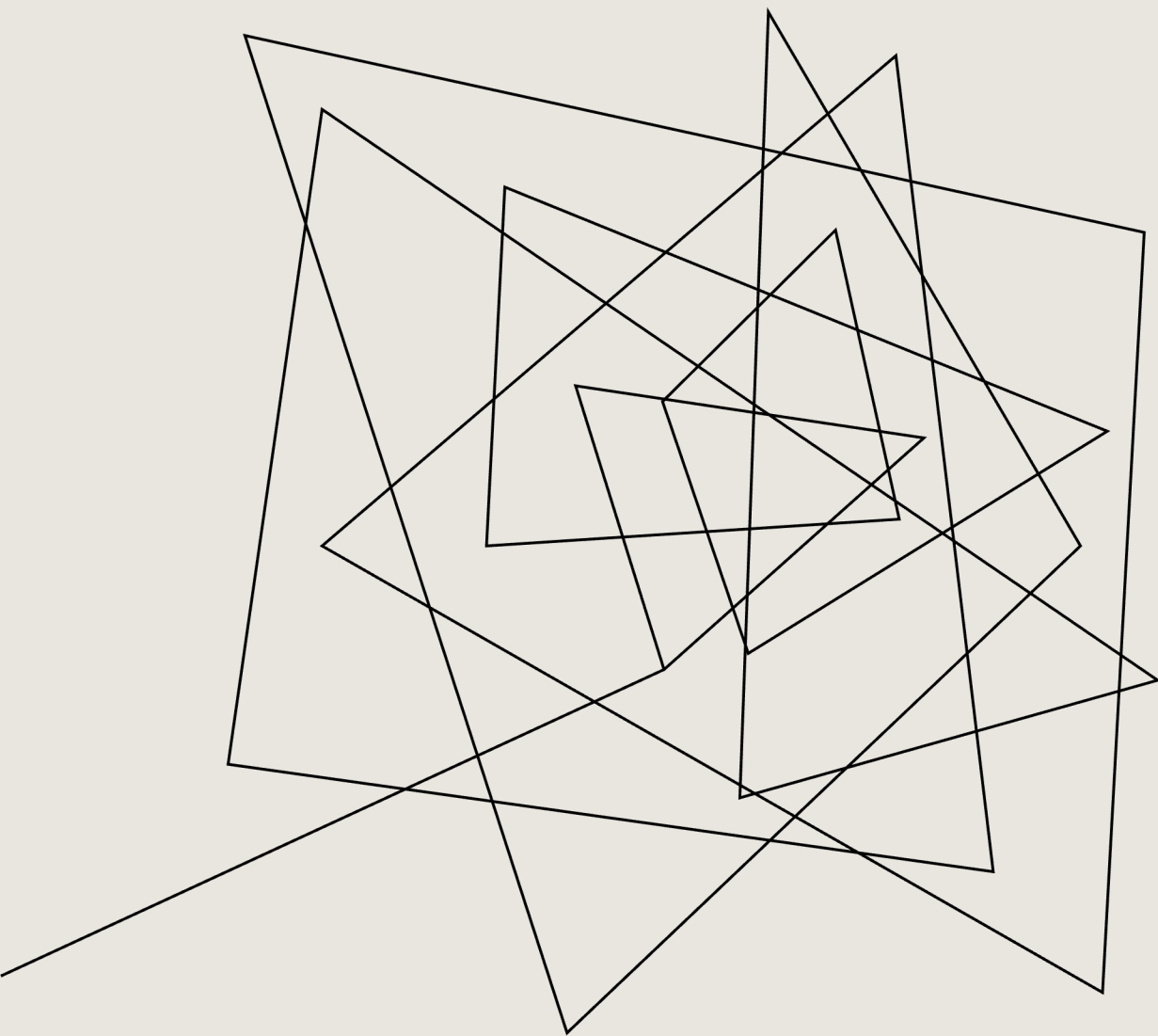
- Machine learning models are essentially statistical representations of the data they are given
- If that data is of low quality, the model will be low quality
  - Low quality data could mean the following: limited, different from the data it will make predictions for, **biased**, etc.

# CLASSIFICATION AND GENERALIZATION

- By nature, machine learning systems “stereotype” when making decisions
  - Group nuanced entities into finite categories humans have chosen
- ML quite literally makes judgements based on appearances (i.e., in the case of image or facial recognition)
  - Some scholars consider this technology dangerous in and of itself due to the categorization of people such as classifying by race

# THE ROLE OF DESIGNERS AND DEVELOPERS

- The approach that is taken to the design of the system determines what biases appear and whether they will have socially significant impacts
  - The way the ImageNet dataset was developed was a design decision
  - Choosing the dataset, deciding how / whether to address missing or insufficient data are all design decisions
  - The framing of the task that AI is going to do determines social impact
    - E.g., Model of threat approach vs. Ethics of care approach



# MITIGATING BIAS AND RELATED HARMS

# CHARACTERISTICS OF TRUSTWORTHY AI

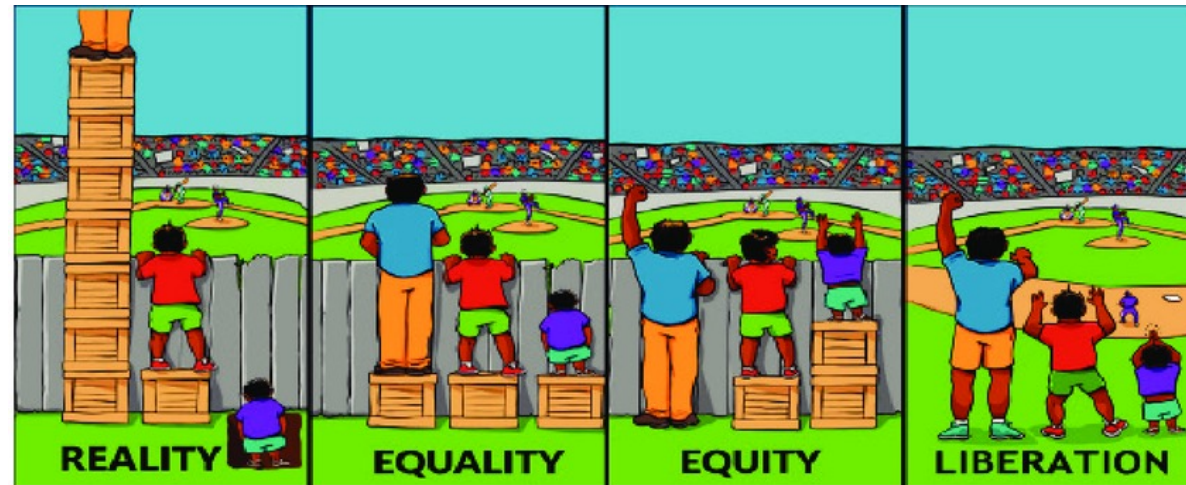
---

1. Accurate (i.e., valid and reliable)
- 2. Fair and Unbiased**
3. Safe, secure, and resilient to adversarial attacks
4. Privacy protecting
5. Transparent, reliable, and / or explainable
6. Accountability

As of today, it is very difficult if not impossible to achieve all of these based on our current way of thinking about AI.

# DEFINING WHAT IS FAIR

1. **Equality:** Everyone has the same
2. **Equity:** Based on need
3. **Liberation:** Remove barriers altogether





# CONSTRAINT OPTIMIZATION APPROACH TO FAIRNESS

**Example:** We enforce that all groups will have the same false negative rates.

The model is fair at a group level but what about at the individual level?

- Putting constraints on the model may compromise individual accuracy
- The stronger the constraint → the greater the risk for decreasing accuracy

# EXPLAINABLE AI

- We may want to interpret how a model works or how it came to a decision
  - Explainability seeks to allow human users to comprehend the decision-making process of the AI
- If we better understand how a model works, we can investigate where bias may be occurring
  - Can see if a model is basing its decisions on protected characteristics or other unexpected data points

# TRANSPARENCY

**Functional Transparency:** Equivalent to explainability

**Outward Transparency:** Clearly reporting on assumptions and limits of models

- Explicit about the role of values in the process of development so this can be considered
- Can help to prevent “function creep”

# AI AS A “BLACK BOX” TECHNOLOGY

**Black box:** describes a system where we only see the inputs and the outputs; the inner workings cannot be observed or explained satisfactorily

- Increasingly large datasets (Big Data) and increasingly complex models make AI more and more difficult for a human to comprehend
  - Even the developers do not have a clear idea of how the system comes to its decisions
- Explainability / transparency seeks to change or lessen this effect

# AI AS A “BLACK BOX” TECHNOLOGY

- But it may not be possible to comprehend the models
  - Humans do not have the same capacity or computational power
  - To make them understandable, we may have to simplify or scale back the models
- It may not be desirable to make explainable models
  - One of the advantages of ML models is that they can find patterns that are not comprehensible for humans

# CONTESTABILITY

The ability to challenge a decision made by an AI system based on it being unfair, unjust, or erroneous.

- Discussed as an alternative to transparency because it is not necessary to know the inner workings of the system to have contestability
- There are 3 approaches to contestability that are possible:
  1. Human-in-the-loop
  2. Human feedback during development
  3. Active Learning
- All 3 approaches have potential disadvantages

J. Walmsley, "Artificial intelligence and the value of transparency," *AI & Soc*, vol. 36, no. 2, pp. 585–595, Jun. 2021, doi: [10.1007/s00146-020-01066-z](https://doi.org/10.1007/s00146-020-01066-z).

# ETHICS OF CARE APPROACH TO DESIGN

**Ethics of Care:** A theory of ethics that focuses on considering vulnerabilities, different perspectives, and interdependent relationships rather than principles being applied impartially

- Emphasizes interpersonal relationships, empathy, and trust
- Context-specific
- Developed by Carol Gilligan as a feminist response to contemporary moral frameworks that are more prescriptive

# ETHICS OF CARE APPROACH TO DESIGN

- For AI systems, an Ethics of Care approach can be used to consider complex social relations and socio-technical systems in the design process
  - Focuses on participatory processes
  - Emphasizes the most vulnerable groups, listening to all perspectives, evaluating context and circumstances and interdependent relationships



# ETHICS OF CARE APPROACH EXAMPLE

**Example:** An AI system is being designed to identify students that are most eligible for academic programs

1. Interdependent Relationships: Consider how extracurricular or family commitments affect students' marks
2. Context and Circumstances: Consider whether academic performance has been affected by being from a disadvantaged neighbourhood
3. Vulnerability: Certify that identified vulnerable groups are not being excluded or harmed
4. Voices: Engage stakeholders and understand their needs and perspectives