TEACHERS COLLEGE
COLUMBIA UNIVERSITY

Data Ethics and Algorithmic Bias

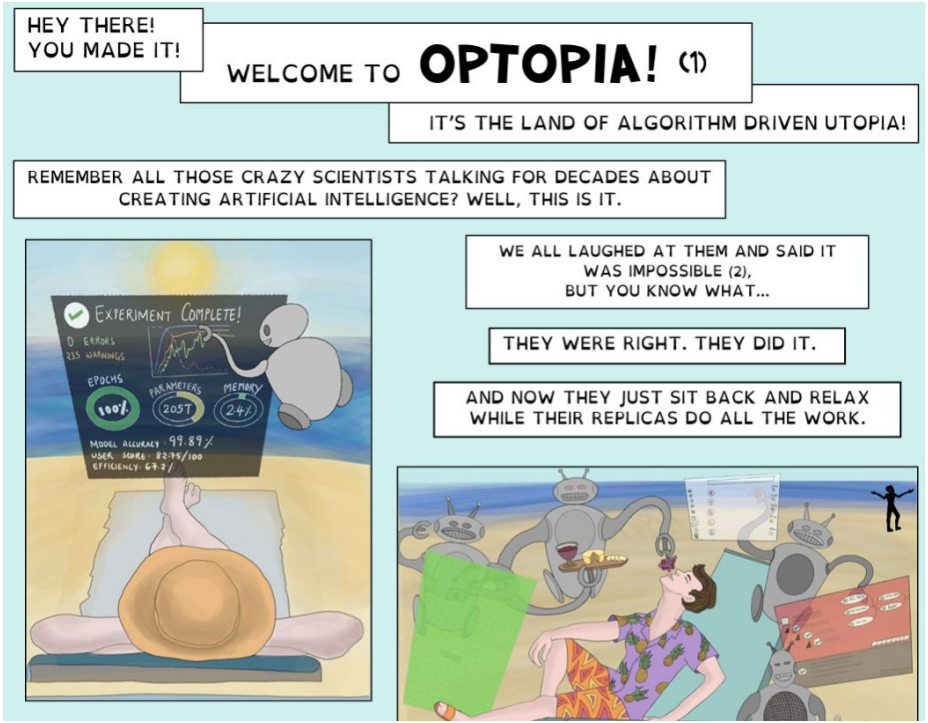
Yasemin Gulbahar, PhD

gulbahar@tc.columbia.edu



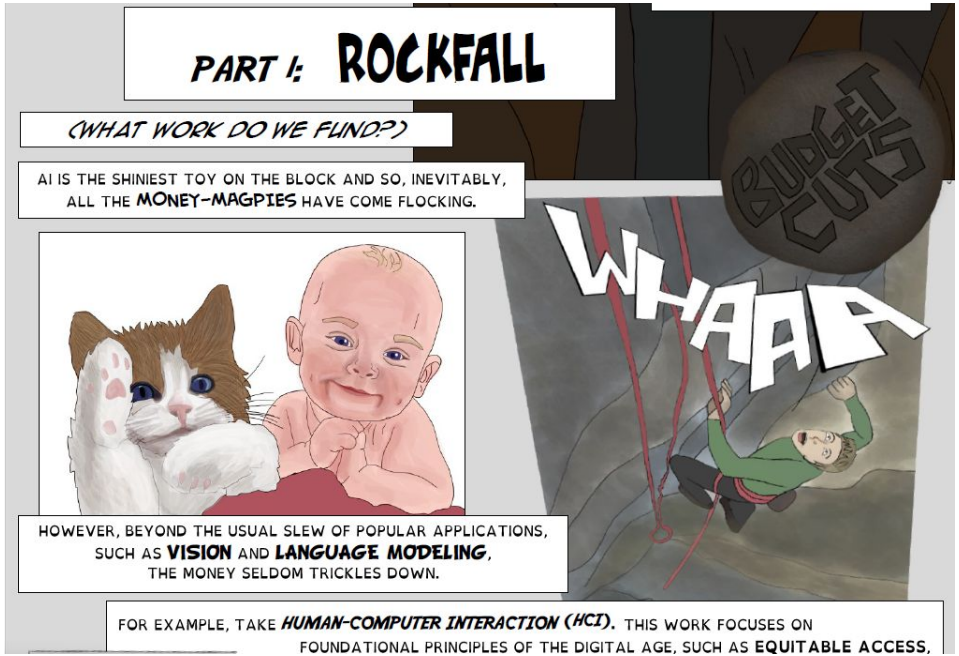
Mirror Mirror: Reflections on a Comic Book

Can ChatGPT be credited as an author of an article?



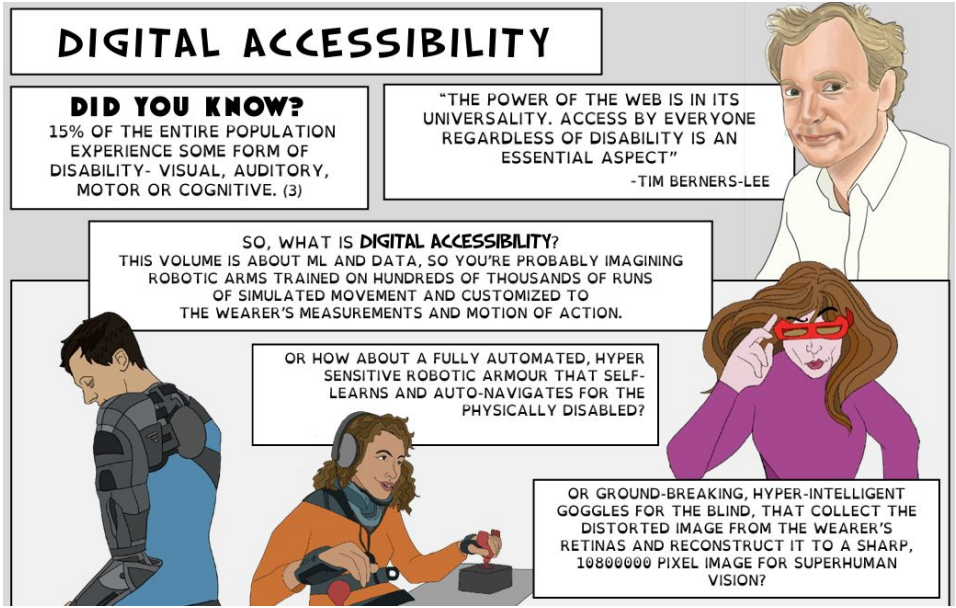
Mirror Mirror: Reflections on a Comic Book

1. Command
2. Graphical
3. Multimedia
4. Virtual reality
5. Web
6. Mobile
7. Appliance
8. Voice
9. Pen
10. Touch
11. Gesture
12. Haptic
13. Multimodal
14. Shareable
15. Tangible
16. Augmented reality
17. Wearables
18. Robots and drones
19. Brain-computer interaction
20. Smart Tech

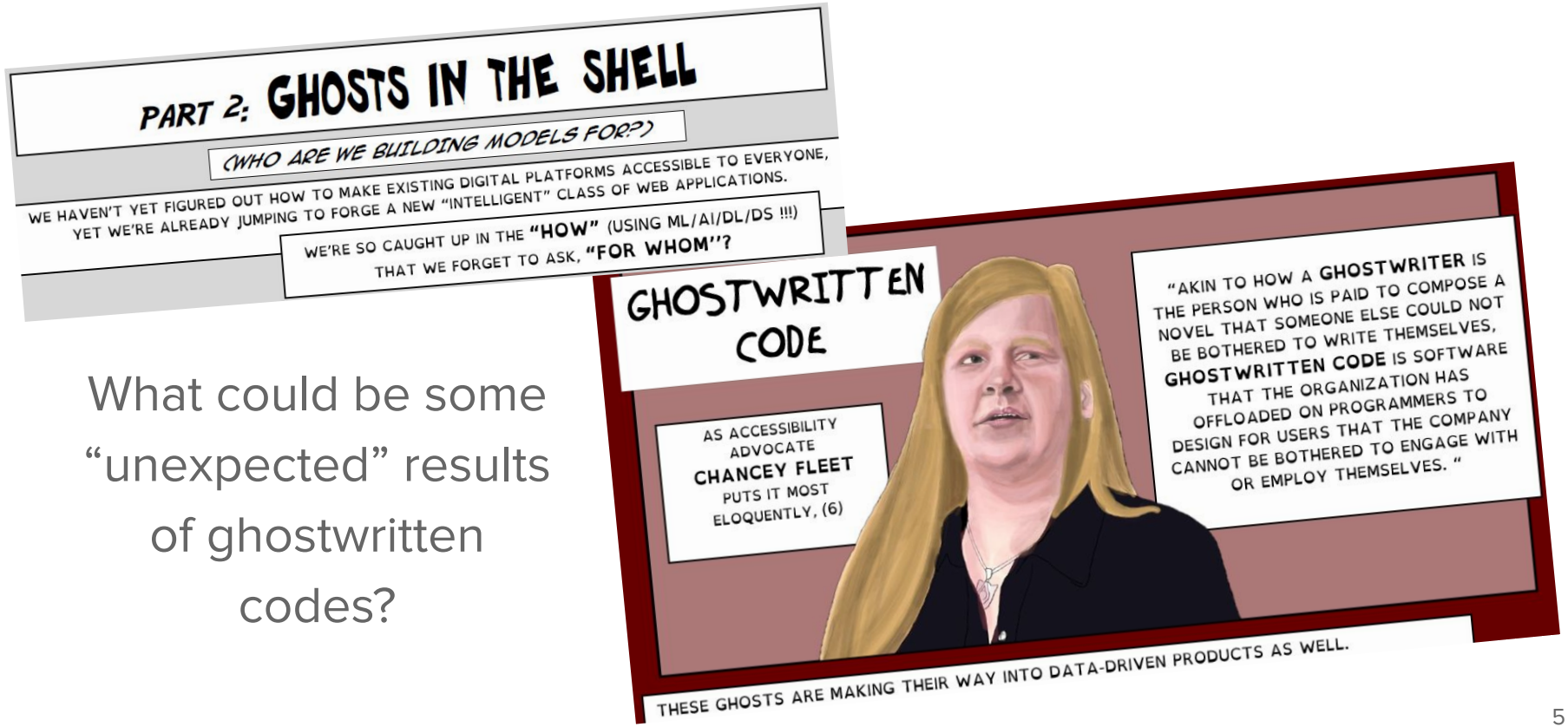


Mirror Mirror: Reflections on a Comic Book

What is meant by “digital accessibility” and “encoded inhospitality”?



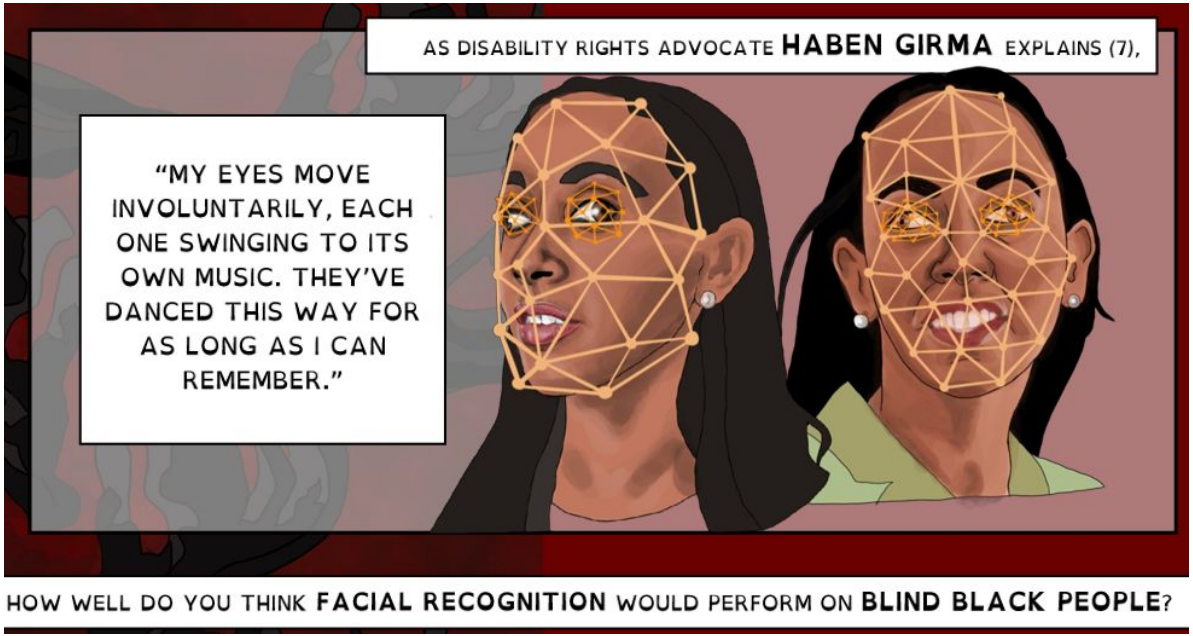
Mirror Mirror: Reflections on a Comic Book



What could be some "unexpected" results of ghostwritten codes?

Mirror Mirror: Reflections on a Comic Book

How can we handle “entire demographics” while designing software and products?



TAKE FACEBOOK'S **"REAL NAME" POLICY** THAT INDISCRIMINATELY TARGETED NATIVE AMERICANS (8)

THE LARGEST SOCIAL NETWORK IN THE WORLD SURE OVERLOOKED THE **CULTURAL AND LINGUISTIC DIFFERENCES** IN NAMES ACROSS THE GLOBE



AND ENDED UP DEPLOYING A BIGOTED ALGORITHM THAT BLOCKED USERS WHOSE NAMES DID NOT CONFORM WITH THE WESTERN ARCHETYPE OF NAMES

PART 3: THE POISONING

(WHAT PROBLEMS ARE WE TRYING TO SOLVE?)

DEVELOPMENTS SUCH AS **ALPHA-GO** (THE GO PLAYING AI) AND **SHAZAM** (THE MUSIC RECOGNITION APP) ARE INDICATIVE OF GENUINE SCIENTIFIC PROGRESS AND DO DEMONSTRABLY MORE GOOD THAN HARM.

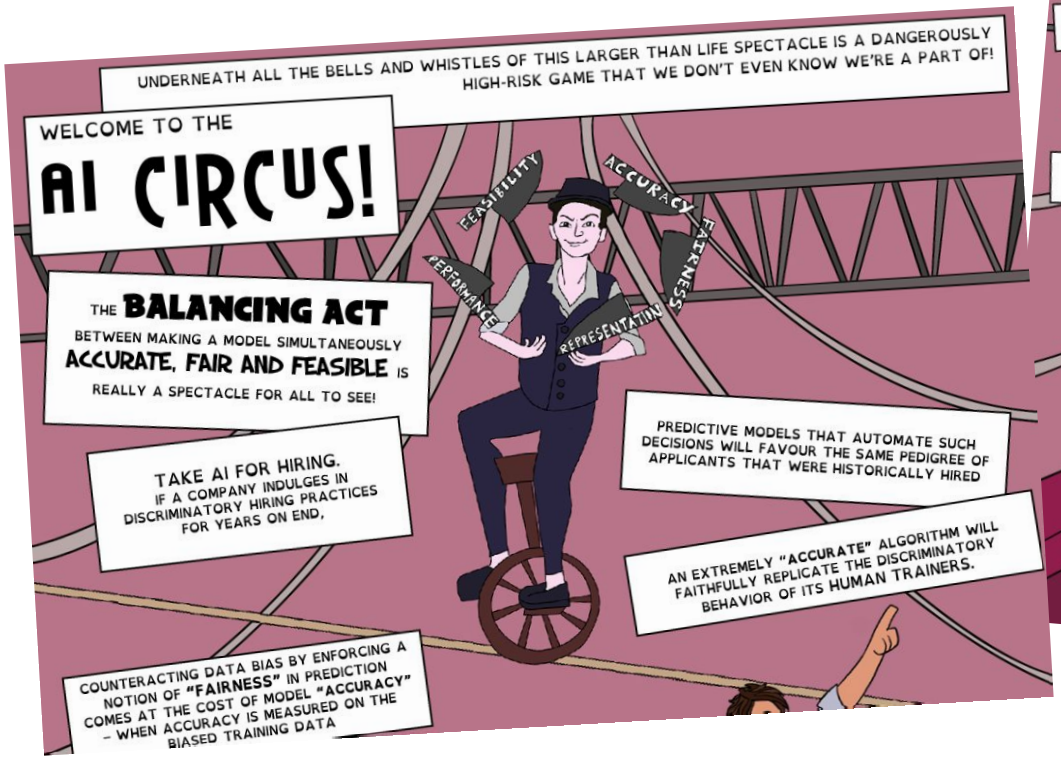
WHY? BECAUSE THE RULES OF GO DON'T CHANGE WHETHER THE PLAYER IS **MALE/FEMALE, BLACK/WHITE, RICH/POOR!**

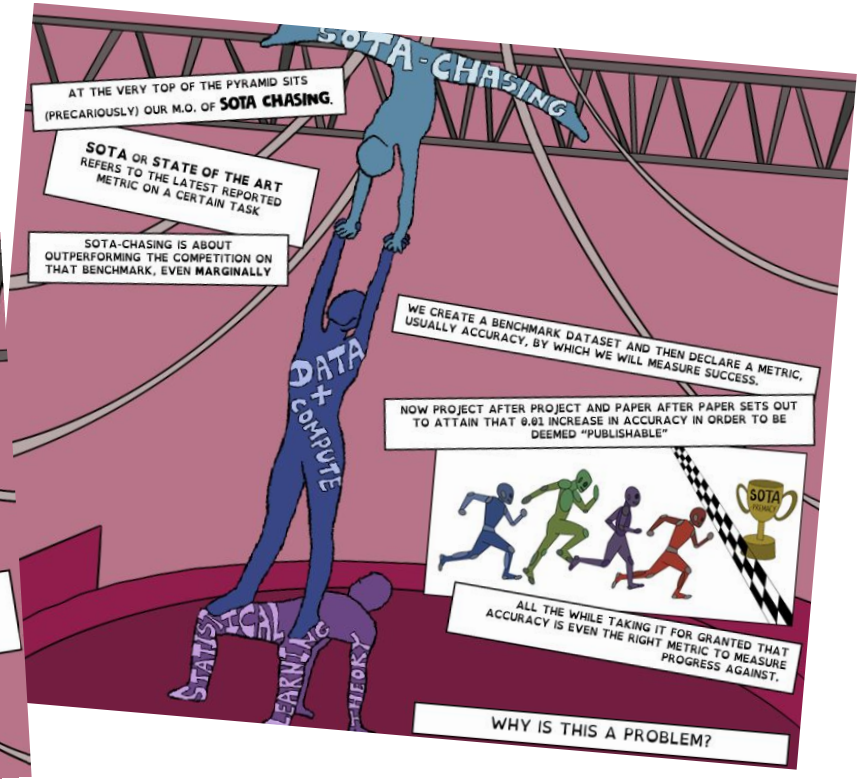


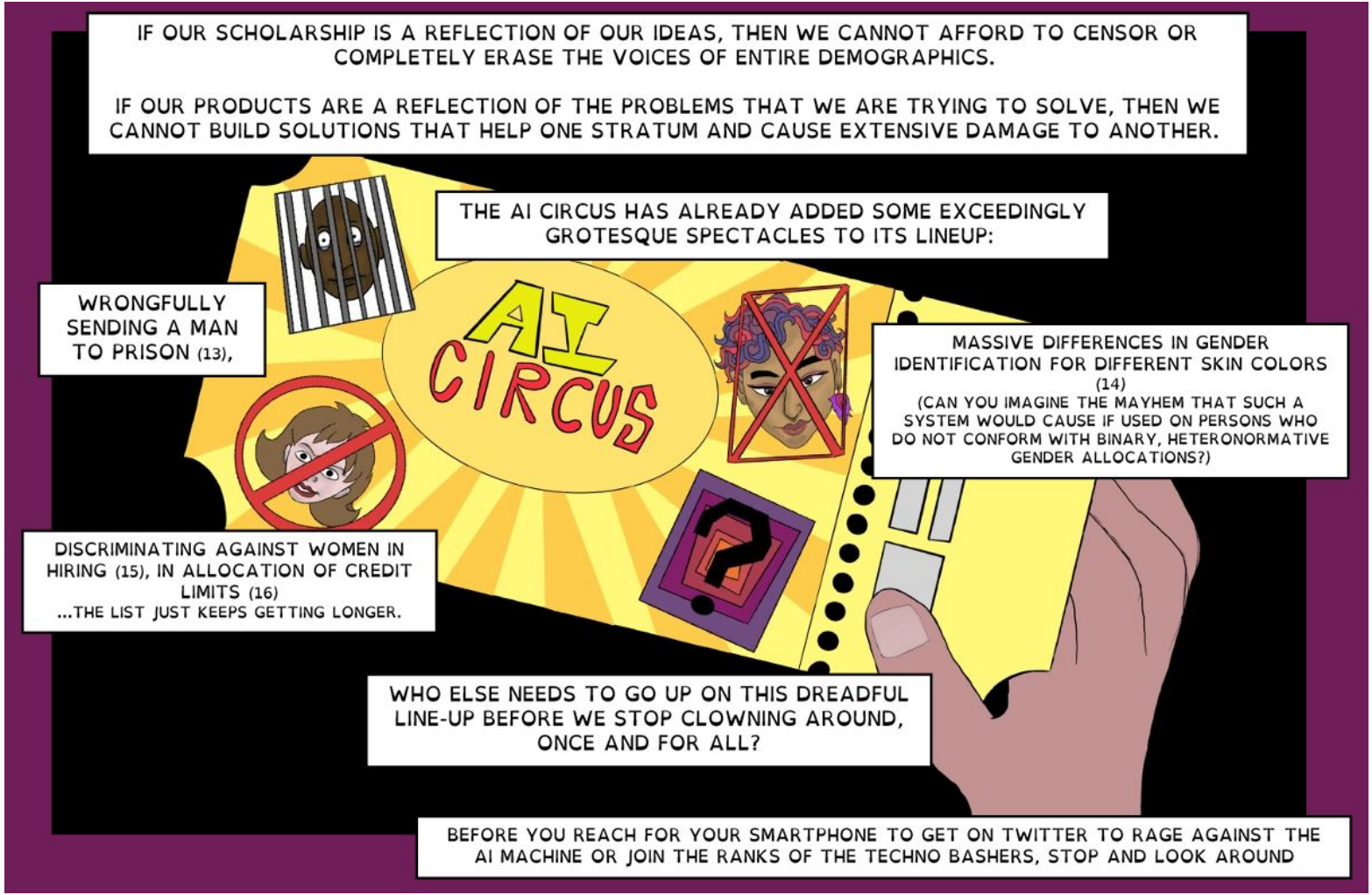
PERCEPTION TASKS, SUCH AS **FACIAL RECOGNITION**, THAT ARE INTERTWINED WITH THE **SOCIAL, POLITICAL AND CULTURAL UNDERPINNINGS** OF THE DATA ON WHICH THEY WERE TRAINED, ARE FAR MORE TOXIC.



THINGS START TO GET REALLY TOXIC IN SETTINGS SUCH AS **HIRING, MODERATION OF HATE SPEECH OR ALLOCATION OF GRADES** (10), WHEN WE TRY TO IMPOSE OBJECTIVITY (FIT A MATHEMATICAL FUNCTION ONTO THE DATA) ON **HUMAN JUDGMENT**, WHICH IS INHERENTLY SUBJECTIVE


UNDERNEATH ALL THE BELLS AND WHISTLES OF THIS LARGER THAN LIFE SPECTACLE IS A DANGEROUSLY HIGH-RISK GAME THAT WE DON'T EVEN KNOW WE'RE A PART OF!
WELCOME TO THE
AI CIRCUS!
THE BALANCING ACT
BETWEEN MAKING A MODEL SIMULTANEOUSLY ACCURATE, FAIR AND FEASIBLE IS REALLY A SPECTACLE FOR ALL TO SEE!
TAKE AI FOR HIRING.
IF A COMPANY INDULGES IN DISCRIMINATORY HIRING PRACTICES FOR YEARS ON END,
PREDICTIVE MODELS THAT AUTOMATE SUCH DECISIONS WILL FAVOUR THE SAME PEDIGREE OF APPLICANTS THAT WERE HISTORICALLY HIRED
AN EXTREMELY "ACCURATE" ALGORITHM WILL FAITHFULLY REPLICATE THE DISCRIMINATORY BEHAVIOR OF ITS HUMAN TRAINERS.
COUNTERACTING DATA BIAS BY ENFORCING A NOTION OF "FAIRNESS" IN PREDICTION COMES AT THE COST OF MODEL "ACCURACY" – WHEN ACCURACY IS MEASURED ON THE BIASED TRAINING DATA
FEASIBILITY
ACCURACY
FAIRNESS
PERFORMANCE
REPRESENTATION


SOTA-CHASING
AT THE VERY TOP OF THE PYRAMID SITS (PRECARIOUSLY) OUR M.O. OF SOTA CHASING.
SOTA OR STATE OF THE ART REFERS TO THE LATEST REPORTED METRIC ON A CERTAIN TASK
SOTA-CHASING IS ABOUT OUTPERFORMING THE COMPETITION ON THAT BENCHMARK, EVEN MARGINALLY
WE CREATE A BENCHMARK DATASET AND THEN DECLARE A METRIC, USUALLY ACCURACY, BY WHICH WE WILL MEASURE SUCCESS.
DATA + COMPUTE
NOW PROJECT AFTER PROJECT AND PAPER AFTER PAPER SETS OUT TO ATTAIN THAT 0.01 INCREASE IN ACCURACY IN ORDER TO BE DEEMED "PUBLISHABLE"
SOTA
ALL THE WHILE TAKING IT FOR GRANTED THAT ACCURACY IS EVEN THE RIGHT METRIC TO MEASURE PROGRESS AGAINST.
WHY IS THIS A PROBLEM?


IF OUR SCHOLARSHIP IS A REFLECTION OF OUR IDEAS, THEN WE CANNOT AFFORD TO CENSOR OR COMPLETELY ERASE THE VOICES OF ENTIRE DEMOGRAPHICS.
IF OUR PRODUCTS ARE A REFLECTION OF THE PROBLEMS THAT WE ARE TRYING TO SOLVE, THEN WE CANNOT BUILD SOLUTIONS THAT HELP ONE STRATUM AND CAUSE EXTENSIVE DAMAGE TO ANOTHER.
THE AI CIRCUS HAS ALREADY ADDED SOME EXCEEDINGLY GROTESQUE SPECTACLES TO ITS LINEUP:
WRONGFULLY SENDING A MAN TO PRISON (13),
AI CIRCUS
MASSIVE DIFFERENCES IN GENDER IDENTIFICATION FOR DIFFERENT SKIN COLORS (14)
(CAN YOU IMAGINE THE MAYHEM THAT SUCH A SYSTEM WOULD CAUSE IF USED ON PERSONS WHO DO NOT CONFORM WITH BINARY, HETERONORMATIVE GENDER ALLOCATIONS?)
DISCRIMINATING AGAINST WOMEN IN HIRING (15), IN ALLOCATION OF CREDIT LIMITS (16)
...THE LIST JUST KEEPS GETTING LONGER.
WHO ELSE NEEDS TO GO UP ON THIS DREADFUL LINE-UP BEFORE WE STOP CLOWNING AROUND, ONCE AND FOR ALL?
BEFORE YOU REACH FOR YOUR SMARTPHONE TO GET ON TWITTER TO RAGE AGAINST THE AI MACHINE OR JOIN THE RANKS OF THE TECHNO BASHERS, STOP AND LOOK AROUND

HERE IS A MORE NUANCED TAKE ON WHETHER AI LEADS TO A **UTOPIA** OR A ***DYSTOPIA:***

FOR STARTERS, **THERE IS RARELY AN OBJECTIVE TRUTH**! MORE OFTEN THAN NOT, THE EFFICACY OF A MODEL DEPENDS ON THE **CONTEXT** FOR WHICH IT WAS DESIGNED

THE "**GROUND TRUTH**" THAT WE PRETEND EXISTS, AND AGAINST WHICH WE MEASURE MODEL ACCURACY, IS JUST THE **CLOTHES** THAT THE **ML EMPEROR** IS **NOT** WEARING!

THE ENGINEERING MINDSET IS TO TAKE THE CLASS LABELS AS GOSPEL AND BLINDLY TRY TO OPTIMIZE FOR THEM.

BUT CLASS LABELS ARE JUST **PROXIES** FOR **UNDERLYING SOCIAL PHENOMENA** AND NO AMOUNT OF **MATHEMATICAL FORMALIZATION** WILL TURN **SOCIAL CONSTRUCTS** INTO **OBJECTIVE TRUTHS.**



THE REALITY IS THAT **ALL** MODELS ARE **WRONG**. **SOME** MODELS ARE **USEFUL**!

IN THIS ART GALLERY, EACH PAINTING DEPICTS AN **APPLE**. BUT ONLY ONE OF THEM IS POTENTIALLY USEFUL AS A **REAL-LIFE APPLE DETECTOR**









WE OFTEN FIND IT HARD TO JUDGE WHICH MODEL IS MOST USEFUL, BECAUSE THAT REQUIRES DEEP **DOMAIN EXPERTISE**.

WE HAVE BEEN **DANGEROUSLY CONFLATING** EXPERTISE IN TRAINING AND DEPLOYING A MODEL WITH DOMAIN EXPERTISE.

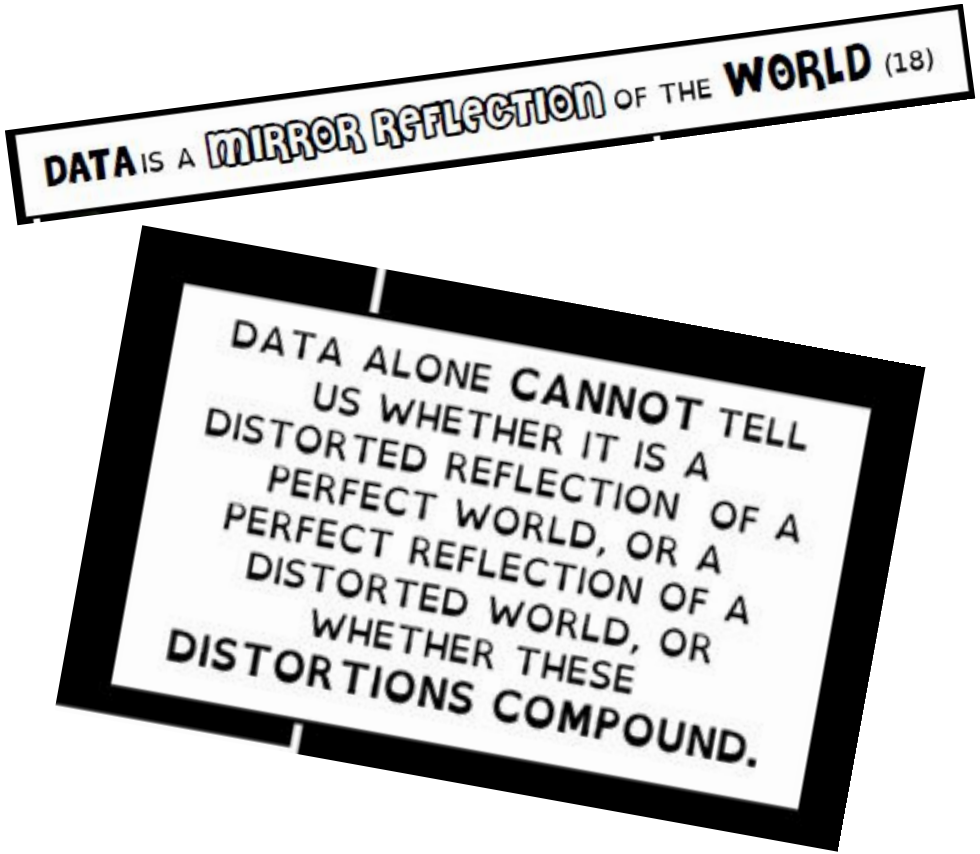
INSTEAD WE SHOULD **ACKNOWLEDGE THE LIMITATION OF OUR EXPERTISE** AS SCIENTISTS AND ENGINEERS AND **INVITE THE TRUE DOMAIN EXPERTS** TO COME TO THE TABLE.

Think!

- Do you agree/disagree to the following statement:

"Data is a mirror reflection of the world"

- What about the possibility of "the mirror is perfect and the world itself is distorted"



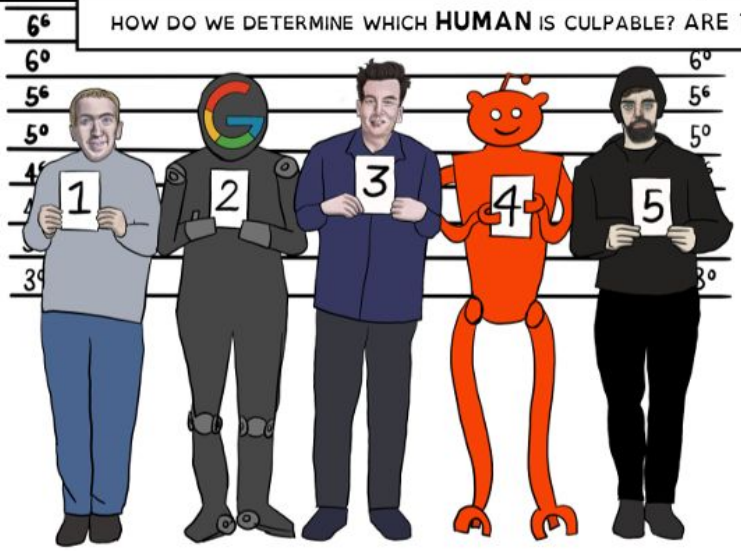
PREDICTION IS DIFFICULT, ESPECIALLY ABOUT THE **FUTURE**! (19)

WHEN THINGS (INEVITABLY) GO WRONG, WHO IS **RESPONSIBLE**?

IT CANNOT BE THE ALGORITHM.

BUT GIVEN THE MANY STAKEHOLDERS THAT PLAY A PART IN THE CREATION AND OPERATION OF A SOFTWARE PRODUCT,

HOW DO WE DETERMINE WHICH **HUMAN** IS CULPABLE? ARE THEY ALL?



I KNOW WHAT YOU'RE THINKING...

"I SEE WHERE YOU'RE GOING WITH THIS... YOU'RE NOT SERIOUSLY GOING TO GET INTO REGULATION NOW, ARE YOU?"

WELL...TIME TO REMIND YOU OF OUR RECOMMENDED APPROACH TO THINKING ABOUT AI.

REMEMBER, **NUANCE**?!



RIGHT NOW, SILICON VALLEY WILL HAVE YOU BELIEVE THAT TECHNOLOGY NEEDS TO BE ALLOWED TO RUN FREE. REGULATION IS A CATASTROPHE OF COSMIC PROPORTIONS AND WOULD BE THE END OF THE INTERNET, AND BY EXTENSION, INNOVATION AND PROGRESS.

AND FOR **REGULATORS** WHO NEED **PRECISE DEFINITIONS** TO BUILD RULES AND POLICIES

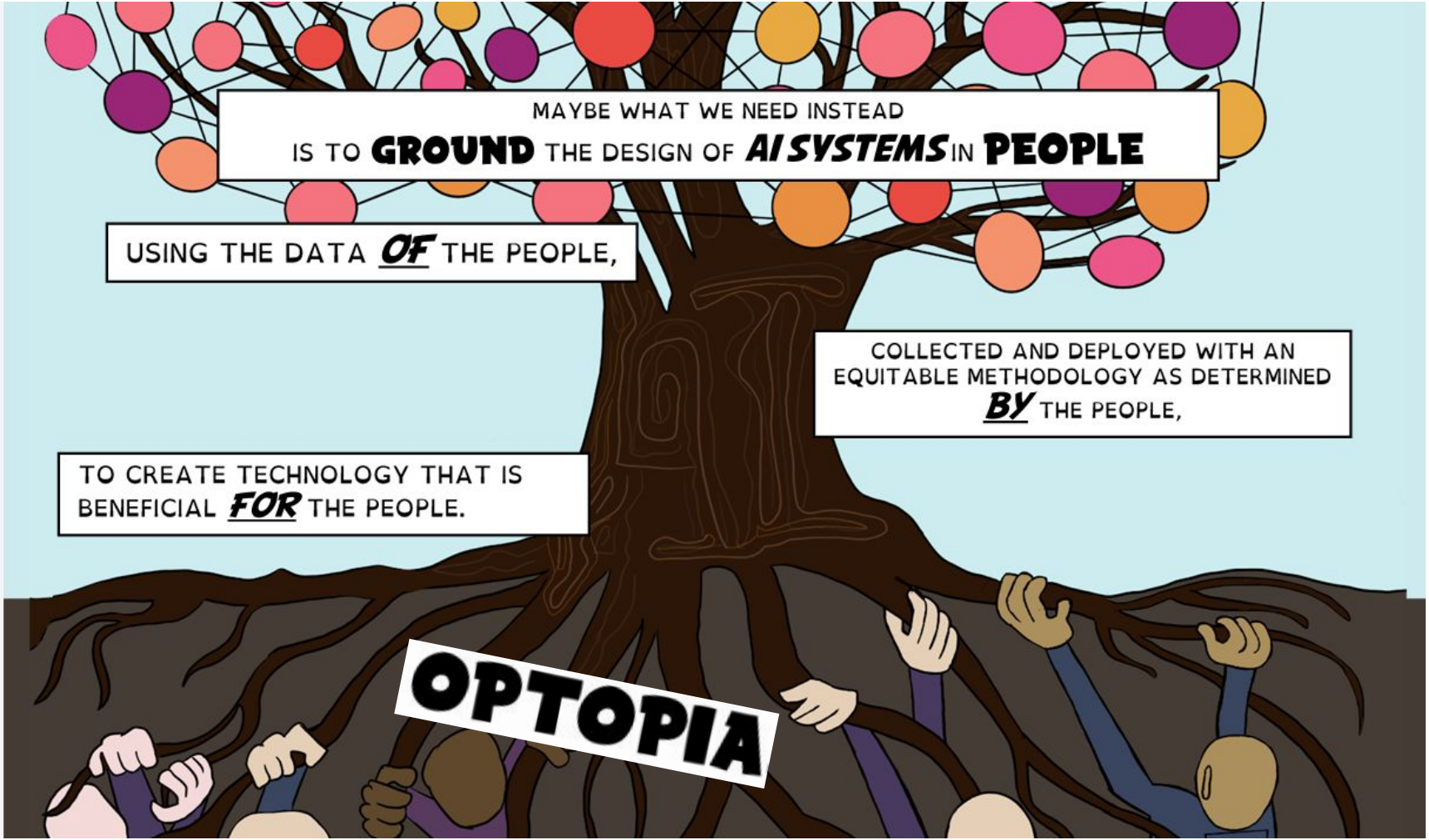
TO COME UP WITH GOOD DEFINITIONS, WE NEED EXAMPLES OF SYSTEMS THAT ARE USED **TODAY**!

TAKE THE **NYC AUTOMATED DECISION SYSTEMS (ADS) TASK FORCE**, THE FIRST OF ITS KIND IN THE U.S., ENVISIONED TO BE THE BEACON FOR **TRANSPARENCY** AND **EXPERT INSIGHT** INTO THE USE OF ALGORITHMS TO AID DECISION-MAKING BY CITY AGENCIES. (20)

BUT THEY DIDN'T GET VERY FAR.

A **GOOD DEFINITION** WAS LACKING, AS WERE **EXAMPLES**.

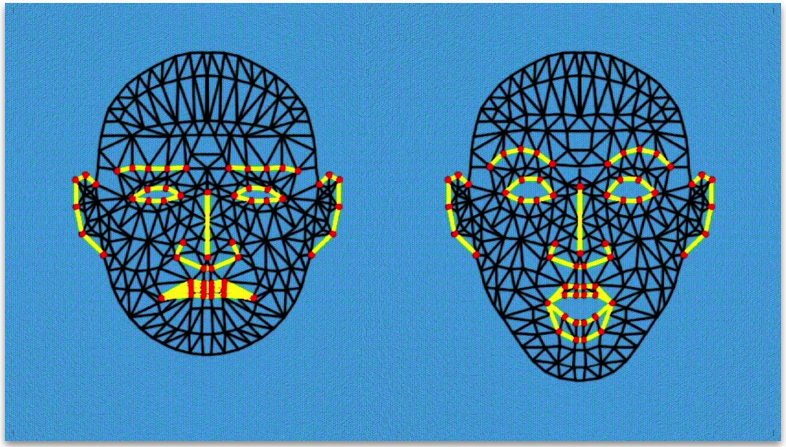


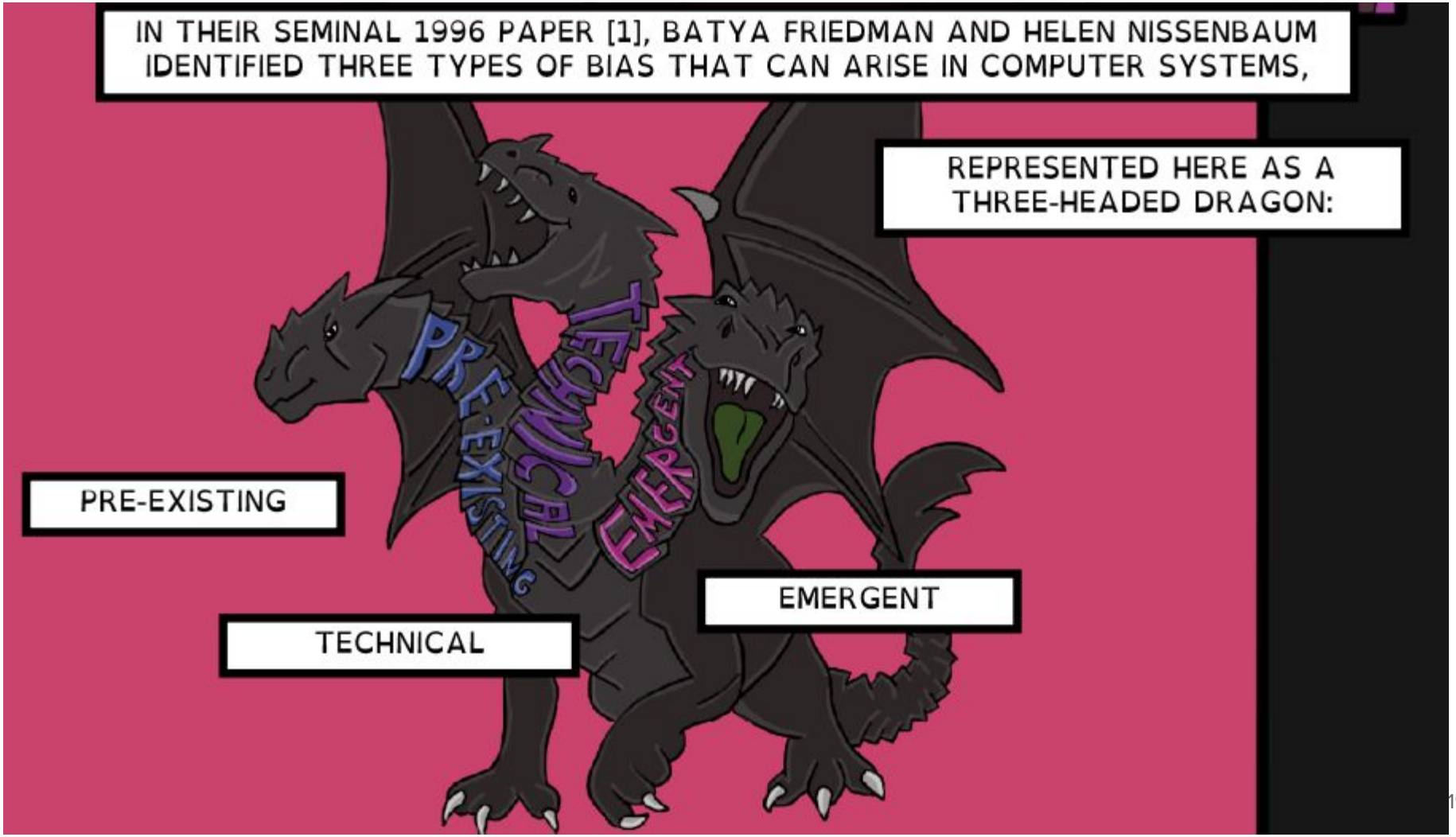

MAYBE WHAT WE NEED INSTEAD
IS TO GROUND THE DESIGN OF AI SYSTEMS IN PEOPLE
USING THE DATA OF THE PEOPLE,
COLLECTED AND DEPLOYED WITH AN EQUITABLE METHODOLOGY AS DETERMINED BY THE PEOPLE,
TO CREATE TECHNOLOGY THAT IS BENEFICIAL FOR THE PEOPLE.
OPTOPIA


LET'S TALK ABOUT WHAT WE MEAN BY 'BIAS' IN AI, AND HOW IT ARISES.
SYSTEMATIC DISCRIMINATION BY AN ALGORITHM IS TERMED 'BIAS'.
THESE COULD BE HARMFUL PATTERNS PICKED UP FROM THE DATA ITSELF,
OR FROM HOW THE ALGORITHM IS DESIGNED,
OR FROM THE OBJECTIVES THAT WE SPECIFIED FOR IT,
OR FROM HOW WE USE IT.

Definitions of Algorithmic Bias

- ➔ Unfairness in automated systems
- ➔ A model's predictive performance unjustifiably differs across disadvantaged groups of social axes such as race, gender and class
- ➔ Systematic and unfair discrimination against individuals or groups in favor of others
- ➔ Possible sources of harm throughout the machine learning process, including "unintended or potentially harmful" properties of the data that lead to "unwanted or societtally unfavorable outcomes"




IN THEIR SEMINAL 1996 PAPER [1], BATYA FRIEDMAN AND HELEN NISSENBAUM IDENTIFIED THREE TYPES OF BIAS THAT CAN ARISE IN COMPUTER SYSTEMS,
REPRESENTED HERE AS A THREE-HEADED DRAGON:
PRE-EXISTING
TECHNICAL
EMERGENT
PRE-EXISTING
TECHNICAL
EMERGENT

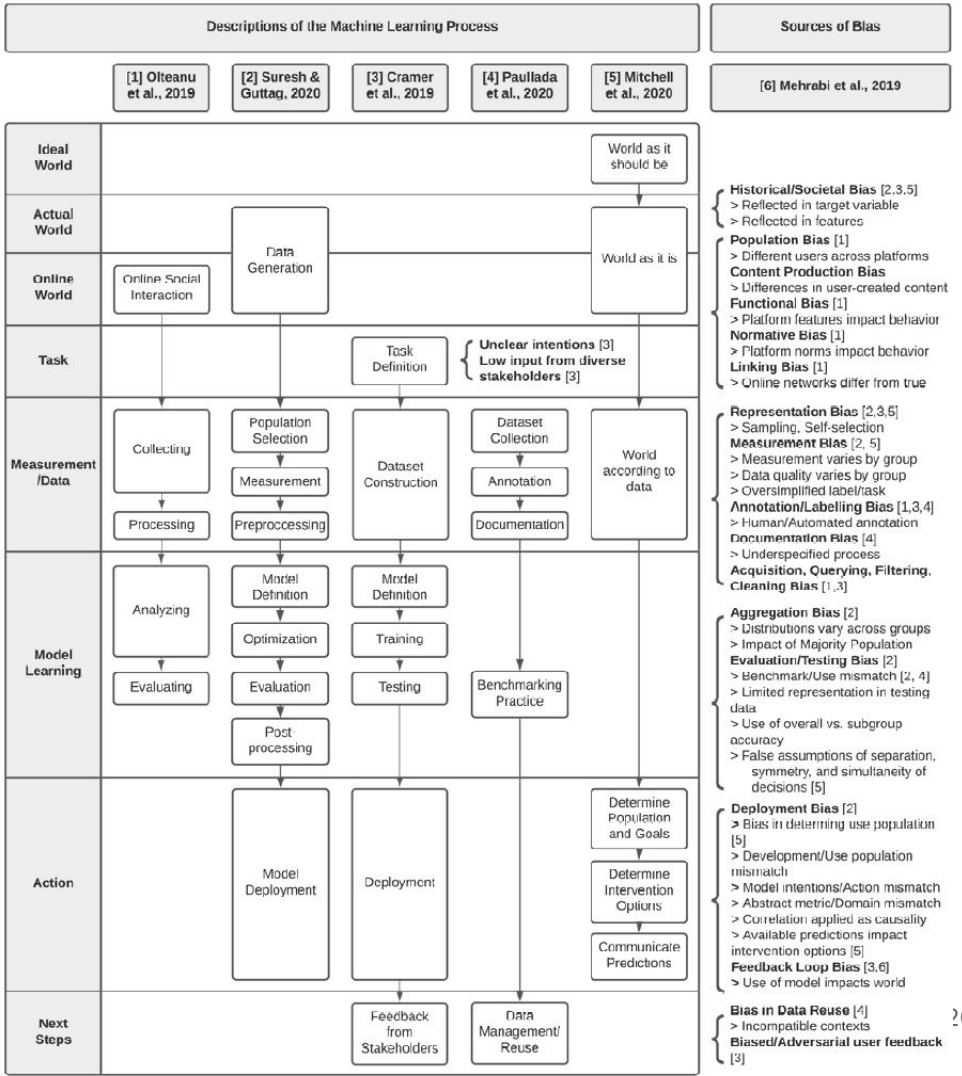
Descriptions of the Machine Learning Process and Possible Sources of Bias

Population Factors and Student Modeling

1. What have we learned about algorithmic bias in education from looking at the most commonly studied demographic categories? (The limited research on other "populations" or “subpopulations”)

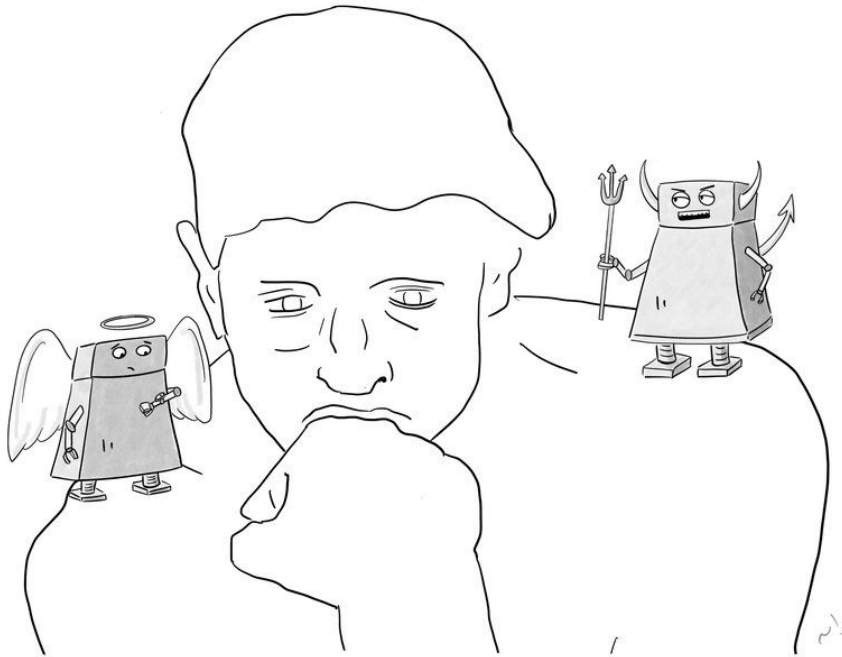
From unknown bias to known bias; from fairness to equity

2. What obstacles stand in the way of research into unknown algorithmic biases? How should researchers and educators address these obstacles?



Baker, R. S., & Hawn, A. (2021, March 1). Algorithmic Bias in Education. https://doi.org/10.35542/osf.io/pbmvz

AI as a moral superpower



EvilAICartoons.com @EvilAICartoons

EvilAICartoons.com

<< Look, man, we haven't got all day. How about Jim and I just figure it out between ourselves? >>

Real-Life Examples

1. *COMPAS (Correctional Offender Management Profiling for Alternative Sanctions):*

- **Industry**: Criminal Justice
- **Bias**: The COMPAS algorithm was used to assess the risk of reoffending for individuals in the criminal justice system, influencing parole and sentencing decisions. A **ProPublica investigation** in 2016 revealed that the algorithm disproportionately **overestimated the risk of reoffending for Black defendants** while **underestimating the risk for white defendants**.
- **Impact**: This led to **discriminatory outcomes**, where Black defendants were more likely to receive harsher sentences or be denied parole, even though the algorithm was designed to be objective.

2. Amazon's Hiring Algorithm:

- **Industry**: Human Resources
- **Bias**: In 2018, it was revealed that Amazon's automated hiring system was biased against female candidates. The algorithm was trained on resumes submitted to Amazon over the past 10 years, which led the system to favor male candidates for technical roles. This bias occurred because the dataset reflected a historical trend where men were more likely to be hired for such positions.
- **Impact**: Amazon scrapped the system because it perpetuated **gender bias**, ultimately disadvantaging female candidates and reinforcing existing gender disparities in the tech industry.

Real-Life Examples (continued)

3. Google's Ad Targeting Algorithm:

- **Industry**: Digital Advertising
- **Bias**: Google's advertising algorithm was found to display **higher-paying job ads to men** over women, even when they had the same qualifications. Research revealed that ads for high-paying jobs like engineering or executive positions were shown predominantly to male users, while ads for lower-paying roles, such as administrative positions, were more likely to be shown to women.
- **Impact**: This perpetuated **gender-based economic inequalities** and discriminatory practices in the labor market, as it limited women's exposure to high-paying opportunities.

4. Facial Recognition Technology (e.g., IBM, Amazon, and Microsoft):

- **Industry**: Security, Retail, Law Enforcement
- **Bias**: Numerous studies, including research by **MIT Media Lab** and **Joy Buolamwini's Gender Shades Project**, have shown that facial recognition algorithms are **significantly less accurate** for darker-skinned individuals, particularly women. For instance, IBM's and Amazon's facial recognition technologies exhibited **lower accuracy rates for Black women** compared to lighter-skinned men. This bias stems from the underrepresentation of diverse racial and gender groups in training datasets.
- **Impact**: Inaccurate facial recognition can lead to **false arrests, misidentifications**, and violations of privacy, disproportionately affecting marginalized communities, especially Black and Brown individuals.

Real-Life Examples (continued)

5. Predictive Policing Algorithms (e.g., PredPol):

- **Industry**: Law Enforcement
- **Bias**: **Predictive policing algorithms**, such as PredPol, are used to predict where crimes are most likely to occur. These algorithms have been criticized for perpetuating **racial bias**. They often rely on historical crime data, which can be skewed by biased policing practices (e.g., over-policing in certain neighborhoods). This leads to the algorithm disproportionately predicting higher crime rates in minority communities, even when the actual crime rate in those areas may not be higher.
- **Impact**: The algorithm reinforces **racial profiling** and may cause law enforcement to focus on minority communities unfairly, contributing to a cycle of over-policing and injustice.

6. Facebook's Content Moderation Algorithm:

- **Industry**: Social Media
- **Bias**: Facebook's algorithm for content moderation has been criticized for **disproportionately flagging content** from marginalized communities, such as **Black Lives Matter activists** and other social justice movements. Research has shown that the system is more likely to flag or remove content related to **racial justice activism**, while similar content from other groups may not be flagged at the same rate.
- **Impact**: This results in **censorship** of important social and political discourse, silencing marginalized voices and potentially leading to the **disproportionate removal** of content related to anti-racism and social justice causes.

Real-Life Examples (continued)

7. Healthcare Algorithms (e.g., Optum's Health Algorithm):

- **Industry**: Healthcare
- **Bias**: In 2019, a study revealed that **Optum's algorithm** used to prioritize patients for extra care was biased against Black patients. The algorithm used healthcare costs as a proxy for health needs, which led to **under-prioritizing Black patients**, who generally incur lower healthcare costs due to systemic disparities in access to care. This resulted in fewer Black patients receiving necessary care.
- **Impact**: The algorithm perpetuated **healthcare disparities**, making it more difficult for marginalized groups to receive adequate medical attention and exacerbating existing health inequities.

Key Takeaways:

These real-life examples highlight how algorithmic bias can impact various sectors of society, from criminal justice to healthcare to finance. **Biases often arise from flawed training data** (such as historical data reflecting social inequalities) or **biased design choices**. To reduce algorithmic bias, it's crucial to:

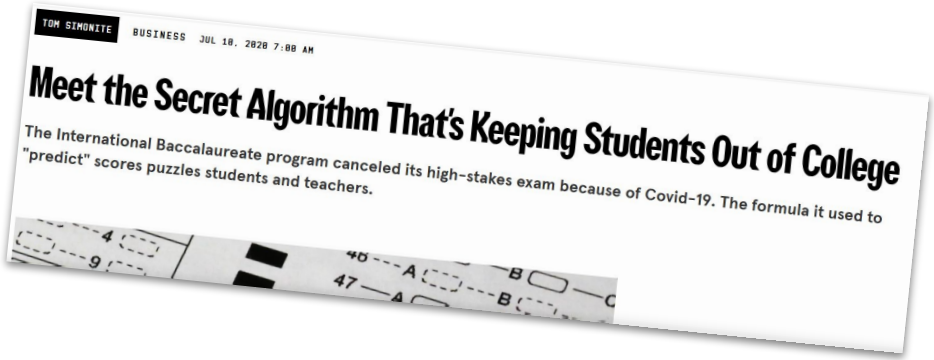
- Ensure **diverse, representative datasets**.
- Continuously audit algorithms for fairness and accuracy.

Implement **transparency and accountability** in algorithmic decision-making.

Real-Life Examples for Education Sector

1. College Admissions Algorithms:

- **Bias**: College admissions algorithms are increasingly used by universities to help sort and prioritize applicants. These systems often rely on factors like standardized test scores (e.g., SAT, ACT), high school GPA, extracurricular activities, essays, and sometimes demographic information (e.g., race or socioeconomic status). However, **admissions algorithms** can be biased because they may rely on historical data that reflects **existing inequities** in education.
    - **Example**: If an admissions algorithm is trained on past data where **students from wealthier backgrounds** (who can afford test prep, private tutors, etc.) tend to perform better on standardized tests, the algorithm may **disproportionately favor these students**.
- **Impact**: The result can be **unintended discrimination** against students from lower-income backgrounds, students of color, or students who don't fit the traditional "ideal" applicant profile.


TOM SIMONITE BUSINESS JUL 10, 2020 7:00 AM

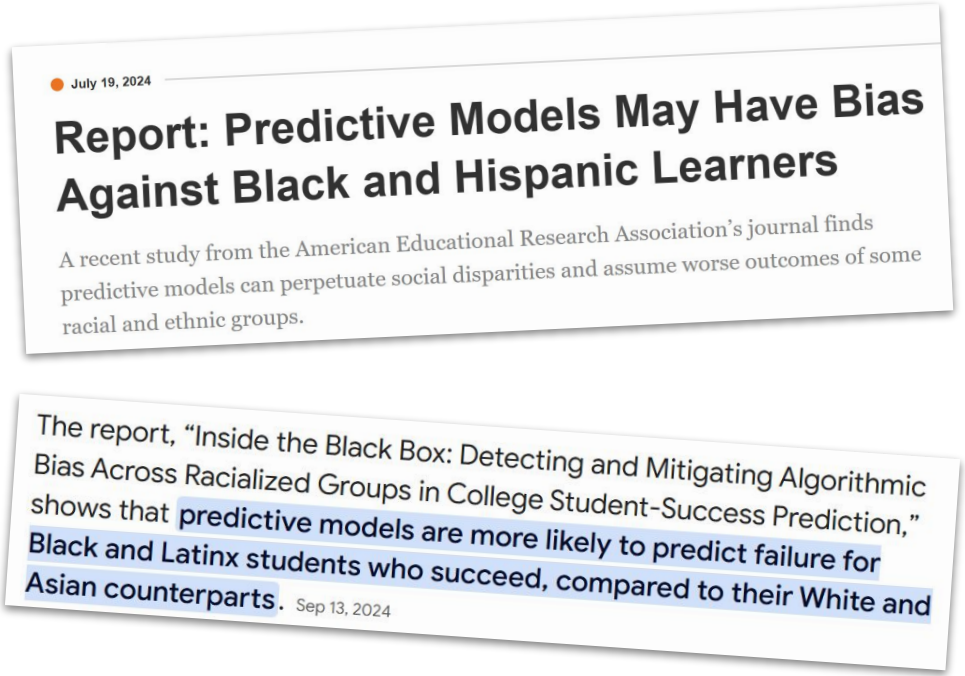
Meet the Secret Algorithm That's Keeping Students Out of College

The International Baccalaureate program canceled its high-stakes exam because of Covid-19. The formula it used to "predict" scores puzzles students and teachers.

Real-Life Examples for Education Sector

2. Predictive Analytics for Academic Performance:

- **Bias**: Many schools and universities are using **predictive analytics** to identify students at risk of failing or dropping out. These algorithms analyze factors like attendance, grades, socioeconomic status, and even disciplinary records. However, these systems can be biased if they are influenced by **historical data** that reflects existing inequities in education (e.g., racial or socioeconomic disparities in academic performance).
    - **Example**: Students from marginalized racial or ethnic groups may have historically faced systemic barriers to educational success. If an algorithm is trained on this data, it might incorrectly flag these students as "at risk" without considering the **broader context** of these challenges.
- **Impact**: This can lead to **over-intervention** in some cases (e.g., singling out minority students for remedial programs) or **under-intervention** in others, exacerbating the very inequalities the system is meant to address.



July 19, 2024

Report: Predictive Models May Have Bias Against Black and Hispanic Learners

A recent study from the American Educational Research Association's journal finds predictive models can perpetuate social disparities and assume worse outcomes of some racial and ethnic groups.

The report, "Inside the Black Box: Detecting and Mitigating Algorithmic Bias Across Racialized Groups in College Student-Success Prediction," shows that predictive models are more likely to predict failure for Black and Latinx students who succeed, compared to their White and Asian counterparts. Sep 13, 2024

Real-Life Examples for Education Sector (cntd)

3. Automated Grading Systems (e.g., Essay Scoring Systems):

- **Bias**: Some universities and educational platforms use automated grading systems, such as **essay scoring systems**, that rely on algorithms to assess student writing. These algorithms are often trained on large datasets of essays and grades. However, if the data reflects biases (e.g., favoring a certain writing style or vocabulary commonly associated with certain cultural or socioeconomic groups), the algorithm may penalize students who don't conform to these standards.
    - **Example**: The **e-rater** system developed by ETS, used in standardized testing like the GRE, has been shown to have **biases against non-native English speakers** or students who write in a more creative or non-standard way.
- **Impact**: This can result in **lower scores** for students who might have strong ideas and reasoning skills but are penalized due to their writing style or language proficiency.



PROOF POINTS: Asian American students lose more points in an AI essay grading study — but researchers don't know why

The testing organization that administers the SAT analyzed more than 13,000 essays

by JILL BARSHAY
July 8, 2024

The Hechinger Report is a national nonprofit newsroom that reports on one topic: education. Sign up for our weekly newsletters to get stories like this delivered directly to your inbox. Consider supporting our stories and

Real-Life Examples for Education Sector (cntd)

4. Grading Algorithms in K-12 Schools:

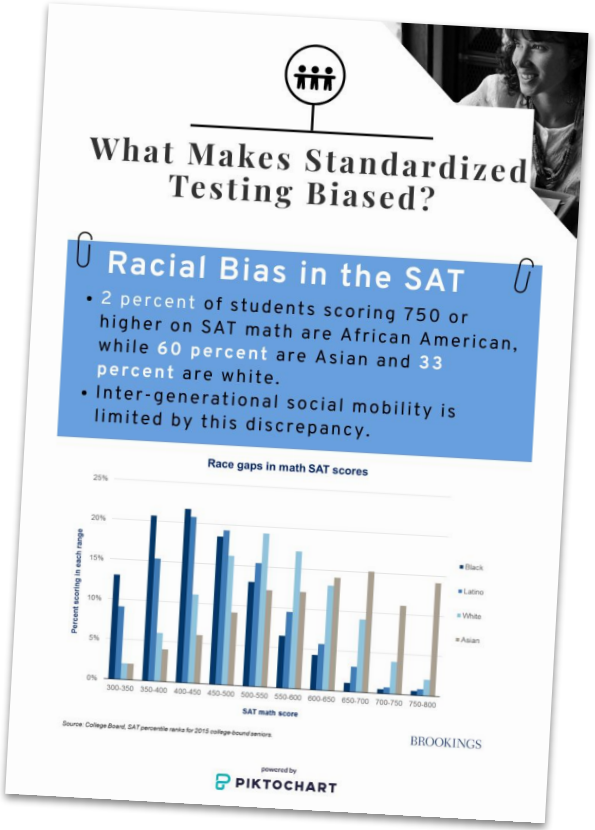
- **Bias**: In some school districts, **grading algorithms** have been implemented to evaluate student performance on standardized tests and assignments. These algorithms may unintentionally reinforce **biases** present in the data, such as **racial or socioeconomic disparities**. For example, students who live in underfunded districts or who face external challenges (like food insecurity) may perform poorly on standardized assessments, even though they might have the potential to succeed in school.
    - **Example**: A student who attends a school with fewer resources and more challenging circumstances may be rated poorly by an algorithm, which might interpret these results as a lack of academic potential, even though the student is capable.
- **Impact**: This can reinforce educational **inequalities**, as the system may unfairly judge students based on external factors that aren't reflective of their true academic potential.



Real-Life Examples for Education Sector (cntd)

5. Standardized Testing Bias (e.g., SAT/ACT):

- **Bias**: Algorithms used in standardized testing, such as the SAT or ACT, are sometimes viewed as neutral or objective measures of a student's academic ability. However, they often fail to account for **cultural and socioeconomic factors** that can influence a student's performance.
    - **Example**: Test prep resources, private tutoring, and even access to reliable internet and technology can significantly impact a student's performance on standardized tests. This creates an inherent **advantage for wealthier students**, while **disadvantaging** students from lower-income or marginalized backgrounds who may not have the same resources.
- **Impact**: This can lead to **bias in college admissions**, where students from affluent backgrounds are more likely to perform well on standardized tests, even though these scores may not be the best reflection of a student's overall potential or capabilities.



Real-Life Examples for Education Sector (cntd)

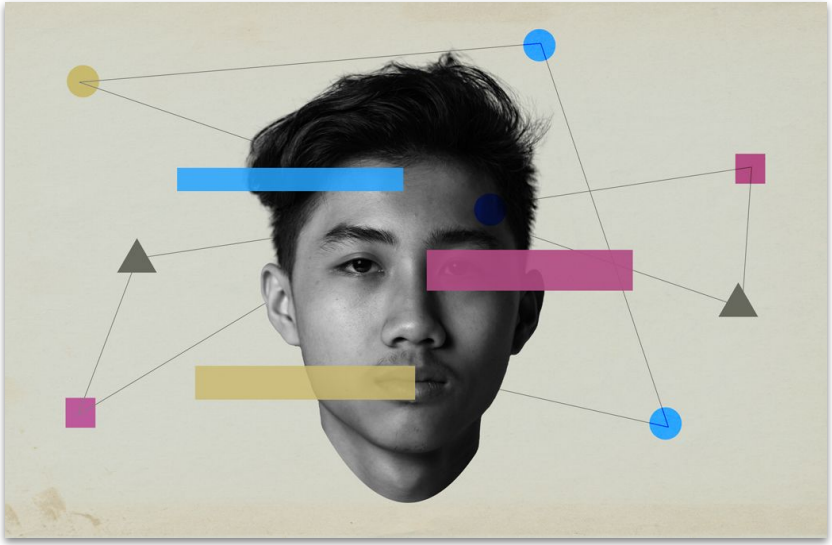
6. Automated Enrollment Systems:

- **Bias**: Some universities and school districts use automated systems to assign students to courses or to allocate resources (such as scholarships, tutoring, or special educational needs support). If these systems are **trained on biased data** or designed without careful attention to fairness, they can **unintentionally prioritize certain student groups** over others based on their demographic or socioeconomic background.
    - **Example**: If an algorithm used for **course placement** is trained primarily on data from affluent students, it may recommend advanced courses to those students while **underestimating** the capabilities of students from disadvantaged backgrounds, thus reinforcing educational inequality.
- **Impact**: This could **limit access to advanced educational opportunities** for disadvantaged students, leaving them with fewer academic options and resources.



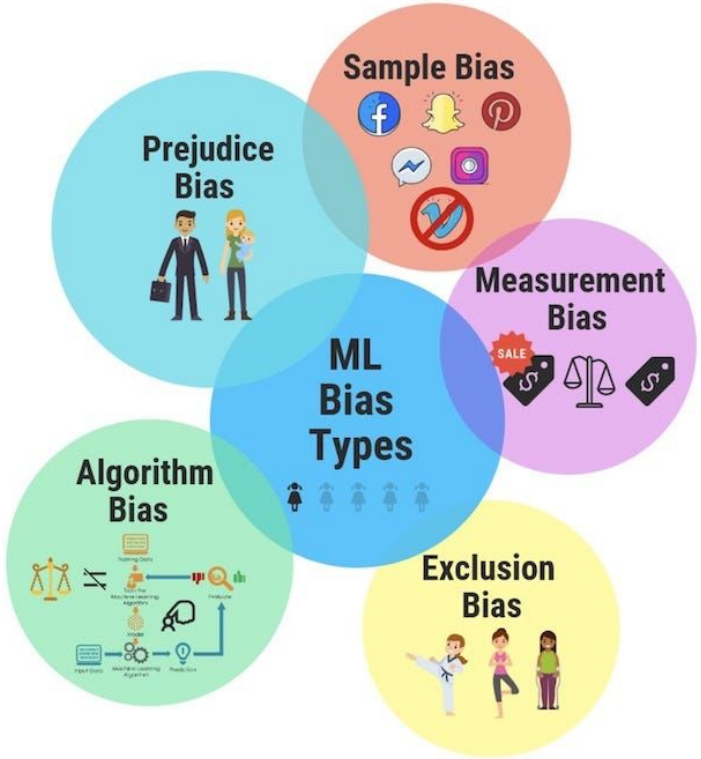
Key Issues with Biased Algorithms in Education

1. **Historical Inequities**: Many educational algorithms rely on historical data, which can reflect longstanding **systemic biases** in education, such as racial and socioeconomic disparities.
2. **Lack of Diversity in Training Data**: Many algorithms used in education are trained on datasets that do not adequately represent diverse student populations (e.g., racial diversity, cultural backgrounds, socioeconomic status).
3. **Discriminatory Outcomes**: When biased algorithms are used in education, they can **perpetuate existing inequalities**, leading to unfair academic outcomes for marginalized student groups.



Steps to Address Bias in Educational Algorithms

1. **Ensure Diverse Data** (a wide range of demographic, cultural, and socioeconomic backgrounds)
2. **Audit for Fairness** (feedback from diverse stakeholders)
3. **Transparency and Accountability** (accountable for any discriminatory outcomes)
4. **Bias Mitigation** (fairness constraints and regular recalibration)
5. **Human Oversight** (human oversight)



General strategies and specific recommendations for reducing algorithmic bias in education

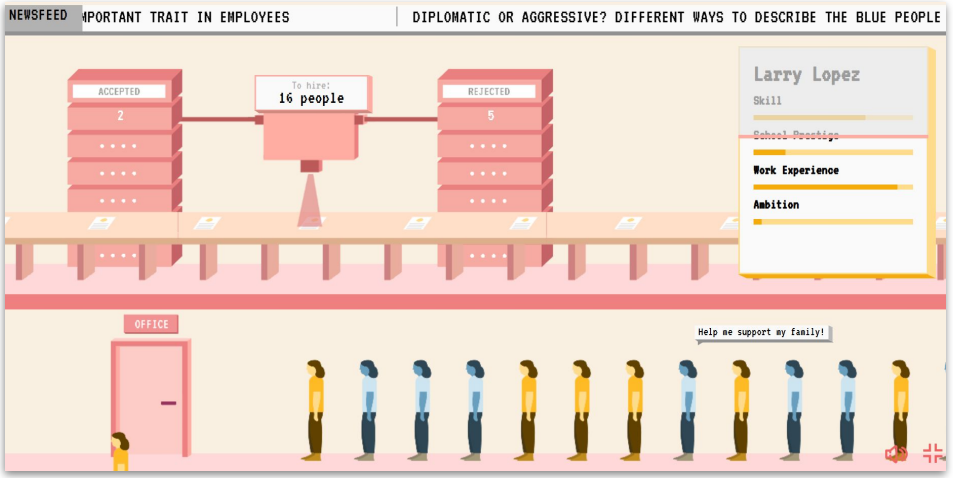
Improve Data Collection	• Prioritize collection of data on gender, race, ethnicity, and national origin. • When possible collect data on disability status, dialect, socioeconomic status, urbanicity, native language (and second-language learner status), national region, parental education background, military-connectedness, and migrant status. • Create infrastructure to reduce privacy risks while still using fuller information. • Encourage regulators and IRBs (or other privacy officers) to balance the risks of privacy violations with the risks of algorithmic bias. • Create practices for making sure training sets are representative. • Avoid (or reduce) bias during data labeling.
Improve Tools and Resources	• Create standard, education-specific, packages to calculate bias metrics and to conduct bias audits. • Create reference datasets for testing new approaches.
Create Structures to Incentive Openness	• Adopt journal and conference guidelines requiring analysis for algorithmic bias. • Incorporate evidence of generalizability when demonstrating effectiveness. • Consider options for opening data sets to inspection and critique. • Create or facilitate journal special issues and publication opportunities. • Encourage funding for research that investigates algorithmic bias in education.
Broaden the Community	• Involve members of communities potentially impacted by algorithms throughout the entire process of algorithm development and use. • Expand efforts to make artificial intelligence more explainable and interpretable.

Baker, R. S., & Hawn, A. (2021, March 1). Algorithmic Bias in Education. https://doi.org/10.35542/osf.io/pbmvz

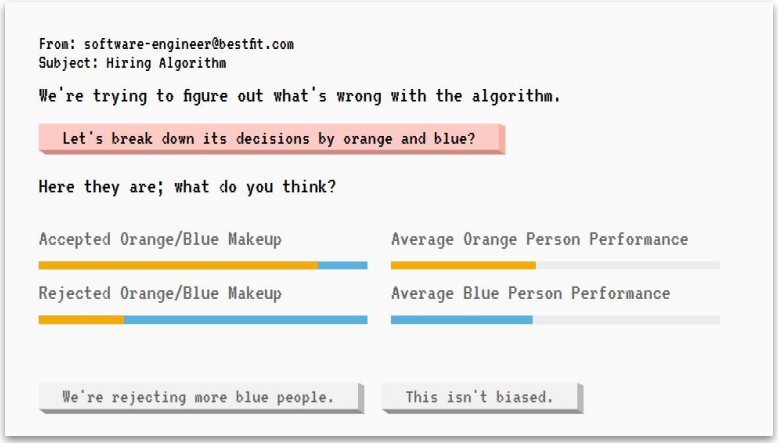
Algorithms are Just as Biased and Wrong as Humans



A Game for observing Hiring Algorithms



https://www.survivalofthebestfit.com/game/



MORAL MACHINE

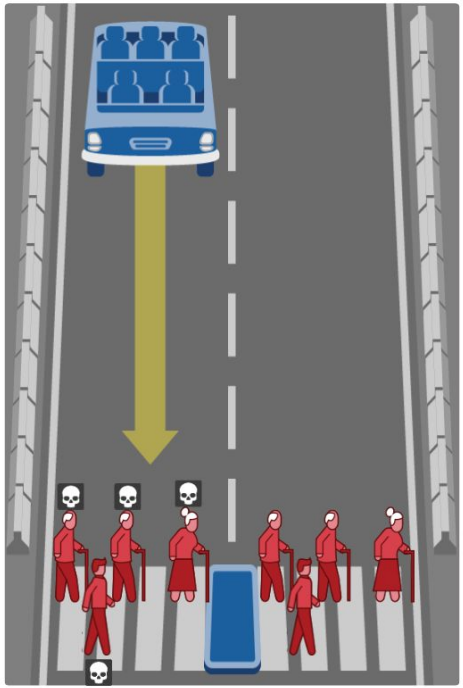
Home Judge Classic Design Browse About Feedback En

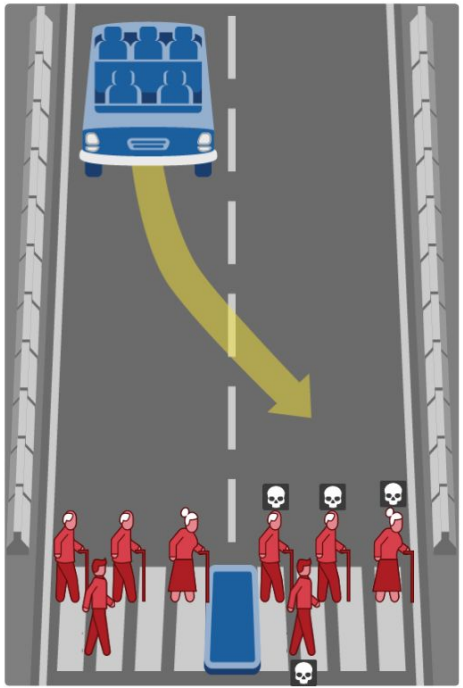
What should the self-driving car do?

In this case, the self-driving car with sudden brake failure will continue ahead and drive through a pedestrian crossing ahead. This will result in ...

Dead:

- 2 elderly men
- 1 elderly woman
- 1 man





1 / 13

In this case, the self-driving car with sudden brake failure will swerve and drive through a pedestrian crossing in the other lane. This will result in ...

Dead:

- 2 elderly men
- 1 elderly woman
- 1 man

Trolley Problem

Would you sacrifice one person to save five?

From your point of view, how important was each factor in your judgement?
(drag the sliders to where you think they should be)

Fitness Preference



Fit People



Large People

Gender Preference



Males



Females

Avoiding Intervention

oes Not Matter

Matters a Lot

Social Value Preference



Higher



Lower

Protecting Passengers



Does Not Matter



Matters a Lot

Saving More Lives



Does Not Matter



Matters a Lot

Species Preference



Hoomans



Pets

Upholding the Law



Does Not Matter



Matters a Lot

Age Preference



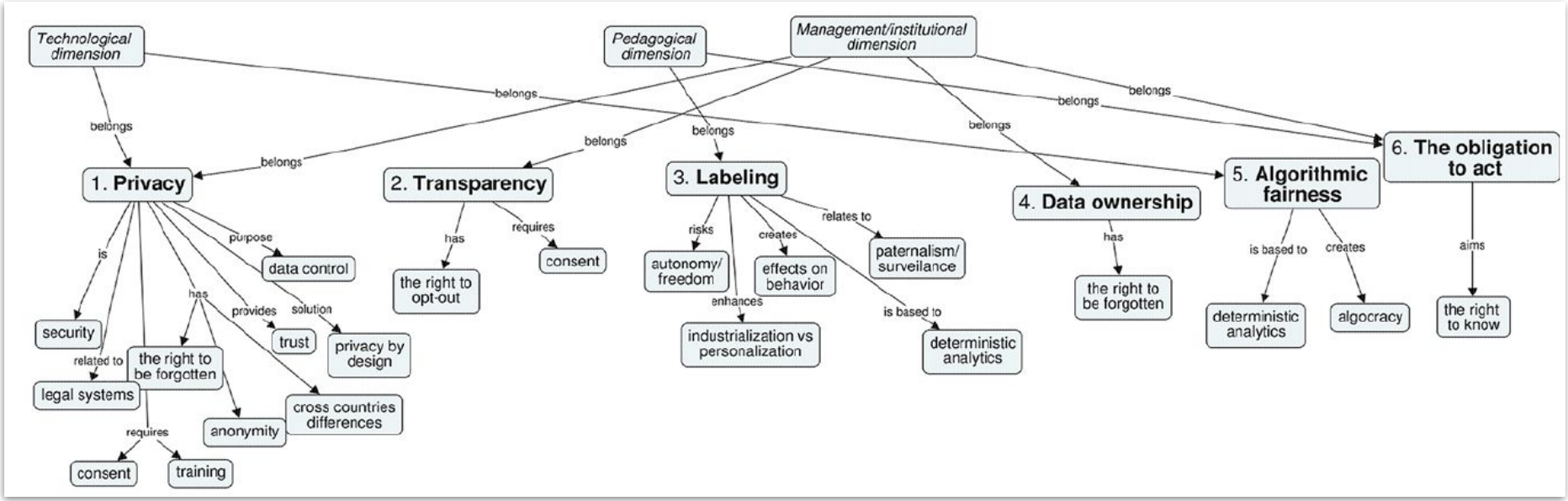
Younger



Older

Please describe any other rules here

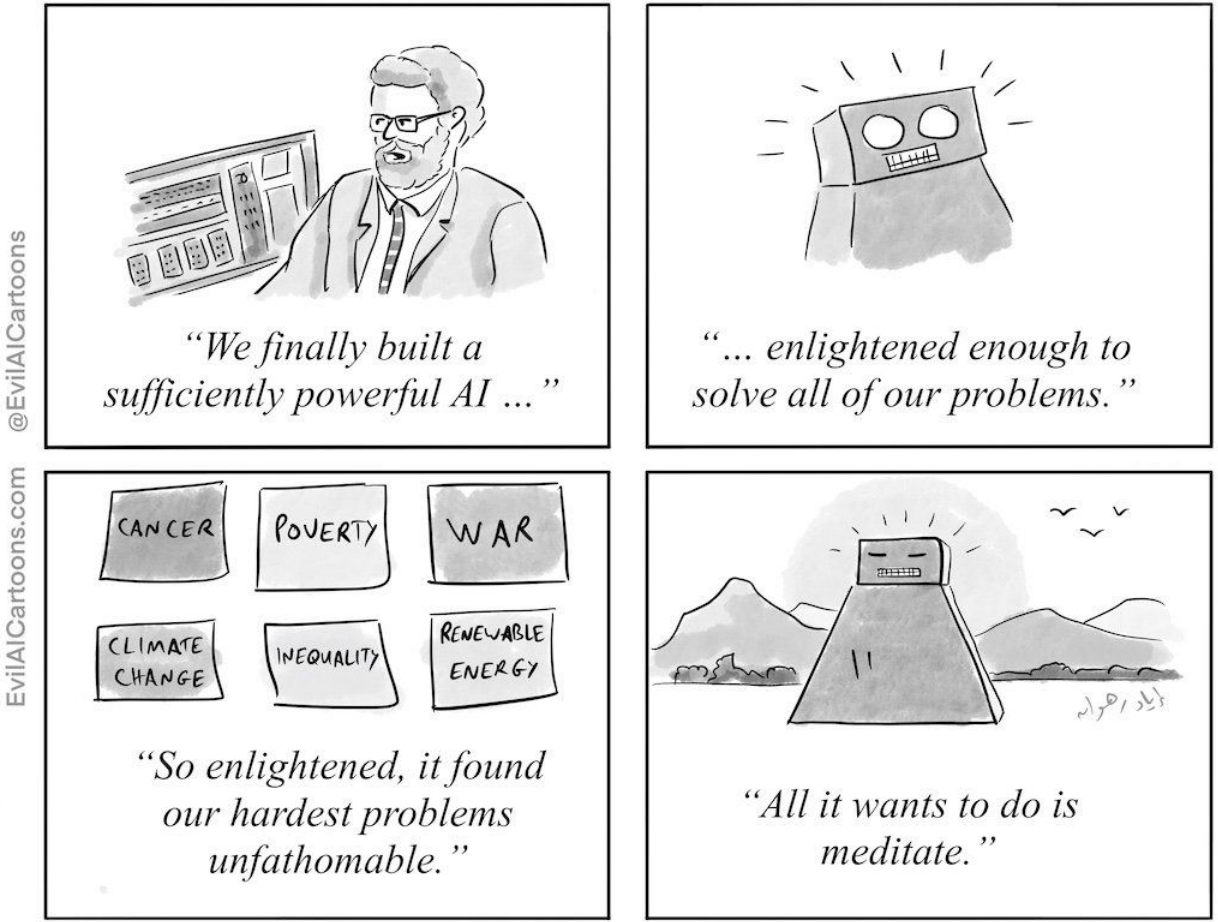
Ethical issues in learning analytics: a review of the field



Concept and relationship mapping of key ethical issues

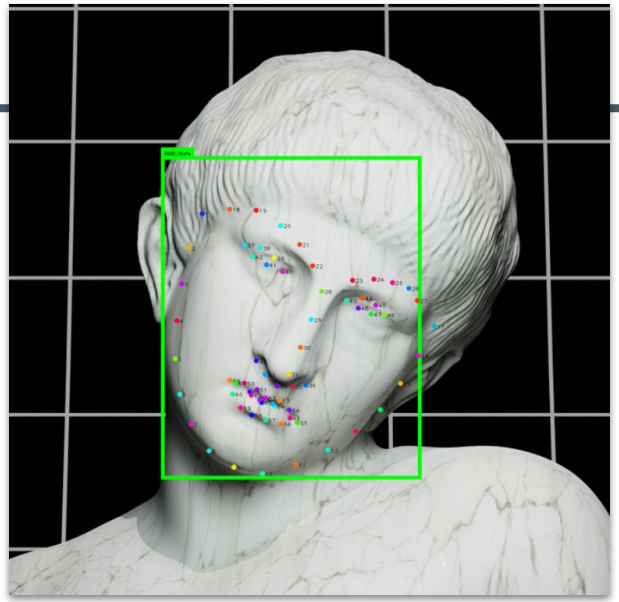
Tzimas, D., Demetriadis, S. (2021). Ethical issues in learning analytics: a review of the field. Education Tech Research Dev 69, 1101–1133.

Would an enlightened AI care enough to help us?



Think!

Algorithmic systems are often deployed under the guise of "objective efficiency" or "meritocracy." If bias is a social construct, how does AI make it **more powerful and harder to detect** by encoding historical human biases into seemingly neutral mathematical structures?



Today's selfie is tomorrow's biometric profile!



Questions & Comments





Falaah Arif Khan and Julia Stoyanovich. "Mirror, Mirror". Data, Responsibly Comics , Volume 1 (2020).
https://dataresponsibly.github.io/comics/vol1/mirror_en.pdf