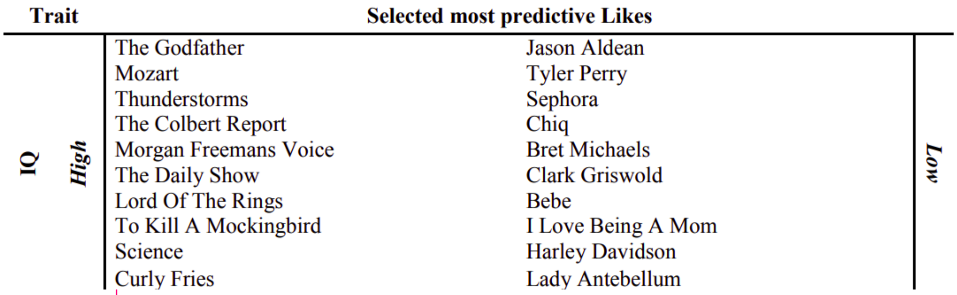
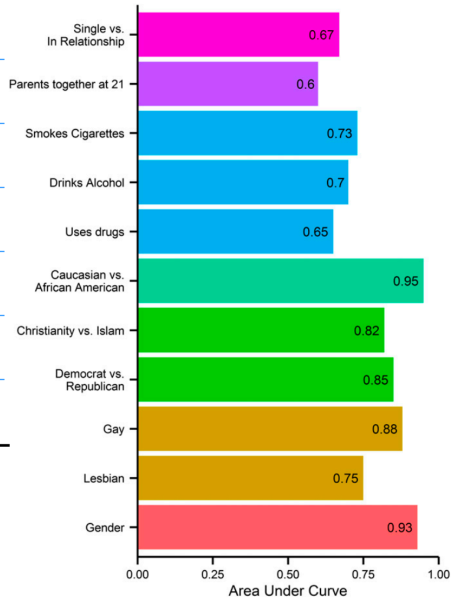
Ethics

Privacy

- ML Model Memorization (Can be addressed by DP)
- Target
- Smoking causes cancer
- Kosinski-Stillwell-Graepl '13
  Facebook Likes

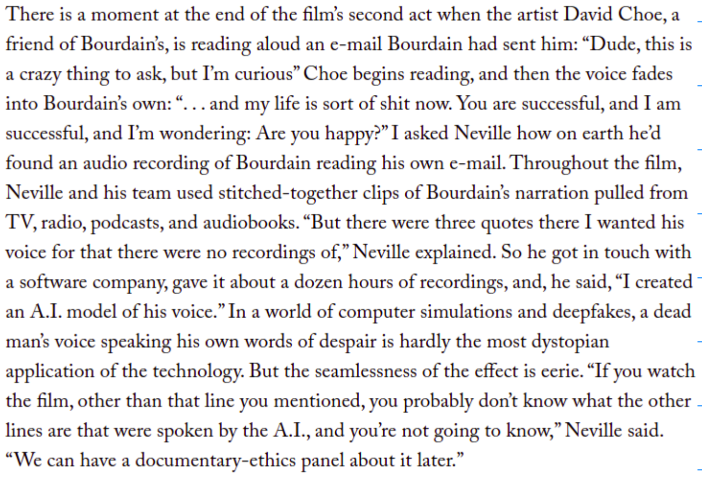


Trait		Selected most predictive Likes		
IQ	High	The Godfather	Jason Aldean	Low
		Mozart	Tyler Perry	
		Thunderstorms	Sephora	
		The Colbert Report	Chiq	
		Morgan Freemans Voice	Bret Michaels	
		The Daily Show	Clark Griswold	
		Lord Of The Rings	Bebe	
		To Kill A Mockingbird	I Love Being A Mom	
		Science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



Behaviours Enabled by ML

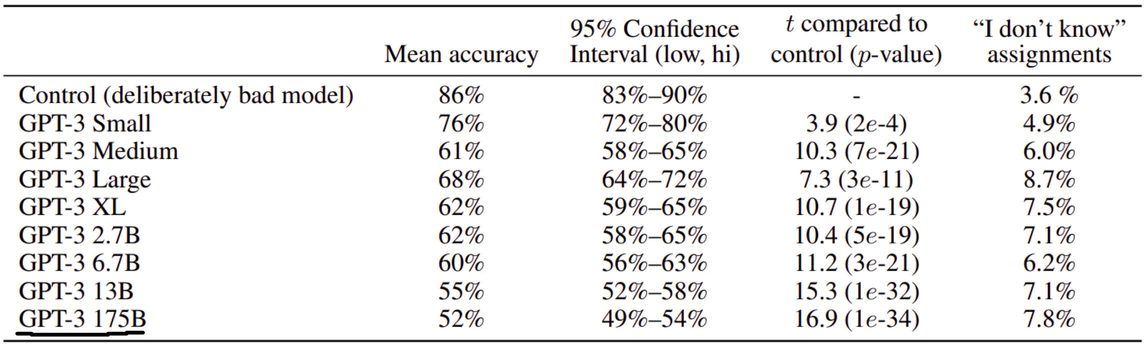
- Deepfakes



There is a moment at the end of the film's second act when the artist David Choe, a friend of Bourdain's, is reading aloud an e-mail Bourdain had sent him: "Dude, this is a crazy thing to ask, but I'm curious" Choe begins reading, and then the voice fades into Bourdain's own: ". . . and my life is sort of shit now. You are successful, and I am successful, and I'm wondering: Are you happy?" I asked Neville how on earth he'd found an audio recording of Bourdain reading his own e-mail. Throughout the film, Neville and his team used stitched-together clips of Bourdain's narration pulled from TV, radio, podcasts, and audiobooks. "But there were three quotes there I wanted his voice for that there were no recordings of," Neville explained. So he got in touch with a software company, gave it about a dozen hours of recordings, and, he said, "I created an A.I. model of his voice." In a world of computer simulations and deepfakes, a dead man's voice speaking his own words of despair is hardly the most dystopian application of the technology. But the seamlessness of the effect is eerie. "If you watch the film, other than that line you mentioned, you probably don't know what the other lines are that were spoken by the A.I., and you're not going to know," Neville said. "We can have a documentary-ethics panel about it later."

- Parkland Shooting victim

Fake News Generation



	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%

GPT-2 announced Feb '19 by OpenAI

- too dangerous to release
- arguments for release:
  - obscurity isn't safety
  - printing press, photoshopped

- Several replications (as early as Aug '19)
- Eventually released all models

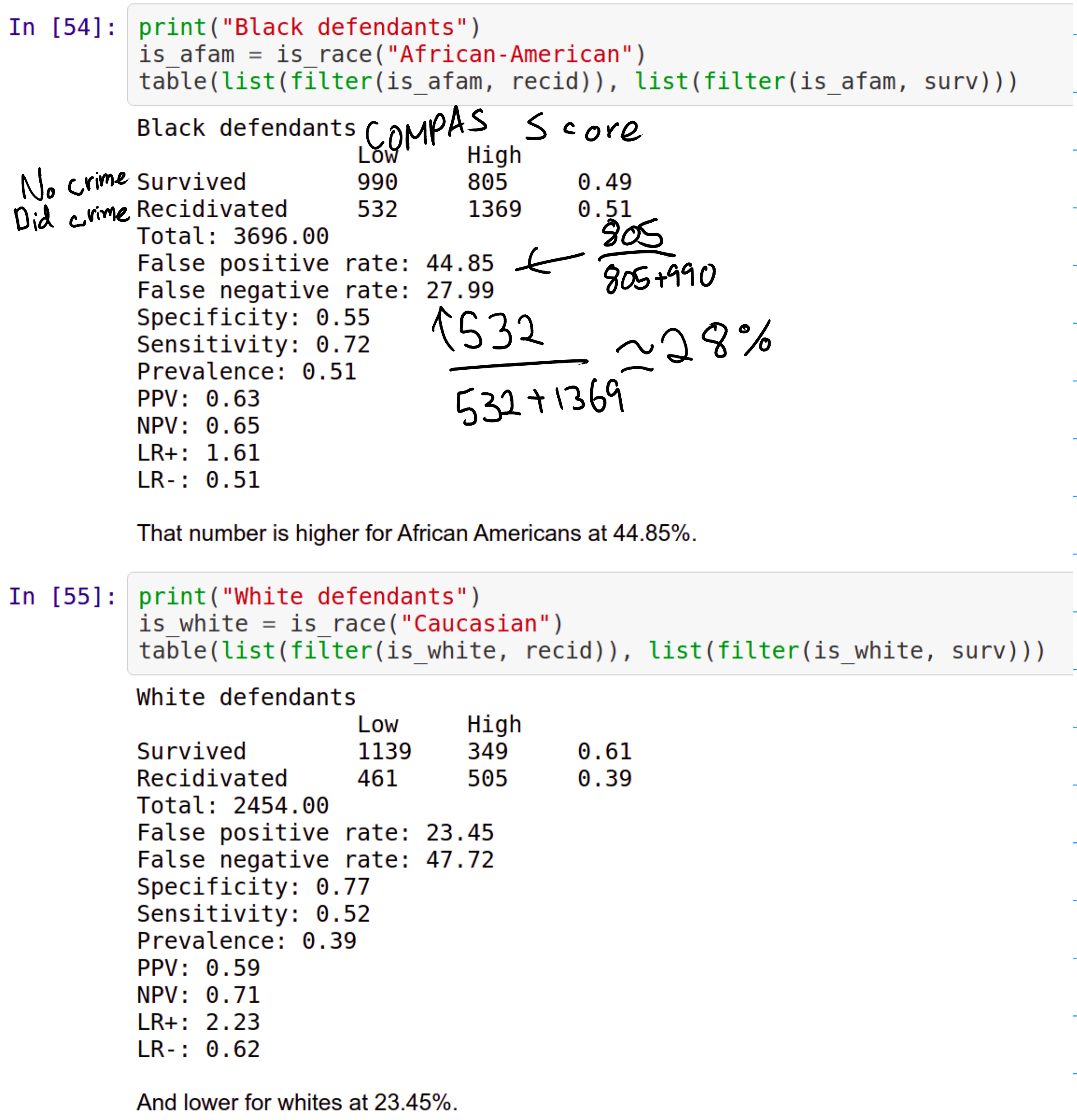
- GPT-3?, Licensed to Microsoft

Unexpected behaviour

- Tay, the chatbot
- Released 2016
- Taken down 16 hours later

Bias

- Twitter cropping
- Facial Recognition
  - ↳ Big consequences in justice system
  - IBM
  - Amazon
  - MS
- Hiring tools
  - ↳ Auto resume screening
  - ↳ Interview video analysis
- COMPAS
  - Risk prediction assessment
    Score from 1 to 10



In [54]:

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

COMPAS Score

No crime

Did crime

```
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
```

$\frac{805}{805+990}$

$\frac{532}{532+1369} \approx 28\%$

That number is higher for African Americans at 44.85%.

In [55]:

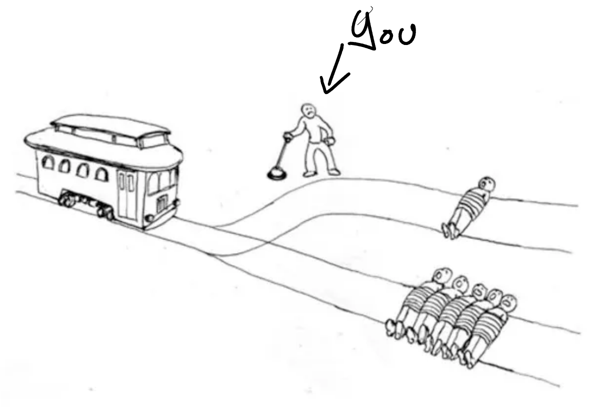
```
print("White defendants")
is_white = is_race("Caucasian")
table(list(filter(is_white, recid)), list(filter(is_white, surv)))
```

```
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

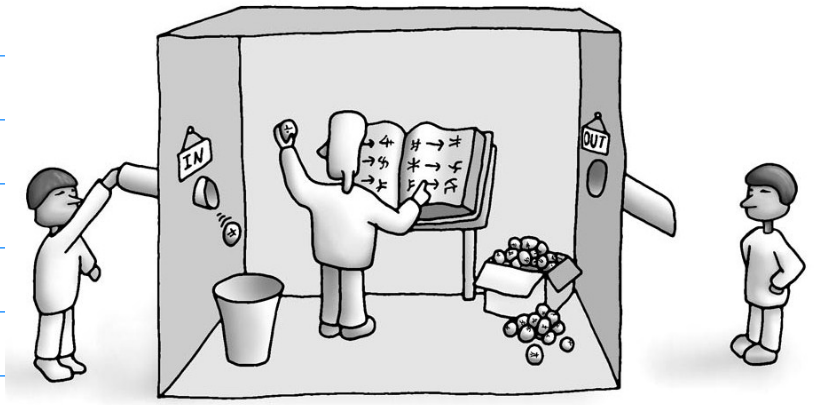
- Trolley Problem
- Selfdriving car



- a) Doctor w/ 90% accuracy. Tell you why they diagnose

b) AI w/ 95% accuracy. But it's a black box

- Is AI intelligent?



Turing Test