CPSC 430
Computers & Society

Class 6B: Bias, Fairness, and Artificial Intelligence

Dr. Firas Moosvi | 2024_S1

Slides courtesy of Dr. Kevin Leyton-Brown

Class Outline

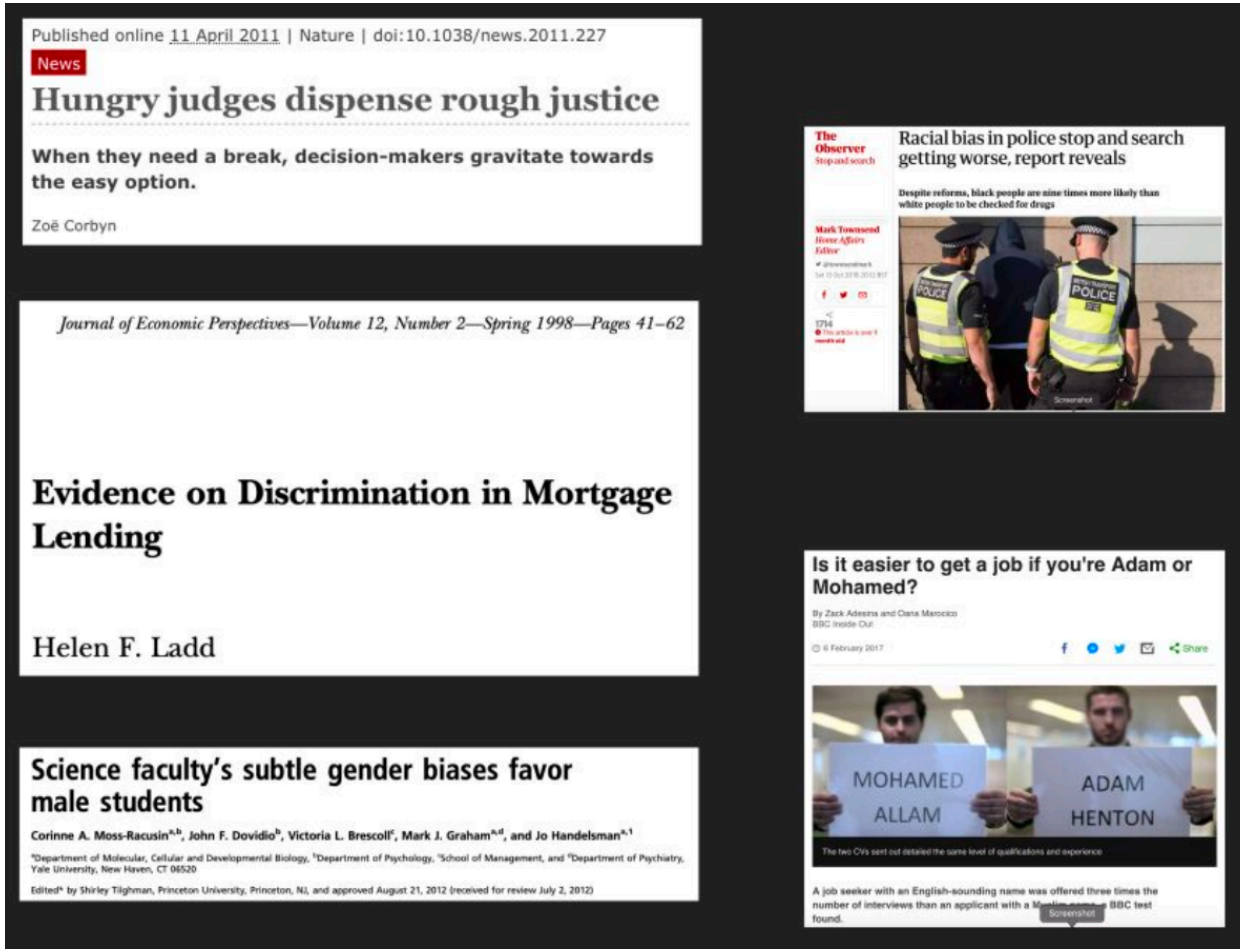
1. Announcements (5 mins)
2. Bias and Fairness (50 mins)
3. Break (10 mins)
4. Peering into the crystal ball of AI (50 mins)
    - Break (10 mins)

Announcements

Bias and Fairness

Human bias

Bias in people refers to our tendency to take quick decisions based on little information

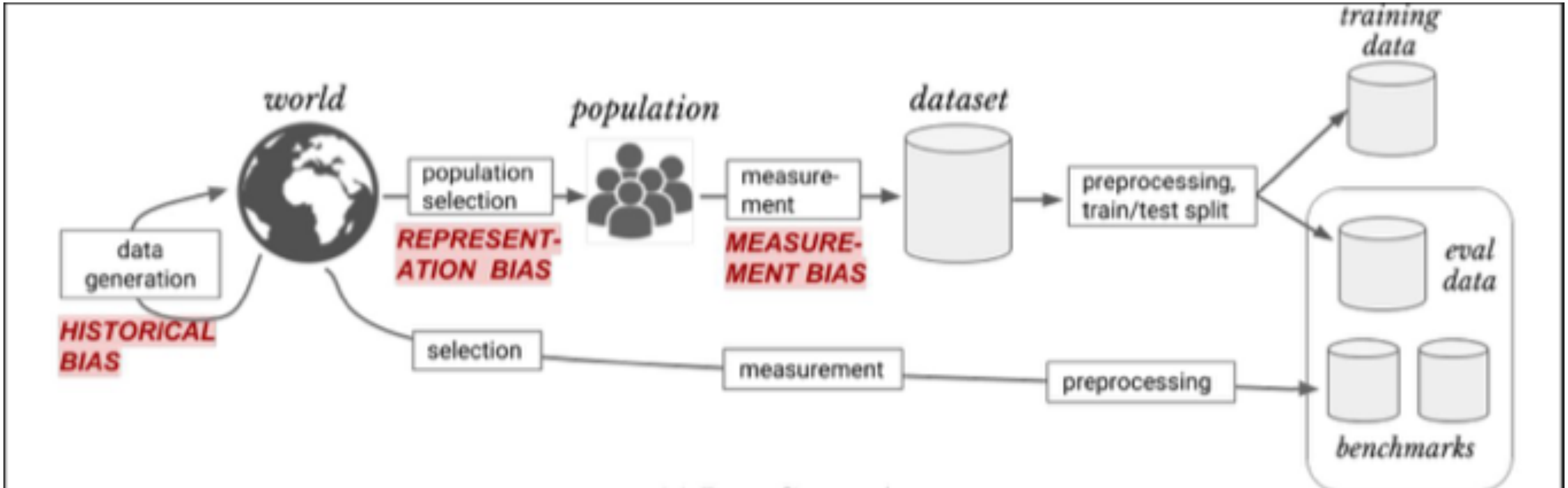


Why worry about bias in algorithms

Decisions made by a ML algorithm are:

- Cheap
- Scalable
- Automated
- Self-reinforcing
- Seemingly objective
- Often lacking appeals processes
- Not just predicting but also causing the future

Sources of bias in ML algorithms



(a) Data Generation

Representation bias

Representation bias arises while defining and sampling a development population. It occurs when the development population under-represents, and subsequently fails to generalize well, for some part of the use population. [1]

1. **The sampling methods only reach a portion of the population.** For example, datasets collected through smartphone apps can under-represent lower-income or older groups, who are less likely to own smartphones. Similarly, medical data for a particular condition may be available only for the population of patients who were considered serious enough to bring in for further screening.

2. **The population of interest has changed or is distinct from the population used during model training.** Data that is representative of Boston, for example, may not be representative if used to analyze the population of Indianapolis. Similarly, data representative of Boston 30 years ago will likely not reflect today's population.

Measurement bias

Measurement Bias arises when choosing and measuring features and labels to use; these are often proxies for the desired quantities. The chosen set of features and labels may leave out important factors or introduce group- or input-dependent noise that leads to differential performance.

1. **The measurement process varies across groups.** For example, if a group of factory workers is more stringently or frequently monitored, more errors will be observed in that group. This can also lead to a feedback loop wherein the group is subject to further monitoring because of the apparent higher rate of mistakes (Barocas and Selbst, 2016;

2. **The quality of data varies across groups.** Structural discrimination can lead to systematically higher error rates in a certain group. For example, women are more likely to be misdiagnosed or not diagnosed for conditions where self-reported pain is a symptom (Calderone, 1990). In this case, "*diagnosed* with condition X" is a biased proxy for "has condition X."

3. **The defined classification task is an oversimplification.** In order to build a supervised ML model, some label to predict must be chosen. Reducing a decision to a single attribute can create a biased proxy label because it only captures a particular aspect of what we really want to measure. Consider the prediction problem of deciding whether a student will be successful (e.g., in a college admissions context). Fully capturing the outcome of 'successful student' in terms of a single measurable attribute is impossible because of its complexity. In cases such as these, algorithm designers resort to some available label such as 'GPA' (Kleinberg et al., 2018), which ignores different indicators of success achieved by parts of the population.

Historical bias

Historical bias arises when there is a misalignment between world as it is and the values or objectives to be encoded and propagated in a model. It is a normative concern with the state of the world, and exists even given perfect sampling and feature selection.

Example: image search In 2018, 5% of Fortune 500 CEOs were women (Zarya, 2018). Should image search results for "CEO" reflect that number? Ultimately, a variety of stakeholders, including affected members of society, should evaluate the particular harms that this result could cause and make a judgment. This decision may be at odds with the available data even if that data is a perfect reflection of the world. Indeed, Google has recently changed their Image Search results for "CEO" to display a higher proportion of women.

Break

Fairness in algorithms

- Increasing attention on algorithms being **fair**, not just accurate
- Fairness can be measured as:
  - demographic (or statistical) parity: population percentage should be reflected in the output classes
  - Equality of false negatives or equalized odds: constant false-negative (or both false-negative and true-negative) rates across groups.
  - Equal opportunity: equal True Positive Rate for all groups
  - Other metrics...
- Accuracy and fairness tend to be at odds with each other.
- Algorithms can be audited to test their fairness.
- *Are we ethically required to sacrifice accuracy for fairness? To what extent?*

When the metric becomes the target (Goodhart’s Law)

"When a measure becomes a target it ceases to be a good measure"

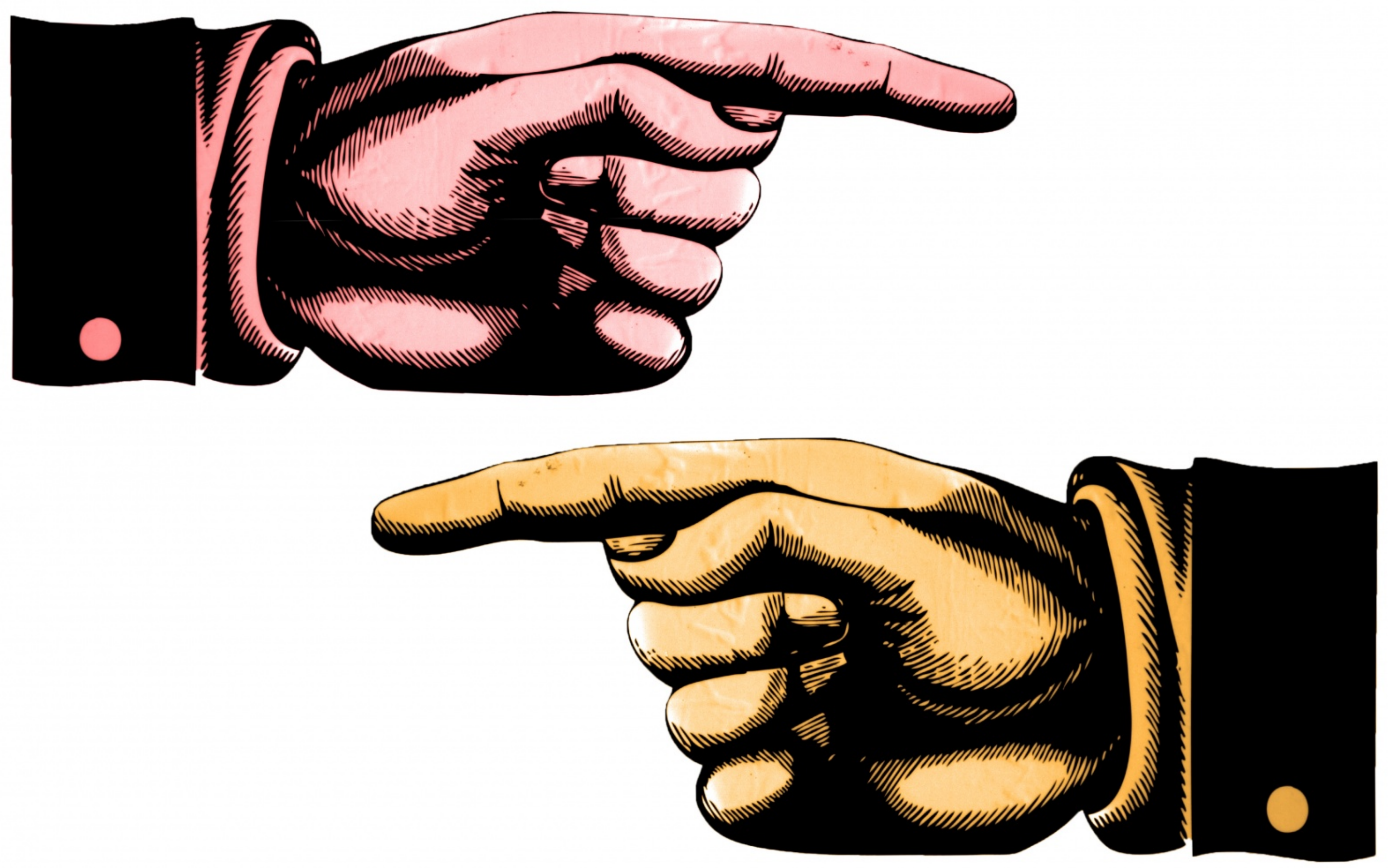
- Metrics introduced in the British public healthcare system (e.g. waiting time in ER) caused people to game it:
  - Cancelled scheduled operations to draft extra staff to ER
  - Required patients to wait outside the ER, e.g. in ambulances
  - Put stretchers in hallways and classified them as "beds"
  - Hospital and patients reported different wait times
- Big Data is significantly changing college applications
  - Universities are given higher ranking for things such as receiving more applications, being more selective, and having more students accept their offers (while tuition is not considered)
  - This even pushed some mid-tier universities to reduce the number of offer letter sent out, especially to good students who they think will not accept. Can affect applications to “safety schools”
- *Can you think of more examples?*
- *Can you think of ways to avoid this trap?*

Ethics of pricing algorithms

- Algorithms are currently used to adjust prices based on:
  - Willingness of buyer
  - Availability
- Uber surge pricing:
  - In 2014, terrorists attacked a café in Sidney, holding 10 customers and 8 employees hostage for 16 hours
  - During this time, people from the surrounding areas were evacuated. Transportation was disrupted.
  - Uber prices adapted by increasing the rate to a minimum of 100$
  - In general, underserved (poorer) areas get worse rates under current pricing policy
  - Drivers are also subjected to different pricing/waiting times
    - https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4331080
- *Is Uber morally obliged to avoid such pricing disparities?*

From trading to gambling addiction?

- Case study: Robinhood Trading app
- Designed to make trading more accessible and equitable. Robinhood's mission statement is "To democratize finance for all"
- Concerns around the most "dopamine-inducing" features (source):
  - Green confetti to celebrate transactions.
  - A constant update of stock related articles.
  - A colorful, eye-catching interface.
  - Emoji phone notifications.
  - One-click trading for instant gratification.
  - Free stocks in the shape of lottery tickets.



Peering into the Crystal Ball

PEERING INTO THE CRYSTAL BALL

Bold predictions about where AI is headed in the longer term

Technology that will be in widespread use

- Tailored solutions for specific tasks, not general intelligence
- Prototypes that work today in labs & narrow deployments
- Some examples:
  - Non-text input modalities (vision; speech)
  - Consumer modeling (recommendation; marketing)
  - Cloud services (translation; question answering; AI-mediated outsourcing)
  - Transportation (automated trucking; some self-driving cars)
  - Industrial robotics (factories; some drone applications)
  - AI knowledge work (logistics planning; radiology; legal research; call centers)
  - Policing & security (electronic fraud; cameras; predictive policing)

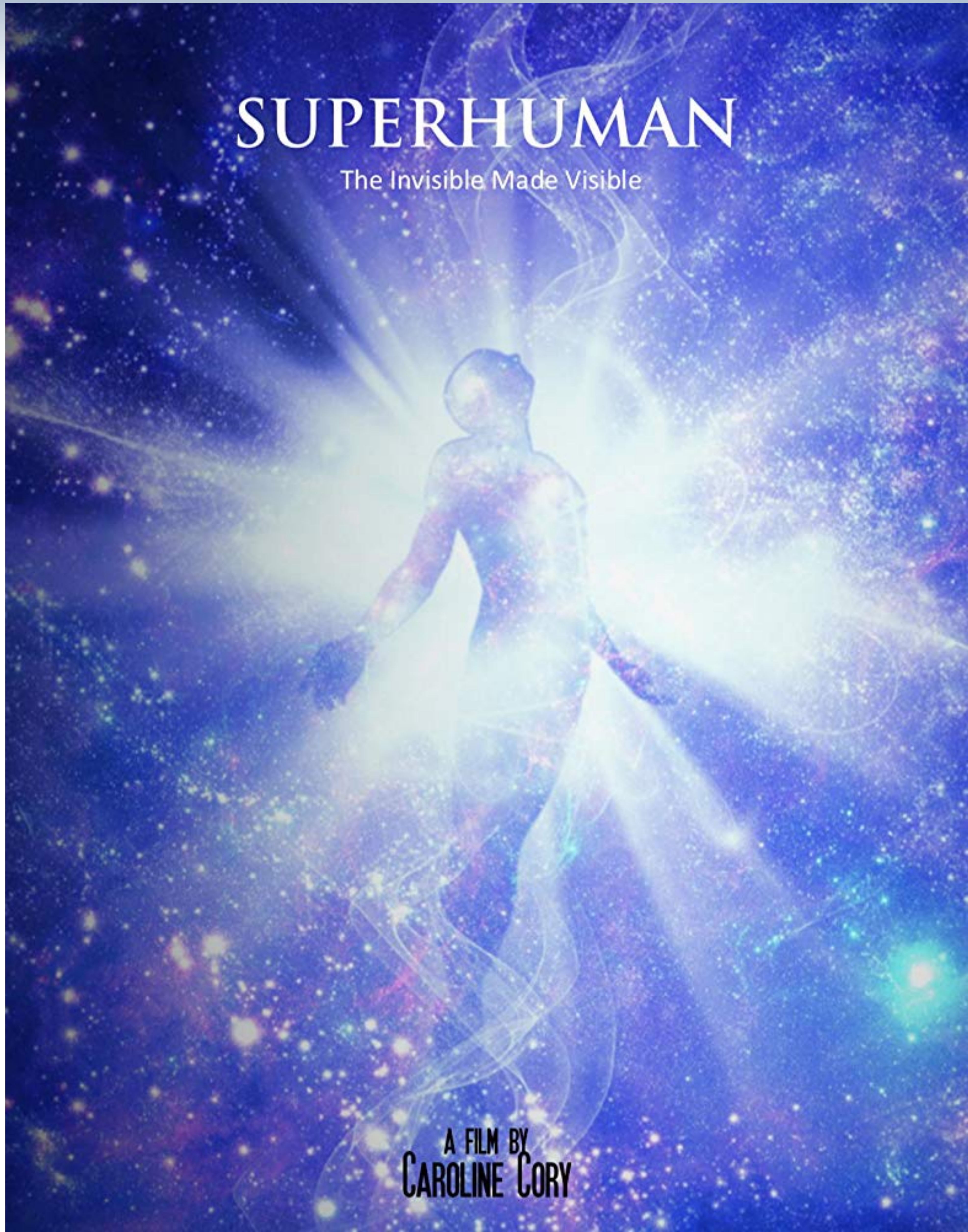
Technologies that won't take off as quickly

- Overall, areas in which
  - major entrenched **regulatory regimes** need to be navigated
  - there exist **social/cultural barriers** to adoption
  - the **human touch** is crucial
  - substantial **new hardware** would need to be developed

- Some **examples**:
  - childcare, healthcare, eldercare
  - education
  - coaching, counselling
  - consumer robots beyond niche applications
  - semantically rich language understanding

Superhuman Intelligence

- AI systems will increasingly be capable of reaching **human-level performance**
- **Superhuman intelligence** isn't such a foreign, scary thing
  - governments, corporations, NGOs exhibit behavior much more sophisticated and complex than that of any individual
- Many important problems need superhuman intelligence; AI can help
  - improved **collective decision making**
  - more efficient use of **scarce resources**
  - addressing **underserved communities**
  - **climate change**; other societal challenges



Break

Ethics of AI

Will a new technology:

- disempower **individuals vs corporations**?

  ⇒ user modeling; data mining; fostering addictive behaviors; developmental effects on children

- disempower **individuals vs governments**?

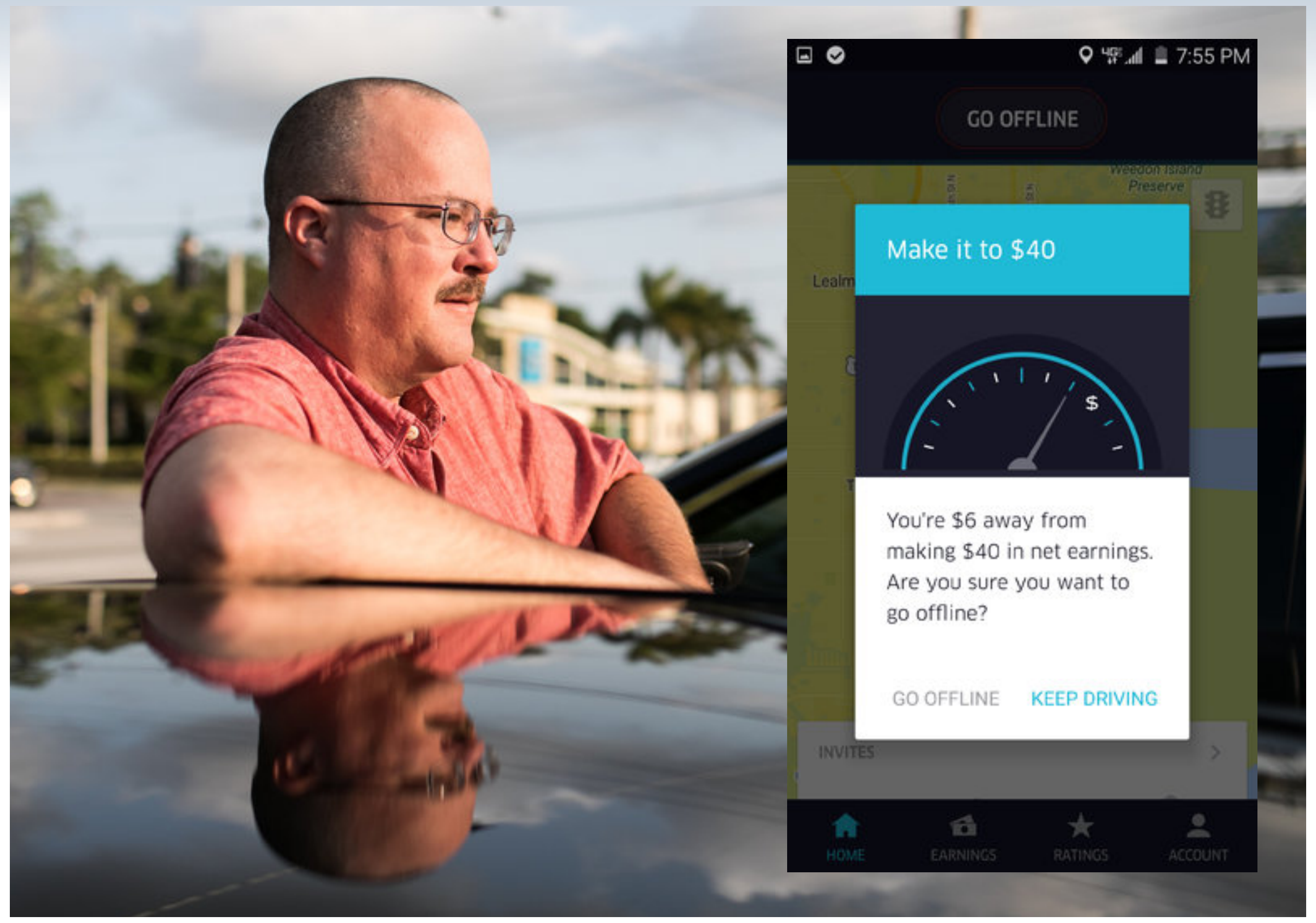
  ⇒ facilitate disinformation (deep fakes; bots masquerading as people; filter bubbles); enable qualitatively new military or security tactics

- take **autonomous actions** in a way that obscures responsibility

  ⇒ autonomous weapons; self-driving cars; loan approval systems

- disproportionately affect **vulnerable/marginalized groups**

  ⇒ automated decision making tools trained in ways that may encode existing biases



Social Impact

- How will AI technologies **transform society**?
- Will there be a **social backlash** against AI?
  - If so, what will be considered AI?
- This **generation of children** will grow up taking for granted many technologies that strike us as magical
- How will **human relationships** change in the presence of always-available social agents?
- As we are increasingly **augmented by AI**, what are our inherent cognitive/emotional/motivational limitations, beyond which augmentation won't help?

Ethical consideration of advances in AI

- *Is it wrong to create machines capable of making human labor obsolete? Will humans become demoralized by the presence of vastly more intelligent robots?*
- *How can we ensure that intelligent robots will not be put to an evil purpose by a malevolent human? How can we ensure they do not adopt malevolent purposes themselves?*
- *Is it morally acceptable to create "personal" (self-conscious) AI?*

Photo by Giftpundits.com from Pexels

Reminders before the Final Exam!