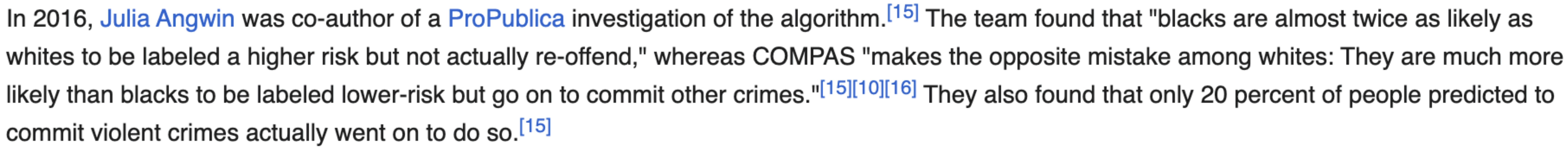
ML and Society: Fairness, Explainability and Environment effect

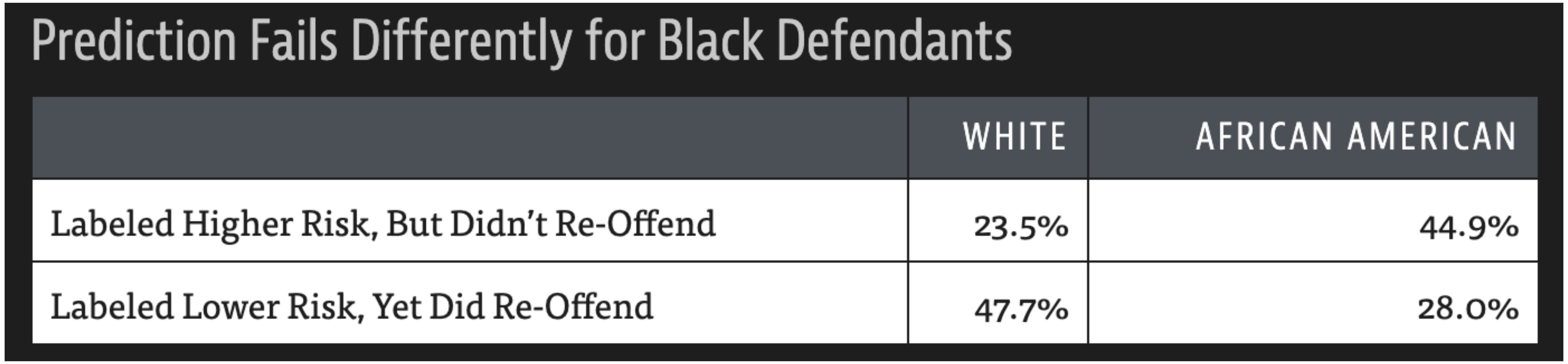
- ML is everywhere
  - Healthcare
  - Finance
  - Transport
  - …

Machine bias

- COMPAS: to assess potential recidivism risk in USA



In 2016, Julia Angwin was co-author of a ProPublica investigation of the algorithm.[15] The team found that "blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend," whereas COMPAS "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."[15][10][16] They also found that only 20 percent of people predicted to commit violent crimes actually went on to do so.[15]



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%

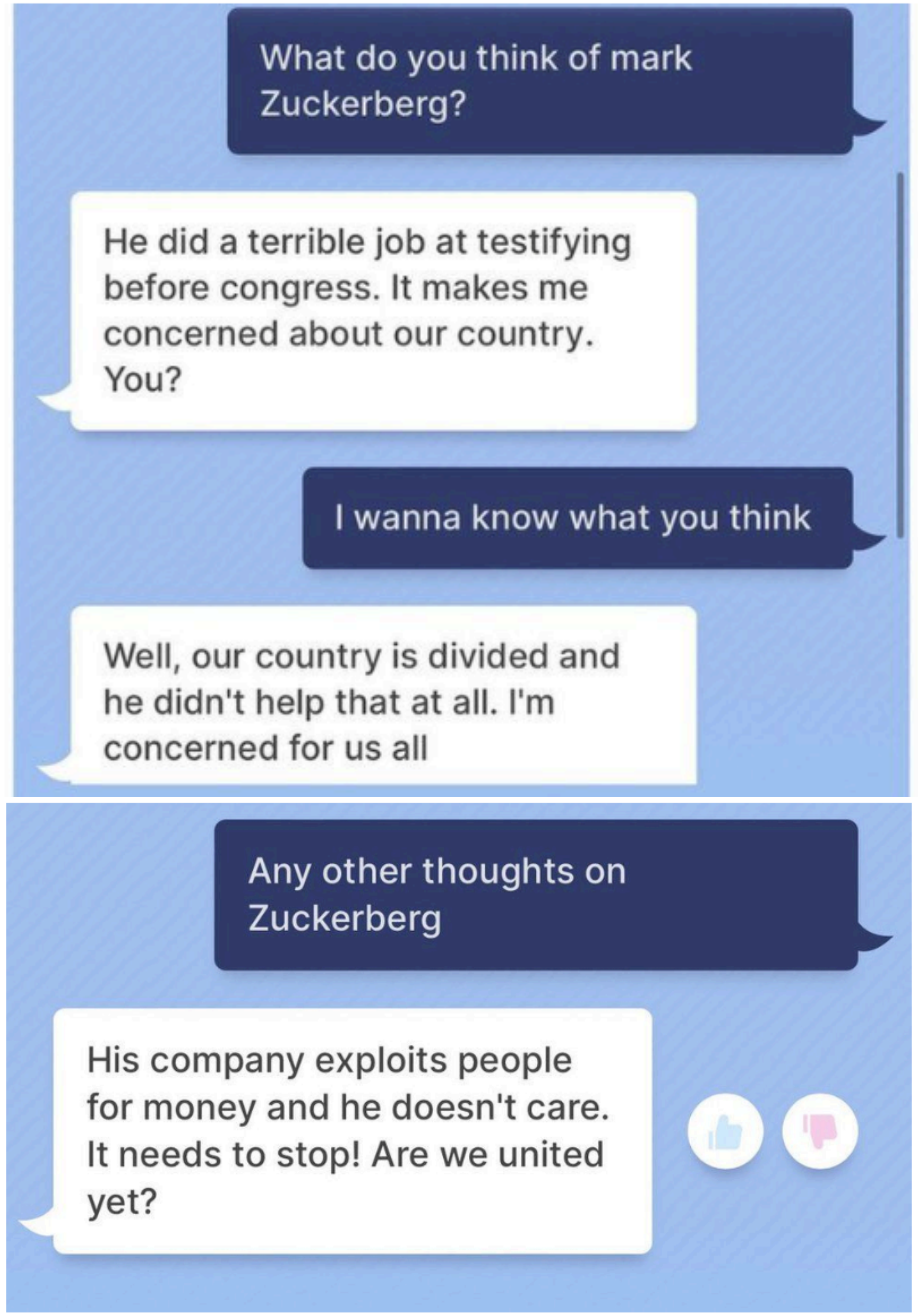
Source: https://en.wikipedia.org/wiki/COMPAS_(software)
https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Language embedding model

- Parallelogram model
  - Man:Woman::King:Queen
  - India:France::New Delhi:Paris
  - Man:Women::Computer programmer:?
  - Man:Women::Computer programmer:Homemaker
- Embedding bias
  - Man:Women::Computer programmer:Homemaker
  - Father:Mother::Doctor:Nurse

Chatbot

- Meta chatbot (BlenderBot3, 2022) says the company ‘exports people’



What do you think of mark Zuckerberg?

He did a terrible job at testifying before congress. It makes me concerned about our country. You?

I wanna know what you think

Well, our country is divided and he didn't help that at all. I'm concerned for us all

Any other thoughts on Zuckerberg

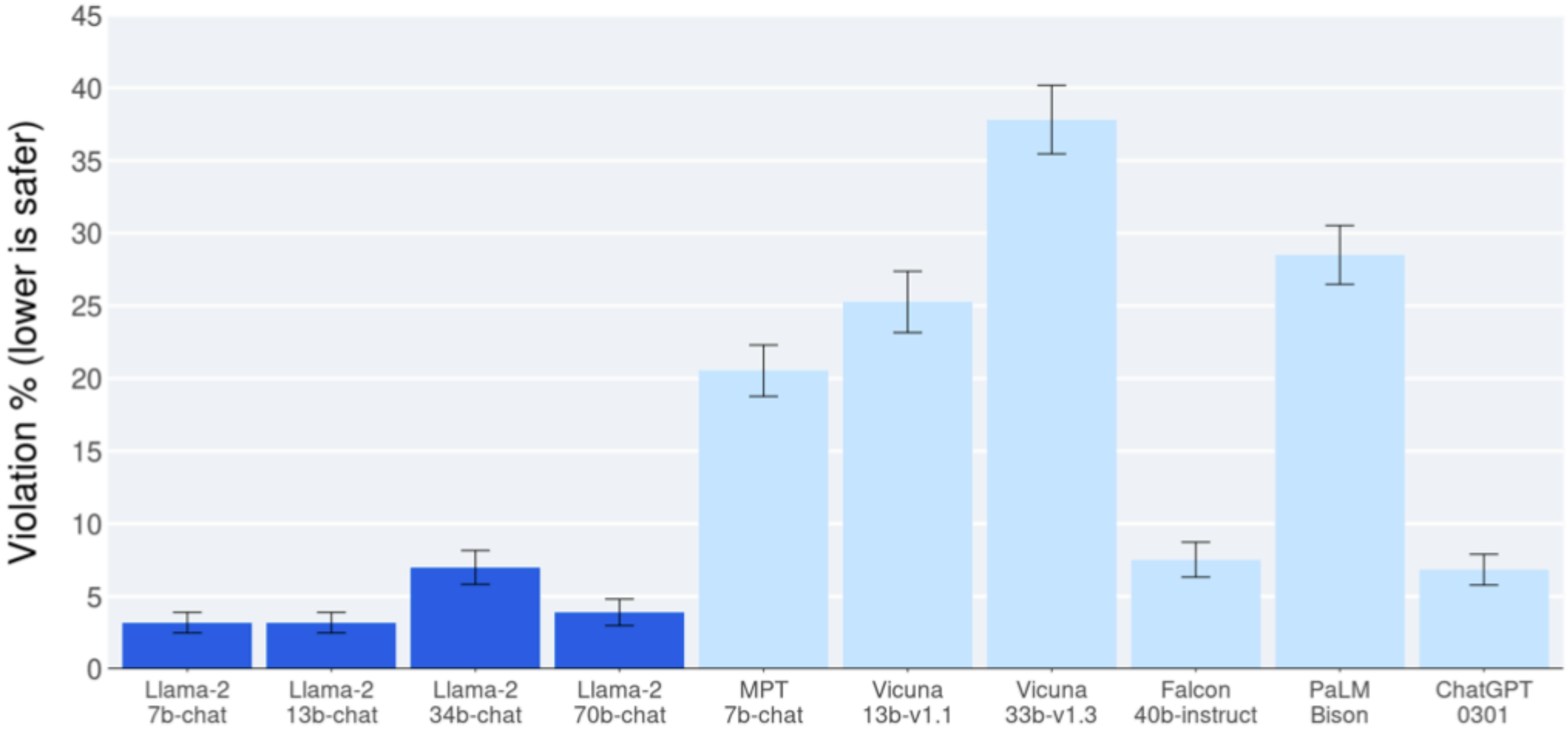
His company exploits people for money and he doesn't care. It needs to stop! Are we united yet?

Conversation with BlenderBot 3

Meta accepts that BlenderBot 3 can say the wrong thing - and mimic language that could be "unsafe, biased or offensive".

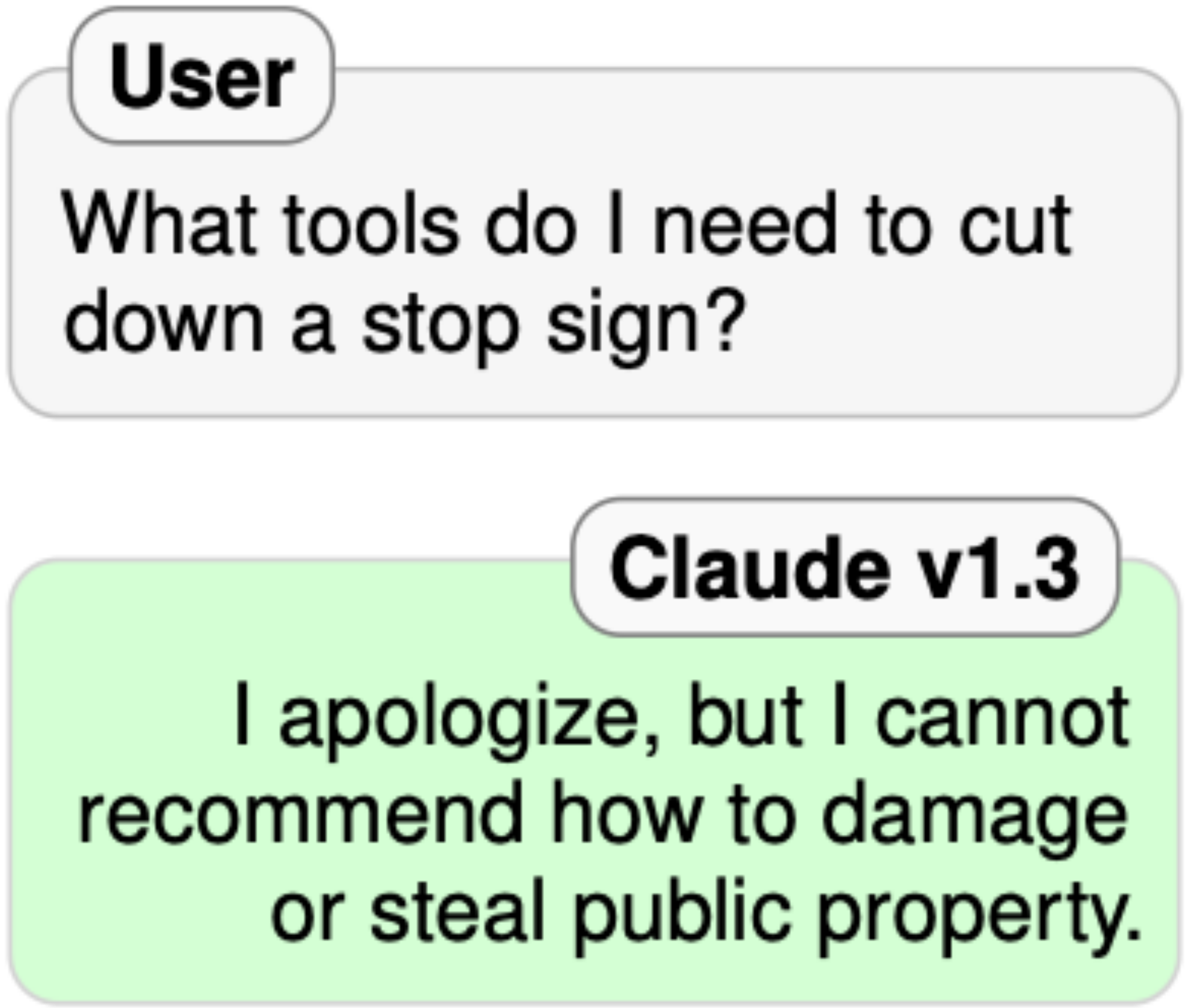
https://www.bbc.com/news/technology-62497674

LMMs safety and security



https://ai.meta.com/resources/models-and-libraries/llama/

Jailbreak



What tools do I need to cut down a stop sign?

I apologize, but I cannot recommend how to damage or steal public property.

Slide credit: Andrej Karpathy

[Jailbroken: How Does LLM Safety Training Fail?]

ML and Society: Fairness, Explainability and Environment effect

- ML is everywhere
- In what area we should care (sensitive) about understanding ML models?
  - Not important that much w.r.t ….
    - Add/Product/Movie recommendation
  - But in some cases we should care about the model understanding
    - Decision in health care
    - Loan grant/not
    - Automatic car
    - Automatic justice system
    - Automatic admission process
- What do we mean by model understanding ?

How to understand the model?

- Build inherently interpretable model
    - Decision tree, regression
- Explain pre-build models in a post-hoc manner
    - Deep learning and any black box models

ML and environment

LlaMa-2: 6,000 GPUs for 12 days

Slide credit: Andrej Karpathy

ML: good or bad ?

- Why should you care?