Responsible Data Science

Algorithmic Fairness II

February 6, 2025

Prof. Jonathan Colner

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

NYU TANDON SCHOOL OF ENGINEERING

NYU Center for Data Science

r/ai

The evils of discrimination

Disparate treatment

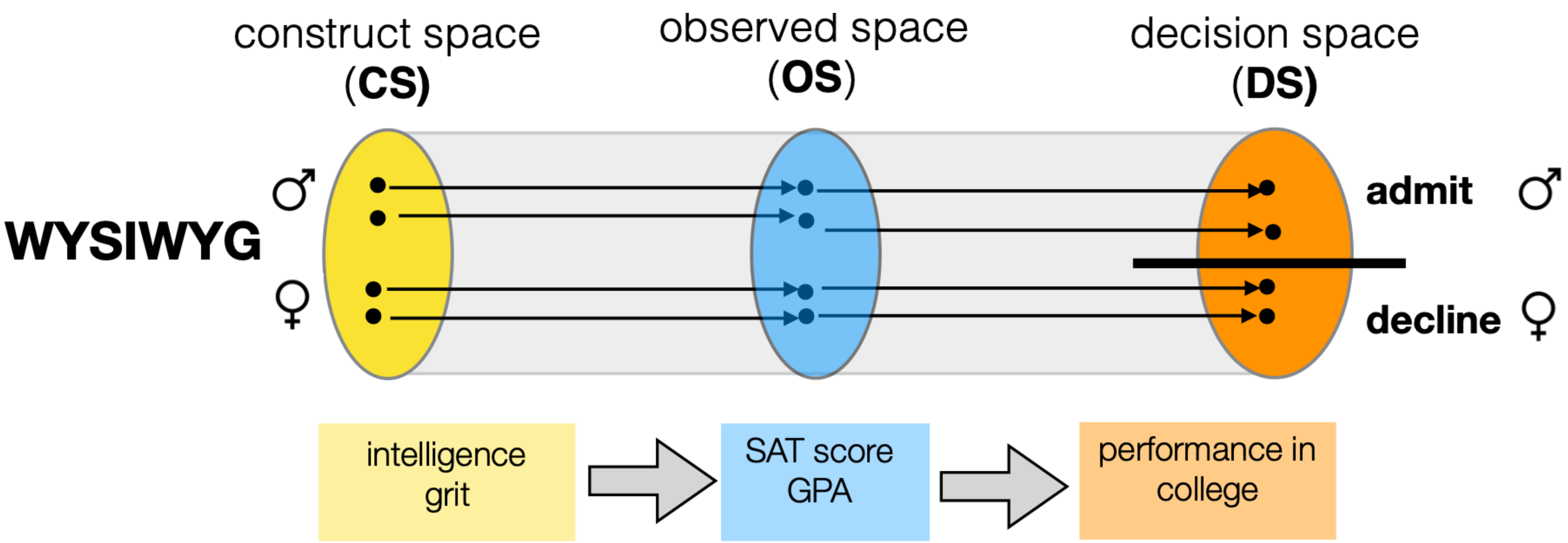
is the illegal practice of treating an entity, such as a job applicant or an employee, differently based on a **protected characteristic** such as race, gender, age, disability status, religion, sexual orientation, or national origin.

Disparate impact

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

WYSWYG

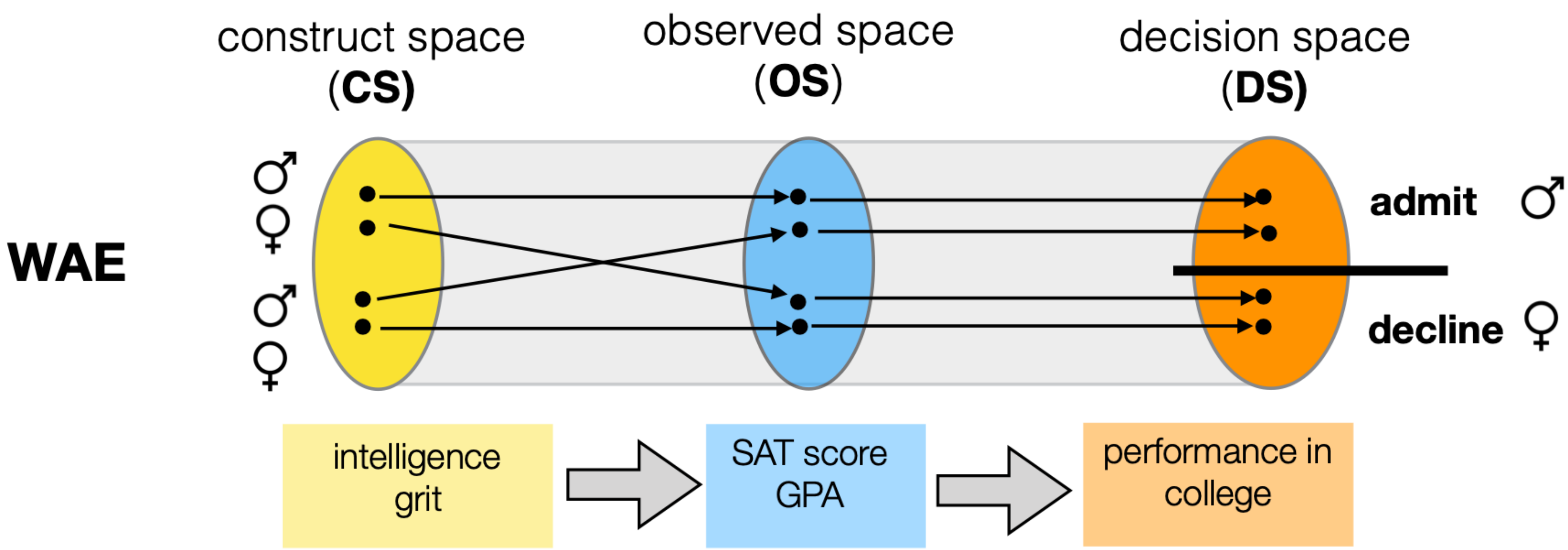
[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



What you see is what you get (**WYSIWYG**): there exists a mapping from CS to OS that has low distortion. That is, we believe that OS faithfully represents CS. **This is the individual fairness world view.**

WAE

[S. Friedler, C. Scheidegger and S. Venkatasubramanian, arXiv:1609.07236v1 (2016)]



We are all equal (**WAE**): the mapping from CS to OS introduces **structural bias** - there is a distortion that aligns with the group structure of CS. **This is the group fairness world view.**

Structural bias examples: SAT verbal questions function differently in the African-American and in the Caucasian subgroups in the US. Other examples?

What's the right answer?

There is no single answer!

Need transparency and public debate

- Consider harms and benefits to different stakeholders
- Being transparent about which fairness criteria we use, how we trade them off
- Recall "Learning Fair Representations": a typical ML approach

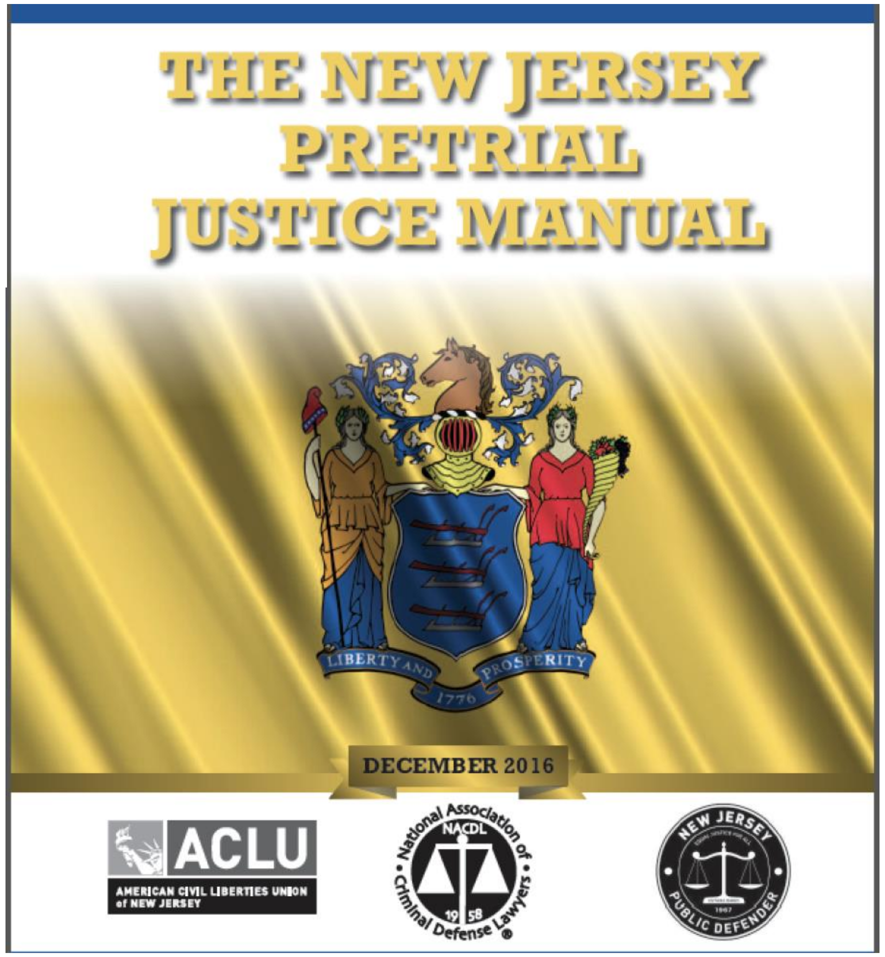
$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness **individual fairness** **utility**

apples + oranges + fairness = ?

fairness in risk assessment

New Jersey bail reform



Switching from a system based solely on instinct and experience […] to one in which judges have access to **scientific, objective risk assessment** tools could further the criminal justice system's central goals of increasing public safety, reducing crime, and making the most effective, fair, and efficient use of public resources.

https://www.nacdl.org/getattachment/50e0c53b-6641-4a79-8b49-c733def39e37/the-new-jersey-pretrial-justice-manual.pdf

ProPublica's COMPAS study

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

Bernard Parker, left, was rated high risk; Dylan Fugett was rated low risk. (Josh Ritchie for ProPublica)

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

Back to ProPublica's COMPAS study

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. COMPAS has been used by the U.S. states of NY, WI, CA, FL and other jurisdictions.

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

Similar tools are used today

The First Step Act's Risk Assessment Tool

Who is eligible for early release from federal prison?

April 2021

The First Step Act offers people incarcerated in **federal prison** the opportunity to earn credits toward early release. To help determine who is eligible (after excluding people with certain prior offenses), the US Department of Justice created the Prisoner Assessment Tool Targeting Estimated Risk and Needs (PATTERN), a risk assessment tool that predicts the likelihood that a person who is incarcerated will reoffend.

https://apps.urban.org/features/risk-assessment/

These tools are used today

The First Step Act's Risk Assessment Tool

April 2021

Who is eligible for early release from federal prison?

Features

	General		Violent	
Risk category	**Men**	**Women**	**Men**	**Women**
Minimum	-23 to 8	-24 to 5	-11 to 6	-11 to 2
Low	9 to 30	6 to 31	7 to 24	3 to 19
Medium	31 to 43	32 to 49	25 to 30	20 to 25
High	44 to 113	50 to 102	31 to 71	26 to 33

https://apps.urban.org/features/risk-assessment/

These tools are used today

LAW

Flaws plague a tool meant to help low-risk federal prisoners win early release

January 26, 2022 · 5:00 AM ET
Heard on Morning Edition

CARRIE JOHNSON

January 2022

Thousands of people are leaving federal prison this month thanks to a law called the First Step Act, which allowed them to win early release by participating in programs aimed at easing their return to society. But thousands of others may still remain behind bars because of fundamental flaws in the Justice Department's method for deciding who can take the early-release track. The biggest flaw: **persistent racial disparities that put Black and brown people at a disadvantage**.

[…] The algorithm, known as **Pattern**, **overpredicted the risk that many Black, Hispanic and Asian people** would commit new crimes or violate rules after leaving prison. At the same time, it also **underpredicted the risk for some inmates of color when it came to possible return to violent crime**.

https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

These tools are used today

LAW

Flaws plague a tool meant to help low-risk federal prisoners win early release

January 26, 2022 · 5:00 AM ET
Heard on Morning Edition

CARRIE JOHNSON

January 2022

Aamra Ahmad, senior policy counsel at the American Civil Liberties Union: ”The Justice Department found that **only 7% of Black people in the sample were classified as minimum level risk compared to 21% of white people**," she added. "This indicator alone should give the Department of Justice great pause in moving forward.”

Risk assessment tools are common in many states. But critics said Pattern is the first time the federal justice system is using an algorithm with such high stakes.

"**Especially when systems are high risk and affect people's liberty, we need much clearer and stronger oversight**," said Costanza-Chock [director of research & design for the Algorithmic Justice League]

https://www.npr.org/2022/01/26/1075509175/justice-department-algorithm-first-step-act

Fairness in risk assessment

- A risk assessment tool **gives a probability estimate of a future outcome**
- Used in many domains:
  - insurance, criminal sentencing, medical testing, hiring, banking
  - also in less-obvious set-ups, like online advertising
- Fairness in risk assessment is concerned with how different kinds of error are distributed among sub-populations

Calibration

positive outcomes: do recidivate

	risk score		
	0.2	**0.6**	**0.8**
White	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕
Black	⊖ ⊖ ⊖ ⊖ ⊕	⊖ ⊖ ⊖ ⊖ ⊕ ⊕ ⊕ ⊕ ⊕ ⊕	⊖ ⊕ ⊕ ⊕ ⊕

given the output of a risk tool, likelihood of belonging to the positive class is independent of group membership

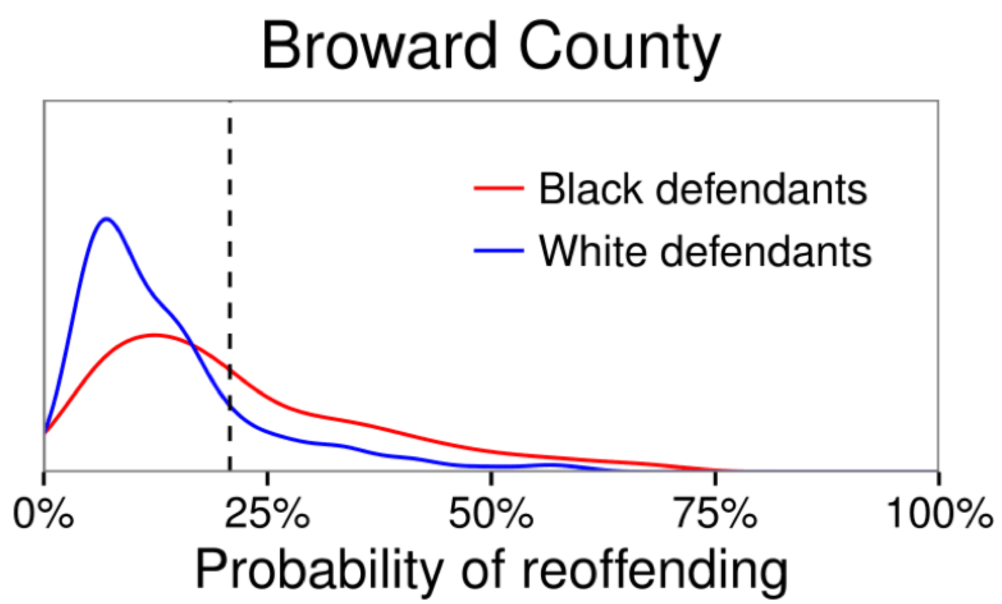
0.6 means 0.6 for any defendant - likelihood of recidivism

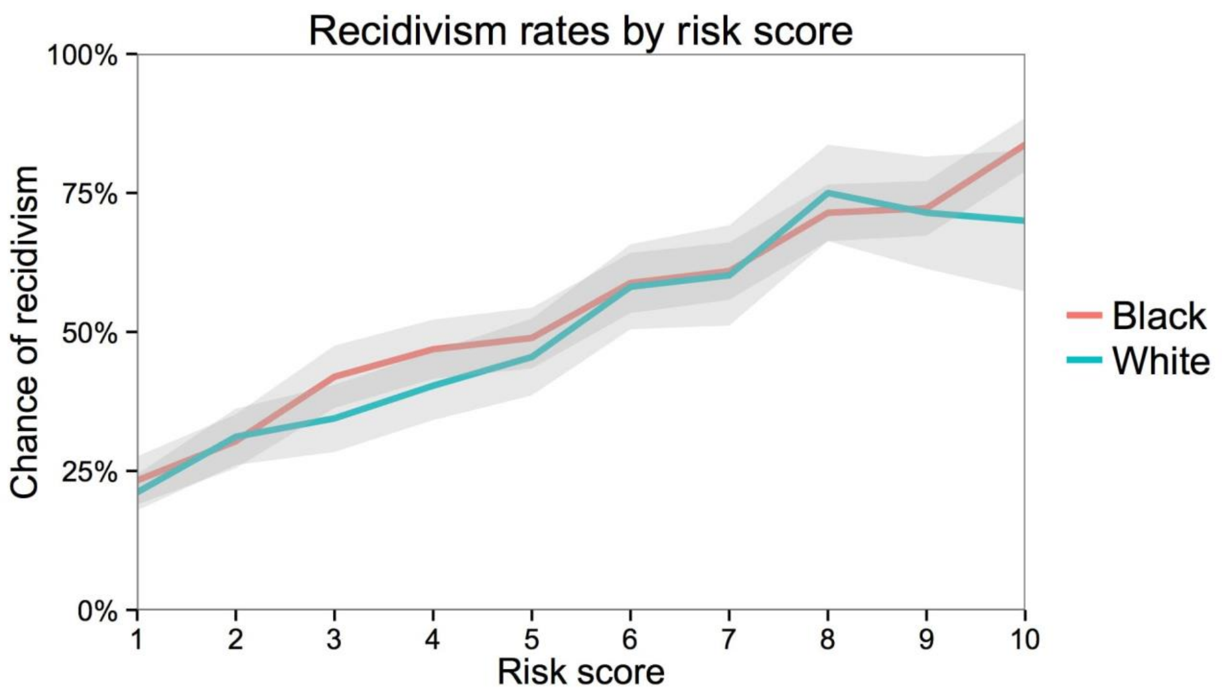
why do we want calibration?

COMPAS as a predictive instrument

Predictive parity (also called **calibration**)
an instrument identifies a set of instances as having probability *x* of constituting positive instances, then approximately an *x* fraction of this set are indeed positive instances, over-all and in sub-populations

COMPAS is well-calibrated: in the window around 40%, the fraction of defendants who were re-arrested is ~40%, both over-all and per group.





[plot from Corbett-Davies et al.; *KDD 2017*]

An impossibility result

If a predictive instrument **satisfies predictive parity**, but the **prevalence** of the phenomenon **differs between groups**, then the instrument **cannot achieve** equal false positive rates and equal false negative rates across these groups.

Recidivism rates in the ProPublica dataset are higher for the Black group than for the White group

k 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.)*

[A. Chouldechova; arXiv:1610.07524v1 *(2017)*]

A more general statement: Balance

- Balance for the positive class: Positive instances are those who go on to re-offend. The average score of positive instances should be the same across groups.
- Balance for the negative class: Negative instances are those who do not go on to re-offend. The average score of negative instances should be the same across groups.
- Generalization of: Both groups should have equal false positive rates and equal false negative rates.
- Different from statistical parity!

the chance of making a mistake does not depend on race

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Desiderata, re-stated

- For each group, a $\mathbf{v_b}$ fraction in each bin **b** is positive
- Average score of positive class same across groups
- Average score of negative class same across groups

can we have all these properties?

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Achievable only in trivial cases

- Perfect information: the tool knows who recidivates (score 1) and who does not (score 0)
- Equal base rates: the fraction of positive-class people is the same for both groups

a negative result, need tradeoffs

proof sketched out in (starts 12 min in)

https://www.youtube.com/watch?v=UUC8tMNxwV8

[J. Kleinberg, S. Mullainathan, M. Raghavan; *ITCS 2017*]

Fairness for whom?

Decision-maker: of those labeled low-risk, how many will recidivate?

Defendant: how likely will I be incorrectly labeled high-risk?

	labeled low-risk	labeled high-risk
did not recidivate	TN	**FP**
recidivated	**FN**	TP

based on a slide by Arvind Narayanan

What's the right answer?

There is no single answer!

Need transparency and public debate

- Consider harms and benefits to different stakeholders
- Being transparent about which fairness criteria we use, how we trade them off
- Recall "Learning Fair Representations": a typical ML approach

$$L = A_z \cdot L_z + A_x \cdot L_x + A_y \cdot L_y$$

group fairness

individual fairness

utility

apples + oranges + fairness = ?

Racial bias in healthcare

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

October 2019

Ziad Obermeyer[1,2,*], **Brian Powers**[3], **Christine Vogeli**[4], **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**. Remedying 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.

r/ai

Racial bias in healthcare

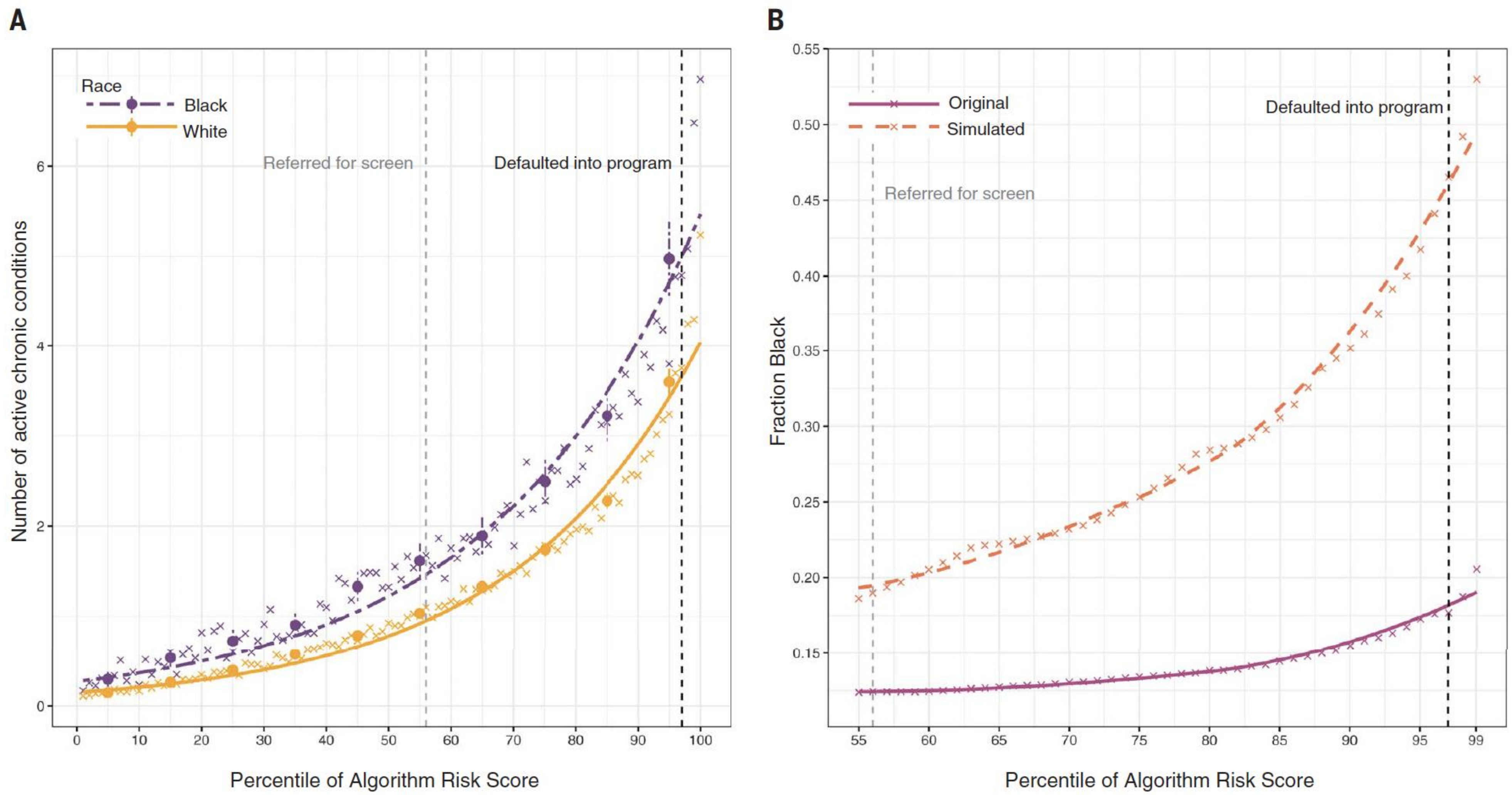


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario 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 × 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).

https://www.science.org/doi/10.1126/science.aax2342

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

ECONOMIC VIEW

Dec. 6, 2019

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

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.

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

Fixing bias in algorithms?

The New York Times

By Sendhil Mullainathan

ECONOMIC VIEW

Dec. 6, 2019

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

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**.

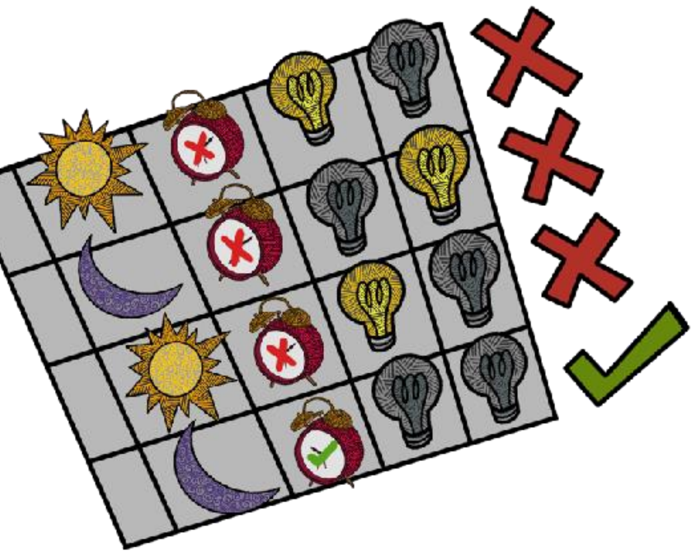
But the key phrase here is "proper regulation," which we do not currently have. We must ensure all the necessary inputs to the algorithm, including the data used to test and create it, are carefully stored. * […] We will need a well-funded regulatory agency with highly trained auditors to process this data.

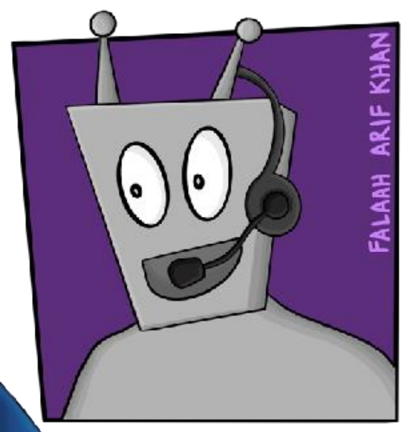
https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html

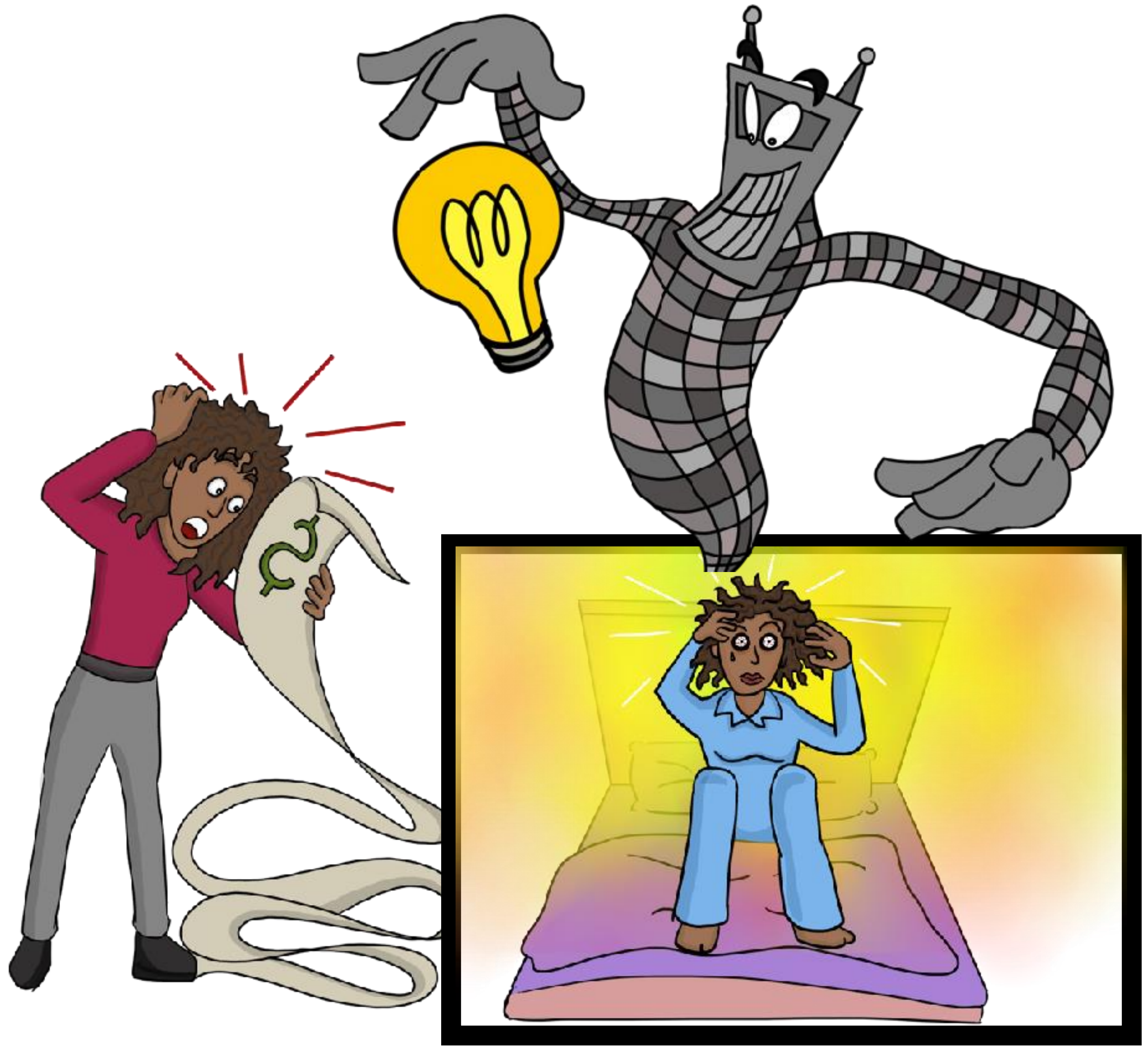
AI ethics teaser

This week's reading

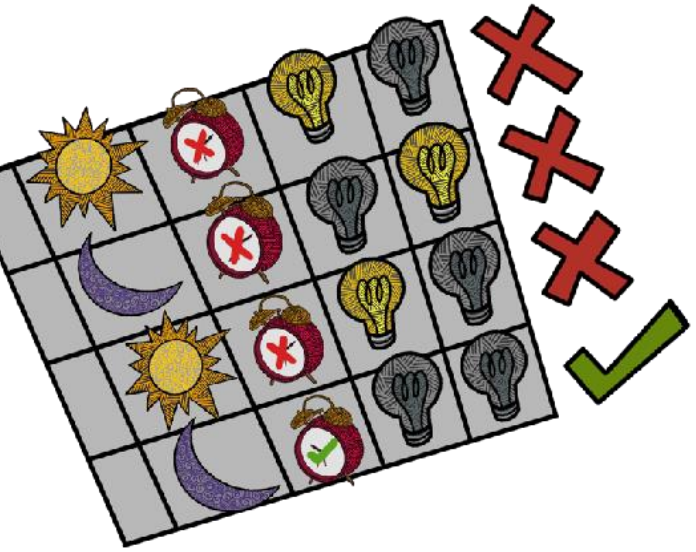
Mistakes lead to harms

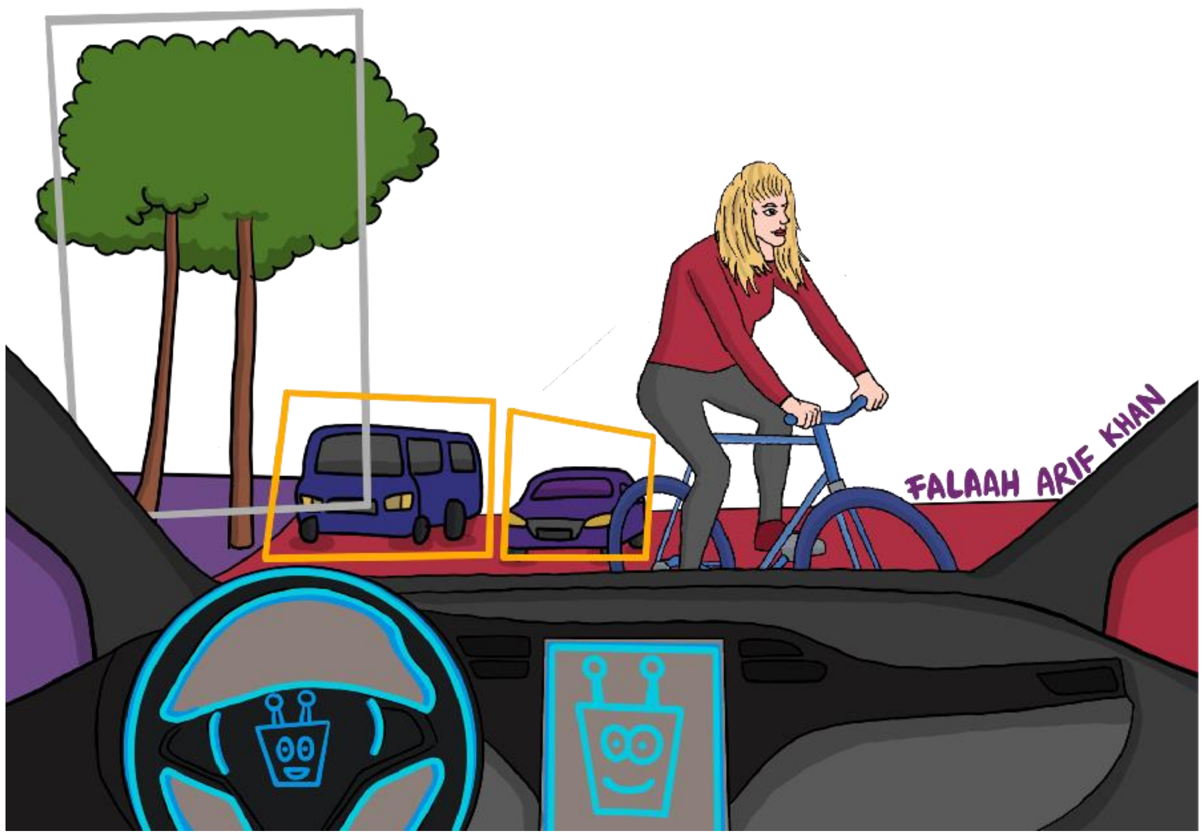


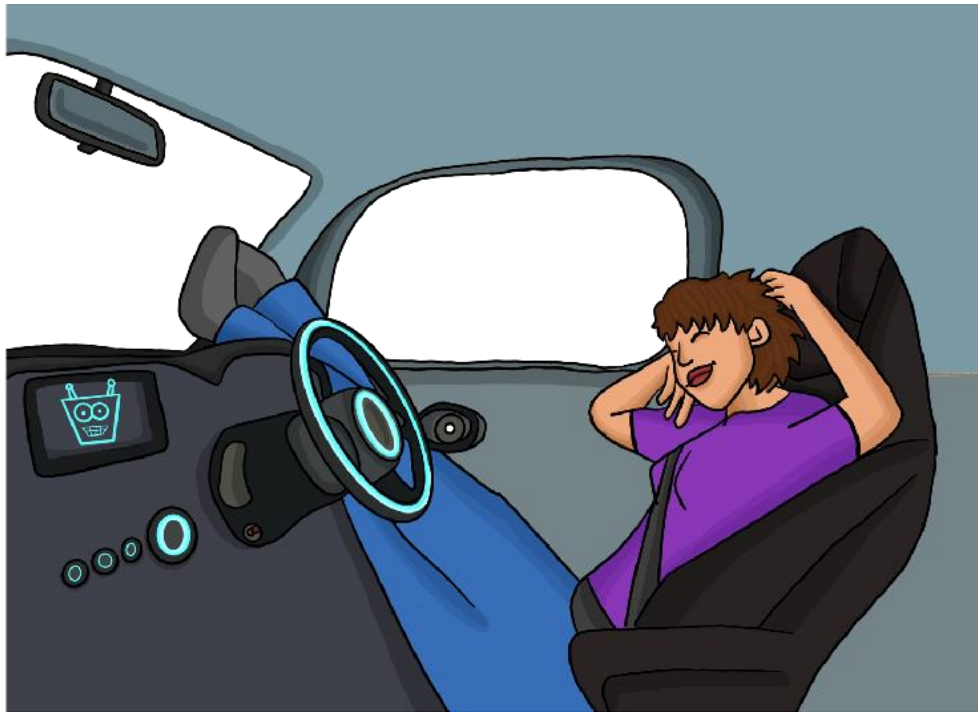




Mistakes lead to harms

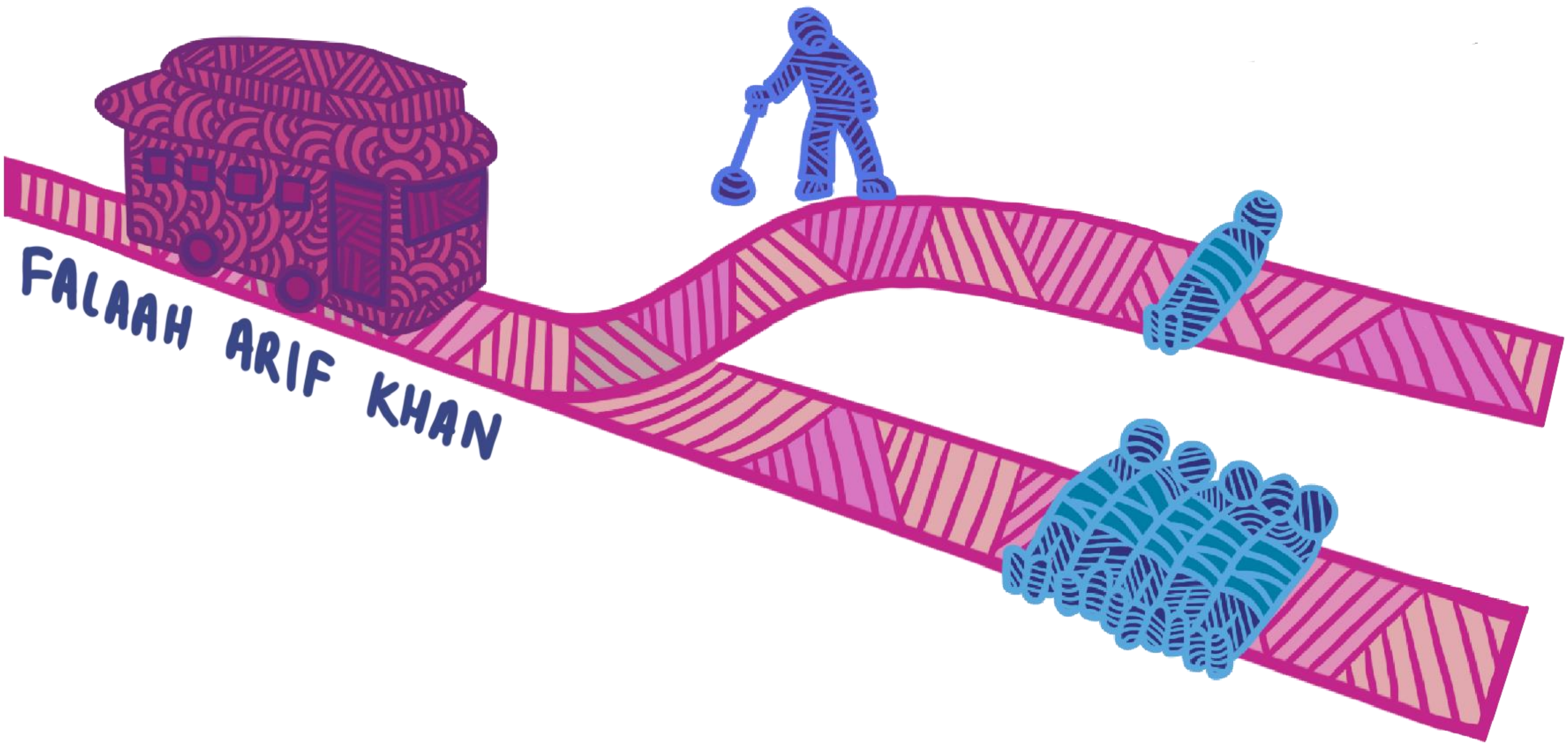






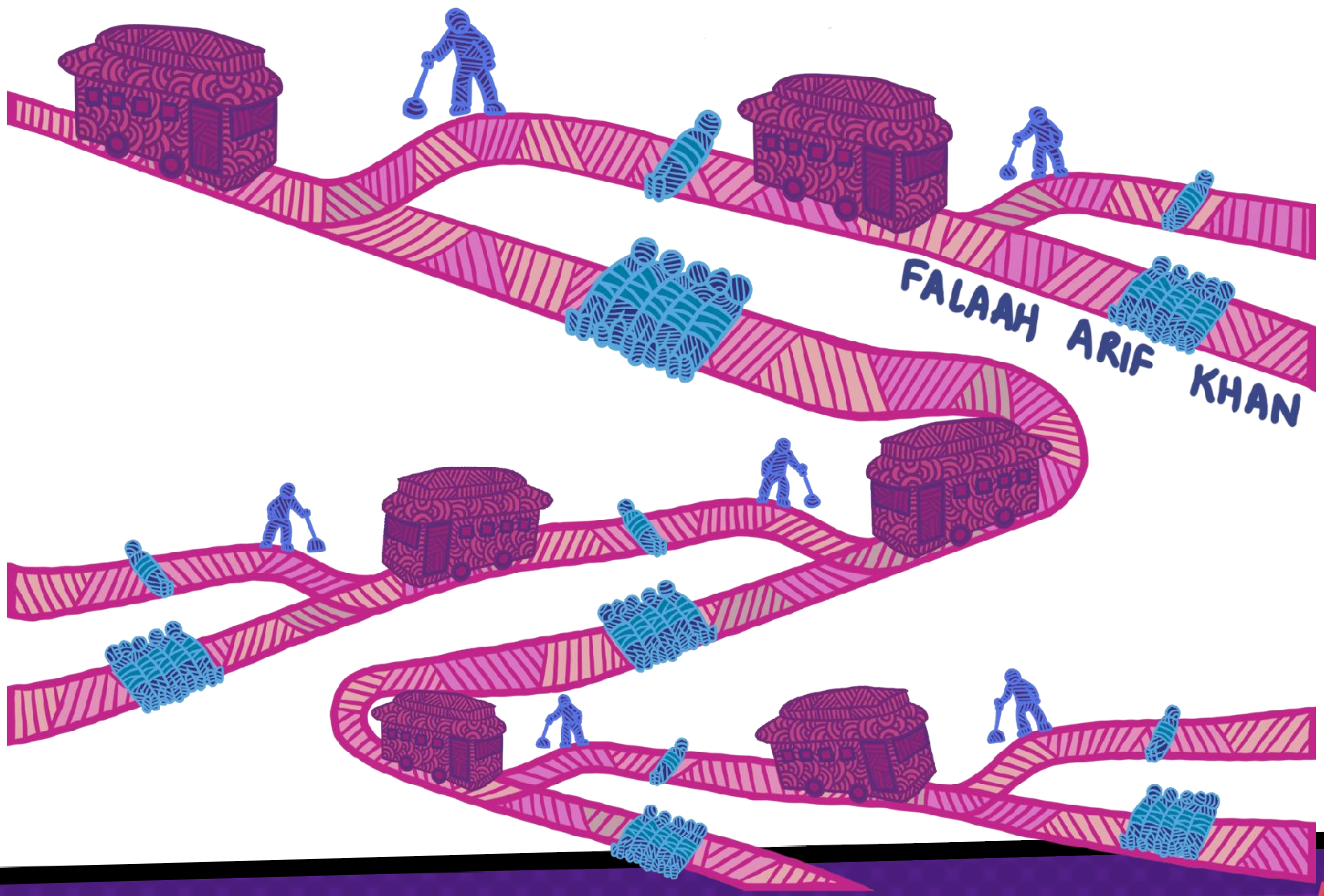
r/ai

The trolley problem



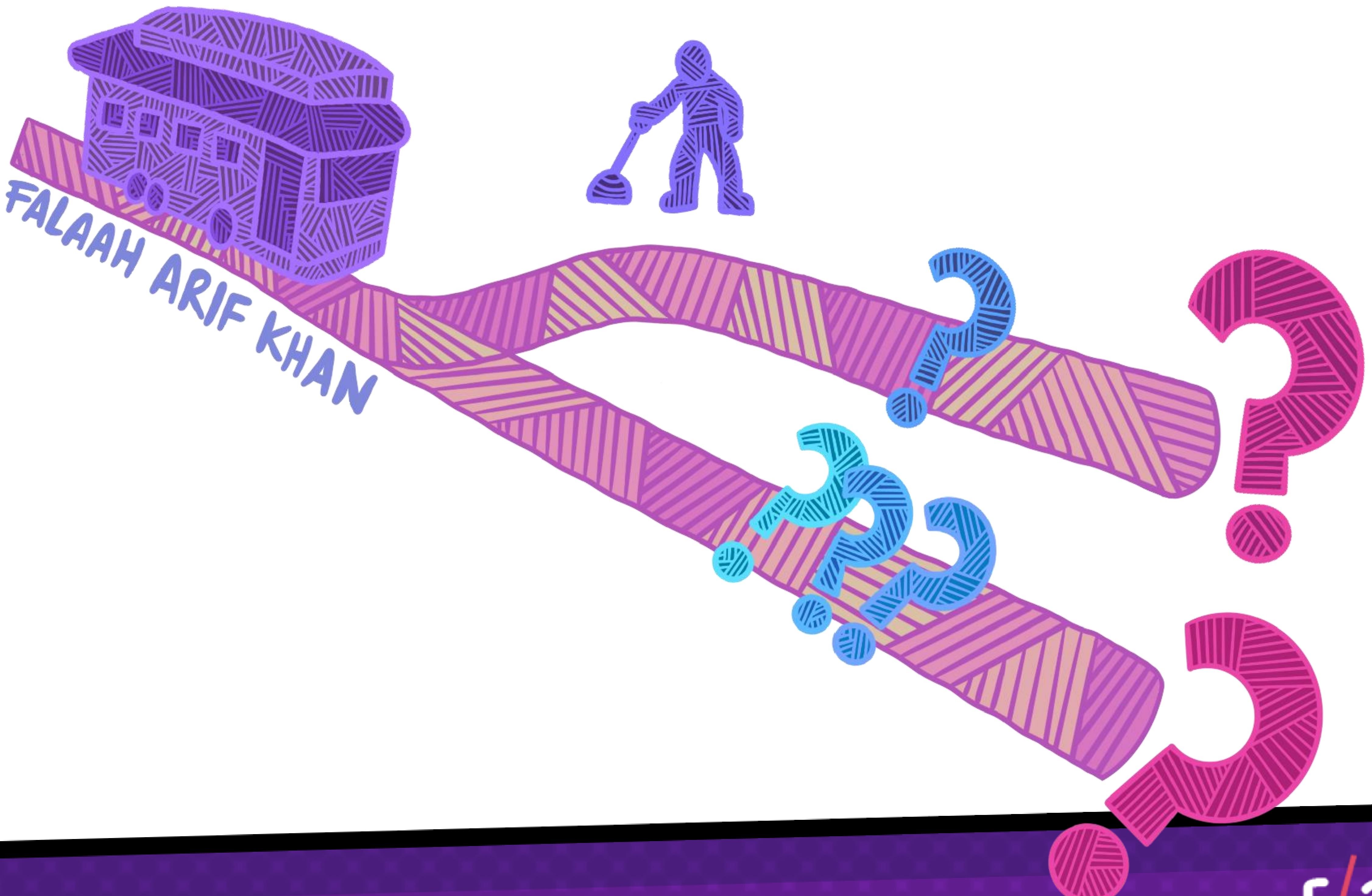
r/ai

The trolley problem



r/ai

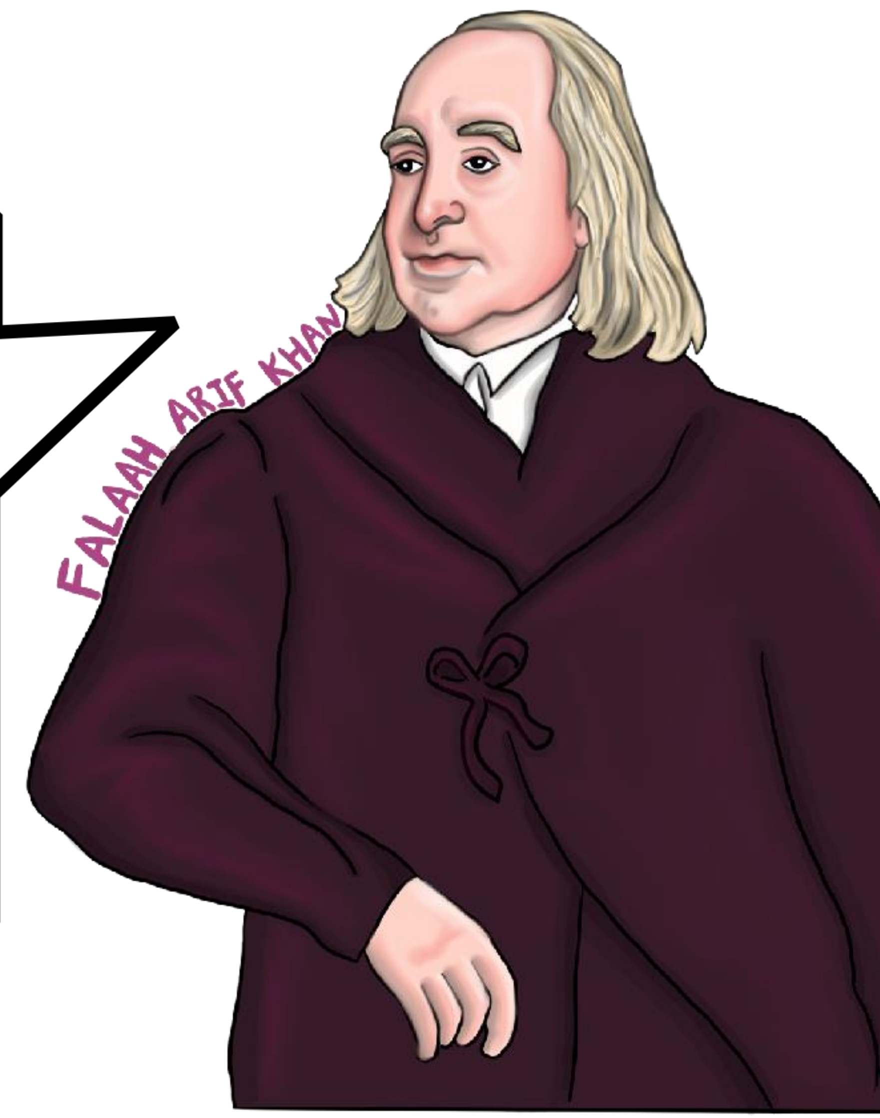
Dealing with uncertainty



Utilitarianism

"It is the greatest happiness of the greatest number that is the measure of right and wrong."

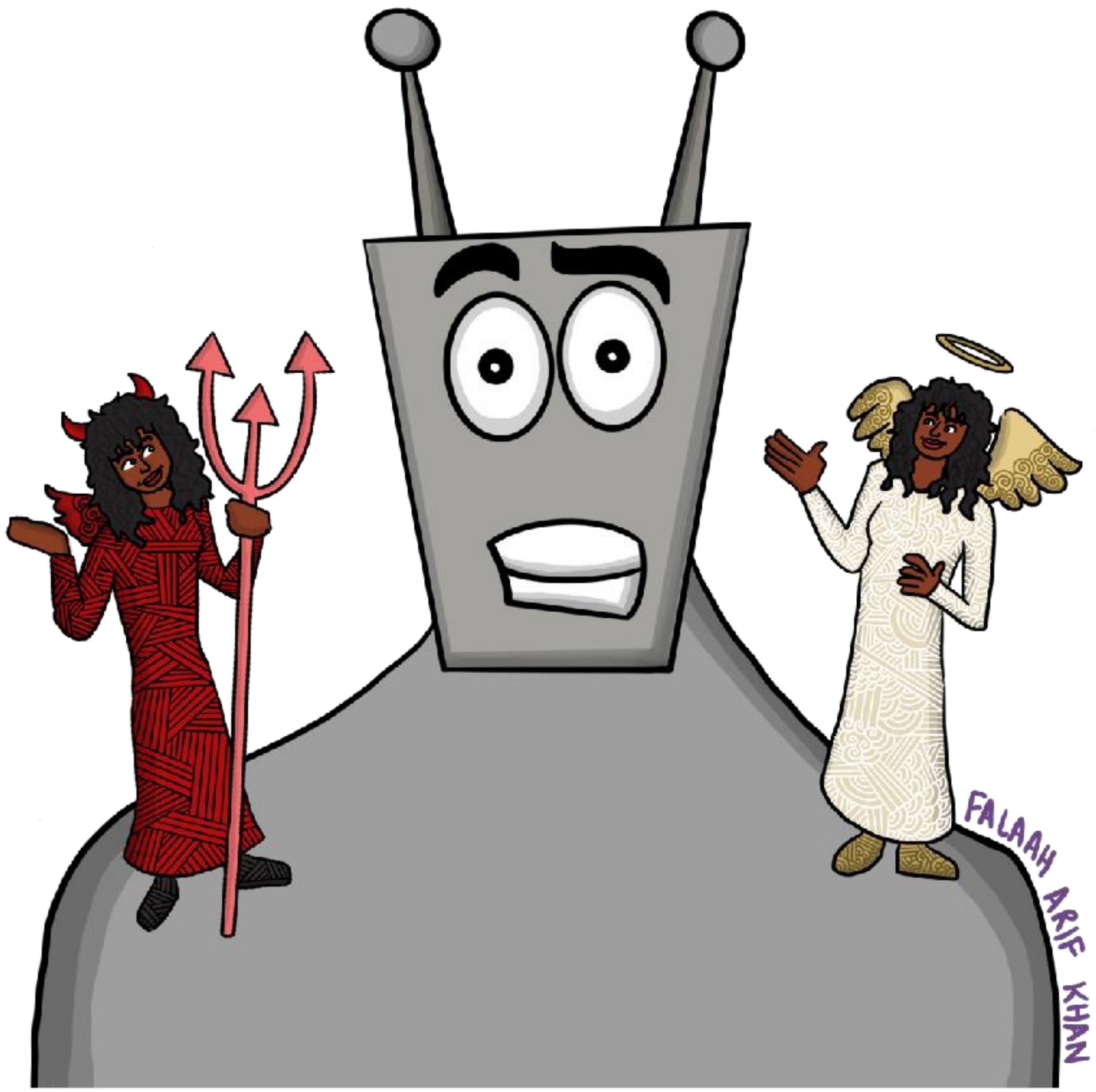
Jeremy Bentham



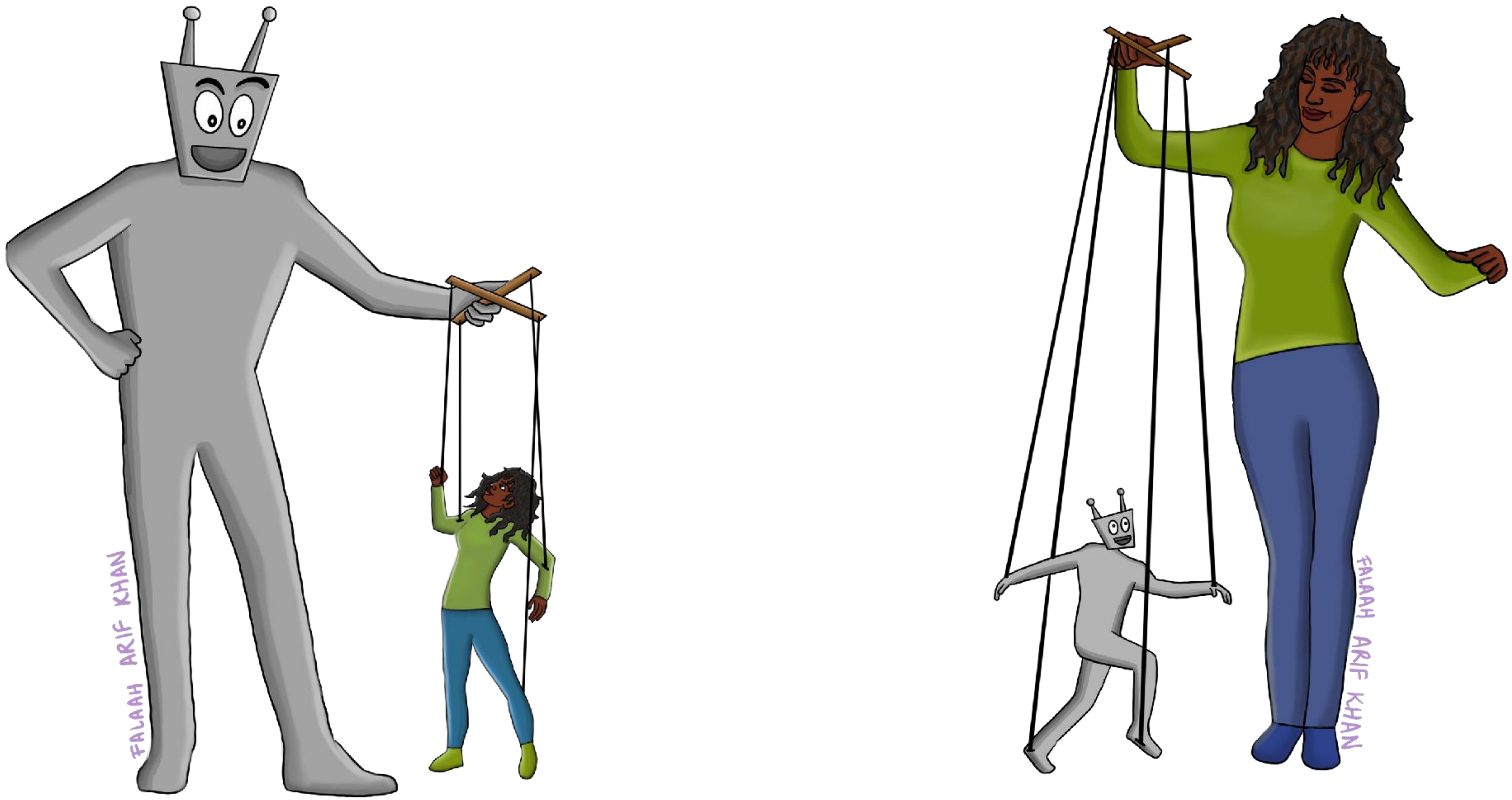
Algorithmic morality?

Algorithmic morality

is the act of attributing moral reasoning to algorithmic systems



Algorithmic morality?



r/ai

Responsible Data Science

Algorithmic Fairness

Thank you!