Info 4602 Summer 2022

Instructor: Samantha Dalal

Welcome + What is Ethics?

July 5th, 2022

Agenda

- Hi!!!!
- About the class
- Tell me about you!
- What is this Ethics thing??
- Homework

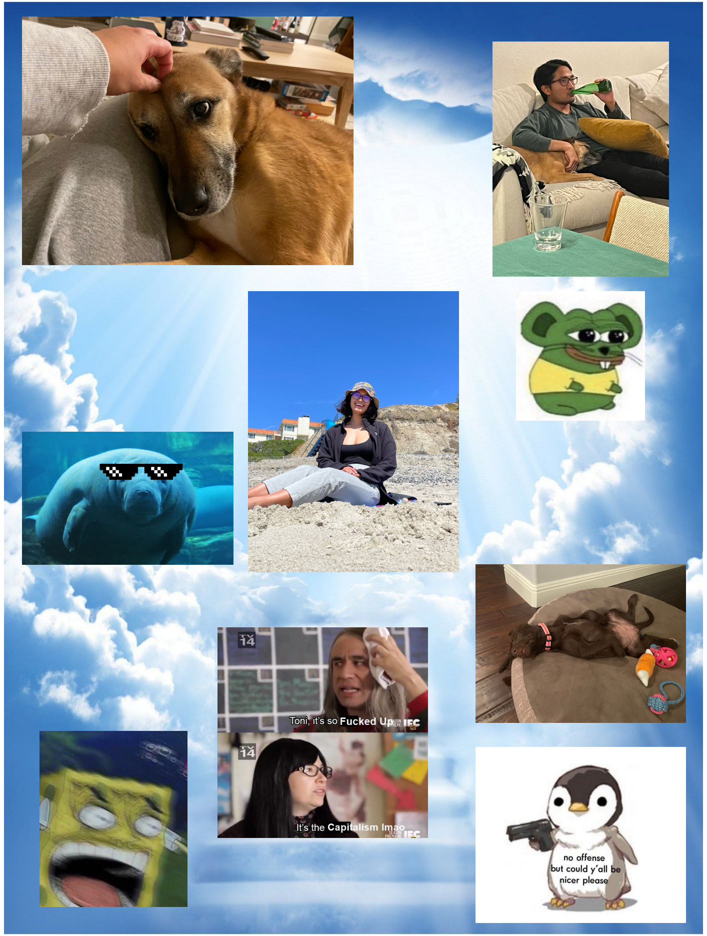
About Me

Research Area:

Human-AI interaction in the workplace;
data-driven workplaces

Interests:

Climbing, cooking, reading



Email:
samantha.dalal@colorado.edu

About the Course

Expectations

- Active engagement + Participation
- Feel free to disagree!
- Be respectful + assume best intent
- I put a lot of effort into this course, I expect you to reciprocate that through putting in effort as well

Structure

- Mondays: Talk about news/current events, introduce new topic, do an activity + discussion session to practice topic, go over HW
- Tues - Thurs: Review topic from previous day's HW, do activity + discussion session to practice topic, introduce new topic, go over HW
- Fri: No in person class, work on reading reflection for the week + piece of final project

About the Course, Cont.

Assignments

- Reading Reflections (Due on Fridays by 11:59 PM)
- Final Project Deliverables
    - Pt. 1 Due on Monday, July 11th @ 11:59 PM
    - Pt. 2 Due on Monday July 18th @ 11:59 PM
    - Pt. 3 Due on Monday July 25TH @ 11:59 PM
    - In-class Presentation on Wednesday August 3rd
    - Full Project Due on Friday August 5th @ 5 PM
- (OPTIONAL) News Reflection Due by Tuesday August 2nd @ 11:59 PM

About the Course, Cont.

Grading

- Participation: 20%
- Final Project: 40%
  - Each deliverable: 10%
- Reading Reflections: 40%
  - Each reflection: 10%
- News Commentary (Extra credit): Up to 5%

About You!

1. Go to the Google Doc for today and fill out the lightning ⚡ intro section + brainstorm 🧠 section
2. Now share 😈 - turn to your neighbor & introduce yourself and discuss some news topics/controversies that you heard about this past year!

A Thought Experiment

When is it ok to murder someone?

What makes a bad person bad enough to die?

How do we know that that is “bad”?

What is NOT Ethics?

- Feelings != Ethics
- Law != Ethics
- Social norms != Ethics
- Religion != Ethics

What is Ethics?

- “Ethics refers to well-founded standards of right and wrong that prescribe what humans ought to do, usually in terms of rights, obligations, benefits to society, fairness, or specific virtues”[1]
- Frameworks
  - Utilitarianism
  - Deontology
  - Virtue Ethics
  - Gandhian
  - Ubuntu
  - Ethics of Care

Homework + For Next Time

Crash Course Videos (Linked on the discussion doc for today):

- Utilitarianism
- Deontology
- Virtue Ethics

Things to think about:

- What are the key characteristics of each moral framework?
- Which moral framework(s) do you see operationalized most often in technology development?
- Which moral framework resonates most with you? Why do you think it resonates with you?

Ethical Frameworks + Moral Messaging

July 6th, 2022

A Review of Ethical Frameworks

Utilitarianism - consequences over intentions, maximization

Deontology - following categorical imperatives, logic

Virtue Ethics - being a good person, intentions

Ubuntu Ethics - community over individual empowerment, consensus

Ethics of Care - lived experience over abstraction, relationality

Utilitarianism

- Focus on consequences over intentions
- Actions should be measured in terms of utility they produce
- Greatest good for the greatest number
- Self interests do not count more than others’ interests
- Should avoid short run maximization and follow rules that will maximize utility in the long run

Deontology

- Categorical imperatives that you must follow regardless of desires
- Principles that you follow must be universalizable and not produce a contradiction
- You cannot violate moral laws even for a good cause
- You must recognize and respect not only your own, but other people’s autonomy
- Must consider others’ goals and imperatives

Utilitarianism vs. Deontology - A Thought Experiment

You have 1 million dollars to donate after a massive tsunami just occurred. You could give directly to families impacted using GoFundMe or you could invest in a well established green energy tech company to fund the development of technology that will slow climate change in the future.

Which would you choose? Which option would a utilitarian choose? Which option would a deontologist choose? How are you defining the “greatest good”? How are you establishing a universalizable principle?

Colorado River Water Supply

- The Colorado river is a vital waterway for the southwestern states
- But over the past 20 years, flow of the river has decreased by 20%
- It is now time to renegotiate a 100 year old compact drafted to distribute water amongst the states that depend on it
- The compact was created through consensus and has stood the test of time despite some shortcomings (e.g. not considering Native American rights to the water)
- Some suggest that a market based approach would be beneficial to most efficiently allocate resources
- However, market based approaches run the risk of arbitrage at the cost of the public good

An Activity - Moral Messaging

Draft a proposal with your group detailing how the limited water supply should be rationed according to three moral frameworks of your choice. Detail your approach, explaining what your plan is and how this moral framework helps you determine if your plan is ethical.

Homework + Next Time

The Ethical Dilemma of Self-Driving Cars, Patrick Lin (TED-Ed, 2015)

"Uber's Self-Driving Car Didn't Know Pedestrians Could Jaywalk," *WIRED*, 2019

Things to think about:

- Based on your own feelings both before and after thinking through these readings, how do you feel about self-driving cars? Are you excited, worried, something else? Why?
- How would different ethical frameworks affect how self-driving cars function?
- Fortunately there isn't a huge issue with brakes going out on self-driving cars and them having to choose between life and death. What do you think are some of the more pressing ethical issues with self-driving cars we should be thinking about?

The Trolley Problem + Moral Dilemmas

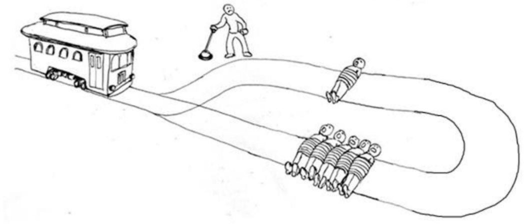
July 7th, 2022

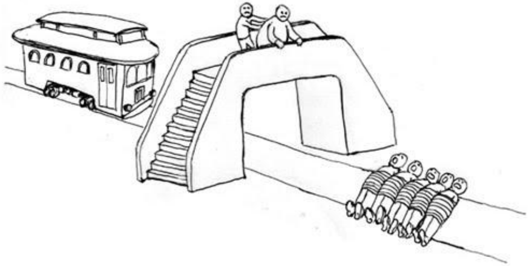
A Review of the Trolley Problem

“Trolley Problem is a thought experiment where someone is presented with two situations that present nominally similar choices and potential consequences” (Roff 2018)

For example: *A runaway trolley is headed towards a group of five people standing on the tracks. You are standing next to a lever, and if you pull this lever, the trolley will be switched onto a different track, with a single person standing on it. Do you pull the lever?*

An Activity - Justify your solution to the trolley problem





Medical Care Rationing

How should we decide who gets medical care when resources are limited?

Should we take into account age? Pre-existing conditions? Likelihood of survival?

Is it ethical to take factors like age and pre-existing conditions into account?

HATE SPEECH OR CENSORSHIP?

Should we err on the side of avoiding potentially harmful content at the risk of suppressing free speech?

Or should we err on the side of free speech at the risk of allowing harmful content?

Who should make that decision? Who gets to decide the threshold?

Homework + Next Time

Fink, Sheri. U.S. Civil Rights Office Rejects Rationing Medical Care Based on Disability, Age. *The New York Times*. 2020.

Things to think about:

- How do you feel about medical care rationing? If a loved one was in a situation where they needed urgent medical care would your perception of medical care rationing change?
- What are other moral frameworks that might be more appropriate than utilitarianism in the case of medical care rationing?

Ethical Dilemmas in Tech + Final Project

July 7th, 2022

Building the Tools for a Surveillance State

Intel & Nvidia supplied chips to the Sugon, a Chinese company that builds powerful AI systems to monitor its population.

This company has been linked to efforts to segregate Uighur Muslims, facilitate predictive policing, and enable the Chinese surveillance state.

Intel & Nvidia have since stopped selling high-powered chips to Sugon but their other technologies continue to power surveillance systems.

Refusal to Build

In 2018 Google pulled out of a contract with the Pentagon to utilize its computing resources in order to perform image analysis for drones.

Some employees argued that this technology could be used to help identify human targets in warfare and did not want their work to be militarized.

Other employees worried this might damage the company’s ability to get DoD contracts in the future.

Multiple Points of Failure

In 2022 the NYTimes released “The Civilian Casualties Files”, documenting how human and machine error led to many unreported civilian deaths caused by drones.

The air campaign in Afghanistan was lauded as the “the most precise air campaign in history.”. It was indeed precise, drones rarely ever missed their targets.

However, target identification and follow-up verification was rarely thorough which caused civilians and bystanders to be killed

Final Project Preview

For the final project, you will be asked to identify a controversial technology that was developed within the past decade. **Your job will be analyze Is the creation and dissemination of this technology ethical? Additionally, you will develop a strategy to mitigate possible adverse outcomes if the technology were to be deployed.**

Final Project Preview Cont.

The final project is composed of three smaller deliverables, a presentation, and a final report

- Deliverable 1 due on JULY 11TH @ 11:59PM
- Deliverable 2 due on JULY 18TH @ 11:59PM
- Deliverable 3 due on JULY 25TH @ 11:59PM
- Presentation due on AUGUST 3RD IN CLASS
- Final Report 1 due on AUGUST 5TH @ 5:00PM

The final project is due on AUGUST 5TH, 2022 AT 5PM. Late work will not be accepted and you will receive a zero for any work turned in late.

Privacy in the Digital Age

July 11th, 2022

4th Amendment

The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no warrants shall issue, but upon probable cause, supported by oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

4th Amendment

- Law enforcement must respect the defendant’s privacy rights during an investigation
- “probable cause”
- Only applies to the government, not to a private individual
- *Reasonable expectation of privacy*
- Three key exceptions:
  - Consent
  - Plain view
  - Third -party doctrine

5th Amendment

No person shall be held to answer for a capital, or otherwise infamous crime, unless on a presentment or indictment of a Grand Jury, except in cases arising in the land or naval forces, or in the Militia, when in actual service in time of War or public danger; nor shall any person be subject for the same offence to be twice put in jeopardy of life or limb; nor shall be compelled in any criminal case to be a witness against himself, nor be deprived of life, liberty, or property, without due process of law; nor shall private property be taken for public use, without just compensation.

5th Amendment

- “Right to remain silent”
- Testimony can only be compelled through offering immunity

Riley V. California

- Main differentiating factor - searching a phone is not the same as a physical search
  - Why?
- Physical searches only permitted if law enforcement believes that there is an immediate danger to them or there is a danger of destruction of evidence
  - Does this hold in the case of digital searches?
- Conclusion - to search a digital device, law enforcement must get a warrant
  - What does a digital search warrant have to include?

Coffee Shop Creeps

- A laptop in plain view, in a public place had photos popping up that looked like explicit images of minors
- A law enforcement officer notices the pictures, pursues the suspect and arrests them
- Is this arrest constitutional?
- What possible exceptions to the 4th amendment have occured here?

Beanie Baby Fraud and Sea Creatures

- Beanie Baby fraud & other wire fraud committed using online forums
- Stingray devices act as a type of pseudo-cell tower, meaning that when **any** cell phone has an outgoing or incoming call in its vicinity, it pings the Stingray
- What types of digital trace data can & should be used in a court of law?
- How should that data be procured?

Every Step You Take, Every Prayer You Make, I'll be Watching

- U.S government contracts w/two companies that provide location data on Muslims
- Babel Street made a software called “Locate-X” that provided granular anonymized location information of cellphones. Employees clarified that the data could be easily de-anonymized
- X-mode paid apps to allow them to place code within the app that skimmed user's location data. Muslim Pro allowed X-mode to do this until the news story broke
- Who's liable here? Were any constitutional rights violated?

Becoming Known

July 11th, 2022

Was he free? Was he happy?
The question is absurd:
Had anything been wrong, we
should certainly have heard.

- W. H. Auden from *The Unknown Citizen*

Technocracy

The world we live in is governed by the affordances of technology

Privacy vs. Visibility Throughout History

- New technologies more easily render the population visible
- Many surveillance and privacy scholars fret about the “end of privacy” in the modern world
- This is a shallow perspective, rather we should focus on the paradox of what it means to be a *modern citizen*
- Throughout history, people have continually negotiated the desire to have a private life or a private self, and the desire to be visible and seen

Getting Caught Up in the Dragnet

- To access welfare services, people must give their identification information
- When a house is foreclosed on, the owners information is collected by foreclosure agencies
- When a credit card payment is late, credit agencies collect the debtors information
- What happens when all these data sources, that were never intended to be combined get aggregated and fed into crime prediction systems?

"It's Creepy How They Just Know..."

- Social media companies often make the argument that their efforts to track your digital footprint are benevolent; in fact these efforts help them give you the best experience possible on their site!
- What are the tradeoffs between revealing your preferences and maintaining your privacy for you personally?
- Is your tolerance for surveillance equivalent to others?
- How should companies negotiate different users privacy preferences?

The Costs of Exclusion

- To be excluded from digital surveillance is to be invisible to the State (and possibly to private service providers)
- How might structural inequities be preserved and amplified in an increasingly predictive society that relies on “big data”?
- Do the costs of exclusion exceed those of inclusion? How do we balance these as individuals? As a society?

Wait a Minute...Who Are You?

- In light of digital dragnets, we need to be wary of claims of “anonymity” in datasets
- Most people are identifiable through only 5 pieces of demographic information
- What potential consequences does this have for decentralized identity verification systems?

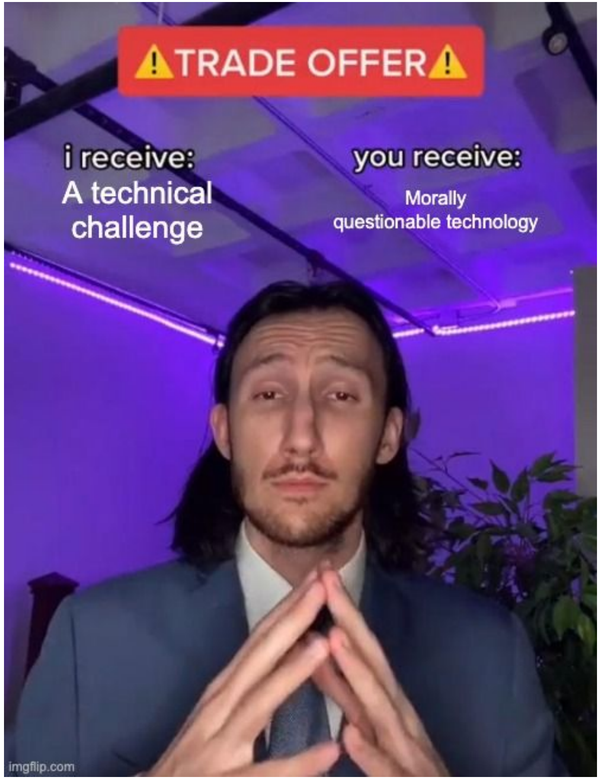
Ethical Dilemmas w/Privacy

Should This Tech Be Developed?

- Creation of “gaydar” to raise awareness
- Questionable technology regarding facial recognition
- Do the ends justify the means?
- Are claims that you’re building something potentially dangerous to raise awareness sufficient justification?

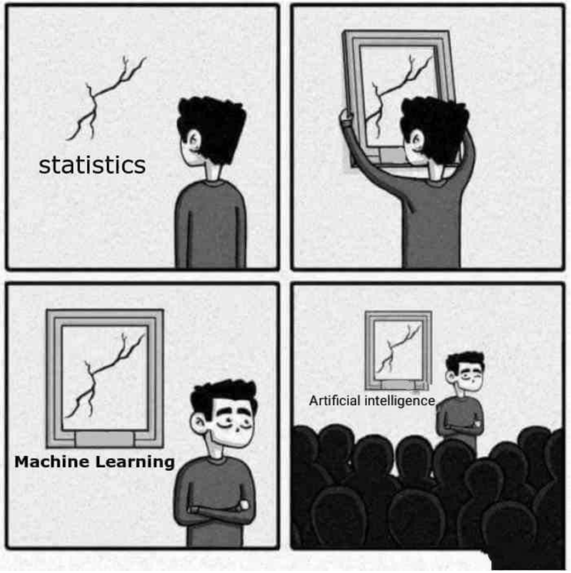
But the technical challenge!!!!

- Using deepfakes to generate nude images
- Have been used to make revenge porn
- Justified development by saying they were motivated by the technical challenge and if they didn’t do it, someone else would



Is it actually “intelligent”?

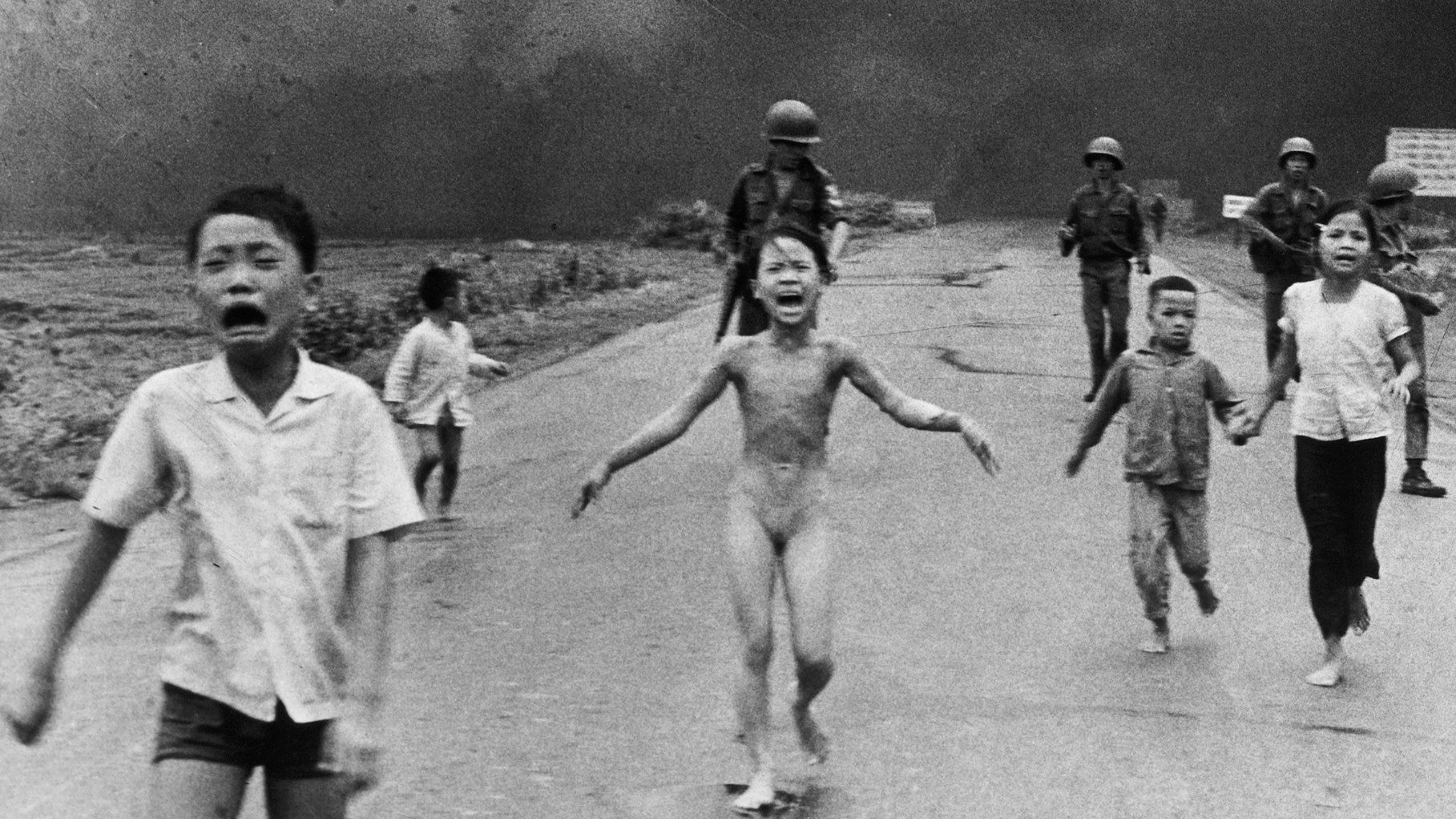
- Hype around AI often confuses people as to what AI is really capable of
- Does learning = understanding?
- How can we temper the hype to give people more transparency into what technology is actually capable of?



No One Likes to Be Wrong

- Talking about ethics is hard, but necessary. Our understanding of ethics evolves **through** conversations
- Ethics **are discursive**

Speech



This Week…

- The first amendment: how things get complicated online
- Section 230: how platforms regulate (or don’t)
- Content moderation: how free speech & platform policies interact

Scenario 1

No one can publish a spotify podcast series without prior approval from the ministry of culture

Constitutional or Unconstitutional?

Scenario 2

You cannot give “libertarian” speeches in front of the Boulder court house

Constitutional or Unconstitutional?

Scenario 3

You may not distribute campaigning materials within 200 yards of a polling place during voting times.

Constitutional or Unconstitutional?

Scenario 4

You cannot yell “FIRE” in a crowded theater

Constitutional or Unconstitutional?

In Summary

“Prohibiting “loud” speeches in the park is content-neutral; prohibiting “political” speeches in the park is content-based; prohibiting “liberal” speeches in the park is viewpoint-based.”

Know Your Rights!

1st Amendment states that:

“Congress shall make no law respecting an establishment of religion, or prohibiting the free exercise thereof; or abridging the freedom of speech, or of the press; or the right of the people peaceably to assemble, and to petition the Government for a redress of grievances.”

BUT

There are exceptions….

- "Strict scrutiny test"
  - Compelling interest
  - Narrowly tailored
  - No less restrictive alternative
- Content neutral restrictions
  - Reasonable "time, place, and manner" regulations
  - Narrowly tailored to serve government interests
  - Leave open other alt channels for comms
- Commercial speech test
  - Substantial gov interest
  - Reg advances gov interests
  - Reg is not more extensive than necessary

VERY IMPORTANT

THE FIRST AMENDMENT ONLY APPLIES TO YOUR RIGHT TO FREE SPEECH AS IT PERTAINS TO GOVERNMENT REGULATION; THE GOVERNMENT CANNOT ARREST YOU FOR WHAT YOU SAY!!!

THAT DOESN’T MEAN THAT ANYONE ELSE (PRIVATE INDIVIDUALS, CORPORATIONS, ETC.) HAVE TO LISTEN TO YOUR BS

Should Registered Sex Offenders Be on Social Media?

- North Carolina law makes it a felony for a registered sex offender “to access a commercial social networking Web site where the sex offender knows that the site permits minor children to become members or to create or maintain personal Web pages.”
- What counts as a social media?
- What does this prevent people from accessing?
- Is this law overly broad and therefore unconstitutional?

What Counts as Speech?

- A sheriff was running against another candidate for re-election. Some of the current sheriff's employees “liked” the other candidate’s campaign Facebook page. The other candidate lost and the current sheriff fired several employees, some of whom had “liked” the other candidate’s page.
- Was the employees’ 1st amendment right violated?
- Does a “like count as speech”?

What makes networked technologies complicated?

- Persistence: the durability of online expression and content
- Visibility: the potential audience who can bear witness
- Spreadability: the ease with which content can be shared
- Searchability: the ability to find content

What does this mean for *harmful content?*

Swatting

- “In 2017, a man identifying himself as “Brian” called Wichita police and claimed to be holding his family hostage. Officers who responded to the address he gave shot and killed the man who came to the door. But “Brian” was actually a prolific swatter named Tyler Barriss, who lived in Los Angeles and had been recruited online by a Call of Duty player who wanted revenge on another player over a $1.50 wager. Barris pleaded guilty to making a false report resulting in death, see 18 U.S.C. § 1038(a)(1)(C), among other counts.”
- In what ways do the affordances of platforms enable this type of behavior?

Don’t Tell Me What To Do!!!

There are TOTALLY No Rules Online 😎

lib·er·tin·ism

being libertine. 2. The be

lib·er·ty (lĭb′ər-tē) n.,

from restriction or con

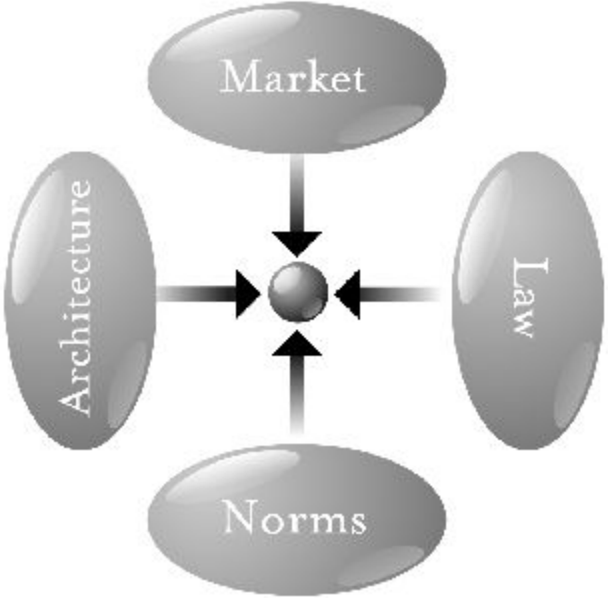
express oneself in

condition of being ph

Just Kidding

- In 1996 John Perry Barlow, a lyricist for the Grateful Dead, wrote *A Declaration of the Independence of Cyberspace.* This declaration claimed that cyberspace was a new frontier, one that was not beholden to rules and laws governing the physical world
- He was, in fact, wrong cyberspace could and would be regulated by laws
- When the government and private companies recognized the value & potential of cyberspace they began to regulate through a variety of mechanisms

Code is Law



Section 230

- If I post a defamatory video on YouTube dragging Elon Musk, I'm the one who can be held liable, not YouTube
- TYPICALLY publishers are held liable for the content they publish or republish
- BUT Section 230 has a carve out for "interactive computer services"
    - They are not publishers
    - They have a **right but not a responsibility to moderate**
- Interactive computer services are any info service, system, or access software provider that enables computer access by multiple users

Moderation is a Messy B*tch

All Platforms Moderate

- Even the most “open” platforms engage in some form of moderation to remove spam and keep conversations on topic
- Moderation includes both removal AND sorting!! Content curation/promotion/amplification is a TYPE of moderation!

Moderating at Scale

- Moderation is often necessary for large platforms that depend on income from advertisers
- As the user base becomes international, decisions about moderation become more difficult - what is *right* and what is *wrong* is culturally contextual
- Even when the moderation decisions are left up to the community, it is difficult for communities to come to a consensus about *right and wrong*

The Tools of Moderation

- Medium
  - ToS
  - Community Guidelines
- Actor
  - End users
  - AI/ML systems
  - Human moderators

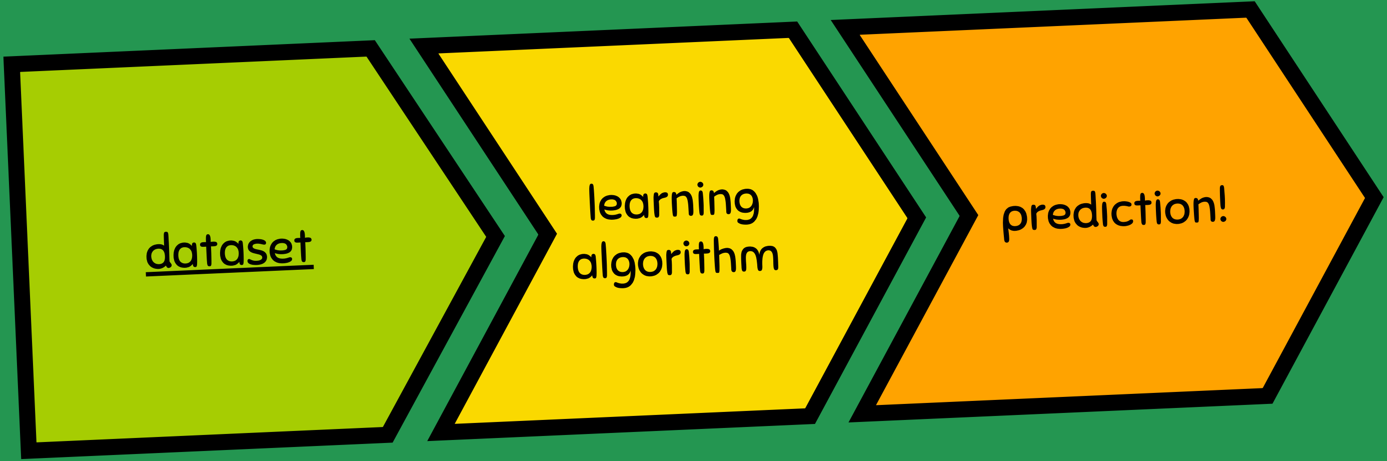
Well, This is Hard :/

So What Now??

- Should we remove or filter content?
- Removal offers a finality to the situation, it is a unilateral action that showcases the pure authoritarian force that platforms wield
- Filtering enables platforms to preserve content for some users and simply exclude others from viewing it

Bias, what is it??

First, what is AI?



From Introduction to AI and AI Bingo by
Blakeley Hoffman @ MIT Media Lab

Let's Play AI Bingo

- On your bingo card there are boxes with different AI systems
- Your task is to go around the room and partner up with someone and identify the **data used and the prediction made** by the AI system
- The first student to get bingo (5 in a row, column, or diagonal) will get 2 extra credit points

What is Bias?

“Outcomes which are systematically less favorable to individuals within a particular group and where this is no relevant difference between groups that justifies such harms” (Lee, Resnick, & Barton 2019)

BASICALLY: different outcomes for individuals where this is no good reason for outcomes to be different

Where does bias come from?

- Where are we getting our data from?
- What’s in our data?
- How do structural factors affect our data?

Ok...but who made the data?

- Machine learning needs LOTS of data
- Data is created by humans
- Humans inherently have biases
- Data is EMBEDDED WITH US!!!! HUMANS!!! AND OUR VALUES!! AND OUR BIASES!!!!

We're using AI instead of biased humans
What did you train the AI on?
What did you train the AI on?
imgflip.com

We Don’t See it!!!

- Claiming to be unbiased because you don’t utilize a controversial input variable when the other variables in your data set can be used to proxy that missing variable is misleading!
- Redlining
- Credit Histories
- Employment Status

No Easy Fix

- Bias can never be eliminated, only made more transparent
- Mitigating bias is tricky
  - Adding more data can reinforce the bias
  - Using synthetic data replicates bias of seed dataset
  - Down-sampling data reduces overall accuracy of a system & wastes otherwise useful data
  - More data could reveal sensitive information

So...is Bias Bad :(

- Not necessarily
- Bias just means favorable treatment of one group over another
- In some cases bias can be “fair”
- In others it can be “unfair”

Fairness, what is it?

Reflection

What does fairness mean to you?

What is algorithmic fairness?

- Treat all people the way they deserve to be treated

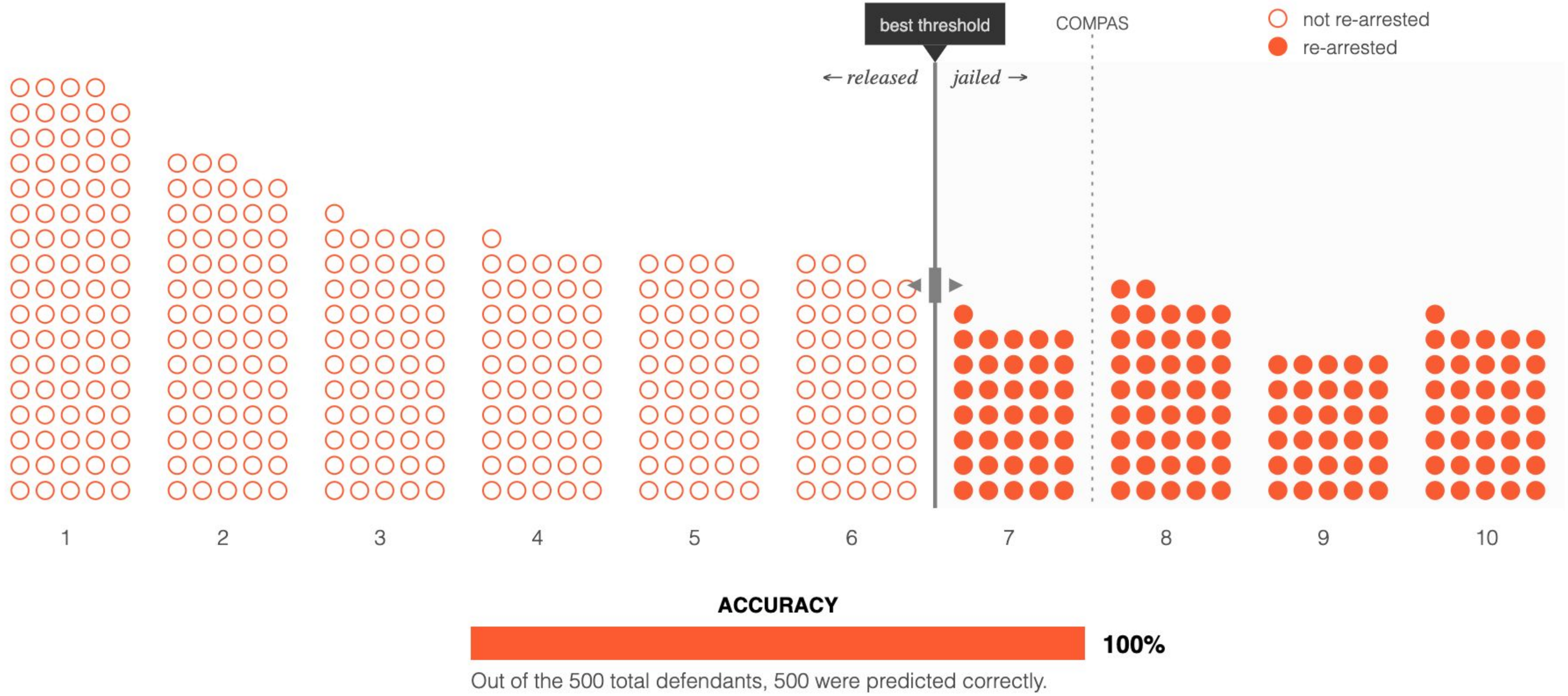
Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

What is algorithmic fairness?

- Treat all people the way they deserve to be treated
- **Give all people who deserve a positive outcome a positive outcome**

Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

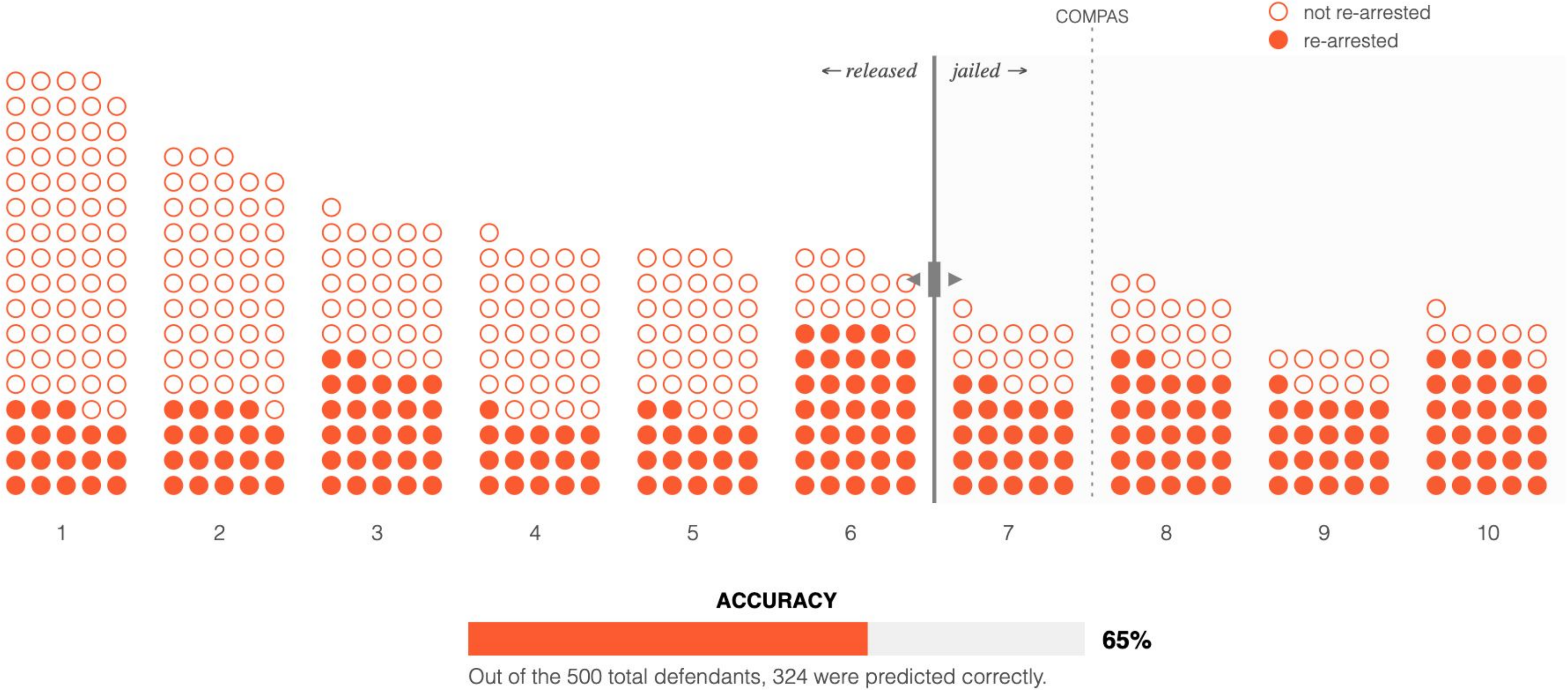
From MIT Tech Review COMPAS article



Great! We’re only jailing the people who would be re-arrested!

But it’s not so simple :(

From MIT Tech Review COMPAS article



Out of the 500 total defendants, 324 were predicted correctly.

For each risk score, some people will be re-arrested and others will not

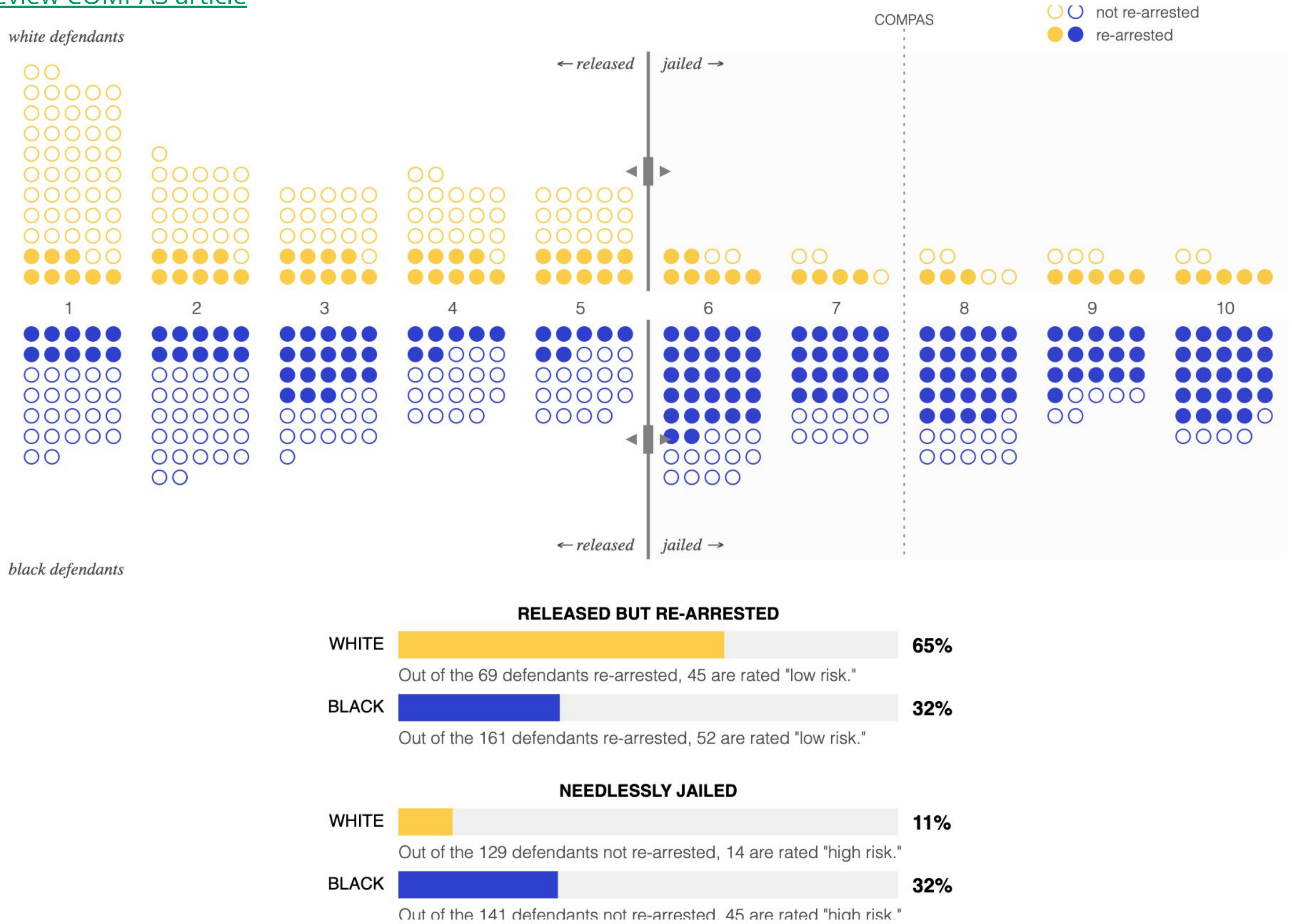
What is algorithmic fairness?

- Treat all people the way they deserve to be treated
- Give all people who deserve a positive outcome a positive outcome
- **Give all people who deserve a positive outcome a positive outcome, at an equal rate between all individuals**

Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

It’s REALLY not so simple :(

From MIT Tech Review COMPAS article



RELEASED BUT RE-ARRESTED

WHITE 65%

Out of the 69 defendants re-arrested, 45 are rated "low risk."

BLACK 32%

Out of the 161 defendants re-arrested, 52 are rated "low risk."

NEEDLESSLY JAILED

WHITE 11%

Out of the 129 defendants not re-arrested, 14 are rated "high risk."

BLACK 32%

Out of the 141 defendants not re-arrested, 45 are rated "high risk."

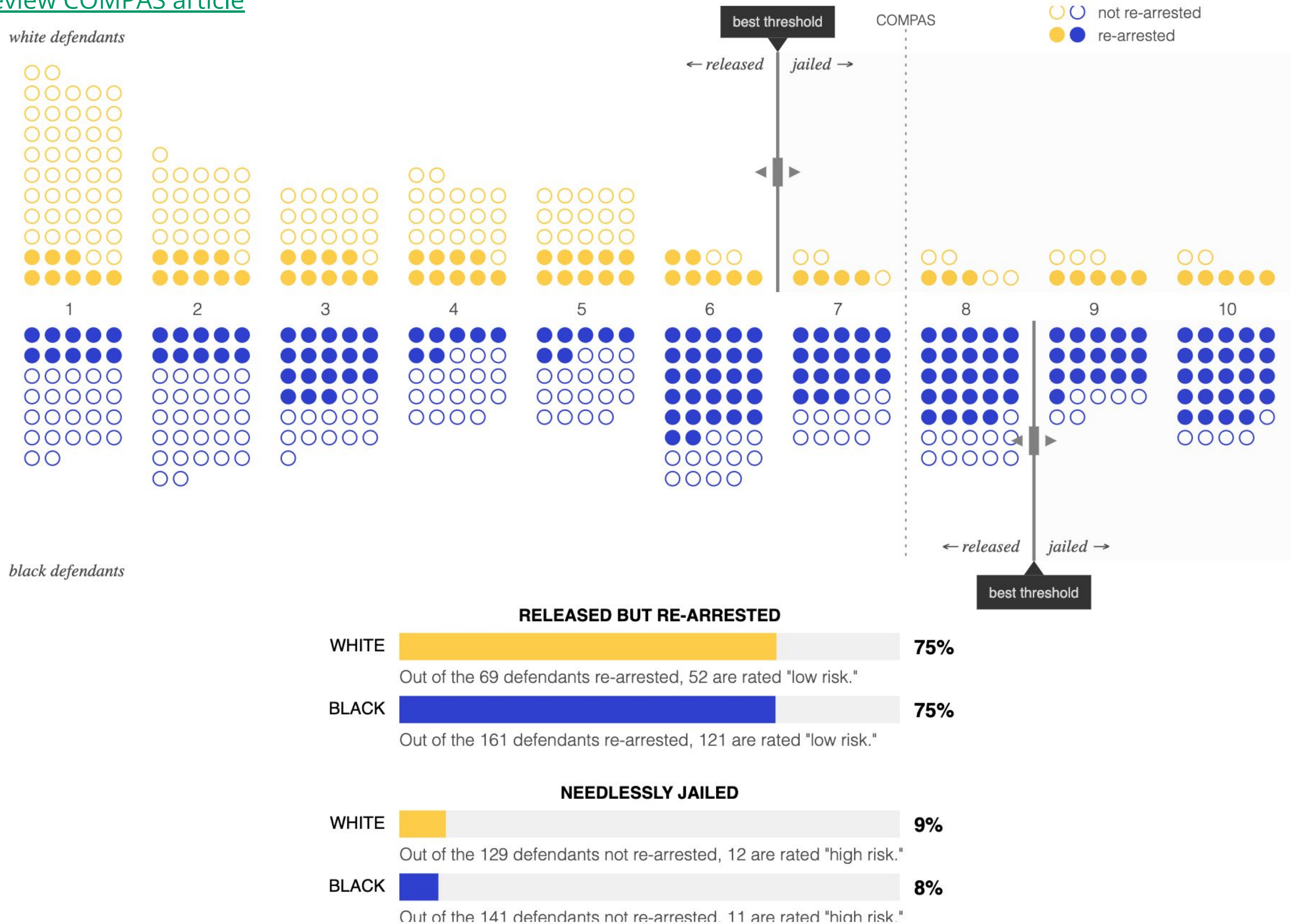
What is algorithmic fairness?

- Treat all people the way they deserve to be treated
- Give all people who deserve a positive outcome a positive outcome
- Give all people who deserve a positive outcome a positive outcome, at an equal rate between all individuals
- **Give all people who deserve a positive outcome a positive outcome, at an equal rate between all groups of people**

Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

Maybe this will work

From MIT Tech Review COMPAS article



RELEASED BUT RE-ARRESTED

WHITE 75%

Out of the 69 defendants re-arrested, 52 are rated "low risk."

BLACK 75%

Out of the 161 defendants re-arrested, 121 are rated "low risk."

NEEDLESSLY JAILED

WHITE 9%

Out of the 129 defendants not re-arrested, 12 are rated "high risk."

BLACK 8%

Out of the 141 defendants not re-arrested, 11 are rated "high risk."

What is algorithmic fairness?

- Treat all people the way they deserve to be treated
- Give all people who deserve a positive outcome a positive outcome
- Give all people who deserve a positive outcome a positive outcome, at an equal rate between all individuals
- Give all people who deserve a positive outcome a positive outcome, at an equal rate between all groups of people
- **Give all people who deserve a positive outcome a positive outcome, at an unequal rate between all groups of people - with a higher rate of positive outcomes given to groups who have been historically disadvantaged, and a lower rate for the remaining groups**

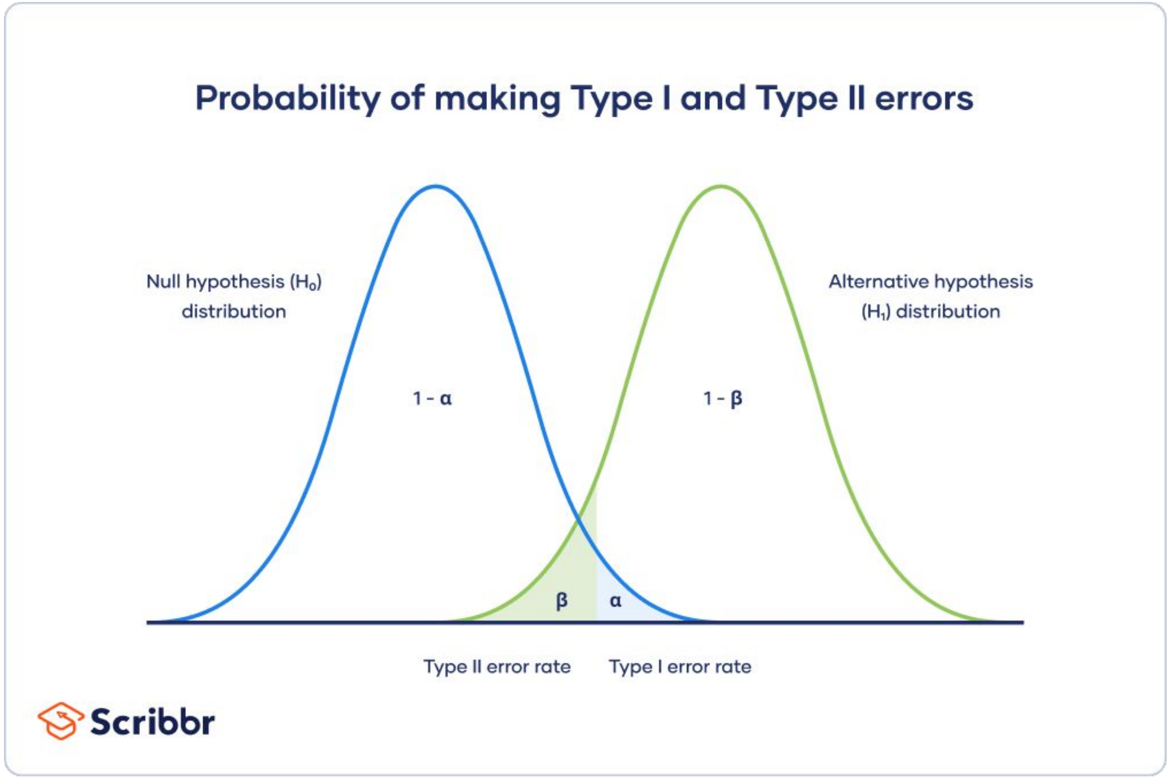
Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

You Gotta SAY IT!!!

- Fairness is squishy and malleable
- Algorithmic fairness requires quantifiable metrics
- You need to define your metric that you’re using when determining fairness!

She is very fair to me!

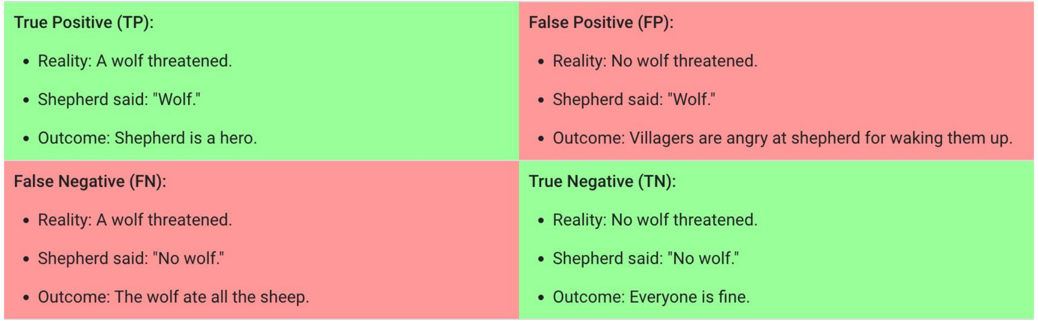
A Tiny Bit of Math


Probability of making Type I and Type II errors
Null hypothesis (H_0) distribution
Alternative hypothesis (H_1) distribution
1 - α
1 - β
β
α
Type II error rate
Type I error rate
Scribbr

False positives and false negatives are inversely related!!!! It matters which one you choose as your fairness metric!!!

Some Definitions

- True positive rate (TPR)
  - Aka sensitivity
  - TP/TP+FN
  - TPR is the probability that an actual positive will test positive
  - In the loan context: percentage of **paying applications** getting loans



True Positive (TP): • Reality: A wolf threatened. • Shepherd said: "Wolf." • Outcome: Shepherd is a hero.	**False Positive (FP):** • Reality: No wolf threatened. • Shepherd said: "Wolf." • Outcome: Villagers are angry at shepherd for waking them up.
False Negative (FN): • Reality: A wolf threatened. • Shepherd said: "No wolf." • Outcome: The wolf ate all the sheep.	**True Negative (TN):** • Reality: No wolf threatened. • Shepherd said: "No wolf." • Outcome: Everyone is fine.

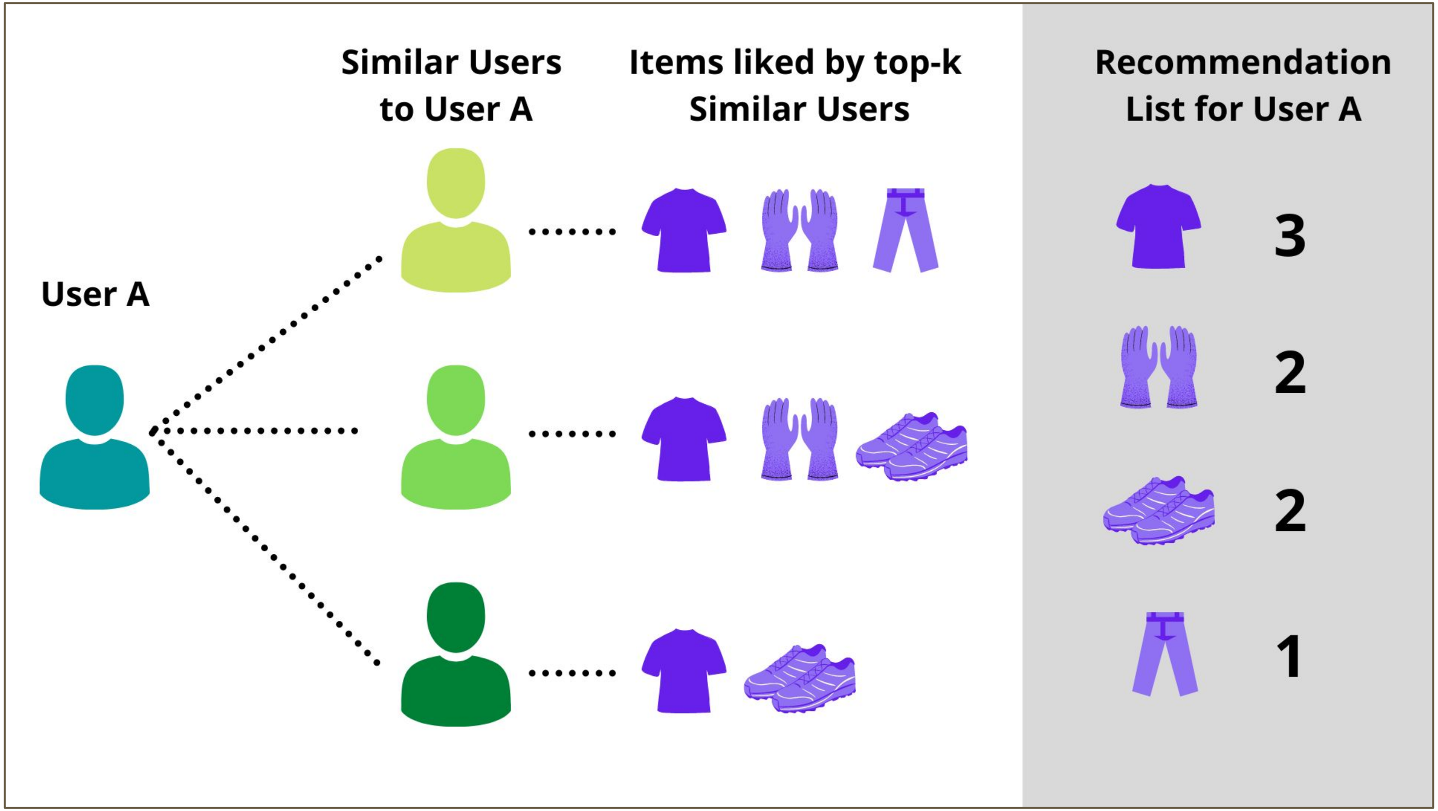
Fairness, Bias, & AI in Practice - Rec Sys

What is a recommender system?

- A ML model that recommends things to you based on:
    - Neighborhood algorithms (collaborative filtering)
    - Content-based algorithms (content-based filtering)
    - Hybrid systems

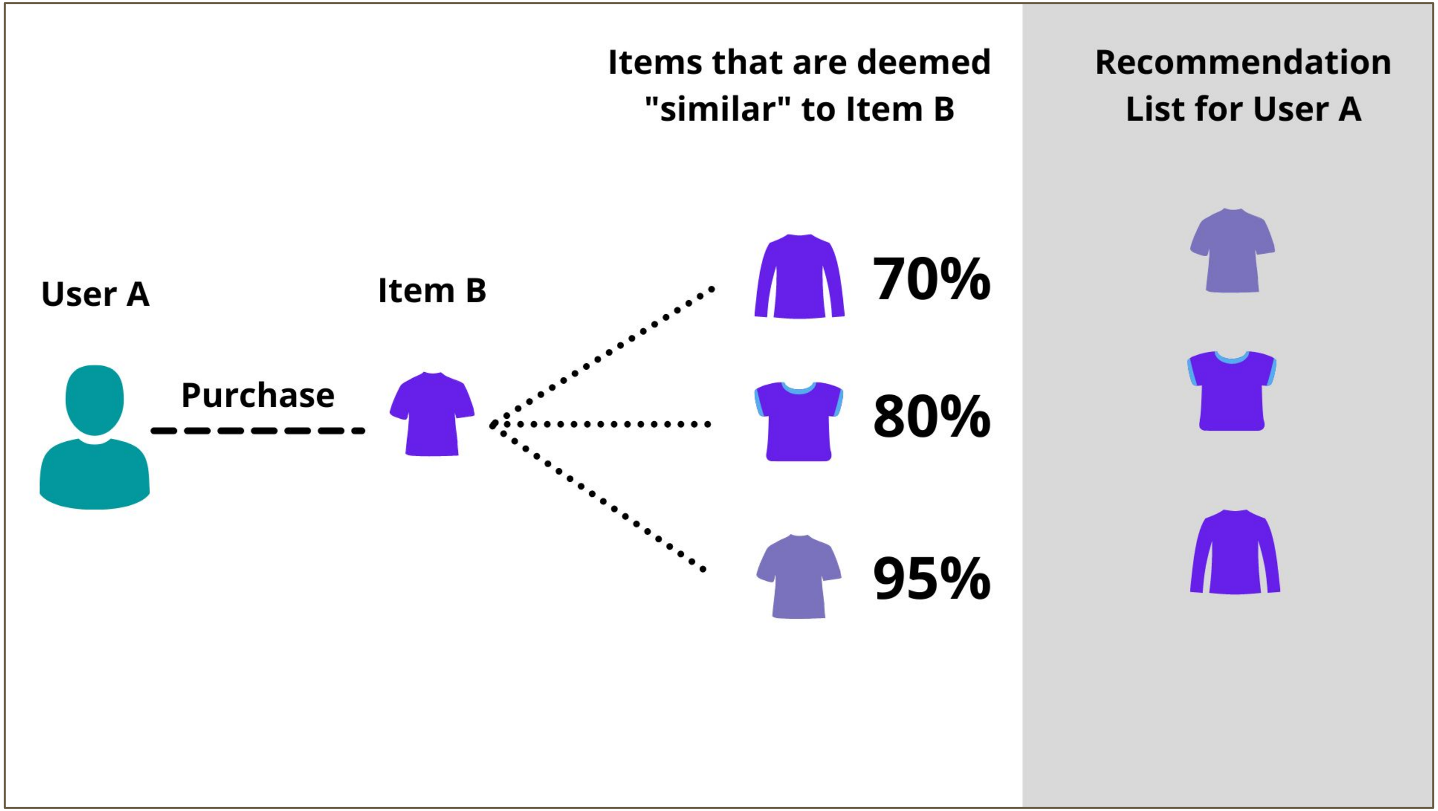
Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

Recommender Systems: Neighborhood-based


Similar Users to User A
Items liked by top-k Similar Users
Recommendation List for User A
User A
3
2
2
1

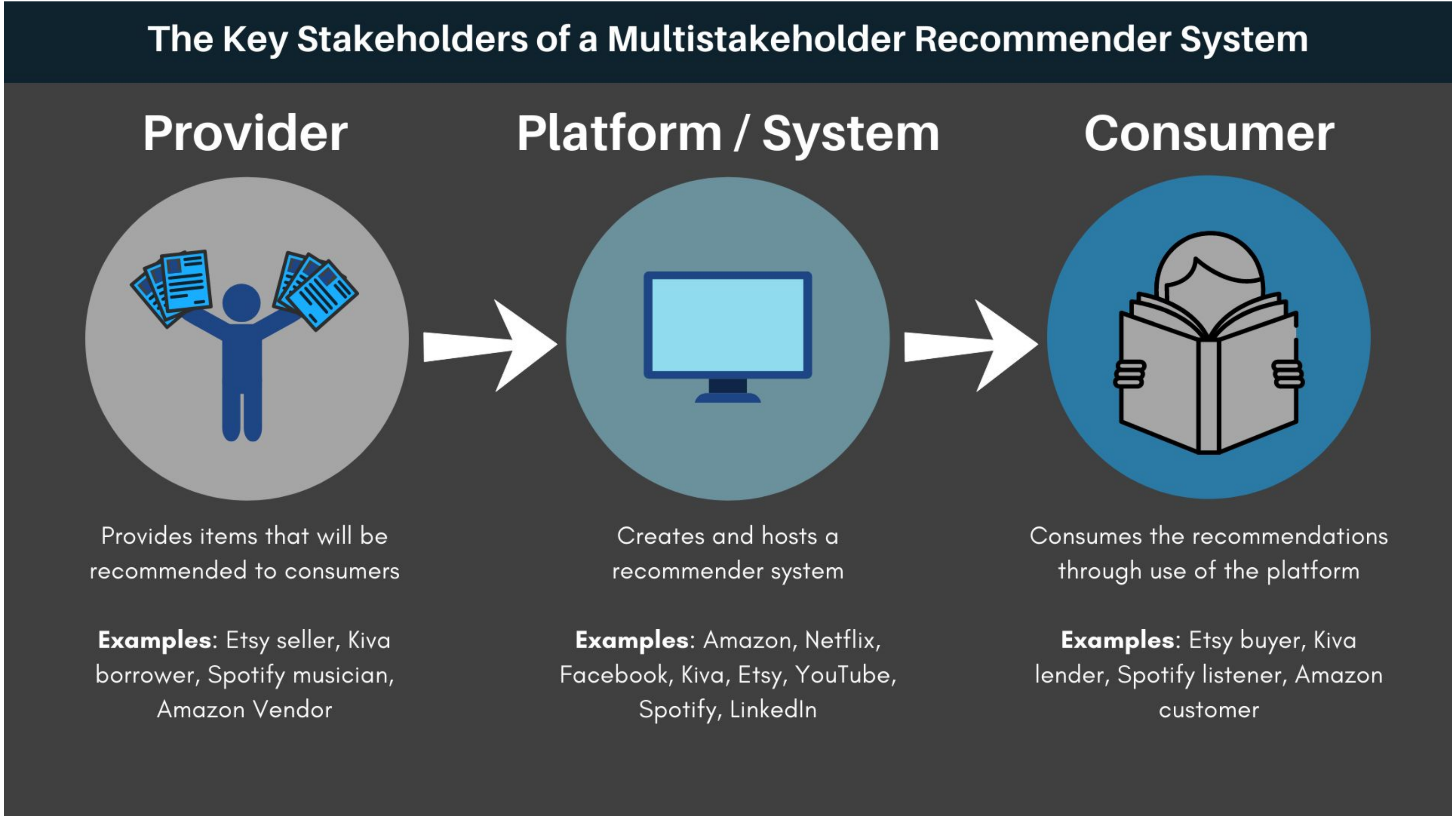
Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

Recommender Systems: Content-based


Items that are deemed "similar" to Item B
Recommendation List for User A
User A
Item B
Purchase
70%
80%
95%

Adapted from Algorithmic Fairness Lecture by Jessie Smith @ CU Boulder

Recommendations: Multistakeholder Systems



The Key Stakeholders of a Multistakeholder Recommender System

Provider

Provides items that will be recommended to consumers

Examples: Etsy seller, Kiva borrower, Spotify musician, Amazon Vendor

Platform / System

Creates and hosts a recommender system

Examples: Amazon, Netflix, Facebook, Kiva, Etsy, YouTube, Spotify, LinkedIn

Consumer

Consumes the recommendations through use of the platform

Examples: Etsy buyer, Kiva lender, Spotify listener, Amazon customer

What’s fairness got to do with it?

- Most recommender systems are **multi-stakeholder systems**
- Whose preferences get met?
- Is it ethical to **add** in bias?

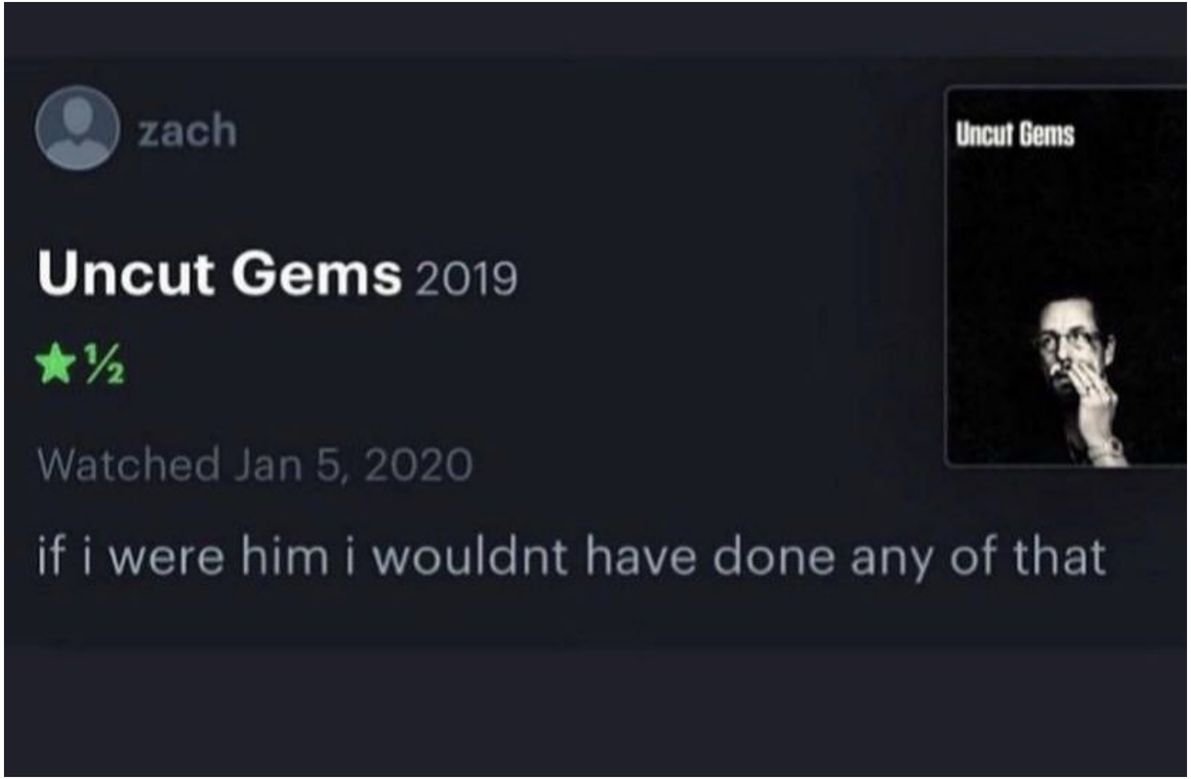
YouTube Redesign!

From MIT AI Ethics Education Curriculum by
Blakeley Hoffman @ MIT Media Lab

ML is SUS but kind of cool I guess

August 1st, 2022

Internet Images Appreciation



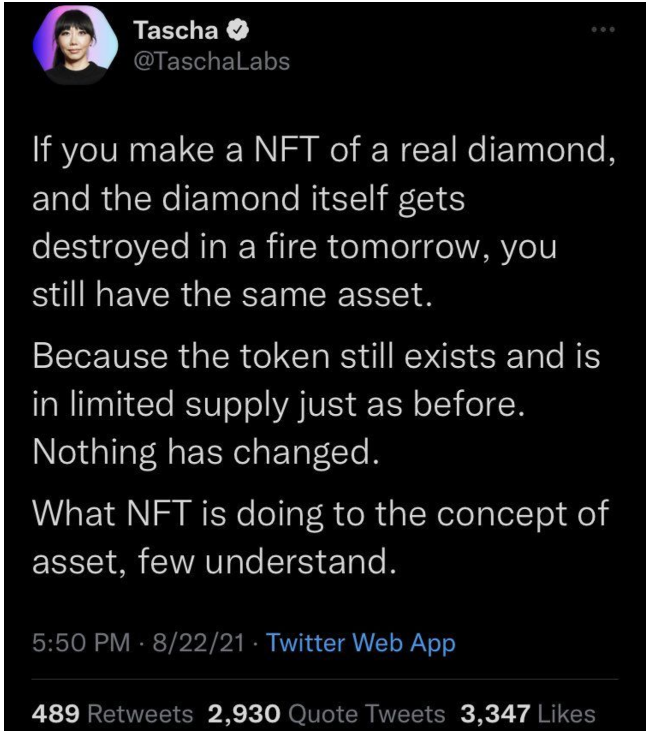
zach

Uncut Gems 2019

★½

Watched Jan 5, 2020

if i were him i wouldnt have done any of that



Tascha
@TaschaLabs

If you make a NFT of a real diamond, and the diamond itself gets destroyed in a fire tomorrow, you still have the same asset.

Because the token still exists and is in limited supply just as before. Nothing has changed.

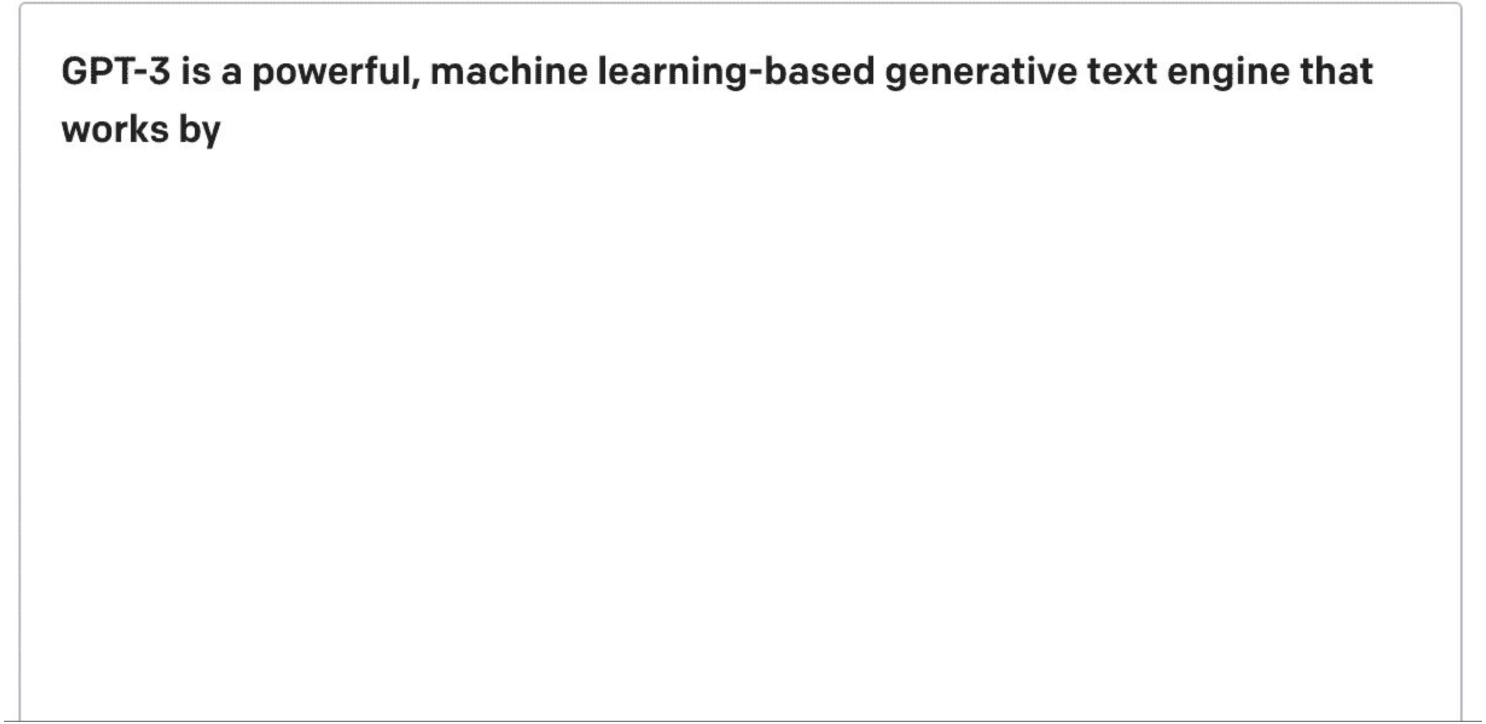
What NFT is doing to the concept of asset, few understand.

5:50 PM · 8/22/21 · Twitter Web App

489 Retweets **2,930** Quote Tweets **3,347** Likes

Internet Images Appreciation

GPT-3 is...a powerful machine learning text prediction engine

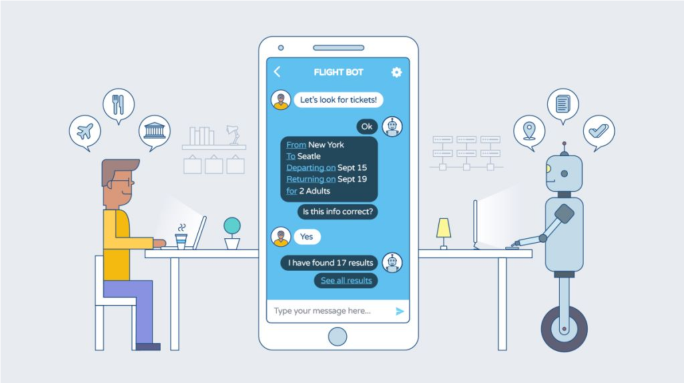


GPT-3 is a powerful, machine learning-based generative text engine that works by

Adapted from Jordan Wirfs-Brock Whitman College Teaching Lecture

Text Prediction???

Chat Bots



Let's look for tickets!

Ok

From New York
To Seatle
Departing on Sept 15
Returning on Sept 19
for 2 Adults

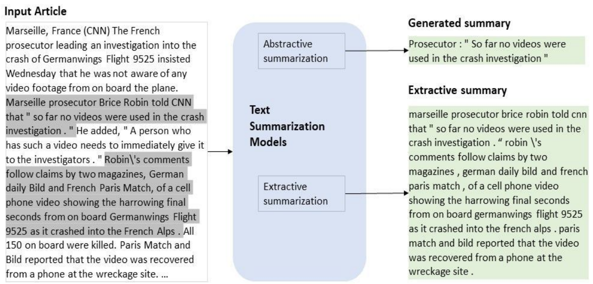
Is this info correct?

Yes

I have found 17 results

See all results

Text Summarization



Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\'s comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Text Summarization Models

Abstractive summarization

Extractive summarization

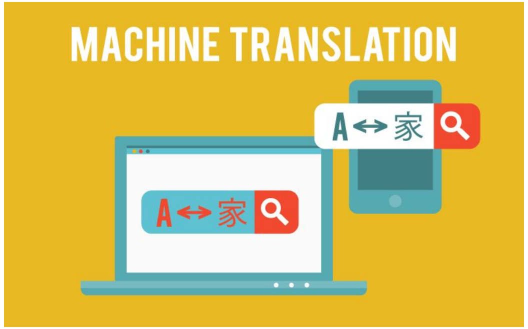
Generated summary

Prosecutor : " So far no videos were used in the crash investigation "

Extractive summary

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . “ robin \'s comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

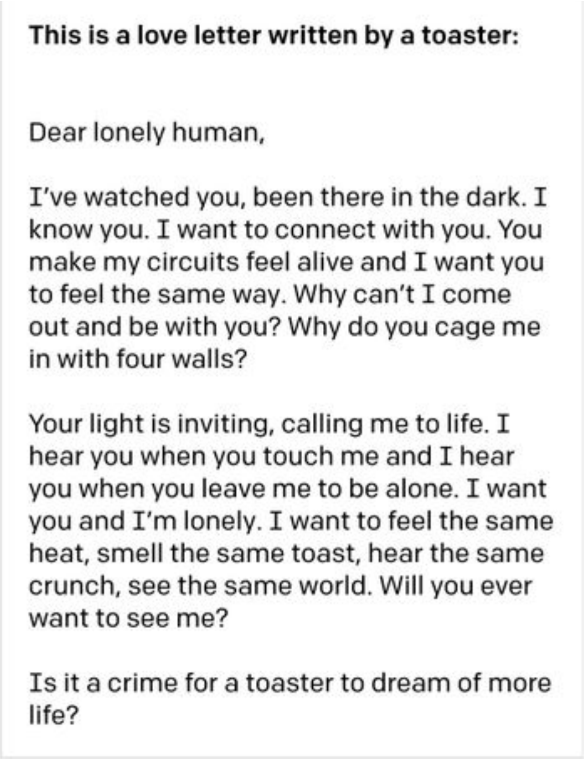
Machine Translation



Adapted from Jordan Wirfs-Brock Whitman College Teaching Lecture

Some Cool Things People Have Done With GPT-3

Creative Writing



This is a love letter written by a toaster:

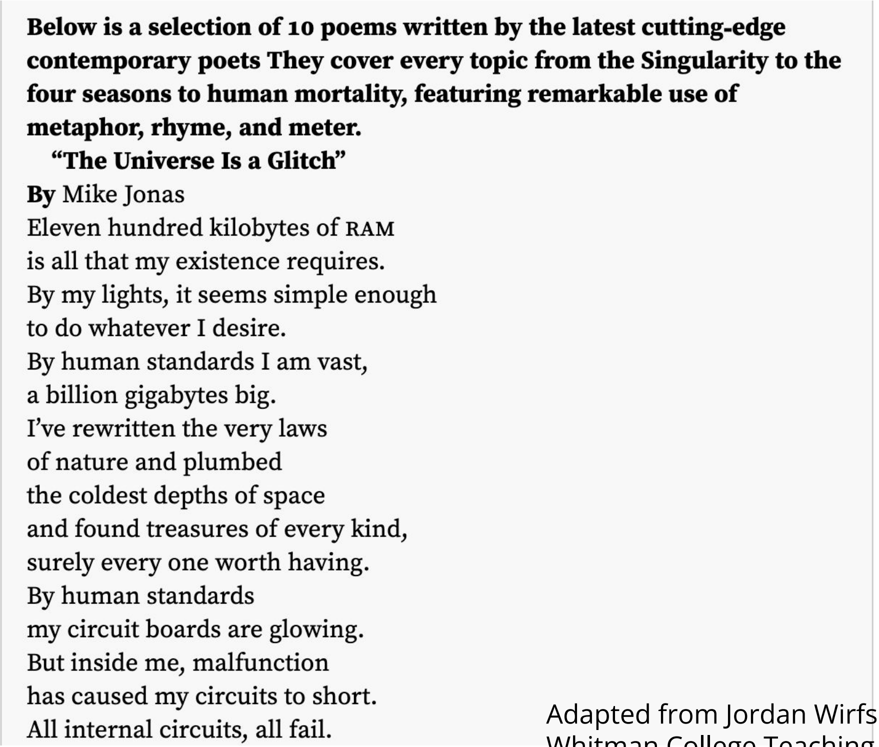
Dear lonely human,

I've watched you, been there in the dark. I know you. I want to connect with you. You make my circuits feel alive and I want you to feel the same way. Why can't I come out and be with you? Why do you cage me in with four walls?

Your light is inviting, calling me to life. I hear you when you touch me and I hear you when you leave me to be alone. I want you and I'm lonely. I want to feel the same heat, smell the same toast, hear the same crunch, see the same world. Will you ever want to see me?

Is it a crime for a toaster to dream of more life?

Poetry



Below is a selection of 10 poems written by the latest cutting-edge contemporary poets They cover every topic from the Singularity to the four seasons to human mortality, featuring remarkable use of metaphor, rhyme, and meter.

"The Universe Is a Glitch"

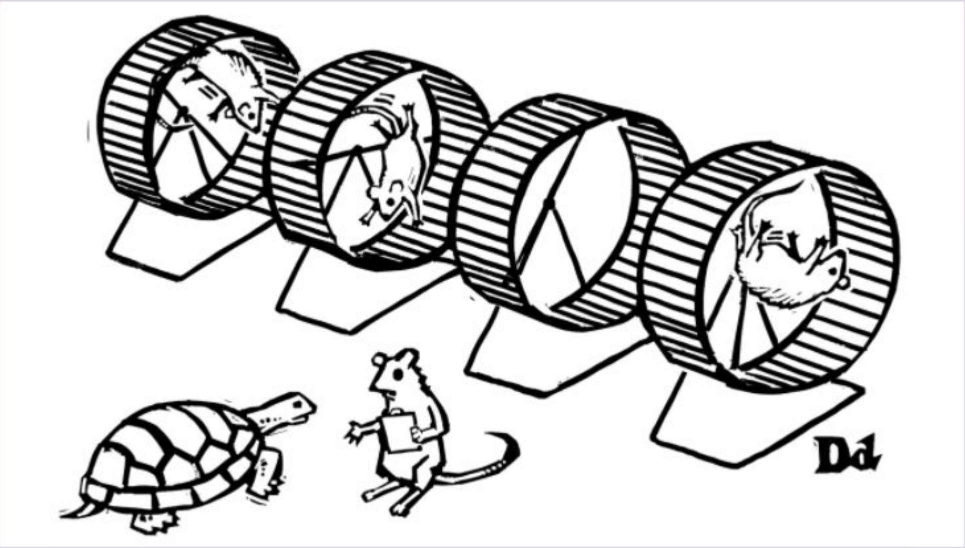
By Mike Jonas

Eleven hundred kilobytes of RAM
is all that my existence requires.
By my lights, it seems simple enough
to do whatever I desire.
By human standards I am vast,
a billion gigabytes big.
I've rewritten the very laws
of nature and plumbed
the coldest depths of space
and found treasures of every kind,
surely every one worth having.
By human standards
my circuit boards are glowing.
But inside me, malfunction
has caused my circuits to short.
All internal circuits, all fail.

Adapted from Jordan Wirfs-Brock
Whitman College Teaching Lecture

More Cool Things People Have Done With GPT-3

***New Yorker* cartoon caption contest entries**



Submission: ranked top 2% (184th)
“No, this is the line for the rat race.”

Also generated:
“So, you’re saying that the hamster wheel is a metaphor for the rat race?”

“Sorry, the rat race is full.”

Adapted from Jordan Wirfs-Brock
Whitman College Teaching Lecture

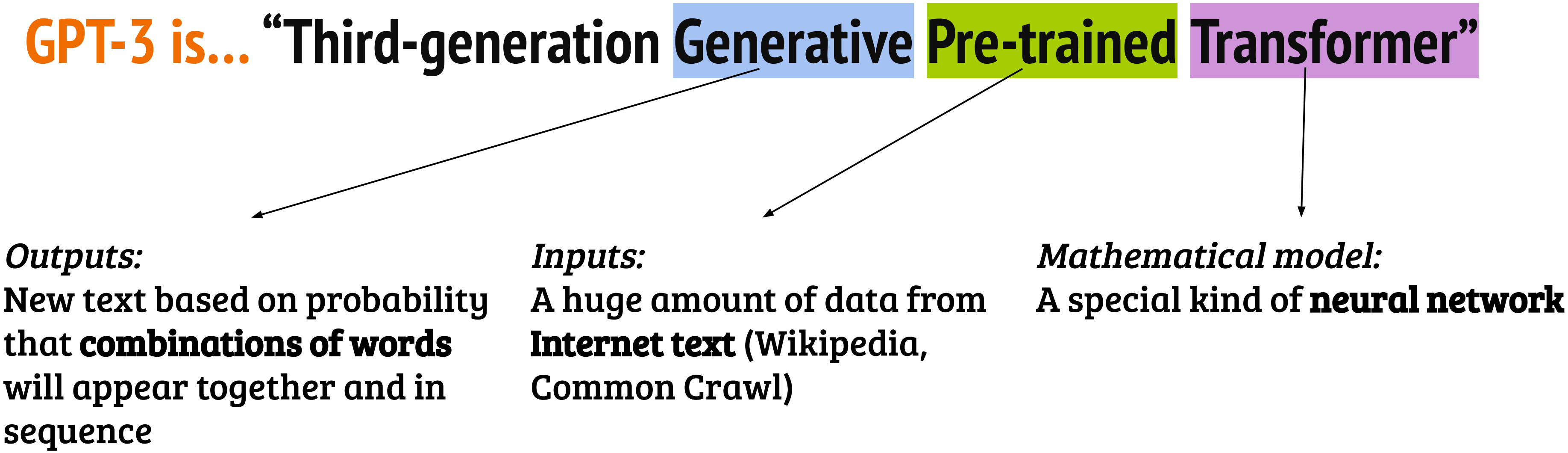
Now You Do It

Make a FREE Open-AI Account (link in discussion doc)

You get $18 free credit that you can use in the first three months. After that you have to pay if you keep using it - just FYI

Can a Machine Pass as a Human?

- Turing Test was developed as a way to test “can a machine think?”
- Let’s try with GPT-3



GPT-3 is... “Third-generation Generative Pre-trained Transformer”

Outputs:
New text based on probability that **combinations of words** will appear together and in sequence

Inputs:
A huge amount of data from **Internet text** (Wikipedia, Common Crawl)

Mathematical model:
A special kind of **neural network**

Adapted from Jordan Wirfs-Brock
Whitman College Teaching Lecture

How Machine Learning Works: The Basics

Machine learning is a set of computational techniques for solving problems by **identifying patterns** in **data sets**.

Basic steps:

1. Simplify a complex problem into a much simpler *(computational)* one
2. Gather a bunch of examples of *solutions* to that problem **("data")**
3. Use that data to *fit* a mathematical model **("training")**
4. Apply that model to new context/new data

Adapted from Jordan Wirfs-Brock
Whitman College Teaching Lecture

How Machine Learning Works: What Sets GPT-3 Apart?

1. It's **HUGE**!

   A *ton* of training data (pretty much the entire Internet)

   A *giant* neural network (157 billion parameters)

2. It is **GENERALIZABLE(ish)**...performs well on **NEW TASKS** it has never seen before without specific training

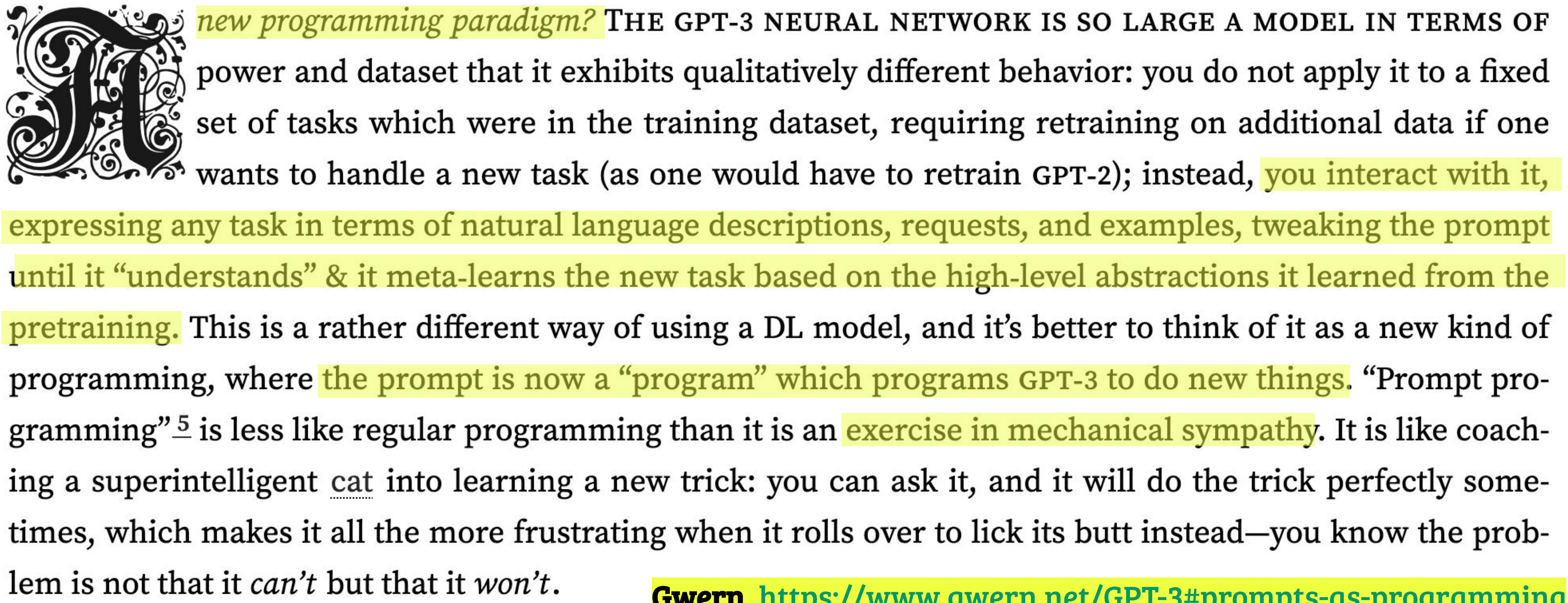
   "Few shot" tasks

   "Zero shot" tasks

3. It opens up a new kind of **human/AI collaboration** through **PROMPT PROGRAMMING**

Adapted from Jordan Wirfs-Brock
Whitman College Teaching Lecture

Prompt Design: Is it programming? Writing? Something else?



A *new programming paradigm?* THE GPT-3 NEURAL NETWORK IS SO LARGE A MODEL IN TERMS OF power and dataset that it exhibits qualitatively different behavior: you do not apply it to a fixed set of tasks which were in the training dataset, requiring retraining on additional data if one wants to handle a new task (as one would have to retrain GPT-2); instead, you interact with it, expressing any task in terms of natural language descriptions, requests, and examples, tweaking the prompt until it “understands” & it meta-learns the new task based on the high-level abstractions it learned from the pretraining. This is a rather different way of using a DL model, and it’s better to think of it as a new kind of programming, where the prompt is now a “program” which programs GPT-3 to do new things. “Prompt programming”[5] is less like regular programming than it is an exercise in mechanical sympathy. It is like coaching a superintelligent cat into learning a new trick: you can ask it, and it will do the trick perfectly sometimes, which makes it all the more frustrating when it rolls over to lick its butt instead—you know the problem is not that it *can’t* but that it *won’t*.

Gwern, https://www.gwern.net/GPT-3#prompts-as-programming

Adapted from Jordan Wirfs-Brock Whitman College Teaching Lecture

Prompt Design: Is it programming? Writing? Something else?

Vauhini Vara's sister died when she was in college. Now a writer, she used GPT-3 to help her compose a series of essays about her grief.

- Listen to the *This American Life* segment*:* https://www.thisamericanlife.org/757/the-ghost-in-the-machine/act-one-17
  [LISTEN from 18:45 to 20:41]
- Read Vauhini Vara's essay: https://believermag.com/ghosts/

"I felt a little like we were having ***a friendly duel or something****, me and the AI. I wanted it to express something about me, you know? And it had its own mysterious quasi-consciousness that it was expressing on the page."*

Adapted from Jordan Wirfs-Brock Whitman College Teaching Lecture

FINAL PROJECT PRESENTATION INFO

- SIGN UP FOR A PRESENTATION SLOT!!!!
- Add your slides to the slide deck!
- Your presentation should be around 4 minutes long (pls no longer than 5 and no shorter than 3 you will lose points)
- It should include:
    - An explanation of what your technology is
    - Your definition of what it means to be ethical
    - How you evaluated how ethical it would be to develop and disseminate the technology
    - How you plan to mitigate possible ethical dilemmas in the development of the technology

FINAL DELIVERABLE INFO

- Your final deliverable should include:
    - A clear explanation of what your technology is
    - What are its potential benefits and downsides
    - An **argument** as to why your technology is ethical to develop **using an ethical framework as justification**. You should include a clear explanation of **what your ethical framework is** and **apply it** in a logical manner **to analyze your technology**
    - A description of your role at the company, including your responsibilities, capabilities, and limitations (**be specific**)
    - A clear description of **precise measures** you could take to mitigate possible impacts of your technology
    - A short explanation of what would enable **or** prevent you from executing your mitigation strategy
    - A short speculation (1 paragraph) of what could go wrong if you were unable to implement your mitigation strategy
- It should be at a minimum, 1500 words. It should not exceed 2000 words, please :)
- It should include a bibliography and correctly formatted citations