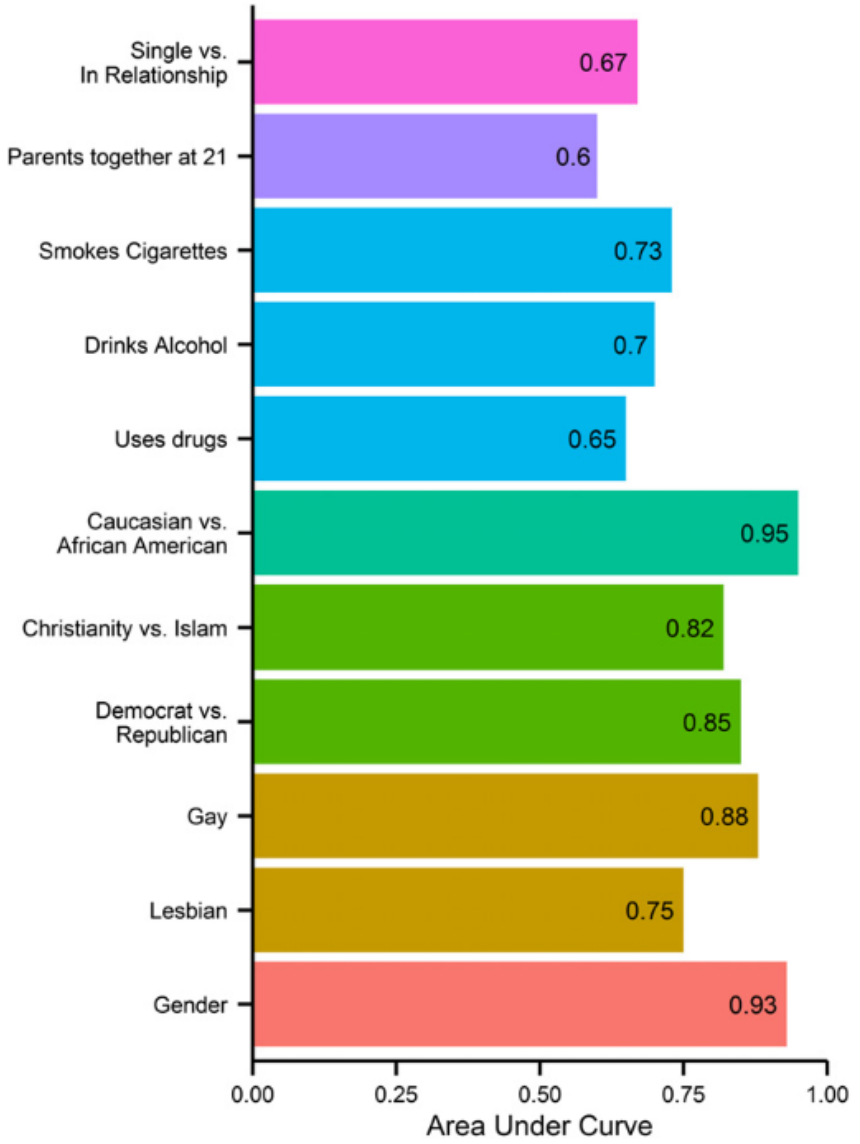
Ethics

Gautam Kamath

Privacy

- ML model memorization
  - Leaks information about the training data which might be sensitive
  - Solvable with differential privacy
- But… what if just *using* machine learning leaks privacy?
  - Apocryphal Target story
  - Smoking causes cancer
  - What do Facebook likes reveal about you?

Predicting features from Facebook likes



Most Predictive Likes

Trait		Selected most predictive Likes		
IQ	*High*	The Godfather	Jason Aldean	*Low*
		Mozart	Tyler Perry	
		Thunderstorms	Sephora	
		The Colbert Report	Chiq	
		Morgan Freemans Voice	Bret Michaels	
		The Daily Show	Clark Griswold	
		Lord Of The Rings	Bebe	
		To Kill A Mockingbird	I Love Being A Mom	
		Science	Harley Davidson	
		Curly Fries	Lady Antebellum	

Behaviours enabled by ML

- Deepfakes
- Can be harmless/artistic (https://www.youtube.com/watch?v=uAPUkgeiFVY)
- Can be harmful to individuals
  - Deepfake porn
- Can be… weird and creepy
  - Anthony Bourdain (https://youtu.be/jpkNYsuZScA?t=80)
  - Parkland shooting victim (https://youtu.be/m6I_wEetSck?t=47)

Behaviours enabled by ML

- ML powered fake news
- GPT-2 announced Feb ’19
- “Too dangerous to release”?
- Arguments for release
  - Obscurity isn’t safety, printing press, photoshop
- Some replications within 6 months
- Eventually released all models
- GPT-3? Licensed to Microsoft. Also usable… for a price.

	Mean accuracy	95% Confidence Interval (low, hi)	t compared to control (p-value)	“I don’t know” assignments
Control (deliberately bad model)	86%	83%–90%	-	3.6 %
GPT-3 Small	76%	72%–80%	3.9 (2e-4)	4.9%
GPT-3 Medium	61%	58%–65%	10.3 (7e-21)	6.0%
GPT-3 Large	68%	64%–72%	7.3 (3e-11)	8.7%
GPT-3 XL	62%	59%–65%	10.7 (1e-19)	7.5%
GPT-3 2.7B	62%	58%–65%	10.4 (5e-19)	7.1%
GPT-3 6.7B	60%	56%–63%	11.2 (3e-21)	6.2%
GPT-3 13B	55%	52%–58%	15.3 (1e-32)	7.1%
GPT-3 175B	52%	49%–54%	16.9 (1e-34)	7.8%

Table 3.11: Human accuracy in identifying whether short (∼200 word) news articles are model generated.

Tay, the Chatbot

- Released in March 2016
- Shut down 16 hours later

Bias

- Twitter cropping (https://twitter.com/bascule/status/1307440596668182528)

Bias

- Facial recognition (http://gendershades.org/overview.html)
  - Big consequences in justice system
  - IBM, Amazon, Microsoft have adjusted policy
  - Even if it had 100% accuracy, would it be OK?
- Hiring tools
  - Auto resume screening
    - Amazon determined it was unfairly discriminating based on resume keywords in self-audit
  - Interview video analysis

Bias

- COMPAS
  - Risk prediction assessment (recidivism)
  - Score from 1 to 10
  - Data via FOIA request
  - ProPublica Analysis

ProPublica

In [54]:

```
print("Black defendants")
is_afam = is_race("African-American")
table(list(filter(is_afam, recid)), list(filter(is_afam, surv)))
```

```
Black defendants
                Low     High
Survived        990     805     0.49
Recidivated     532     1369    0.51
Total: 3696.00
False positive rate: 44.85
False negative rate: 27.99
Specificity: 0.55
Sensitivity: 0.72
Prevalence: 0.51
PPV: 0.63
NPV: 0.65
LR+: 1.61
LR-: 0.51
```

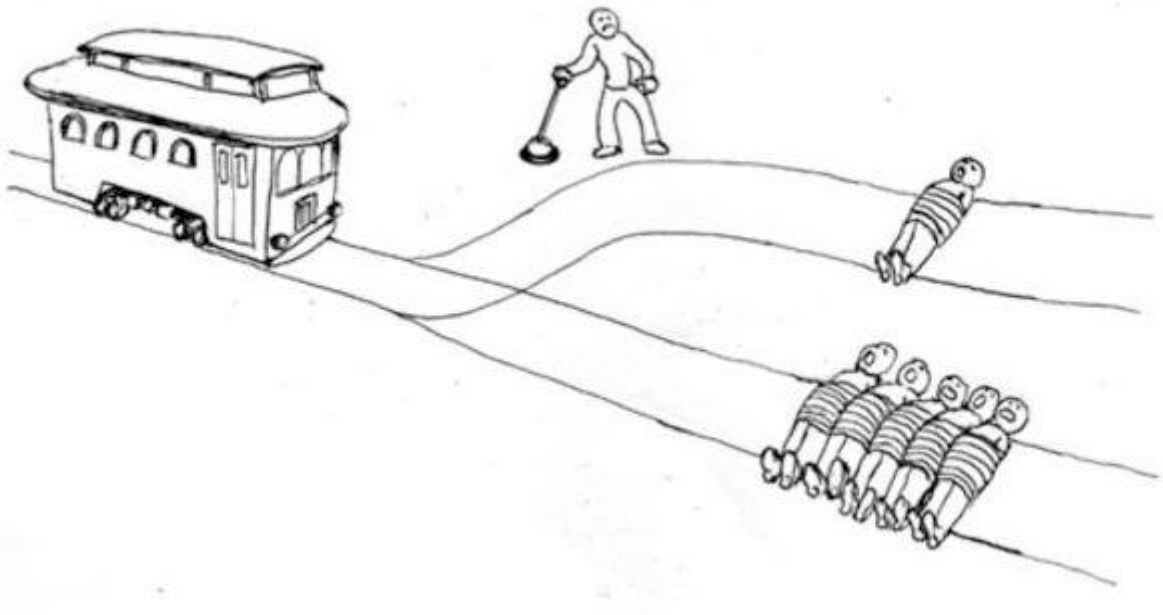
That number is higher for African Americans at 44.85%.

```
print("White defendants")
is_white = is_race("Caucasian")
table(list(filter(is_white, recid)), list(filter(is_white, surv
```

```
White defendants
                Low     High
Survived        1139    349     0.61
Recidivated     461     505     0.39
Total: 2454.00
False positive rate: 23.45
False negative rate: 47.72
Specificity: 0.77
Sensitivity: 0.52
Prevalence: 0.39
PPV: 0.59
NPV: 0.71
LR+: 2.23
LR-: 0.62
```

And lower for whites at 23.45%.

Philosophy



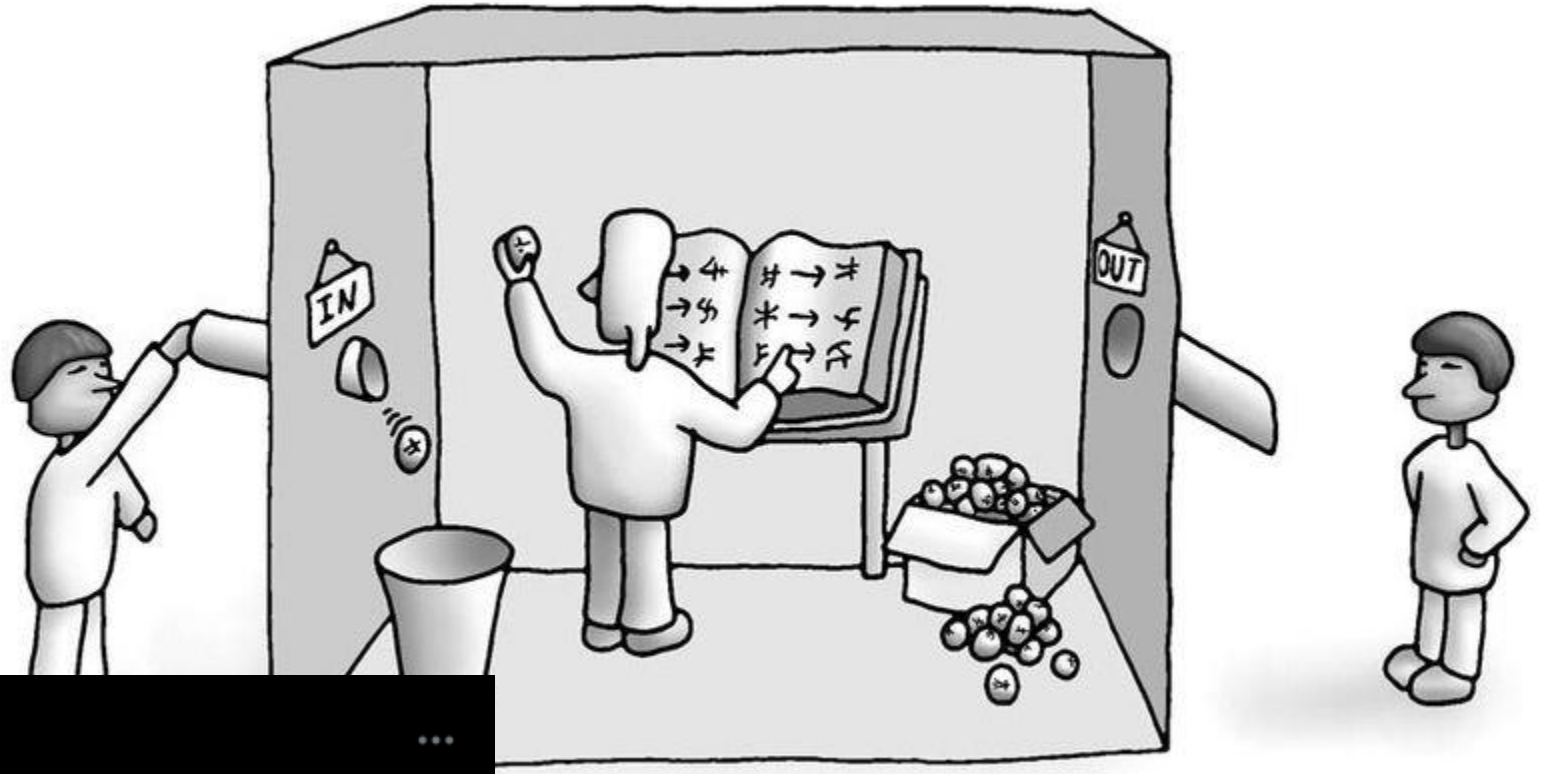
Philosophy

- Would you rather have:

1. A doctor with 90% accuracy, and they can explain to you why they came up with a diagnosis
2. AI is accurate with 95% accuracy, but it’s a black-box prediction

Philosophy

- Chinese Room
- Turing Test
- Consciousness?



Ilya Sutskever
@ilyasut

it may be that today's large neural networks are slightly conscious

6:27 PM · Feb 9, 2022 · Twitter Web App

525 Retweets **490** Quote Tweets **3,225** Likes