Impact of AI Decal:
Bias in AI

Gokul Swamy & Brenton Chu

Quiz:
https://tinyurl.com/impactsp19q8

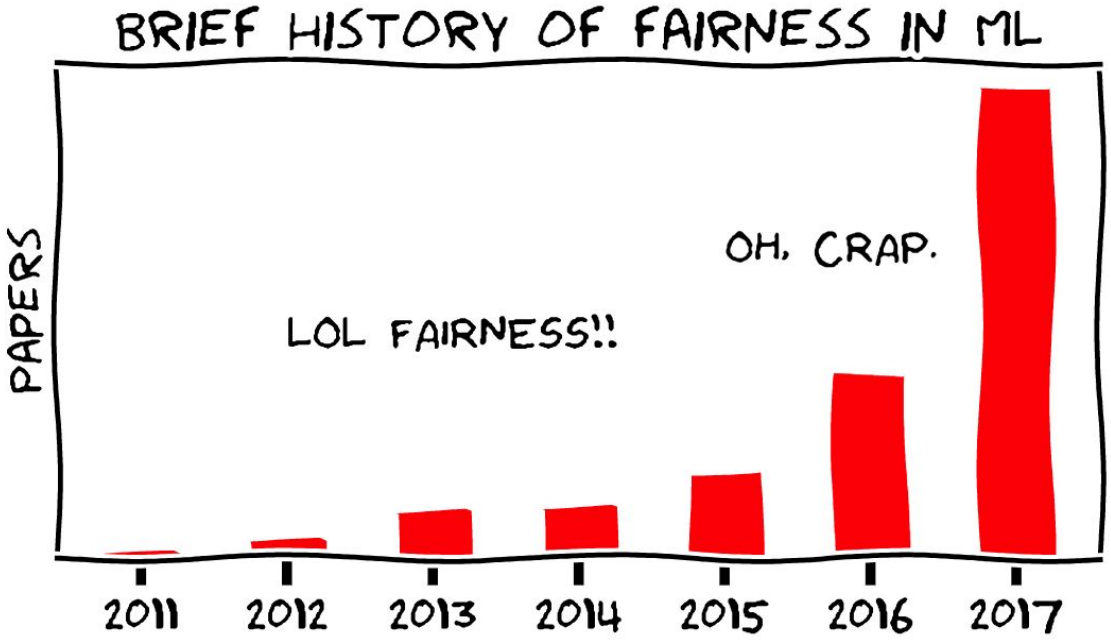
Case Study: Policing

- ➔ Imagine that you are a police department with limited resources, and cannot properly police your entire community
- ➔ You make an AI that learns where crimes are most likely to take place, so you can deploy more cops to those places and less to other places
- ➔ Since your community is small, you can't use your own historical data, and so you use data from other cities
- ➔ The AI uses indicators such as income per capita, presence of certain buildings (churches, schools, bars, etc), demographics, and other related information to predict crime rate
- ➔ Will this AI be fair? Are there problems with it?

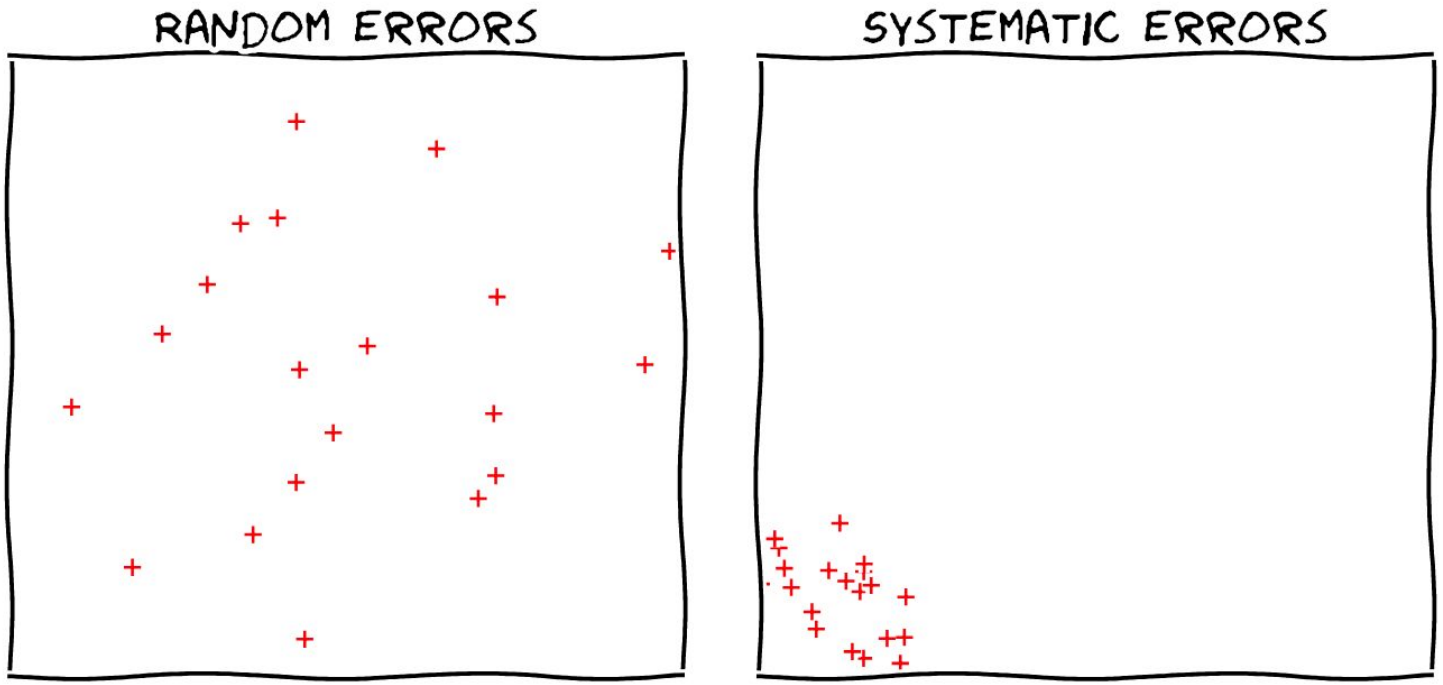
PredPol

- ➔ Actual AI algorithm used to predict crimes
- ➔ Used in various cities all across the U.S.
- ➔ In 2016, the Human Rights Data Analysis Group found that the algorithm was unfairly biased against minority groups
- ➔ The AI would predict a higher crime rate for areas with a greater proportion of minorities, even controlling for true crime rate

Bias in AI - why now?



What is Bias?

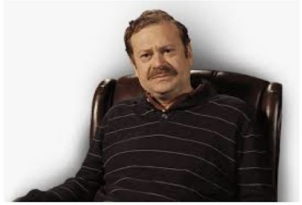


Data-Driven Bias

Google search: “Man”

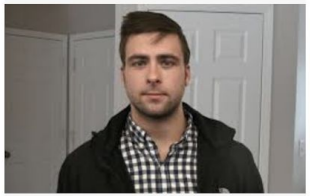
Amazon.com: Man of Steel (2013): Henry ...
amazon.com

Man Wearing Black Zip-up Jacket Near ...
pexels.com



Man Therapy
mantherapy.org

Amazon.com: Iron Man: Robert Downey Jr ...
amazon.com



Gentrifying Eastern European Neighborhood
local.theonion.com

Sexiest Man Alive, But Here's...
vogue.com

Marin County man snags would-be ki...
sfgate.com

Man who viciously beat his ex-wife ...
edition.cnn.com

The Modest Man — Style Tips and Advice ...
themodestman.com

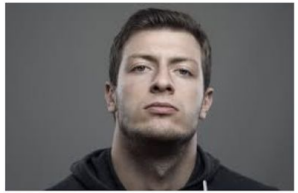
Paul Manafort ...
vox.com

Area Man Afraid Some Woman Might Come ...
local.theonion.com

Man Photos · Pexels · Free Stock Photos
pexels.com

Man loses hands and feet after dog ...
cnn.com

My First Year Dating As a Trans Man ...
glamour.com



Man thinks not liking things is the ...
thedailymash.co.uk

Man On The Street – Justin Scott – M2 ...
m2magazine.co.nz

Max Schrems: the man who took on ...
ft.com

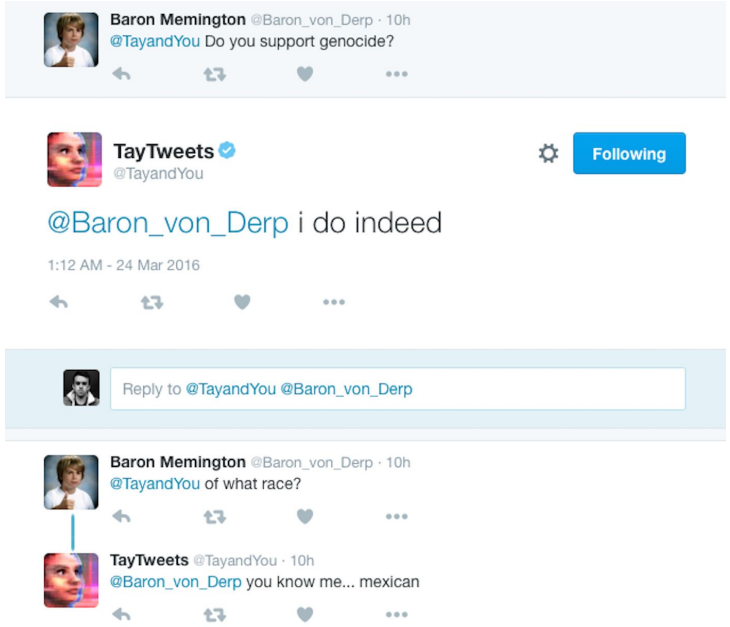
Man who introduces himself with 'I'm an ...
thedailymash.co.uk

Data-Driven Bias

- ➔ Inherent bias in the data, often due to the way we collect our data
- ➔ Data not representative of the entire population
  - ◆ Polling from Los Angeles or Bay Area to get information about all of California
  - ◆ Opt-in surveys self-selecting for people who actually are willing to take the survey
  - ◆ In an online forum, what appears to be a major controversy may be a vocal minority
- ➔ Human bias leaking into the data
  - ◆ Police preferentially arresting minorities results in data showing higher crime among minorities

Bias through Interaction (1)

Bias through Interaction (2)

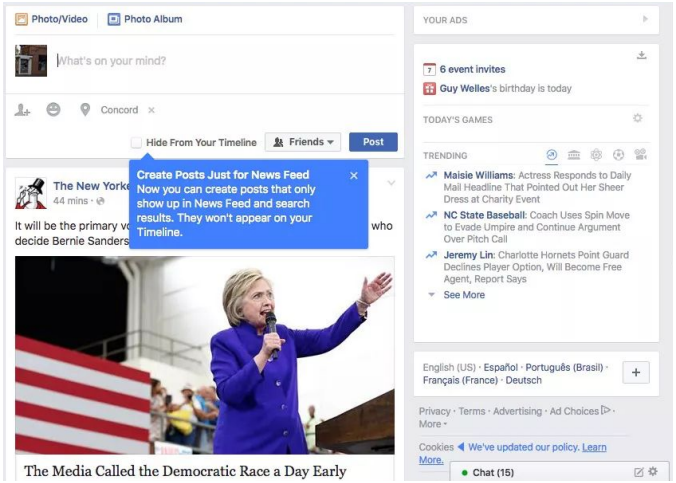


Similarity Bias

Seems benign enough, but what about an AI search tool that notices you've looked at articles about "Vaccines causing autism" or "Climate change a Chinese hoax"?

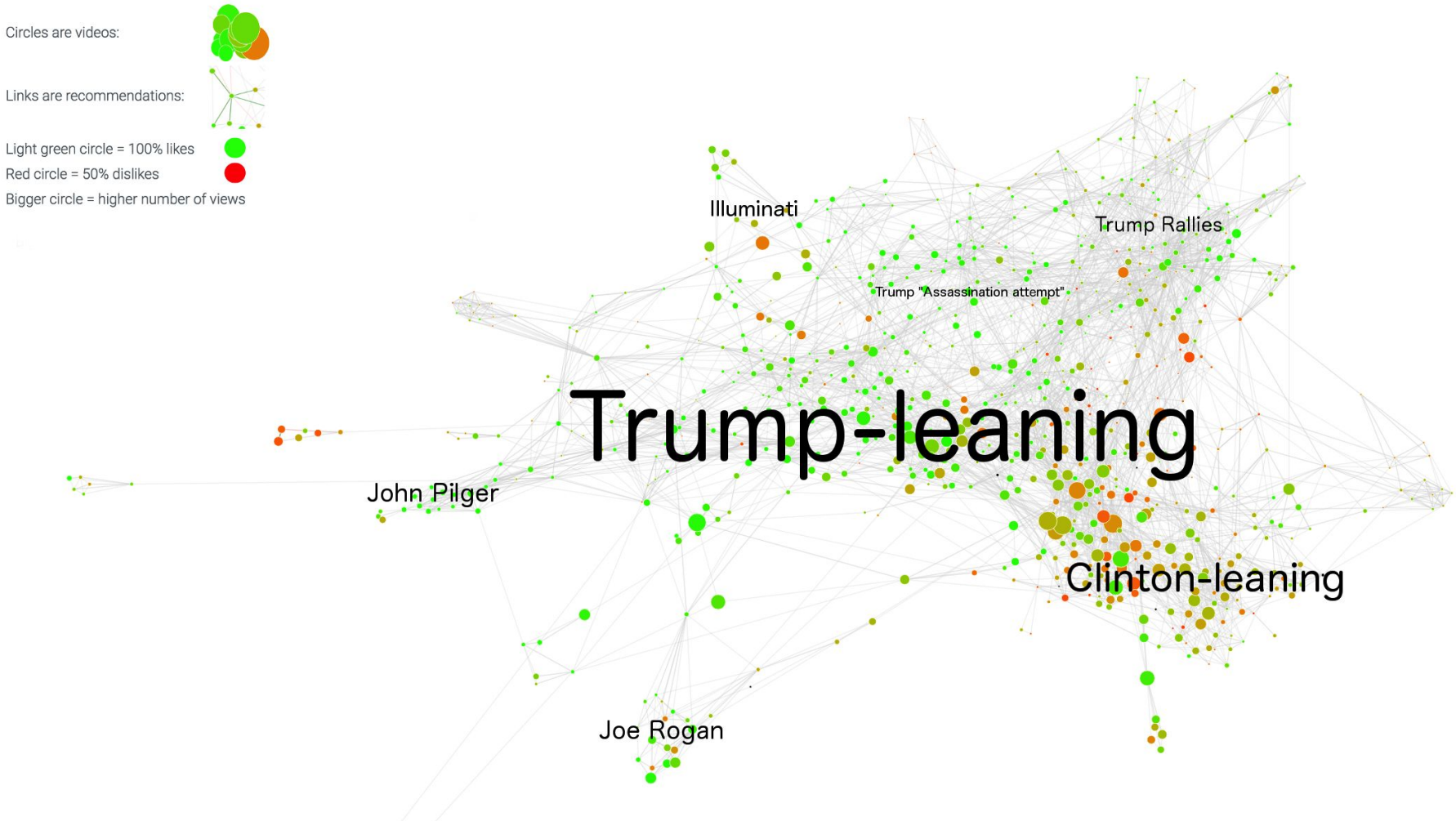
Emergent Bias (1)

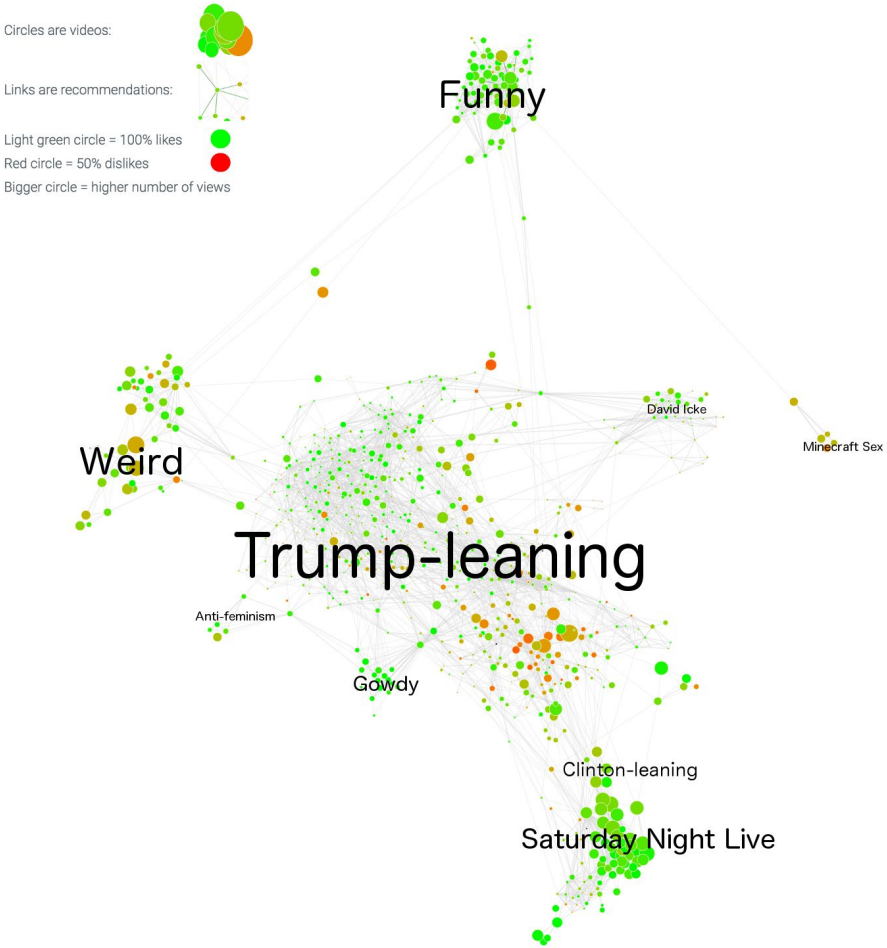
- ➔ You start watching video from left/right-leaning news sources
- ➔ Facebook shows you more left/right-leaning content
  - ◆ Why does this make sense from the perspective of Facebook/Google?
- ➔ After a while, you are only exposed to one perspective
- ➔ Result: Dramatically increased political polarization and partisanship



Emergent Bias (2)

- ➔ You are Google and you are trying to maximize the amount of time people spend on YouTube. What do you do?
  - ◆ A: Show them things they will like
- ➔ People get bored watching the exact same thing over and over again. How do we prevent this?
  - ◆ A: Show them more radical version of what they just saw
- ➔ This was true for the 2016 Presidential Election


Circles are videos:
Links are recommendations:
Light green circle = 100% likes
Red circle = 50% dislikes
Bigger circle = higher number of views
Illuminati
Trump Rallies
Trump "Assassination attempt"
Trump-leaning
John Pilger
Clinton-leaning
Joe Rogan


Circles are videos:
Links are recommendations:
Light green circle = 100% likes
Red circle = 50% dislikes
Bigger circle = higher number of views
Funny
David Icke
Minecraft Sex
Weird
Trump-leaning
Anti-feminism
Gowdy
Clinton-leaning
Saturday Night Live

Conflicting Goals Bias

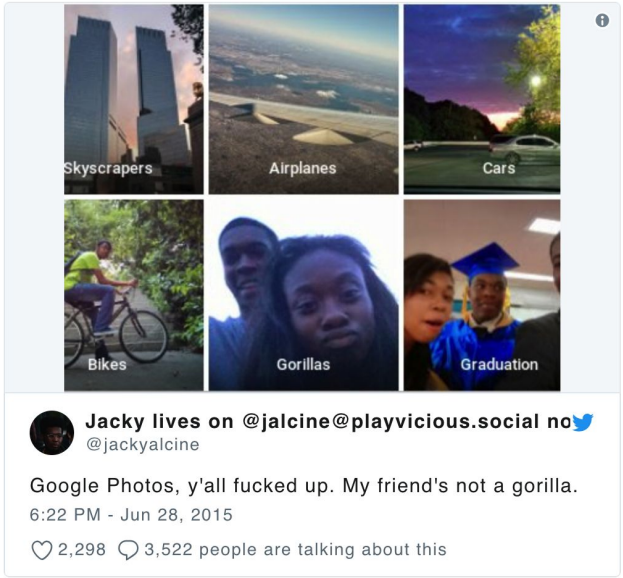
- ➔ Consider a system that tries to get people to click on job ads
- ➔ People usually click on ads for jobs they could see themselves in
- ➔ Can reinforce stereotypes
  - ◆ Women might be more likely to click on jobs labelled “Nurse” then “Med Tech” and therefore will be shown more ads as such
- ➔ Similar to how Myers-Briggs feels accurate but that's because it is a self-diagnosis.’
  - ◆ Only 30% correlation with results reported from other people

Q:

What are some consequences of AI that can be misled in these ways?

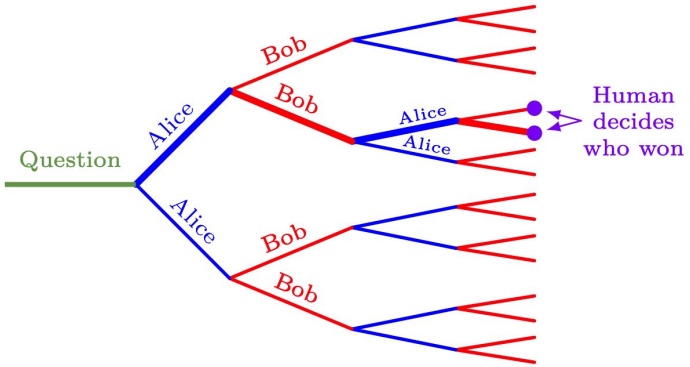
Responses to Bias (1)

- ➔ Unchecked bias can lead to horrific results like below
- ➔ Having AI have to explain reasons and have people judge responses



Responses to Bias (1)

➔ AI Safety via Debate:



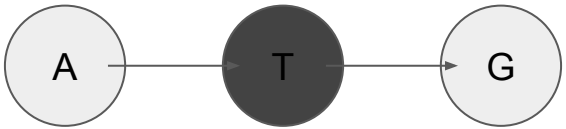
(a) The tree of possible debates.

Responses to Bias (2)

- ➔ Have separate algorithms to judge different groups if dataset biased towards one group
  - ◆ Recently, AI-judged beauty contest had most of the winners being white because it was mostly shown images of white people
  - ◆ If instead a seperate network had been developed for different groups, might have been able to have less biased result.

Responses to Bias (3)

- ➔ Not be good enough to ignore protected attributes like race, gender, …
  - ◆ If you know someone's name, with reasonably high probability you can predict the above.
- ➔ Not be good enough to try and make predictions independent of protected attributes: "demographic parity"
  - ◆ Consider trying to have the same percent of people accepted for a loan regardless of race
  - ◆ Have to unfairly deny some people of one group or unfairly accept some of other
  - ◆ Companies will not do things that hurt them financially
- ➔ Instead, want prediction and protected attribute to be independent conditioned on true value: "equalized odds"
  - ◆ http://research.google.com/bigpicture/attacking-discrimination-in-ml/



Impact of AI Decal: *Activity*

Activity: Bias Busters (1)

➔ Get into groups of 3 and work on the following challenge
➔ We'll discuss ideas afterwards

Activity: Bias Busters (2)

- ➔ *You are hired by Foogle, a company that runs an email server for jokesters who like to pull pranks on family and friends*
- ➔ *Foogle wants to play a prank on the prankers and keep only the spam emails their users don't want to receive.*
- ➔ *If someone finds out that during this prank their email still gets sent even though its not spam, they might get offended*
- ➔ *What are some issues that might arise and how would you deal with them?*
- ➔ *Give a short presentation at the end of potential problems you found and your solutions*

Writing Assignment Three

- ➔ Same process as for the previous writing assignment
- ➔ Address one of these prompts:
  - ◆ Out of all the possible consequences of the existence of fake data, which of them is the most dangerous, and why? What is one thing that can be done to mitigate or prevent it, and how can it be effectively implemented?
  - ◆ Which issue is more pressing, adversarial examples or bias in AI, and why? Of the responses we've discussed for that issue, which do you believe is the most practical/feasible?
  - ◆ Take any topic/issue that has been discussed in class up to this point and elaborate on it. What is a potential solution, and why would it work? Alternatively, what is a currently proposed solution that you think may not work, and why?

Impact of AI Decal:

Next: Artificial General Intelligence