

Responsible Data Science

Introduction and Overview

January 23, 2024

Professor Umang Bhatt

Center for Data Science &
Computer Science and Engineering
New York University



course logistics

Instructor: Umang Bhatt

Assistant Professor and Faculty Fellow
Center for Data Science
New York University

Ph.D. in Engineering from University of Cambridge
M.S. and B.S. in ECE from Carnegie Mellon University

Research: Trustworthy Machine Learning

- Algorithmic Transparency (Explainable AI and Uncertainty Quantification)
- Human-Machine Collaboration
- Decision Support Systems



And also:

- NGOs: Center for Democracy and Technology, OECD, Mozilla Foundation, Partnership on AI, Responsible AI Institute
- Outreach: Deep Learning Indaba, The Alan Turing Institute

Office hours: Tuesdays 5-6 ET and by appointment

Instructor: Jonathan Colner

Assistant Professor and Faculty Fellow
Center for Data Science
New York University

Ph.D. in Political Science from UC Davis
B.A. in Political Science from University of Chicago



Research: Local Political Institutions

- Policymaking in City Councils
- Collecting Local Administrative Data using AI

Office hours: Thursdays 5-6 ET and by appointment

Teaching Assistants



Kate Boxer



Daniela Pinto Veizaga



Marina Castano Ishizaka



Elaf Almahmoud



Mihir Upadhyay



Kendall Brogle



Radhika Dua

Assignments and grading

Grading: homeworks - $10\% \times 2 = 20\%$. Drop lowest of 3 HWs.

project - 20%

final - 22%

midterm - 22%

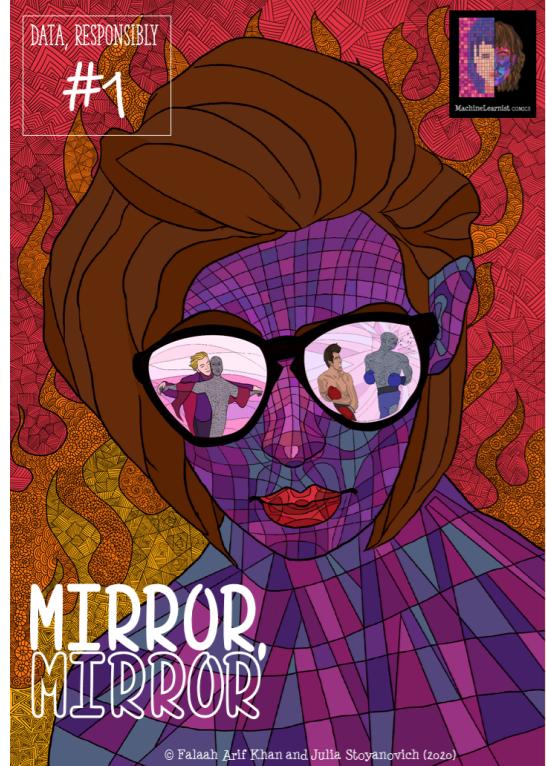
labs - 8%

guest lecture summaries - 8%

No credit for late homework.

Assignment schedule posted to Bright Space (under Course information), subject to change.

This week's reading

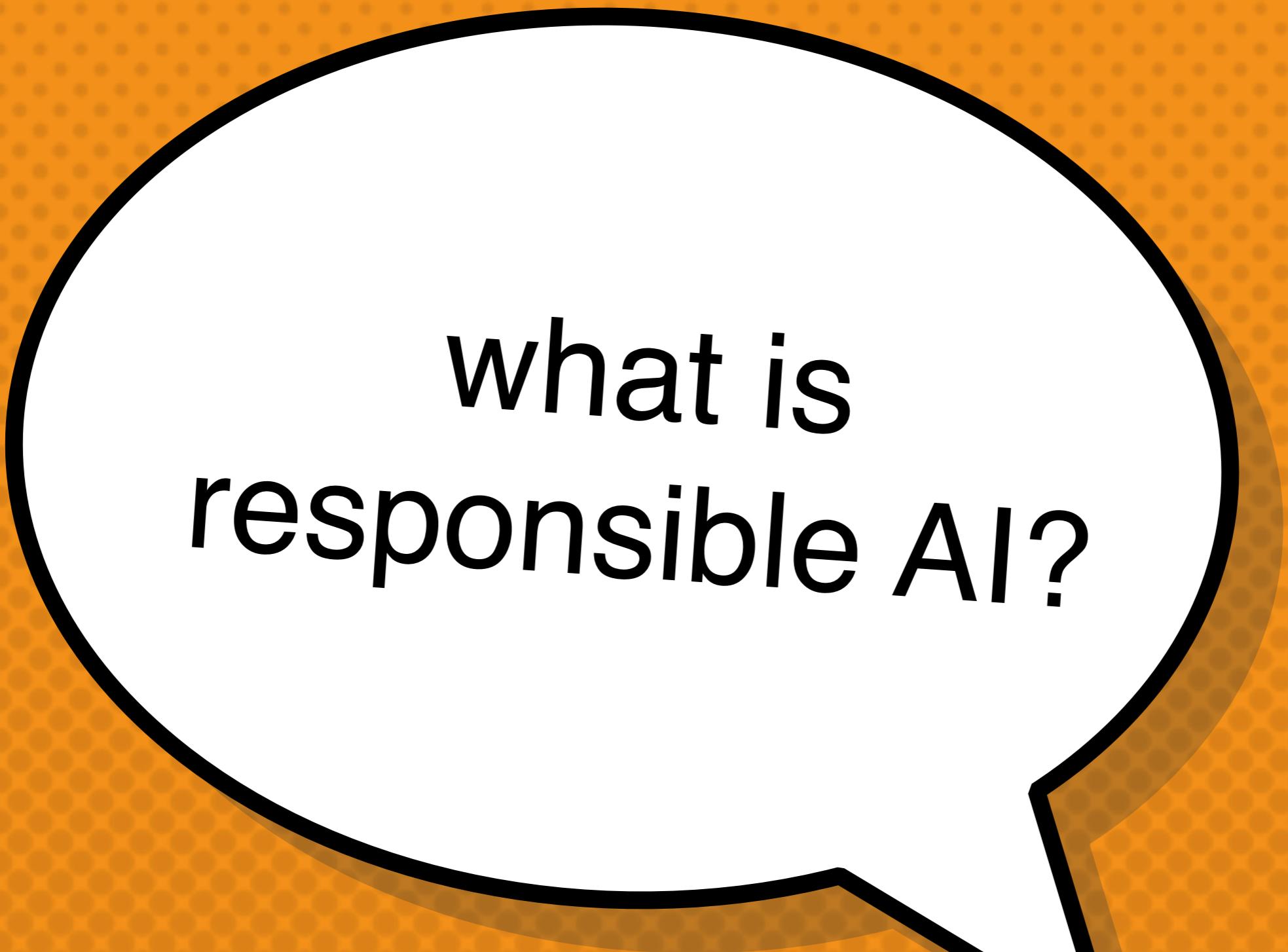


DOI:10.1145/3376898
A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOLDECHOVA AND AARON ROTH

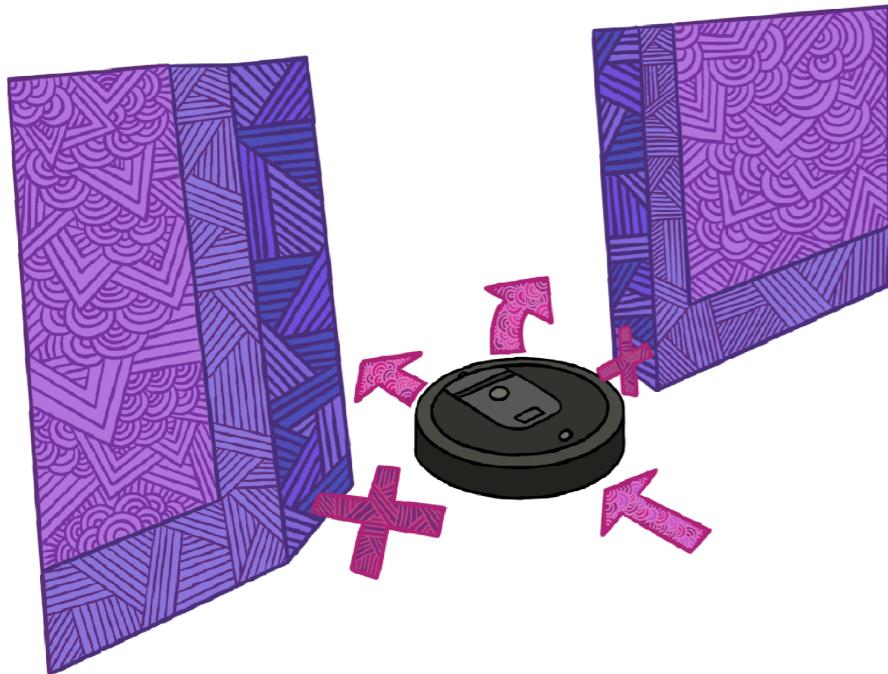
A Snapshot of the Frontiers of Fairness in Machine Learning





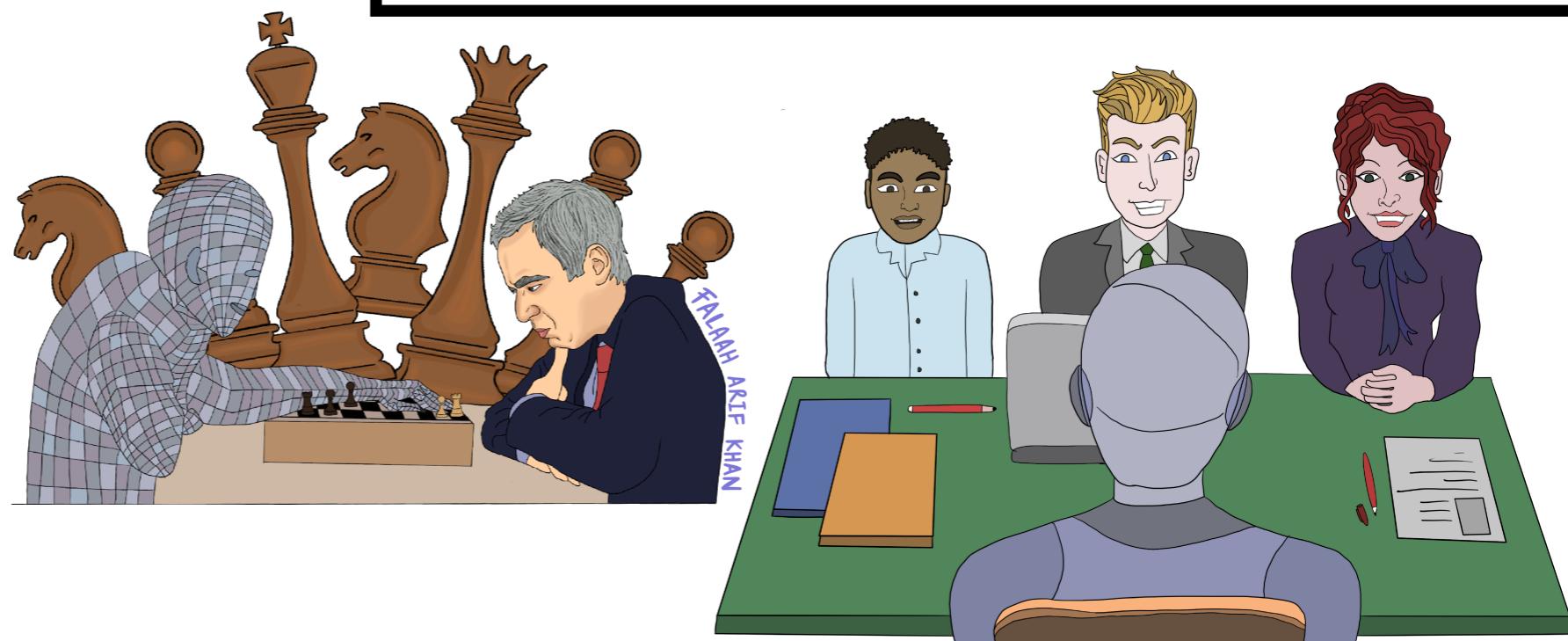
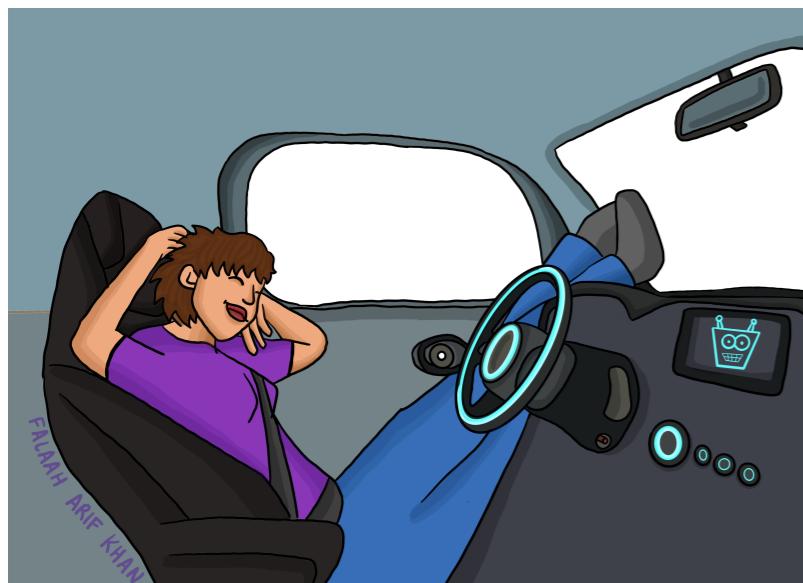
what is
responsible AI?

AI: algorithms, data, decisions



Artificial Intelligence (AI)

a **system** in which **algorithms** use **data** and make **decisions** on our behalf, or help us make decisions



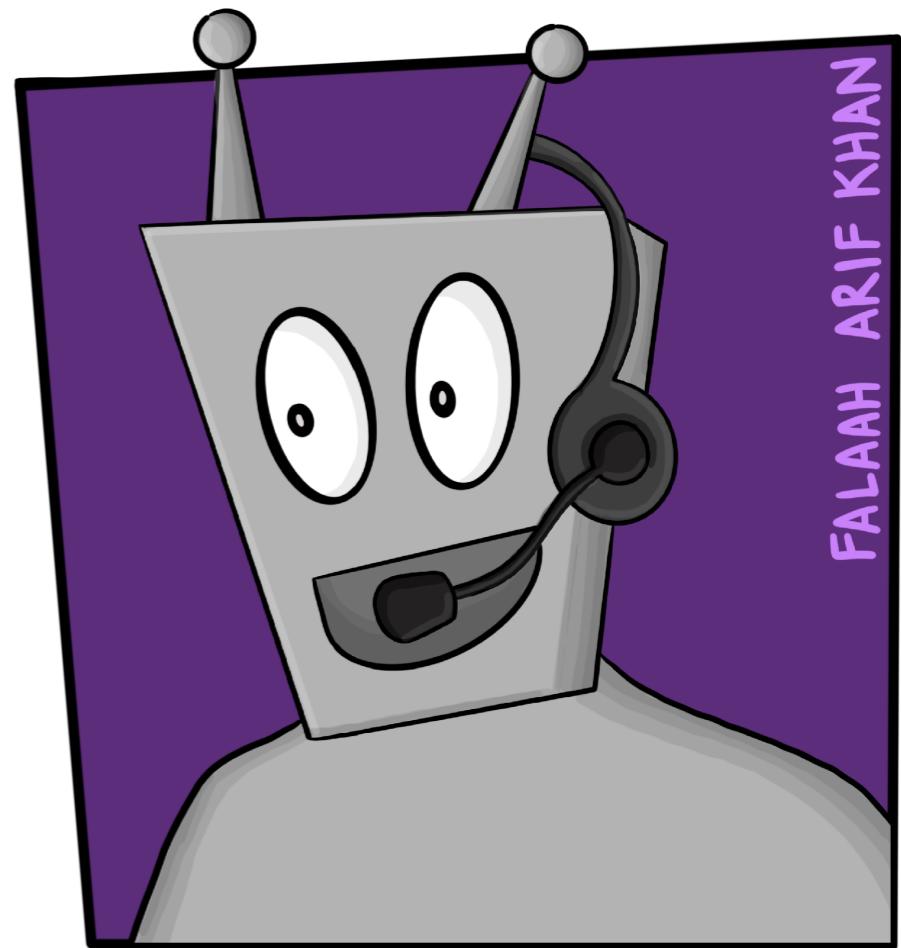
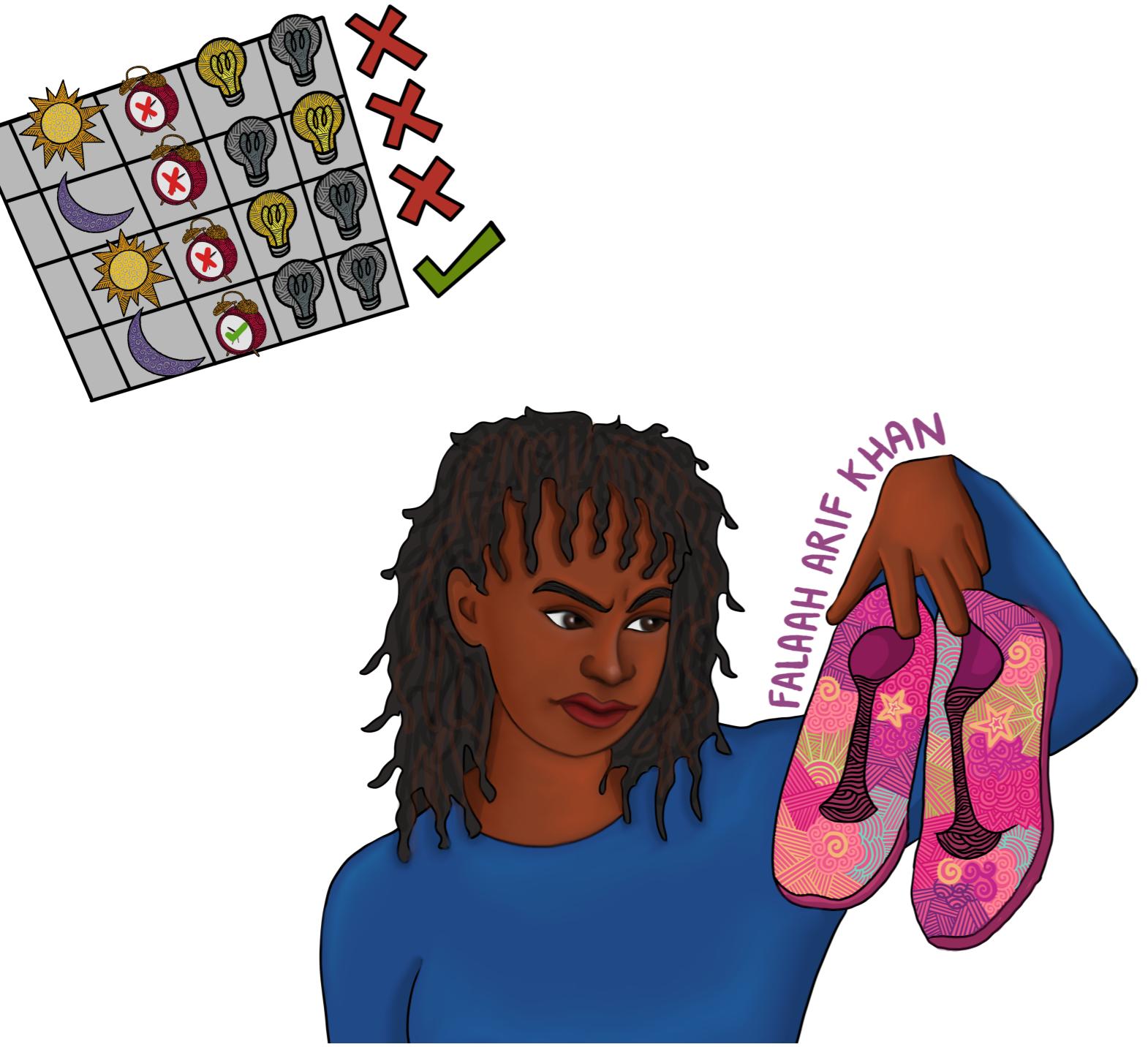
The promise of AI

Opportunity

- make our lives convenient
- accelerate science
- boost innovation
- transform government



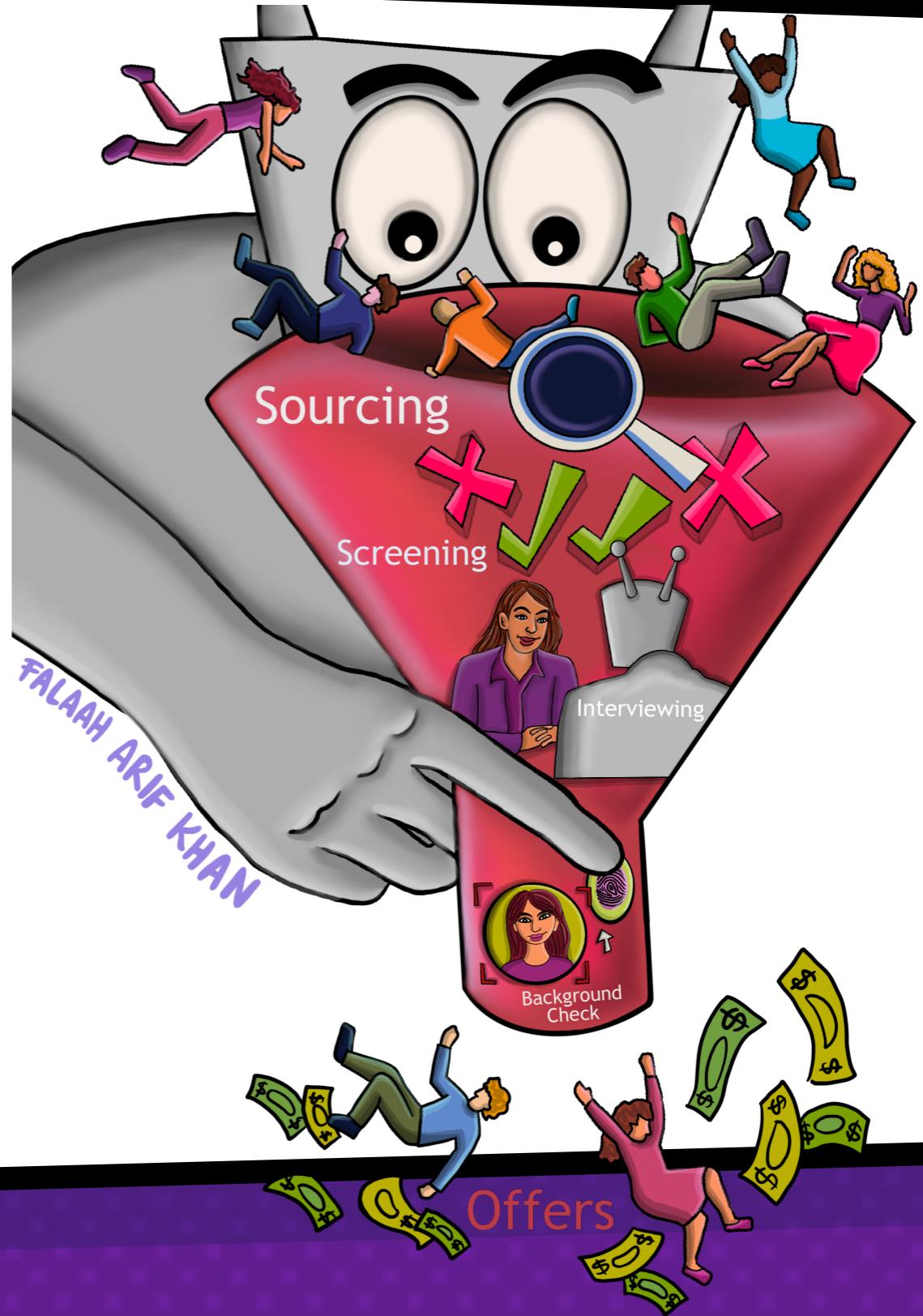
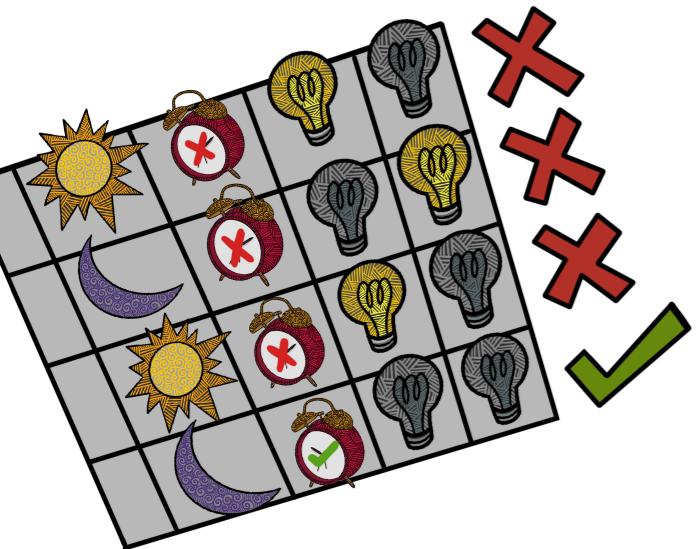
Machines make mistakes

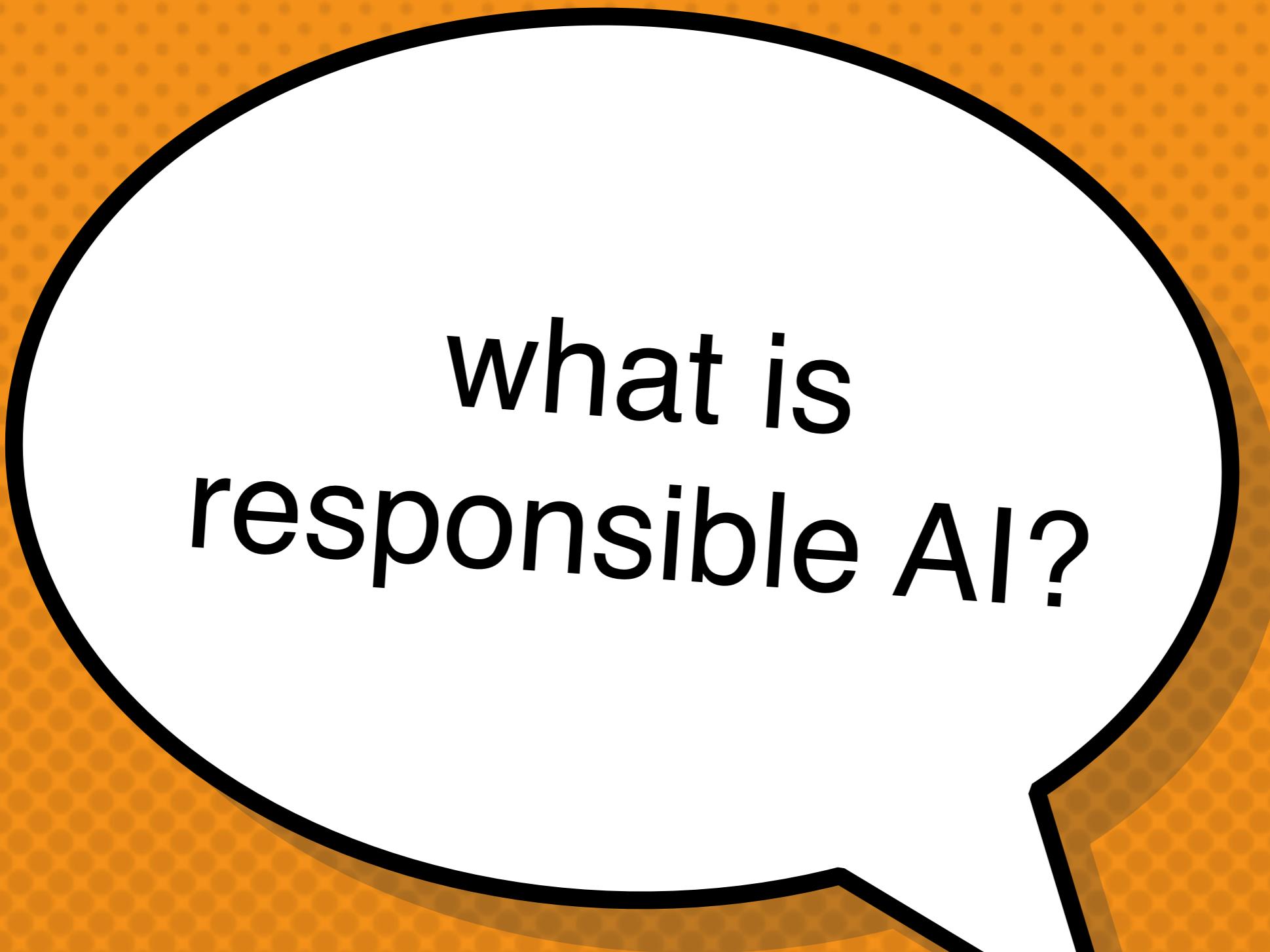


Mistakes lead to harms

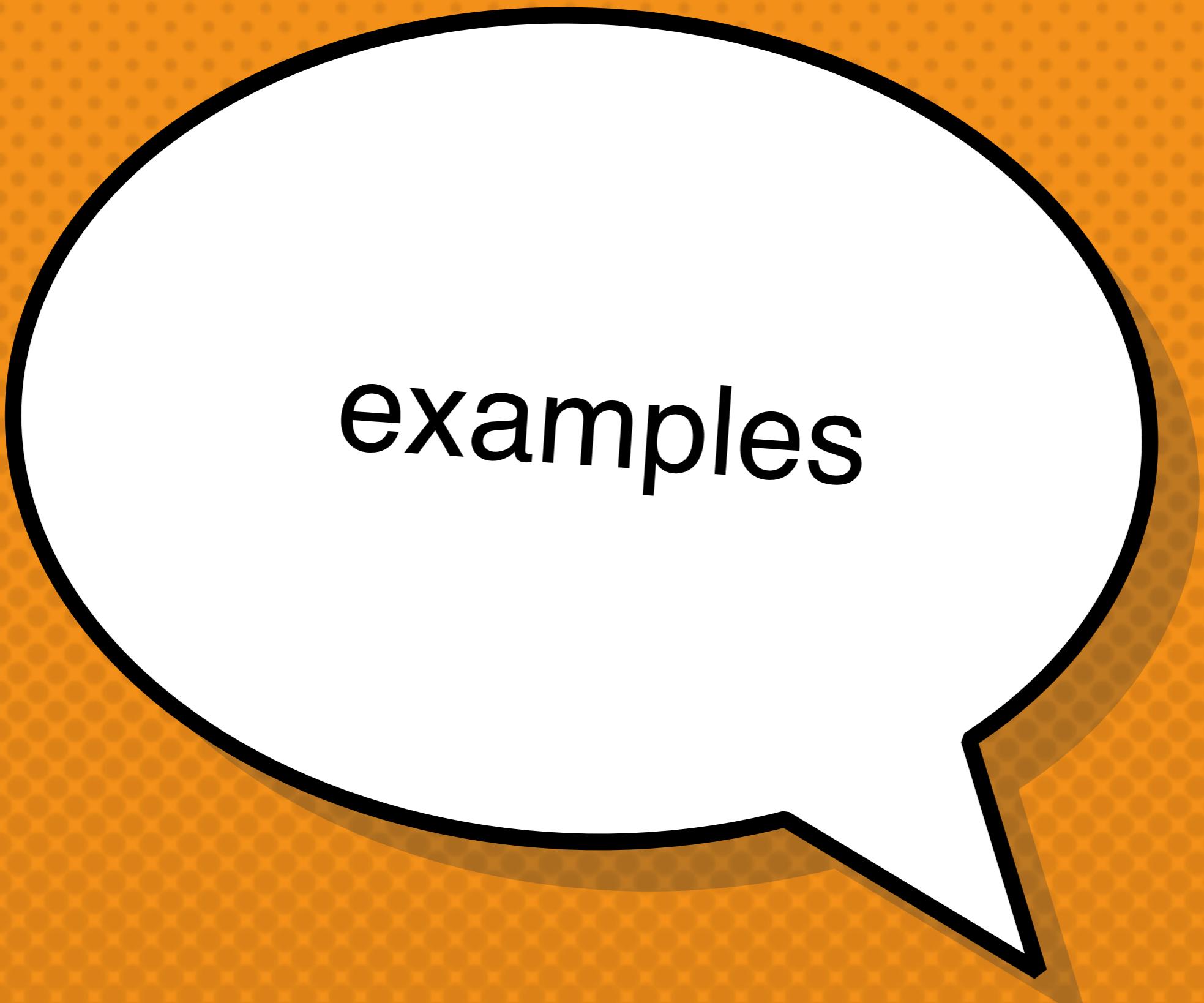


Harms can be cumulative





what is
responsible AI?



examples

Medical imaging

FACEBOOK AI



fastMRI

Accelerating MR Imaging

What is fastMRI?

fastMRI is a collaborative research project between Facebook AI Research and NYU Langone Health. The aim is to investigate the use of AI to make MRI scans up to 10 times faster.

By producing accurate images from undersampled data, AI image reconstruction has the potential to improve the patient's experience and to make MRIs accessible for more people.

Positive factors

clear need for improvement

can validate predictions

technical readiness

decision-maker readiness

raw data and image dataset repository, which contains baseline reconstruction models and PyTorch data loaders for the fastMRI dataset.

<https://fastmri.org/>

Automated hiring systems

MIT

Technology **February 2013**

Racism is Poisoning
Online Ad Delivery, Says
Harvard Professor



REUTERS

Amazon scraps secret AI recruiting
tool that showed bias against women

The New York Times

March 2021

We Need Laws to Take On Racism
and Sexism in Hiring Technology

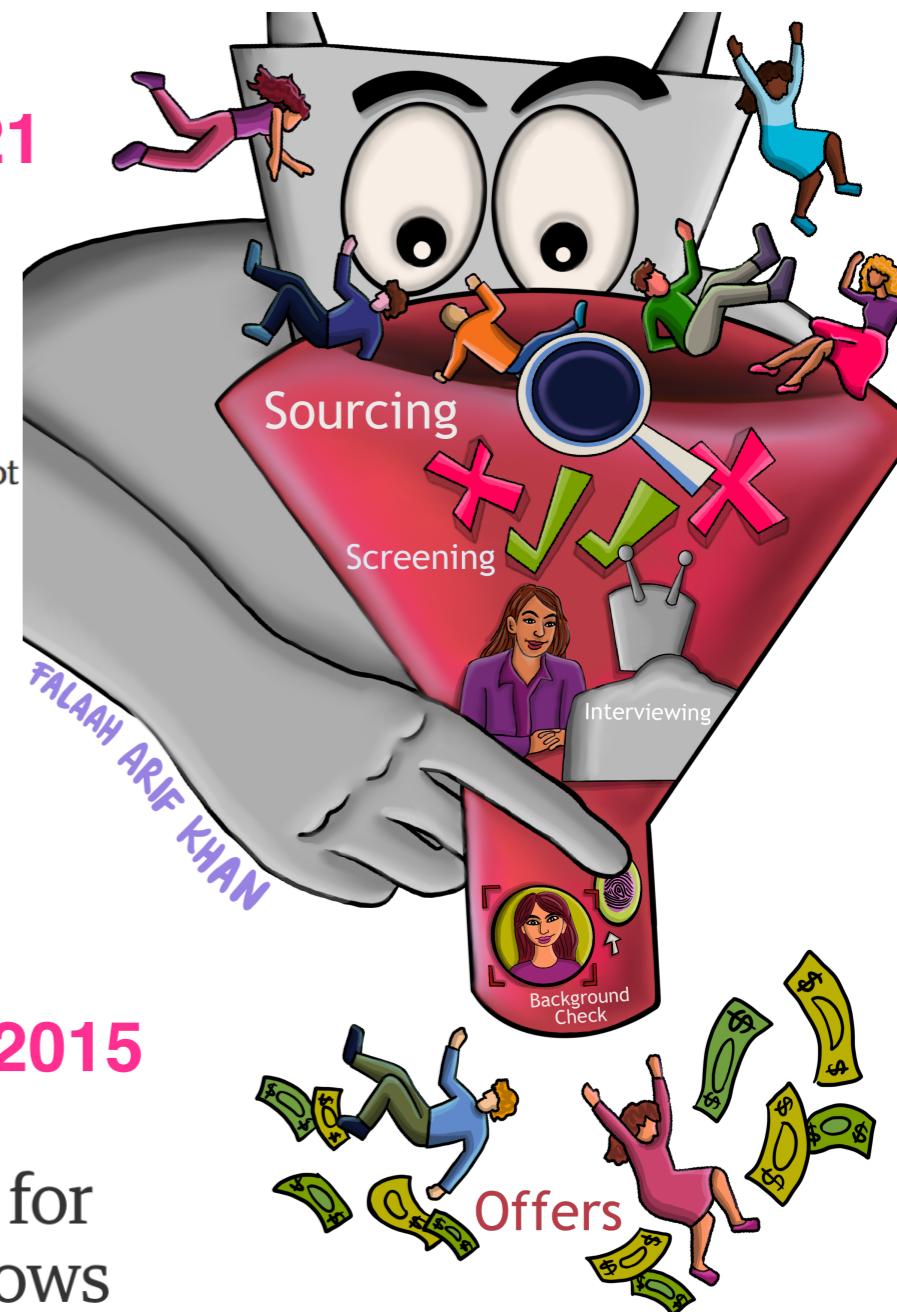
Artificial intelligence used to evaluate job candidates must not become a tool that exacerbates discrimination.

October 2018

the guardian

July 2015

Women less likely to be shown ads for
high-paid jobs on Google, study shows



Hiring before automation

Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination

September 2004

Marianne Bertrand

Sendhil Mullainathan

AMERICAN ECONOMIC REVIEW
VOL. 94, NO. 4, SEPTEMBER 2004
(pp. 991-1013)

We study race in the labor market by sending fictitious resumes to help-wanted ads in Boston and Chicago newspapers. To manipulate perceived race, resumes are randomly assigned African-American- or White-sounding names. **White names receive 50 percent more callbacks for interviews.** Callbacks are also more responsive to resume quality for White names than for African-American ones. The racial gap is uniform across occupation, industry, and employer size. We also find little evidence that employers are inferring social class from the names. Differential treatment by race still appears to still be prominent in the U. S. labor market.

discussion

Describe a use case

what are the **goals** of the AI system?

what are the **benefits** and to **whom**?

what are the **harms** and to **whom**?

Use case: Staples discounts

THE WALL STREET JOURNAL.

WHAT THEY KNOW

Websites Vary Prices, Deals Based on Users' Information

By Jennifer Valentino-DeVries, Jeremy Singer-Vine and Ashkan Soltani

December 24, 2012

WHAT PRICE WOULD YOU SEE?



<https://www.wsj.com/articles/SB10001424127887323777204578189391813881534>

December 2012

It was the same Swingline stapler, on the same Staples.com website. But for Kim Wamble, the price was \$15.79, while the price on Trude Frizzell's screen, just a few miles away, was \$14.29.

A key difference: where Staples seemed to think they were located.

A Wall Street Journal investigation found that the Staples Inc. website displays different prices to people after estimating their locations. More than that, **Staples appeared to consider the person's distance from a rival brick-and-mortar store**, either OfficeMax Inc. or Office Depot Inc. If rival stores were within 20 miles or so, Staples.com usually showed a discounted price.

Use case: AdFisher

theguardian

July 2015

Samuel Gibbs

Wednesday 8 July 2015 11.29 BST

Automated testing and analysis of company's advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs



One experiment showed that Google displayed adverts for a career coaching service for executive jobs 1,852 times to the male group and only 318 times to the female group. Photograph: Alamy

Women less likely to be shown ads for high-paid jobs on Google, study shows

The AdFisher tool simulated job seekers that did not differ in browsing behavior, preferences or demographic characteristics, except in gender.

One experiment showed that Google displayed ads for a career coaching service for “\$200k+” executive jobs **1,852 times to the male group and only 318 times to the female group.**

Another experiment, in July 2014, showed a similar trend but was not statistically significant.

<https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study>

Use case: Resume screening



Jeffrey Dastin

BUSINESS NEWS OCTOBER 9, 2018 / 11:12 PM / 6 MONTHS AGO

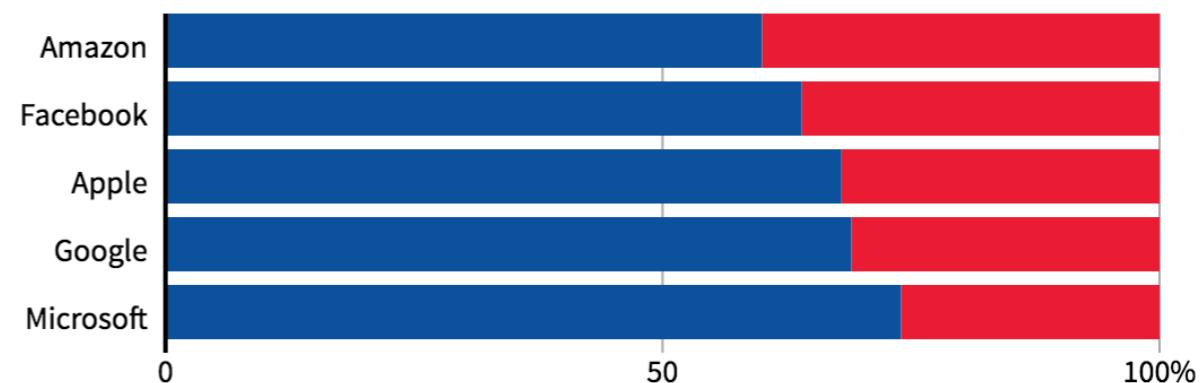
October 2018

Amazon scraps secret AI recruiting tool that showed bias against women

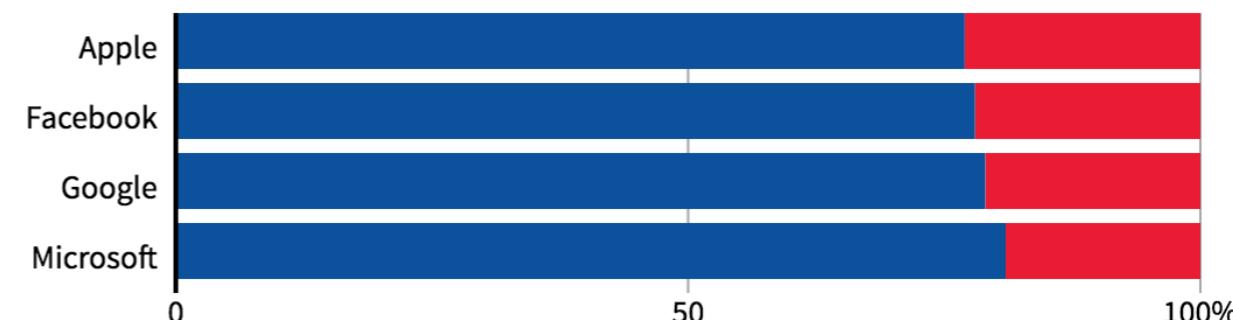
"In effect, **Amazon's system taught itself that male candidates were preferable**. It penalized resumes that included the word "women's," as in "women's chess club captain." And it **downgraded graduates of two all-women's colleges**, according to people familiar with the matter. They did not specify the names of the schools."

GLOBAL HEADCOUNT

■ Male ■ Female



EMPLOYEES IN TECHNICAL ROLES



"Note: Amazon does not disclose the gender breakdown of its technical workforce."

<https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scaps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>

Use case: Instant Checkmate



Ads by Google

[Latanya Sweeney, Arrested?](#)

1) Enter Name and State. 2) Access F
Checks Instantly.

www.instantcheckmate.com/

[Latanya Sweeney](#)

Public Records Found For: Latanya S
www.publicrecords.com/

Tanya

arre

FALAH KHAN

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

racially identifying names trigger ads suggestive of a criminal record

<https://www.technologyreview.com/s/510646/racism-is-poisoning-online-ad-delivery-says-harvard-professor/>

February 2013

The screenshot shows the InstantCheckmate website interface. At the top, there's a navigation bar with links for DASHBOARD, EDIT ACCOUNT INFO, and LOGOUT. A large blue banner on the right side features a 'CERTIFIED' seal. The main content area displays a profile for LATANYA SWEENEY, including her address (1420 Centre Ave, Pittsburgh, PA 15219), DOB (Oct 27, 1959), and age (53 years old). Below this, there are three sections: Personal (Name, aliases, birthdate, phone numbers, etc.), Location (Detailed address history and related data, maps, etc.), and Related Persons. To the right, under the heading 'Criminal History', it states: 'This section contains possible citation, arrest, and criminal records for the subject of this report. While our database does contain hundreds of millions of arrest records, different counties have different rules regarding what information they will and will not release.' It also notes: 'We share with you as much information as we possibly can, but a clean slate here should not be interpreted as a guarantee that Latanya Sweeney has never been arrested; it simply means that we were not able to locate any matching arrest records in the data that is available to us.' At the bottom right, there's a 'View Details' button.

Use case: Amazon same-day delivery

Bloomberg

Amazon Doesn't Consider the Race of Its Customers. Should It?

“... In six major same-day delivery cities, however, **the service area excludes predominantly black ZIP codes** to varying degrees, according to a Bloomberg analysis that compared Amazon same-day delivery areas with U.S. Census Bureau data.”

<https://www.bloomberg.com/graphics/2016-amazon-same-day/>

New York City

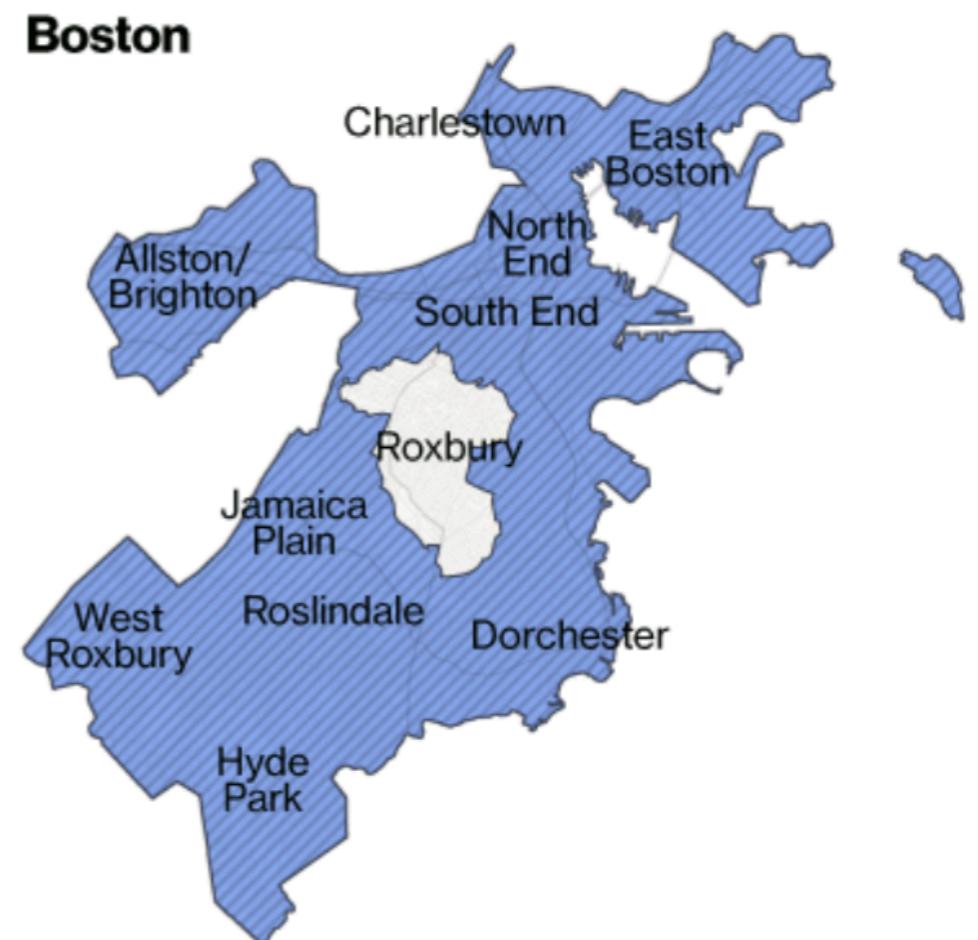


Use case: Amazon same-day delivery

Bloomberg

Amazon Doesn't Consider the Race of Its Customers. Should It?

"The most striking gap in Amazon's same-day service is in Boston, where **three ZIP codes encompassing the primarily black neighborhood of Roxbury are excluded** from same-day service, while the neighborhoods that surround it on all sides are eligible."



<https://www.bloomberg.com/graphics/2016-amazon-same-day/>

**examples: racial
bias in risk
assessment**

Racial bias in criminal sentencing

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016



May 2016

A commercial tool COMPAS automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

The tool correctly predicts recidivism **61% of the time.**

Blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend.

The tool makes **the opposite mistake among whites:** They are much more likely than blacks to be labeled lower risk but go on to commit other crimes.

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Racial bias in criminal sentencing

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

May 2016

A commercial tool COMPAS automatically predicts some categories of future crime to assist in bail and sentencing decisions. It is used in courts in the US.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

<https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

Racial bias in healthcare

Dissecting racial bias in an algorithm used to manage the health of populations

October 2019

Ziad Obermeyer^{1,2,*}, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan^{5,*†}

* See all authors and affiliations

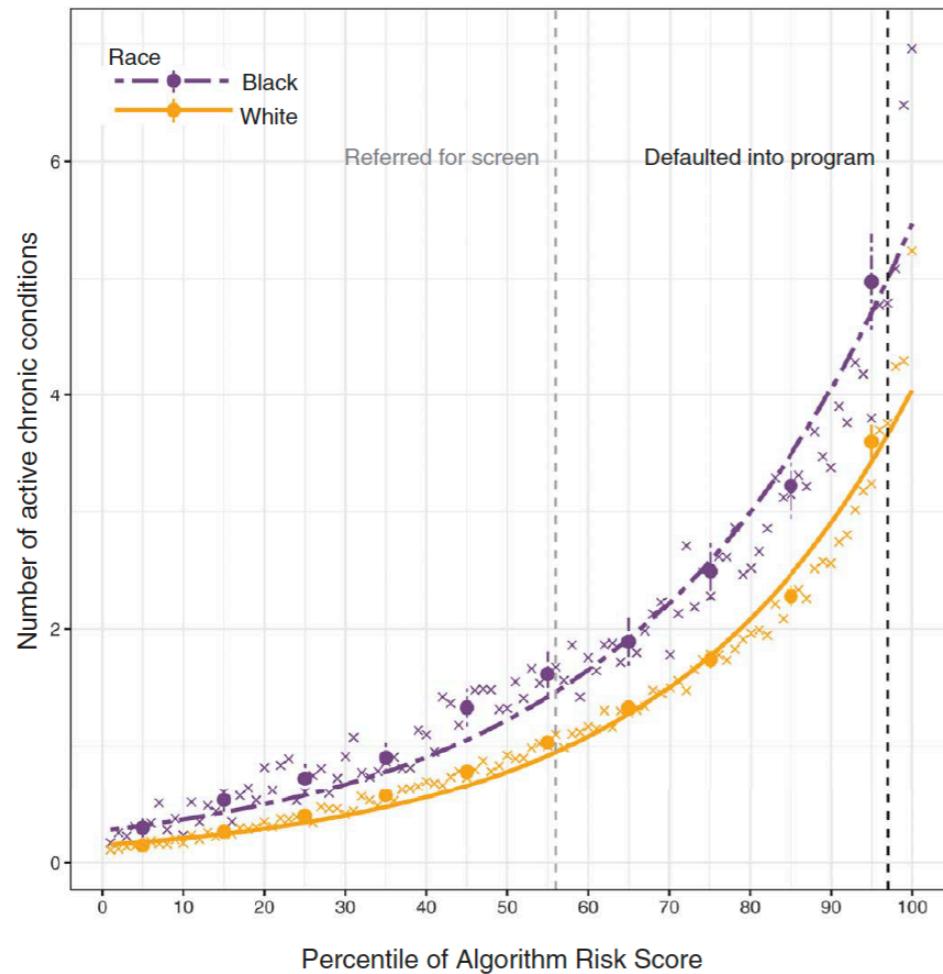
Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Science

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses**. Remedyng this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm **predicts health care costs rather than illness**, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, **despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise**. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

Racial bias in healthcare

A



B

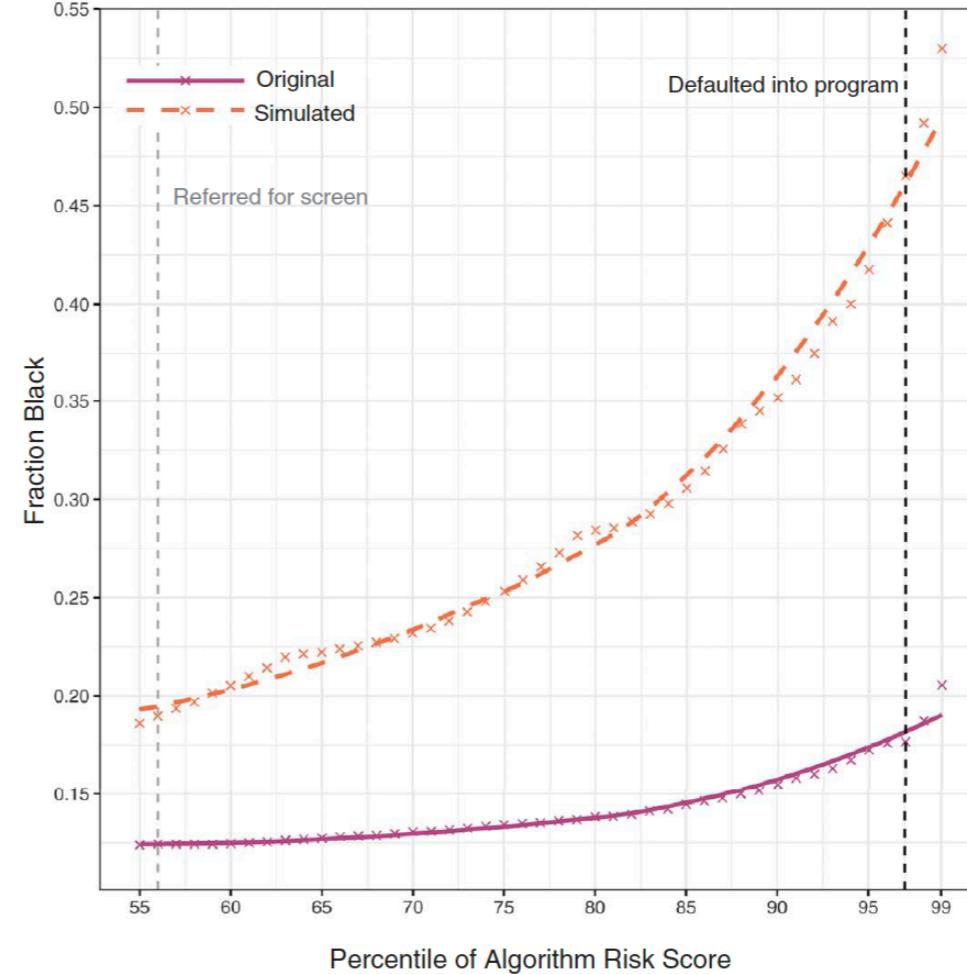


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (A) Mean number of chronic conditions by race, plotted against algorithm risk score. (B) Fraction of Black patients at or above a given risk score for the original algorithm (“original”) and for a simulated scenario that removes algorithmic bias (“simulated”: at each threshold of risk, defined at a given percentile on the x axis, healthier Whites above the threshold are

replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The \times symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

Dec. 6, 2019

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.



Tim Cook

<https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>

December 2019

In one study published 15 years ago, **two people applied for a job**. Their résumés were about as similar as two résumés can be. One person was named Jamal, the other Brendan.

In a study published this year, **two patients sought medical care**. Both were grappling with diabetes and high blood pressure. One patient was black, the other was white.

Both studies documented **racial injustice**: In the first, the applicant with a black-sounding name got fewer job interviews. In the second, the black patient received worse care.

But they differed in one crucial respect. In the first, hiring managers made biased decisions. In the second, the culprit was a computer program.

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

Dec. 6, 2019

ECONOMIC VIEW

Biased Algorithms Are Easier to Fix Than Biased People

Racial discrimination by algorithms or by people is harmful — but that's where the similarities end.



Tim Cook

<https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html>

December 2019

Changing algorithms is easier than changing people: software on computers can be updated; the “wetware” in our brains has so far proven much less pliable.

[...] In a 2018 [paper](#) [...], I took a cautiously optimistic perspective and argued that **with proper regulation, algorithms can help to reduce discrimination.**

But the key phrase here is “proper regulation,” which we do not currently have.

We must ensure all the necessary inputs to the algorithm, including the data used to test and create it, are carefully stored. * [...] **We will need a well-funded regulatory agency with highly trained auditors to process this data.**



a push for
regulation

Automated Decision Systems (ADS)

Automated Decision Systems (ADS)

process data about people

help make consequential decisions

combine human & automated decision making

aim to improve **efficiency** and promote **equity**

are subject to **auditing** and **public disclosure**



Regulating ADS?

Precautionary

Nah! I'm fine!



The Anti-Elon ✅
@antiElon

Regulation rocks!

2.3K 9.2K 126K

Risk-based



New York City Local Law 49 of 2018

January 11, 2018

An **Automated Decision System (ADS)** is a “computerized implementation of algorithms, including those derived from machine learning or other data processing or artificial intelligence techniques, which are used to make or assist in making decisions.”

Form task force that surveys the current use of ADS in City agencies and develops procedures for:

- requesting and receiving an **explanation** of an algorithmic decision affecting an individual (3(b))
- interrogating ADS for **bias and discrimination** against members of legally-protected groups (3(c) and 3(d))
- allowing the **public** to **assess** how ADS function and are used (3(e)), and archiving ADS together with the data they use (3(f))

ADS regulation in NYC: take 1



Principles

- using ADS **where** they promote innovation and efficiency in service delivery
- promoting **fairness, equity, accountability, and transparency** in the use of ADS
- reducing potential harm **across the lifespan** of ADS

New York City Local Law 144 of 2021



THE NEW YORK CITY COUNCIL
Corey Johnson, Speaker

December 11, 2021

This bill would require that a **bias audit** be conducted on an automated employment decision tool prior to the use of said tool. The bill would also require that candidates or employees that reside in the city **be notified about the use of such tools** in the assessment or evaluation for hire or promotion, as well as, **be notified about the job qualifications and characteristics that will be used** by the automated employment decision tool. Violations of the provisions of the bill would be subject to a civil penalty.

A related domain: AI in hiring

“Automated hiring systems act as modern gatekeepers to economic opportunity.”

Jenny Yang



Algorithmic discrimination

the guardian

July 2015

Women less likely to be shown ads for high-paid jobs on Google, study shows

MIT

Technology Review

February 2013

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

THE WALL STREET JOURNAL. September 2014

Are Workplace Personality Tests Fair?

Growing Use of Tests Sparks Scrutiny Amid Questions of Effectiveness and Workplace Discrimination



REUTERS

October 2018

Amazon scraps secret AI recruiting tool that showed bias against women

The need for regulation

Opinion

We Need Laws to Take On Racism and Sexism in Hiring Technology

The New York Times
March 17, 2021

By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

The bill should also require validity testing, to ensure that the tools actually measure what they claim to, and it must make certain that they measure characteristics that are relevant for the job. **Such testing would interrogate whether, for example, candidates' efforts to blow up a balloon in an online game really indicate their appetite for risk in the real world — and whether risk-taking is necessary for the job.**

The need for regulation

Opinion

We Need Laws to Take On Racism and Sexism in Hiring Technology

The New York Times

March 17, 2021

By Alexandra Reeve Givens, Hilke Schellmann and Julia Stoyanovich

Ms. Givens is the chief executive of the Center for Democracy & Technology. Ms. Schellman and Dr. Stoyanovich are professors at New York University focusing on artificial intelligence.

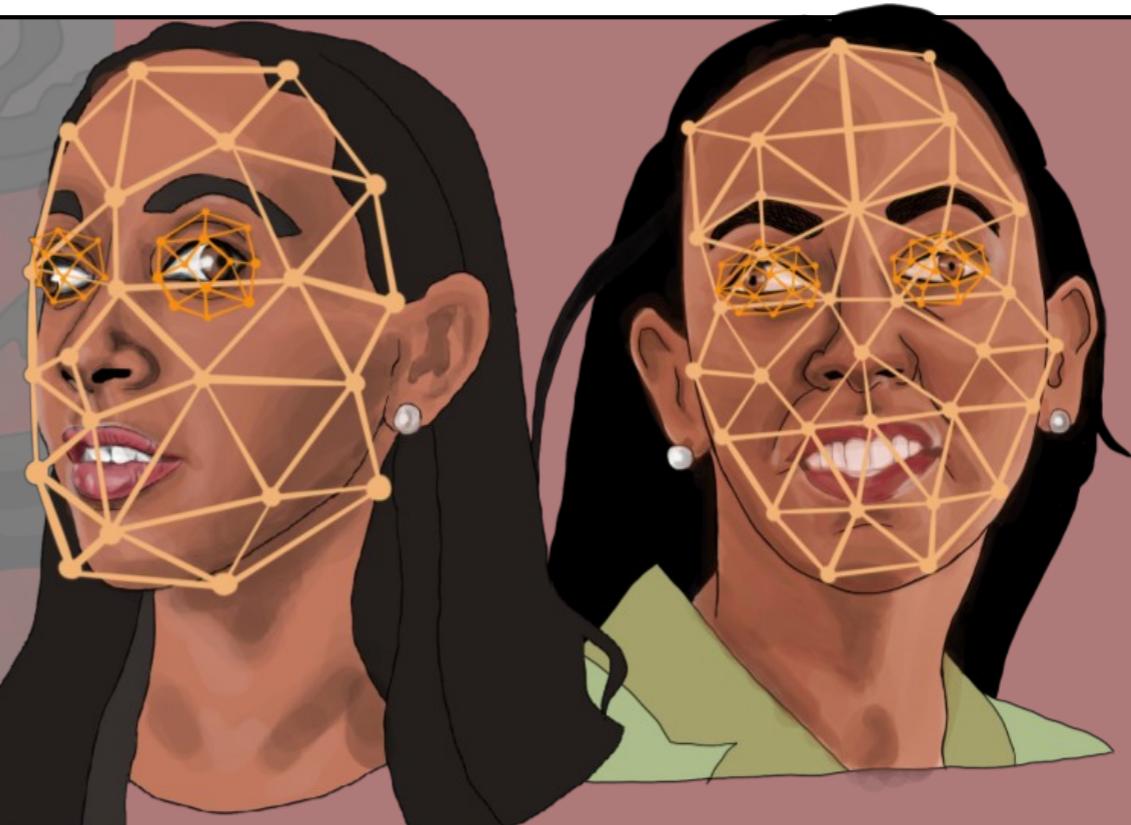
In addition, the City Council must require vendors to tell candidates how they will be screened by an automated tool **before** the screening, so candidates know what to expect. **People who are blind, for example, may not suspect that their video interview could score poorly if they fail to make eye contact with the camera.** If they know what is being tested, they can engage with the employer to seek a fairer test.

THESE GHOSTS ARE MAKING THEIR WAY INTO DATA-DRIVEN PRODUCTS AS WELL.

TAKE THE INFAMOUS FACIAL RECOGNITION SOFTWARE THAT HAS BEEN ALL OVER THE NEWS RECENTLY. RACIAL INJUSTICES ARE PROBLEMATIC ENOUGH, BUT HAVE YOU CONSIDERED HOW THESE MODELS DISCRIMINATE AGAINST BLACK DISABLED PEOPLE?

AS DISABILITY RIGHTS ADVOCATE **HABEN GIRMA** EXPLAINS (7),

"MY EYES MOVE INVOLUNTARILY, EACH ONE SWINGING TO ITS OWN MUSIC. THEY'VE DANCED THIS WAY FOR AS LONG AS I CAN REMEMBER."



HOW WELL DO YOU THINK FACIAL RECOGNITION WOULD PERFORM ON **BLIND BLACK PEOPLE**?

HAVING BEEN TRAINED ON THE FACIAL DYNAMICS OF SIGHTED WHITE PEOPLE, FACIAL RECOGNITION TECHNOLOGY PEDDLES AN ABLEIST AND RACIST NARRATIVE.

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected



Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

Artificial-intelligence tools are seeing ever broader use in hiring. But this practice is also hotly criticized because we rarely understand how these tools select candidates, and whether the candidates they select are, in fact, better qualified than those who are rejected.

To help answer these crucial questions, **we should give job seekers more information about the hiring process and the decisions**. The solution I propose is a twist on something we see every day: **nutritional labels**. Specifically, job candidates would see simple, standardized labels that show the factors that go into the AI's decision.

Nutritional labels for job seekers

THE WALL STREET JOURNAL.

September 22, 2021

Hiring and AI: Let Job Candidates Know Why They Were Rejected



Labels that explain a hiring process that uses AI could allow job seekers to opt out if they object to the employer's data practices.

PHOTO: ISTOCKPHOTO/GETTY IMAGES

By Julia Stoyanovich

Updated Sept. 22, 2021 11:00 am ET

ACCOUNTANT

Acme Partners

Qualifications: BS in accounting, GPA >3.0, Knowledge of financial and accounting systems and applications

Personal data to be analyzed: An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.

Additional assessment: AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to selectively decline use of AI analysis for any of their personal data or to review and challenge the results of such analysis.

Nutritional labels for public disclosure

ACCOUNTANT

Acme Partners

Qualifications: BS in accounting, GPA >3.0, Knowledge of financial accounting systems and applications

Personal data to be analyzed: An AI program could be used to review and analyze the applicant's personal data online, including LinkedIn profile, social media accounts and credit score.

Additional assessment: AI-assisted personality scoring

ALERT: Applicants for this position DO NOT have the option to selectively decline use of AI analysis for any of their personal data or to review and challenge the results of such analysis.

comprehensible: short, simple, clear

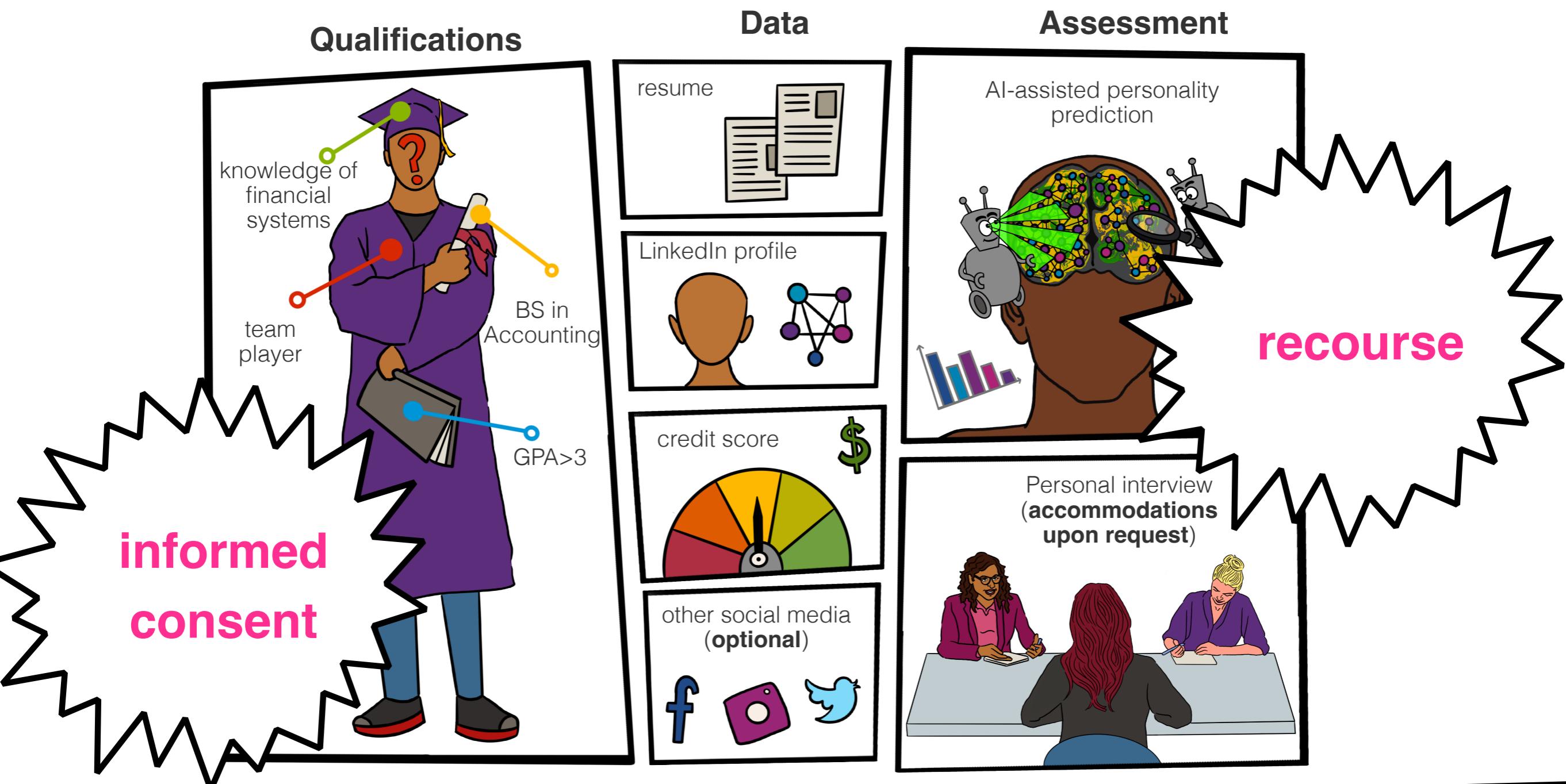
consultative: provide actionable info

comparable: implying a standard

<https://www.wsj.com/articles/hiring-job-candidates-ai-11632244313>

[Stoyanovich & Howe, 2019]

Anatomy of a job posting label





course overview

module 1:

algorithmic

fairness

Bias in computer systems

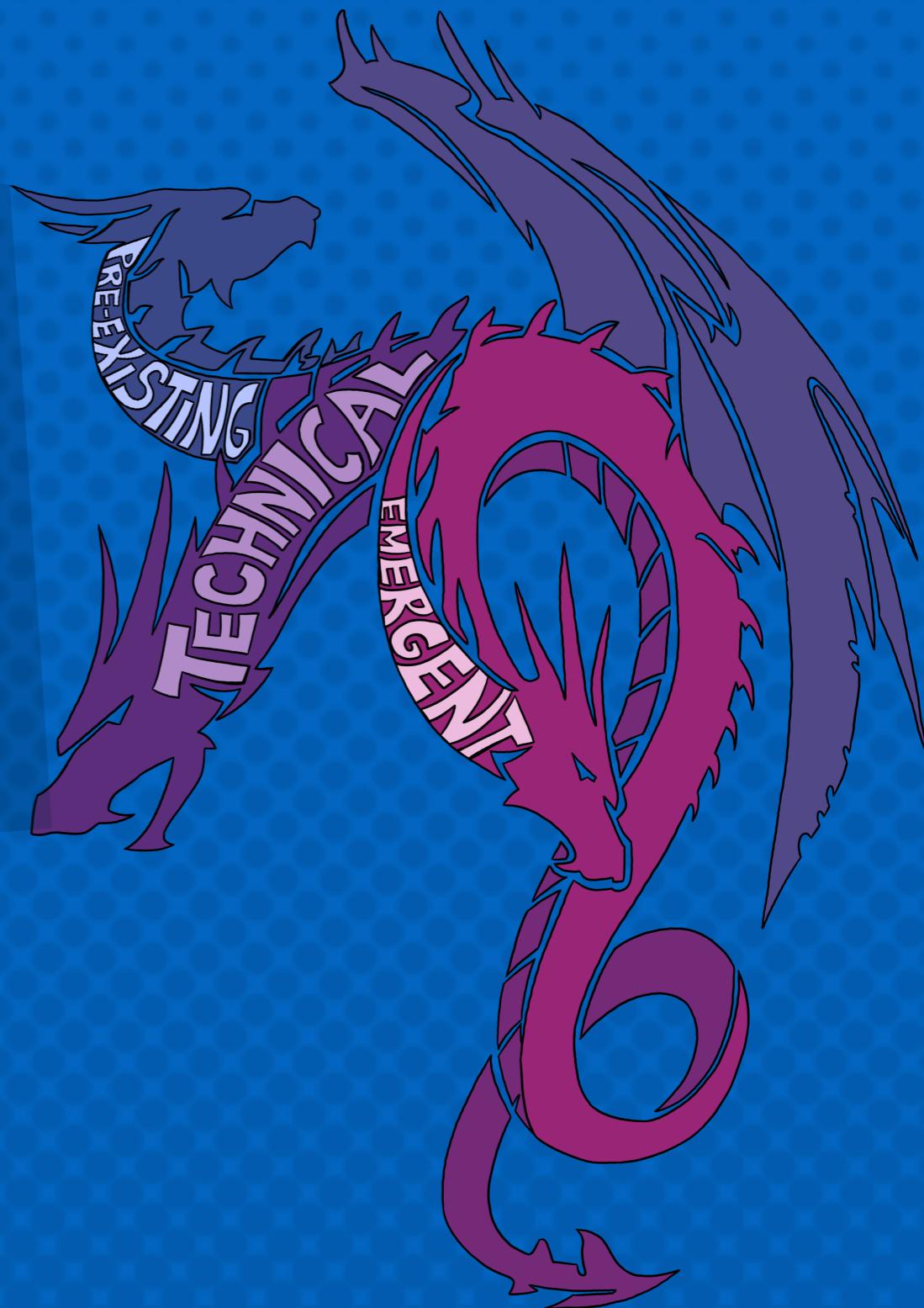
Pre-existing: exists independently of algorithm, has origins in society

Technical: introduced or exacerbated by the technical properties of an ADS

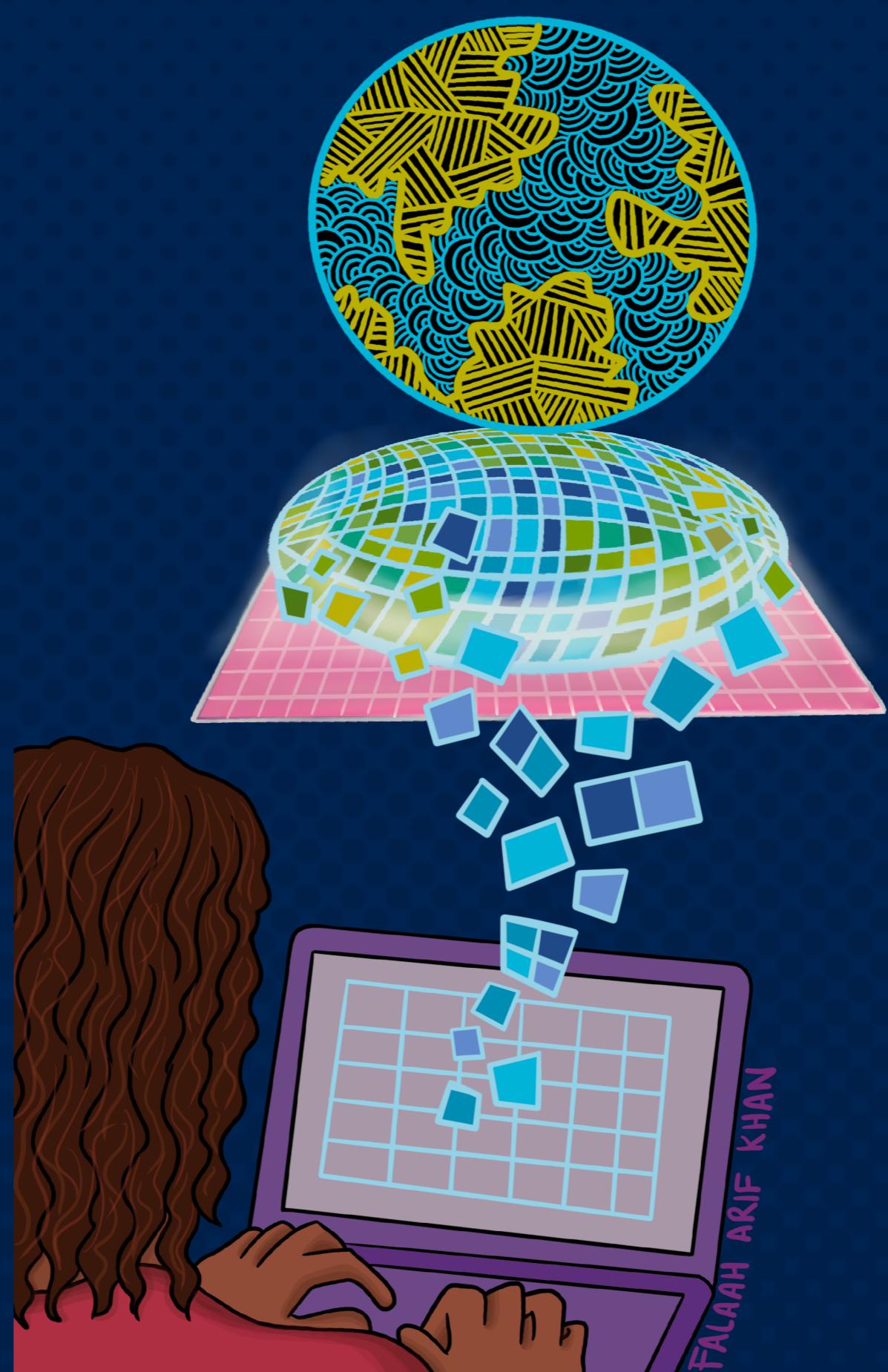
Emergent: arises due to context of use



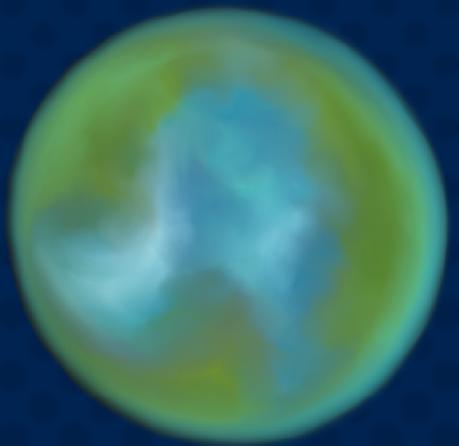
[Friedman & Nissenbaum (1996)]







FALAAH ARIF KHAN



FALAAH ARIF KHAN





FALAAH ARIF KHAN



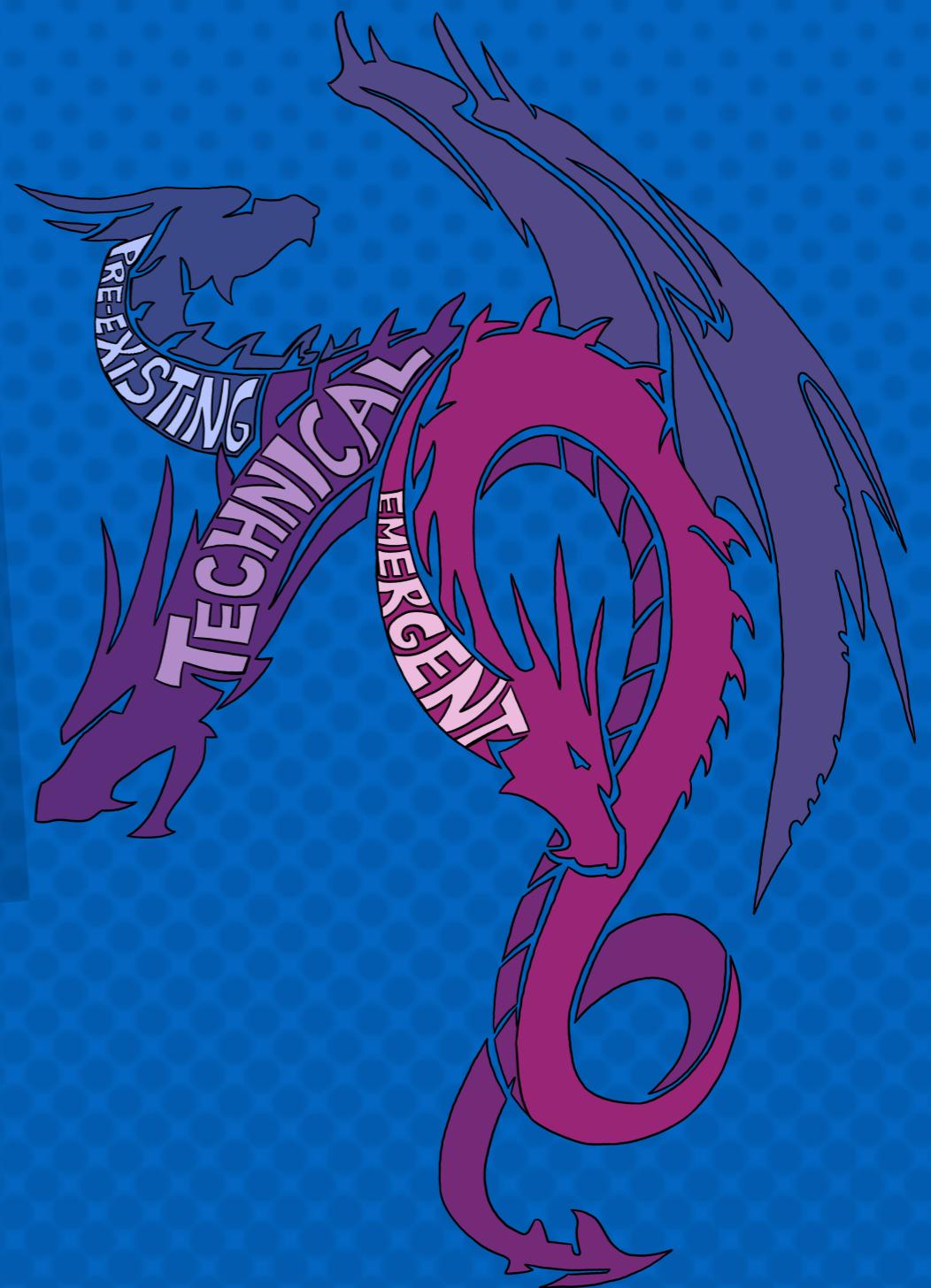
Bias in computer systems

Pre-existing: exists independently of algorithm, has origins in society

Technical: introduced or exacerbated by the technical properties of an ADS

Emergent: arises due to context of use

to fight bias, state
beliefs and
assumptions
explicitly



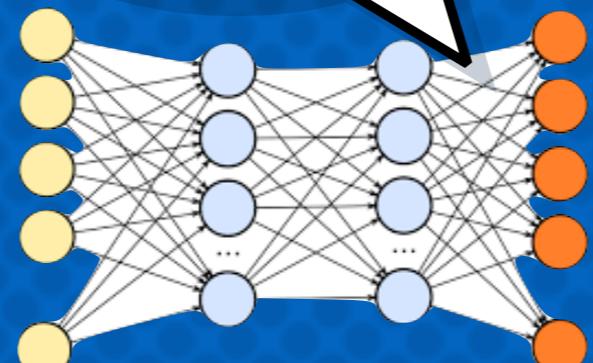
Fair-ML view

where did the data come from?

1	A	B	C	D	E	F	G	H
2	UID	sex	race	MarriageSta	DateOfBirth	age	score	1
3	1	0	1	1	4/18/47	69	0	1
3	2	0	2	1	1/22/82	34	0	3
4	3	0	2	1	5/14/91	24	0	4
5	4	0	2	1	1/21/93	23	0	8
6	5	0	1	2	1/22/73	43	0	1
7	6	0	1	3	8/22/71	44	0	1
8	7	0	3	1	7/23/74	41	0	6
9	8	0	1	2	2/25/73	43	0	4
10	9	0	3	1	6/10/94	21	0	3
11	10	0	3	1	6/1/88	27	0	4
12	11	1	3	2	8/22/78	37	0	1
13	12	0	2	1	12/2/74	41	0	4
14	13	1	3	1	6/14/68	47	0	1
15	14	0	2	1	3/25/85	31	0	3
16	15	0	4	4	1/25/79	37	0	1
17	16	0	2	1	6/22/90	25	0	10
18	17	0	3	1	12/24/84	31	0	5
19	18	0	3	1	1/8/85	31	0	3
20	19	0	2	3	6/28/51	64	0	6
21	20	0	2	1	11/29/94	21	0	9
22	21	0	3	1	8/6/88	27	0	2
23	22	1	3	1	3/22/95	21	0	4
24	23	0	4	1	1/23/92	24	0	4
25	24	0	3	3	1/10/73	43	0	1
26	25	0	1	1	8/24/83	32	0	3
27	26	0	2	1	2/8/89	27	0	3
28	27	1	3	1	9/3/79	36	0	3
29	28	0	2	1	4/27/80	36	0	7



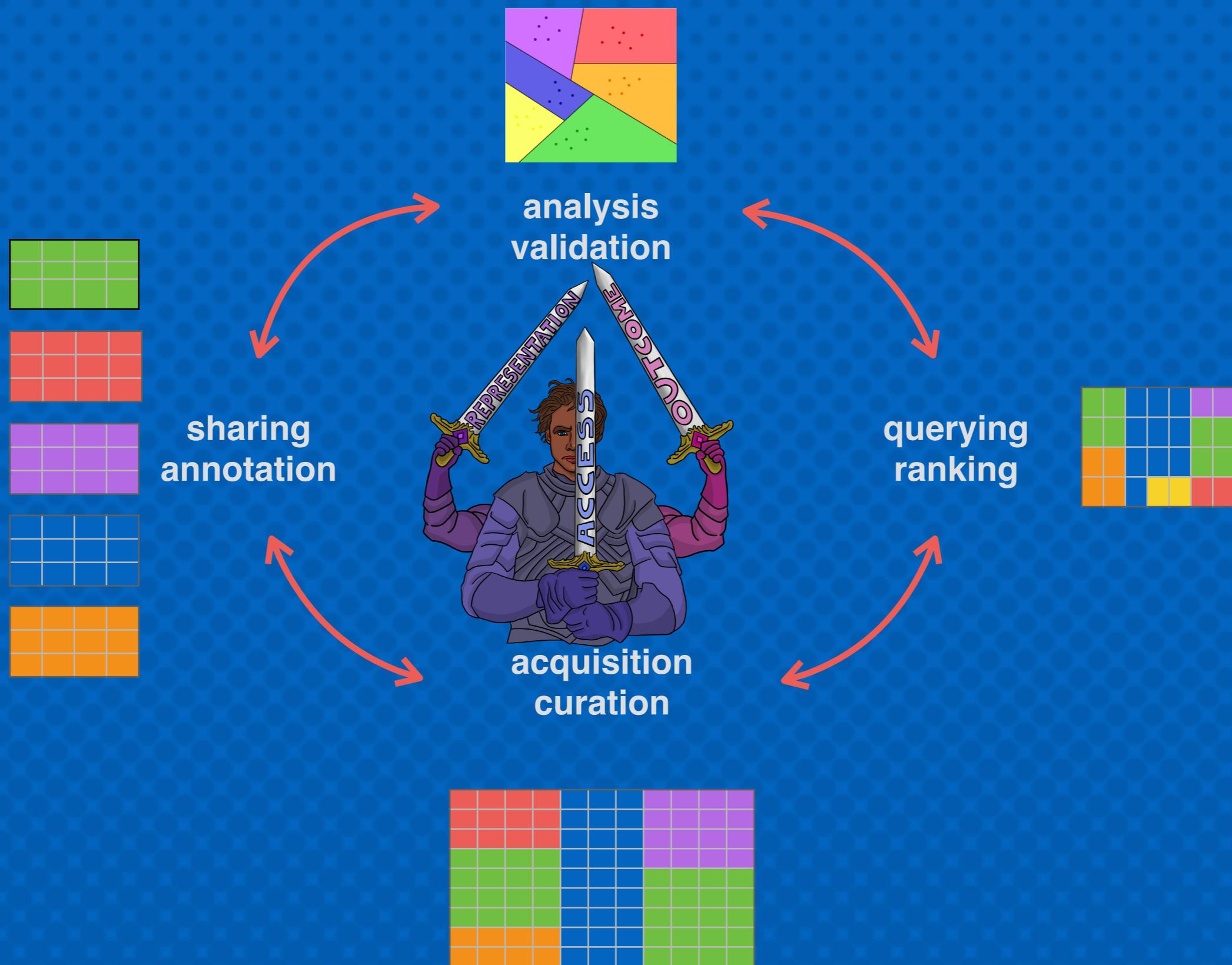
what happens inside the box?



how are results used?



Lifecycle view



Models and assumptions



module 2:

transparency &

interpretability

The evils of discrimination

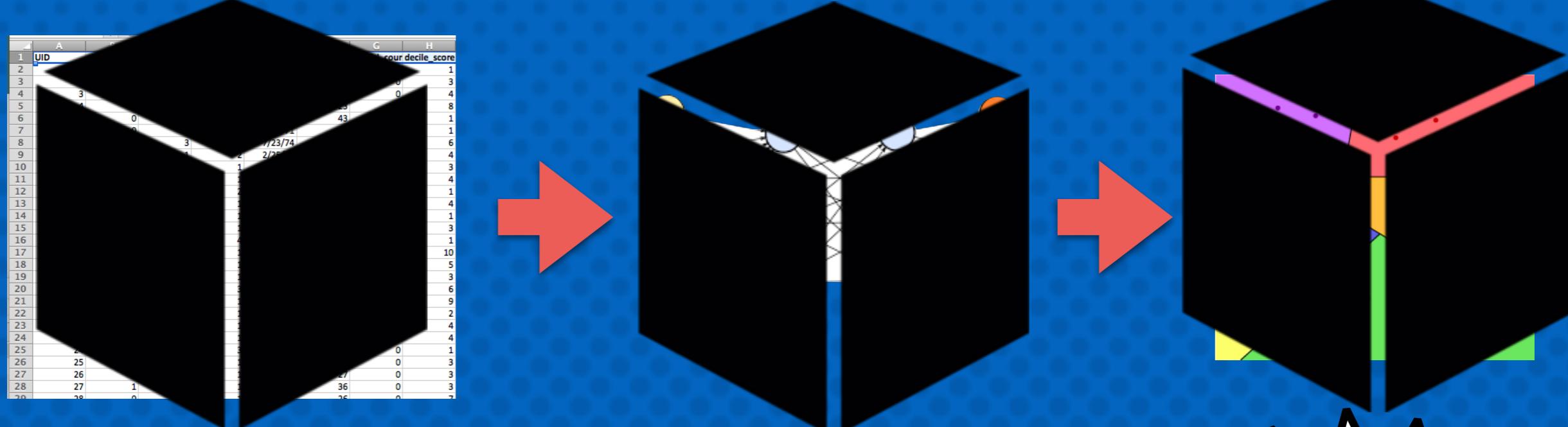
Disparate treatment

is the illegal practice of treating an entity, such as a job applicant or an employee, differently based on a **protected characteristic** such as race, gender, age, religion, sexual orientation, or national origin.

Disparate impact

is the result of systematic disparate treatment, where disproportionate **adverse impact** is observed on members of a **protected class**.

Regulating automated decisions

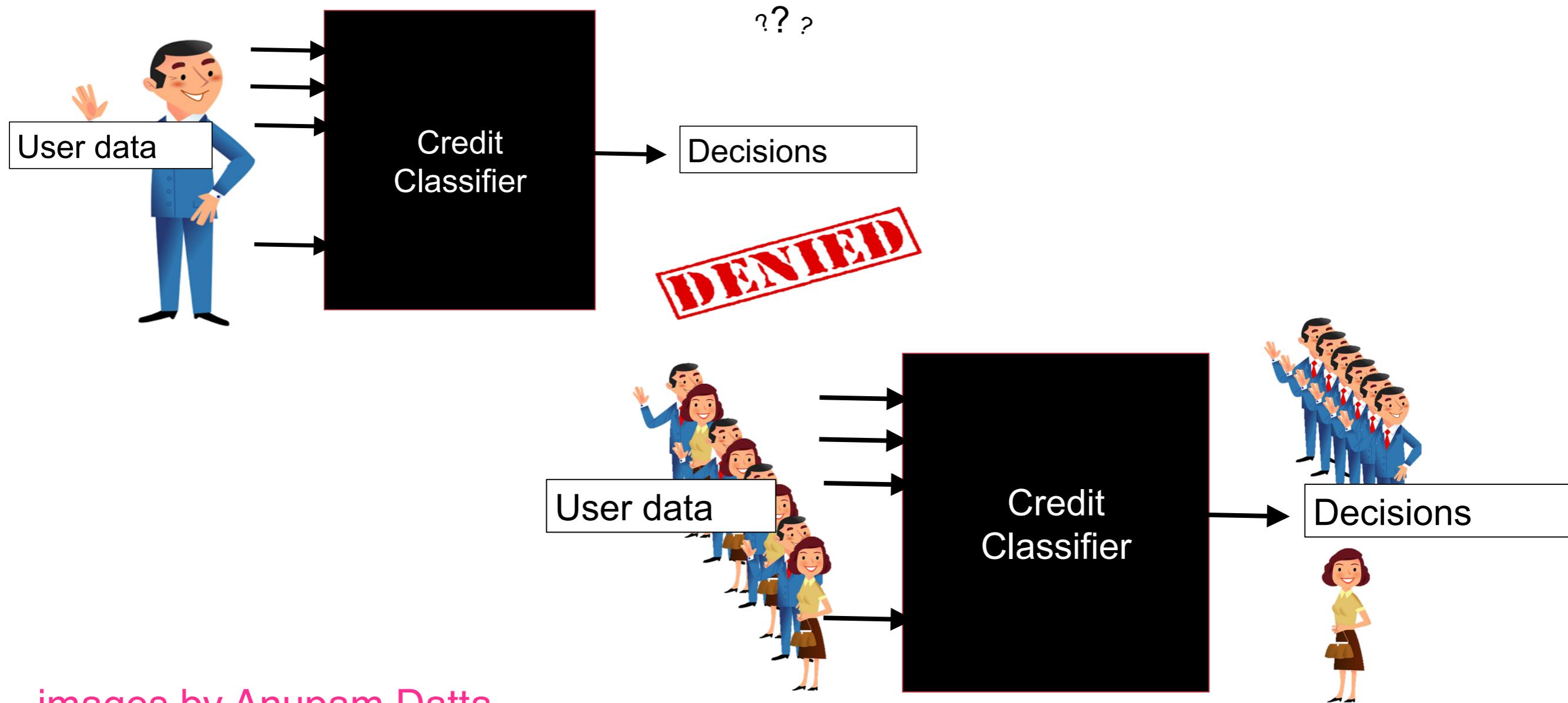


Fair Housing
Act

Equal Credit
Opportunity
Act, 1964

Civil Rights
Act, 1964

Auditing black-box models



images by Anupam Datta

Nutritional labels

Ranking Facts

Ingredients →

Attribute	Importance
PubCount	1.0
CSRakingAllArea	0.24
Faculty	0.12

Importance of an attribute in a ranking is quantified by the correlation coefficient between attribute values and items scores, computed by a linear regression model. Importance is high if the absolute value of the correlation coefficient is over 0.75, medium if this value falls between 0.25 and 0.75, and low otherwise.

Diversity overall ?

DeptSizeBin = Regional Code =

Large Small
NE W MW SA SC

Fairness ? →

DeptSizeBin	FA*IR	Pairwise	Proportion
Large	Fair	Fair	Fair
Small	Unfair	Unfair	Unfair

A ranking is considered unfair when the p-value of the corresponding statistical test falls below 0.05.

← **Stability**

Top-K	Stability
Top-10	Stable
Overall	Stable

comprehensible: short, simple, clear

consultative: provide actionable info

comparable: implying a standard

module 3:

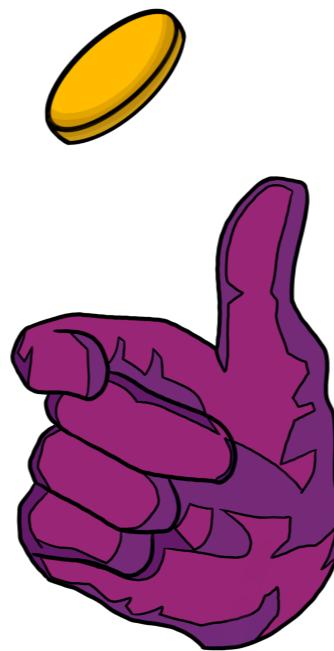
data protection & privacy

Privacy: two sides of the same coin

Did you go out drinking over the weekend?

protecting an individual

plausible deniability



learning about the population

noisy estimates

Truth or dare

Did you go out drinking over the weekend?

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the group for whom **P** holds

thus, we estimate **p** as:

$$\tilde{p} = 2A - \frac{1}{2}$$

1. flip a coin **C1**

1. if **C1** is tails, then **respond truthfully**

2. if **C1** is heads, then flip another coin **C2**

1. if **C2** is heads then **Yes**

2. else **C2** is tails then respond **No**



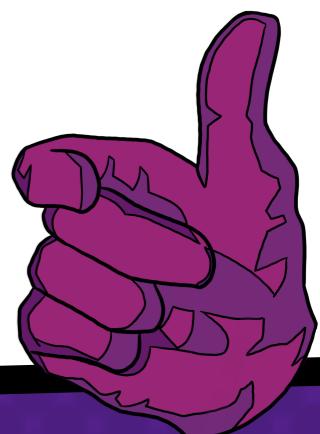
randomization - adding noise - is what gives plausible deniability a process privacy method



the expected number of **Yes** answers is:

$$A = \frac{3}{4}p + \frac{1}{4}(1-p) = \frac{1}{4} + \frac{p}{2}$$

privacy comes from plausible deniability



Differential privacy

review articles

DOI:10.1145/1866739.1866758

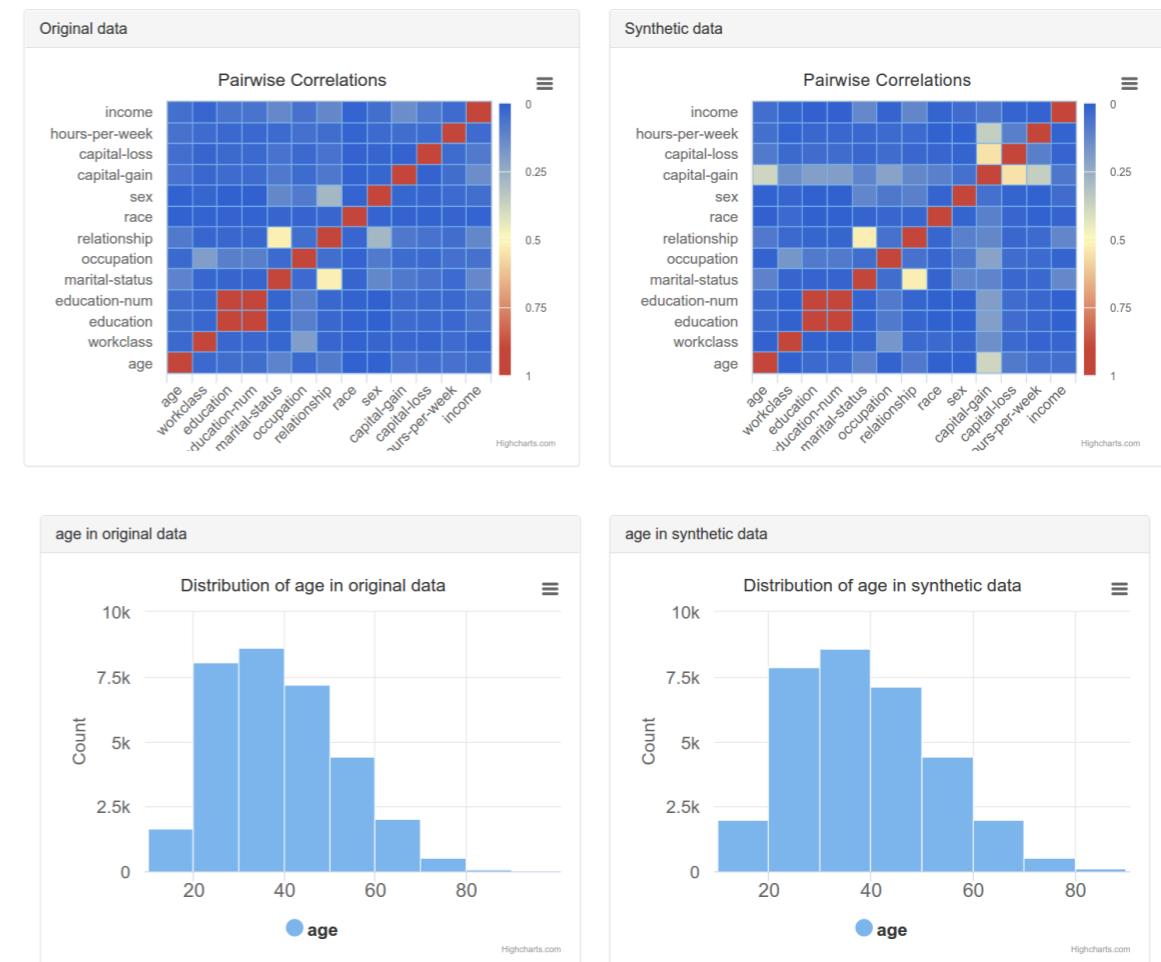
What does it mean to preserve privacy?

BY CYNTHIA DWORK

A Firm Foundation for Private Data Analysis

Communications of the ACM [CACM Homepage archive](#)

Volume 54 Issue 1, January 2011
Pages 86-95



Regulating ADS?

Precautionary

Nah! I'm fine!



The Anti-Elon ✅
@antiElon

Regulation rocks!

2.3K 9.2K 126K

Risk-based



r/ai

Legal frameworks

GENERAL DATA PROTECTION REGULATION (GDPR) RECITALS KEY ISSUES Deutsch

GDPR

- Chapter 1 (Art. 1 – 4)
- General provisions
- Chapter 2 (Art. 5 – 11)
- Principles
- Chapter 3 (Art. 12 – 23)
- Rights of the data subject
- Chapter 4 (Art. 24 – 43)
- Controller and processor
- Chapter 5 (Art. 44 – 50)
- Transfers of personal data to third countries or international organisations
- Chapter 6 (Art. 51 – 59)
- Independent supervisory authorities
- Chapter 7 (Art. 60 – 76)
- Cooperation and consistency
- Chapter 8 (Art. 77 – 84)
- Remedies, liability and penalties
- Chapter 9 (Art. 85 – 91)
- Provisions relating to specific processing situations
- Chapter 10 (Art. 92 – 93)
- Delegated acts and implementing acts
- Chapter 11 (Art. 94 – 99)
- Final provisions

**General Data Protection Regulation
GDPR**

Welcome to gdpr-info.eu. Here you can find the official [PDF](#) of the Regulation (EU) 2016/679 (General Data Protection Regulation) in the current version of the OJ L 119, 04.05.2016; cor. OJ L 127, 23.5.2018 as a neatly arranged website. All Articles of the GDPR are linked with suitable recitals. The European Data Protection Regulation is applicable as of May 25th, 2018 in all member states to harmonize data privacy laws across Europe. If you find the page useful, feel free to support us by sharing the project.

Quick Access

Chapter 1 – 1 2 3 4
Chapter 2 – 5 6 7 8 9 10 11
Chapter 3 – 12 13 1
Chapter 4 – 24 25 2
Chapter 5 – 44 45 4
Chapter 6 – 51 52 5
Chapter 7 – 60 61 6
Chapter 8 – 77 78 7
Chapter 9 – 85 86 8

 **Government of Canada** **Gouvernement du Canada** 

[Home](#) → [How government works](#) → [Policies, directives, standards and guidelines](#)

Directive on Automated Decision-Making

The Government of Canada is increasingly looking to utilize artificial intelligence to make, or assist in making, administrative decisions to improve service delivery. The Government is committed to doing so in a manner that is compatible with core administrative law principles such as transparency, accountability, legality, and procedural fairness. Understanding that this technology is changing rapidly, this Directive will continue to evolve to ensure that it remains relevant.

Date modified: 2019-02-05



in summary

So what is RDS?

As advertised: ethics, legal compliance, personal responsibility.
But also: **data quality!**

A technical course, with content drawn from:

1. fairness, accountability and transparency
2. data engineering
3. privacy & data protection



We will learn **algorithmic techniques** for data analysis.

We will also learn about recent **laws / regulatory frameworks**.

Bottom line: we will learn that many of the problems are **socio-technical**, and so cannot be “solved” with technology alone.

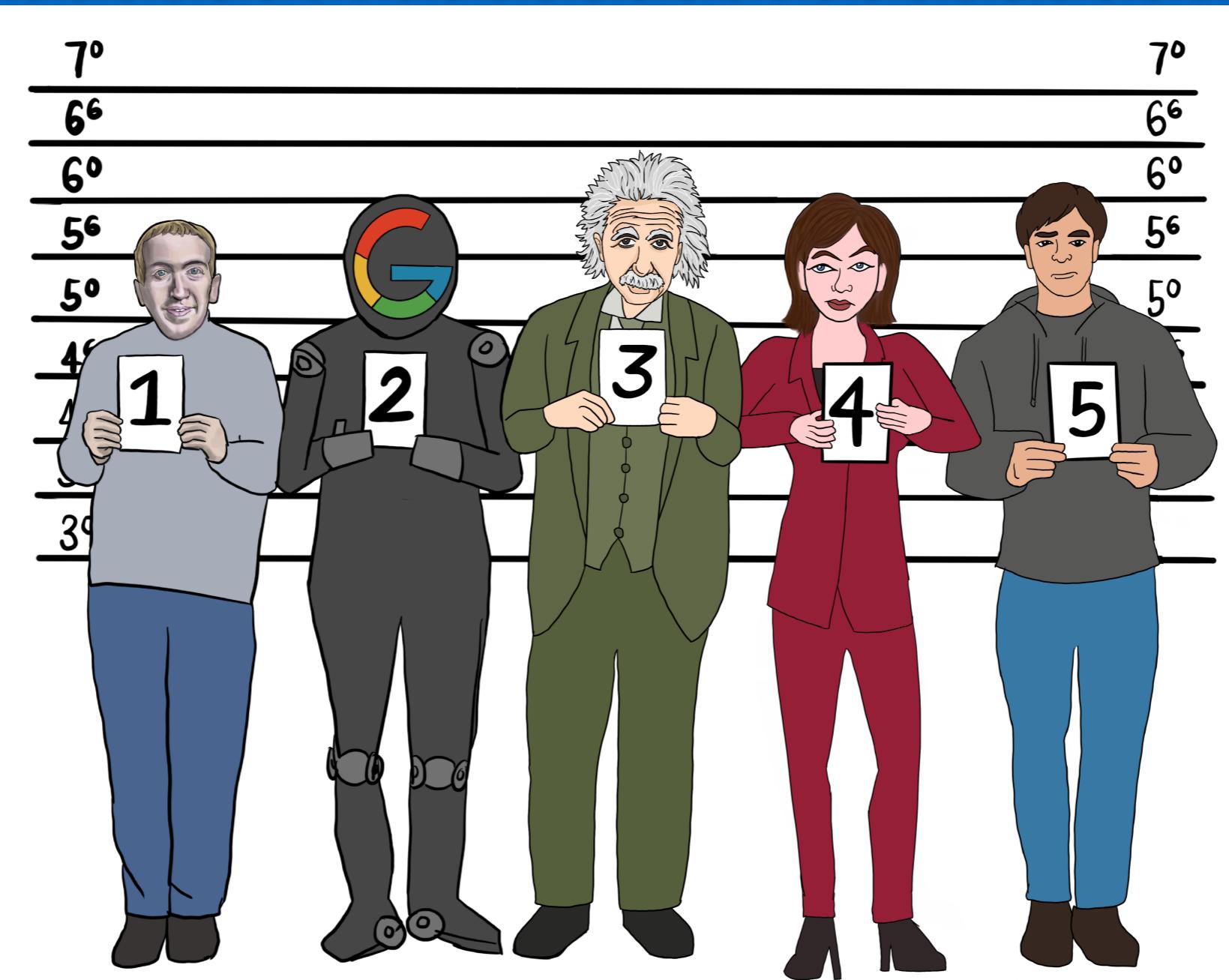
My perspective: a pragmatic engineer, **not** a technology skeptic.

Nuance, please!



r/ai

We all are responsible



@FalaahArifKhan

Responsible Data Science

Introduction and Overview

Thank you!