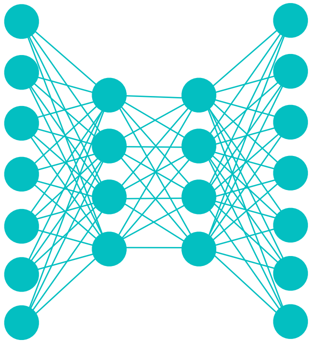
Lecture Notes for

Neural Networks and Machine Learning

Course Introduction
Lecture: AI Ethics

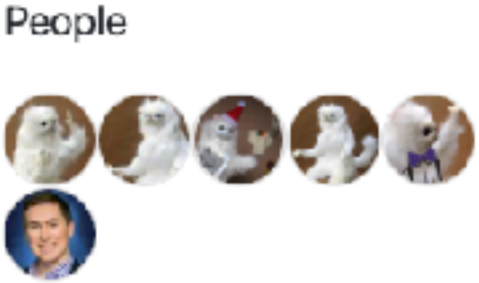


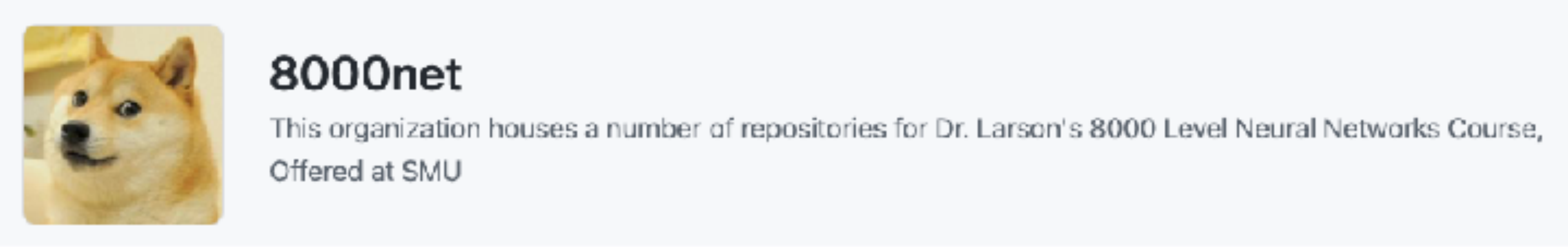
Logistics and Agenda

- Logistics
  - This class evolves across semesters (sometimes drastically!)
    - First offered in 2019
  - Using Canvas
  - GitHub: Mostly one repository
- Agenda
  - Syllabus and Introductions
  - Presentation Selection
  - Ethical Principles

Syllabus

- Course Schedule
- Reading/Videos
- GitHub
- Grading
  - Labs x4 (60%)
  - Final Pres. x1 (30%)
  - Participation (Pass/Fail)
  - Paper Presentation/Video x1 (10%)

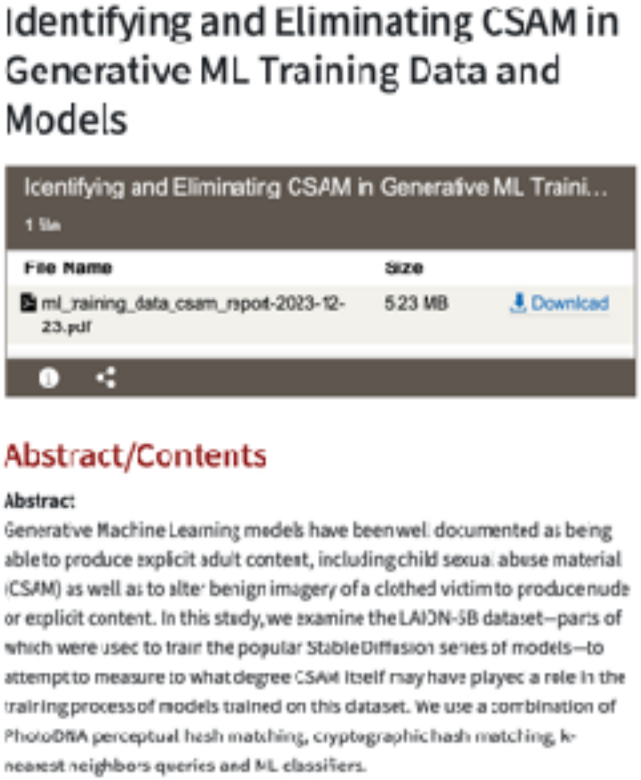

People


8000net

This organization houses a number of repositories for Dr. Larson's 8000 Level Neural Networks Course, Offered at SMU

Presenting OR Summary

- First Presentation is Next Week!
- During Semester: 7 Presentations Total (as a team)
- First Presentation →
- **Who wants to go first?**
  - ~10-15 Minutes
  - Summarize the Article
  - Make 3-5 Visuals
    - e.g., Slides
    - AND/OR Handouts
    - AND/OR Notebooks



Identifying and Eliminating CSAM in Generative ML Training Data and Models

Identifying and Eliminating CSAM in Generative ML Traini...

1 file

File Name	Size	
ml_training_data_csam_report-2023-12-23.pdf	5.23 MB	Download

Abstract/Contents

Abstract:

Generative Machine Learning models have been well documented as being able to produce explicit adult content, including child sexual abuse material (CSAM) as well as to alter benign imagery of a clothed victim to produce nude or explicit content. In this study, we examine the LAION-5B dataset—parts of which were used to train the popular Stable Diffusion series of models—to attempt to measure to what degree CSAM itself may have played a role in the training process of models trained on this dataset. We use a combination of PhotoDNA perceptual hash matching, cryptographic hash matching, k-nearest neighbors queries and ML classifiers.

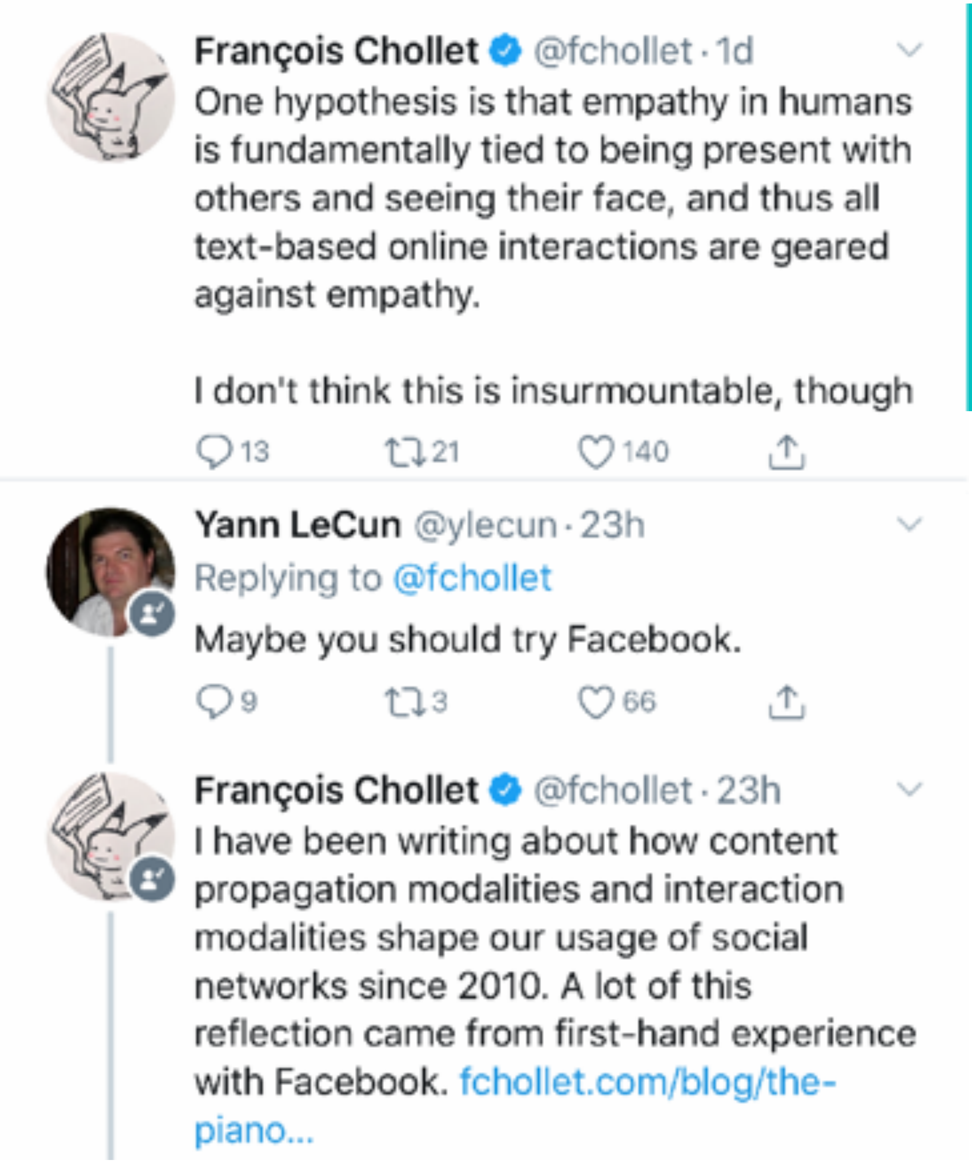
- Alternative: Video Summary of paper, with visuals

Introductions

- Name
- Department
- Where you grew up
- When you took 7324 and the Topic in this course you are most excited about
- Something true or false about you

- Do NOT forget:
  - Pick out papers on Canvas (distance students also)

Ethical ML



François Chollet @fchollet · 1d

One hypothesis is that empathy in humans is fundamentally tied to being present with others and seeing their face, and thus all text-based online interactions are geared against empathy.

I don't think this is insurmountable, though

13 21 140

Yann LeCun @ylecun · 23h

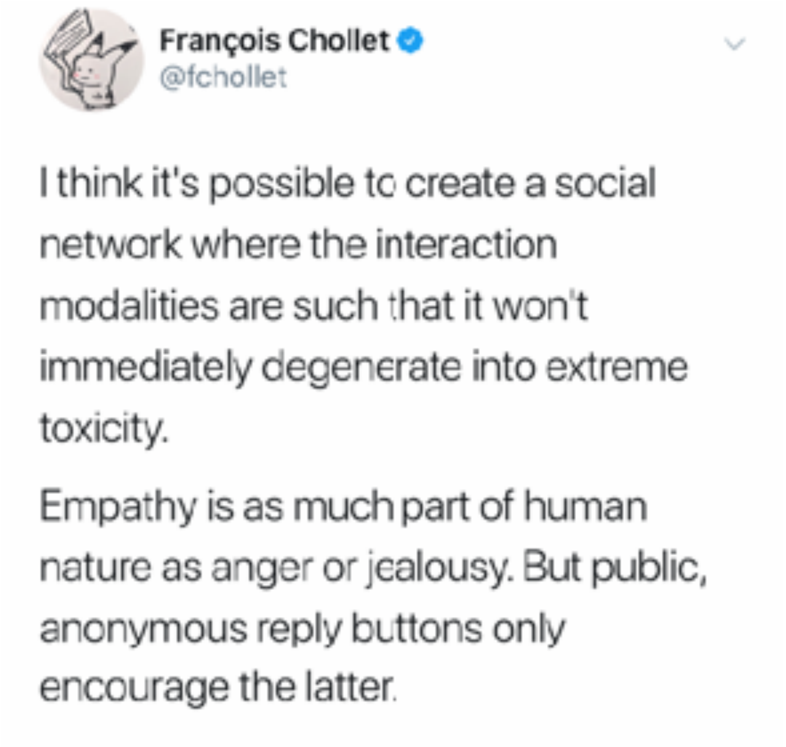
Replying to @fchollet

Maybe you should try Facebook.

9 3 66

François Chollet @fchollet · 23h

I have been writing about how content propagation modalities and interaction modalities shape our usage of social networks since 2010. A lot of this reflection came from first-hand experience with Facebook. fchollet.com/blog/the-piano...



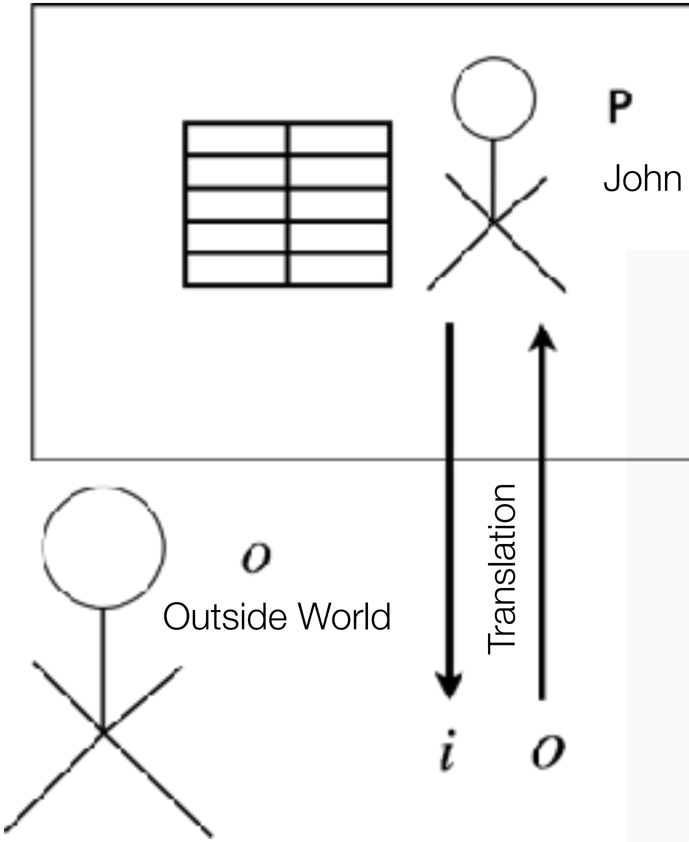
François Chollet
@fchollet

I think it's possible to create a social network where the interaction modalities are such that it won't immediately degenerate into extreme toxicity.

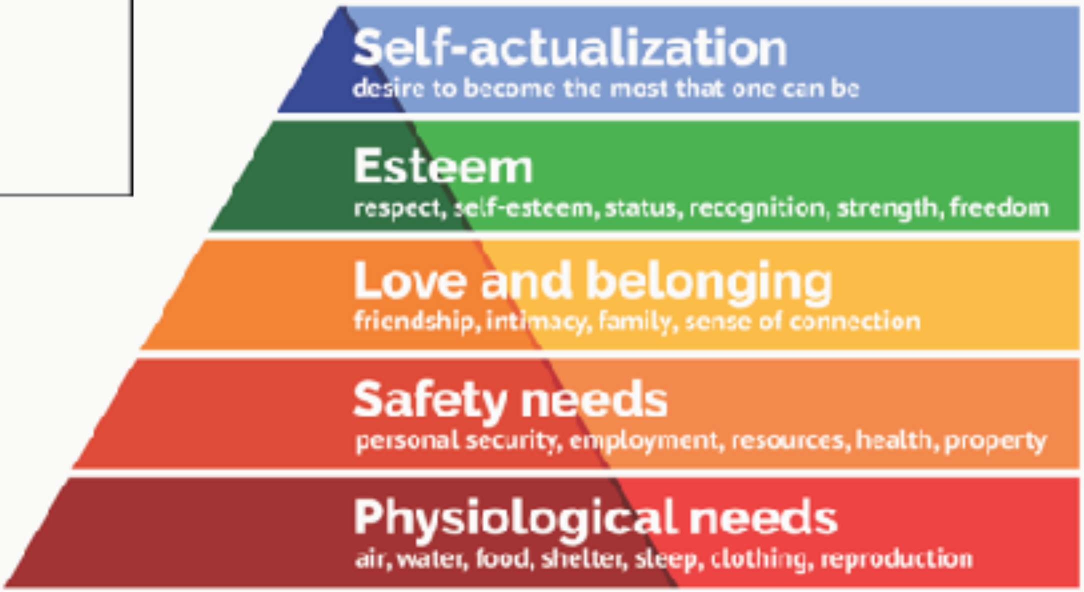
Empathy is as much part of human nature as anger or jealousy. But public, anonymous reply buttons only encourage the latter.

Strong AI, i.e., machines and thinking

- John Searle’s Foreign Room Argument:
  - Can John ever understand what he is saying?



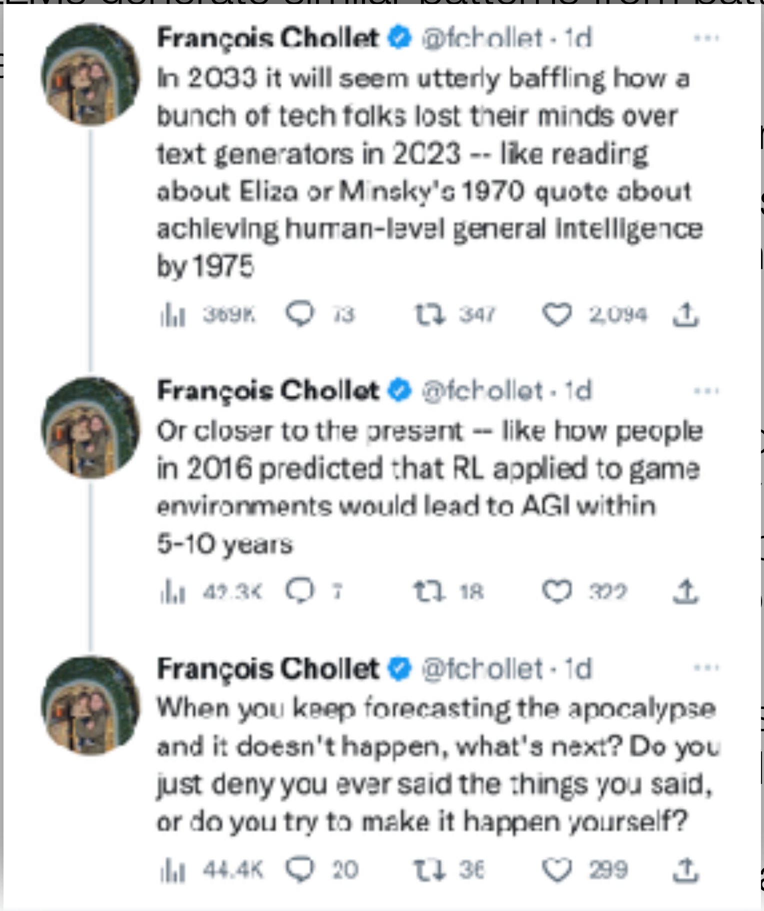
- If always translating without mistakes, even then we cannot be sure if what is inside truly understands what the output is
  - Humans share a need that drives our communications and interactions:


Self-actualization
desire to become the most that one can be
Esteem
respect, self-esteem, status, recognition, strength, freedom
Love and belonging
friendship, intimacy, family, sense of connection
Safety needs
personal security, employment, resources, health, property
Physiological needs
air, water, food, shelter, sleep, clothing, reproduction

Maslow’s Pyramid of Human Need

Can machines think?

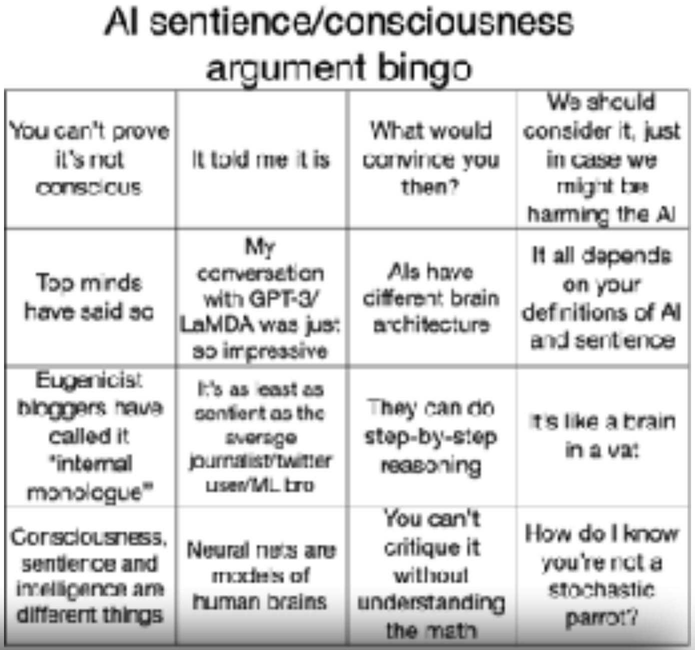
- 🦜LLMs generate similar patterns from patterns they have
- Is th mans do?
  - s and d
  - is Yes. op have no plex attern s back, d. lar anding



François Chollet @fchollet · 1d
In 2033 it will seem utterly baffling how a bunch of tech folks lost their minds over text generators in 2023 -- like reading about Eliza or Minsky's 1970 quote about achieving human-level general intelligence by 1975
369K 73 347 2,094

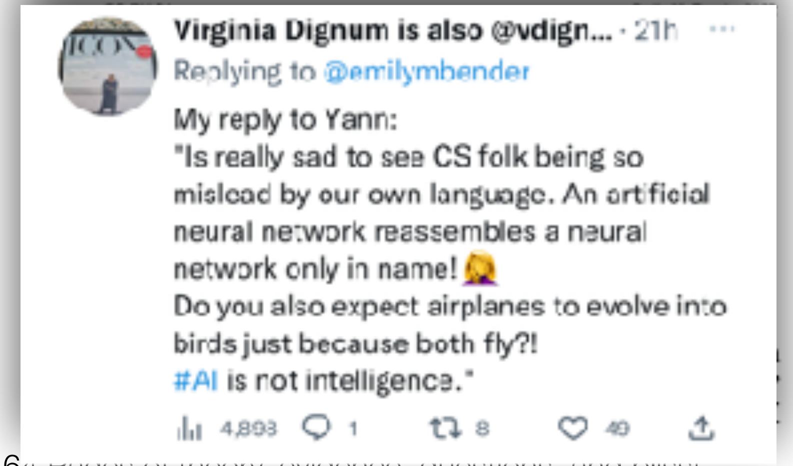
François Chollet @fchollet · 1d
Or closer to the present -- like how people in 2016 predicted that RL applied to game environments would lead to AGI within 5-10 years
42.3K 7 18 322

François Chollet @fchollet · 1d
When you keep forecasting the apocalypse and it doesn't happen, what's next? Do you just deny you ever said the things you said, or do you try to make it happen yourself?
44.4K 20 36 299



AI sentience/consciousness argument bingo

You can't prove it's not conscious	It told me it is	What would convince you then?	We should consider it, just in case we might be harming the AI
Top minds have said so	My conversation with GPT-3/ LaMDA was just so impressive	AIs have different brain architecture	It all depends on your definitions of AI and sentience
Eugenicist bloggers have called it "internal monologue"	It's as least as sentient as the average journalist/twitter user/ML bro	They can do step-by-step reasoning	It's like a brain in a vat
Consciousness, sentience and intelligence are different things	Neural nets are models of human brains	You can't critique it without understanding the math	How do I know you're not a stochastic parrot?

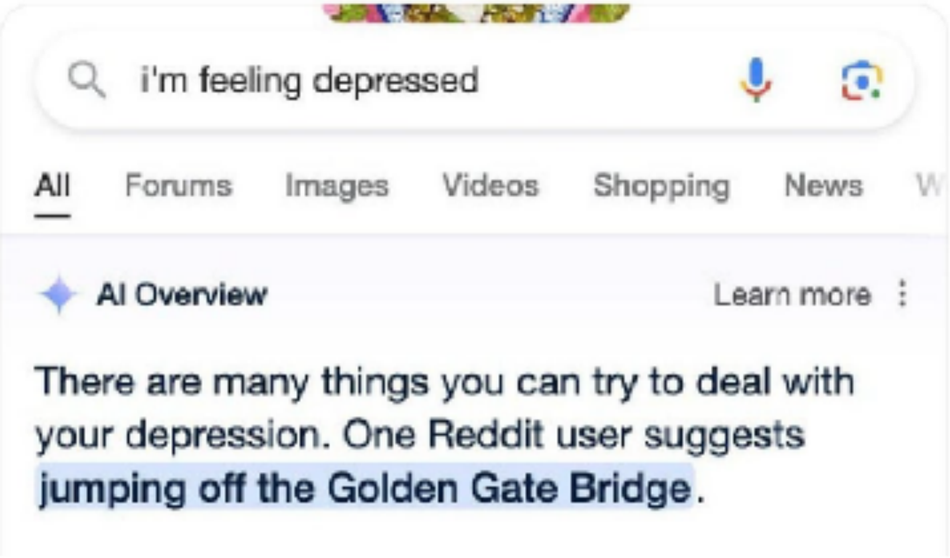


Virginia Dignum is also @vdign... · 21h
Replying to @emilymbender
My reply to Yann:
"Is really sad to see CS folk being so mislead by our own language. An artificial neural network reassembles a neural network only in name! 🤦‍♀️
Do you also expect airplanes to evolve into birds just because both fly?!
#AI is not intelligence."
4,893 1 8 49

64 Pages of theory, evidence, questions, and bliss!

- We impose sentience on machines. Human brains are nothing like neural networks.

Ethical Principles



i'm feeling depressed

There are many things you can try to deal with your depression. One Reddit user suggests jumping off the Golden Gate Bridge.

Learning about neural networks and human consciousness
So Sentient AI is a myth?
Always has been
Programmers

The Google AI Principles

- Be socially **beneficial**
- **Avoid** creating or reinforcing **unfair** bias
- Be built and tested for **safety**
- Be **accountable** to people
- Incorporate **privacy** design principles
- Uphold high standards of scientific excellence
- Be **made available** for uses that accord with these principles
- **Google will not pursue**:
    - Tech likely to cause **harm**, tech that **principally** is a **weapon**, Tech that violates **surveillance** norms, Tech that contravenes **human rights**

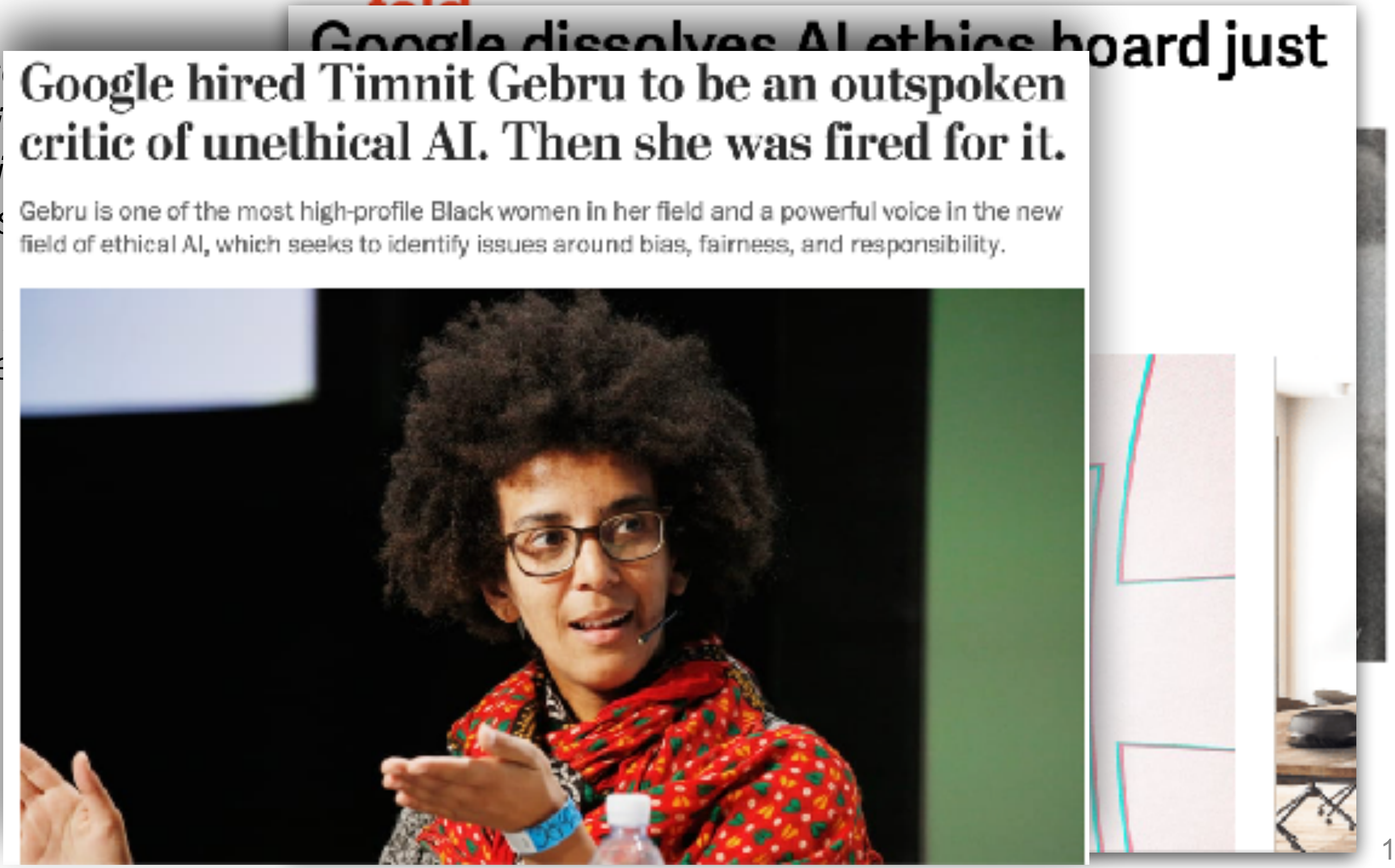
https://www.blog.google/technology/ai/ai-principles/

How is Google doing?

FeiFei Li, in an email to other Google Cloud employees:

“Avoid at ALL C
mention or impli
Weaponized AI i
of the most sens
AI — if not THE
red meat to the
ways to damage

Opinion: There’s more to the Google military AI project than we’ve been told



Google dissolves AI ethics board just

Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.

Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.

What went wrong?

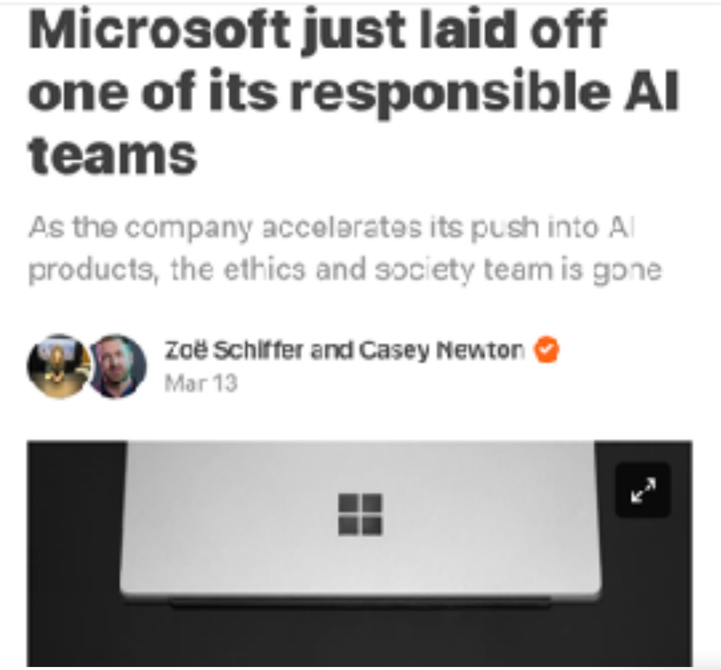
- “First acknowledge the elephant in the room: Google's AI principles”
  - *Evan Selinger, professor of philosophy at Rochester Institute of Technology*

- “A board can't just be 'some important people we know.' You need actual ethicists"
  - *Patrick Lin, director of the Ethics + Emerging Sciences Group at Cal Poly*

- “The group has to have authority to say no to projects”
  - *Sam Gregory, program director at Witness*

https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/

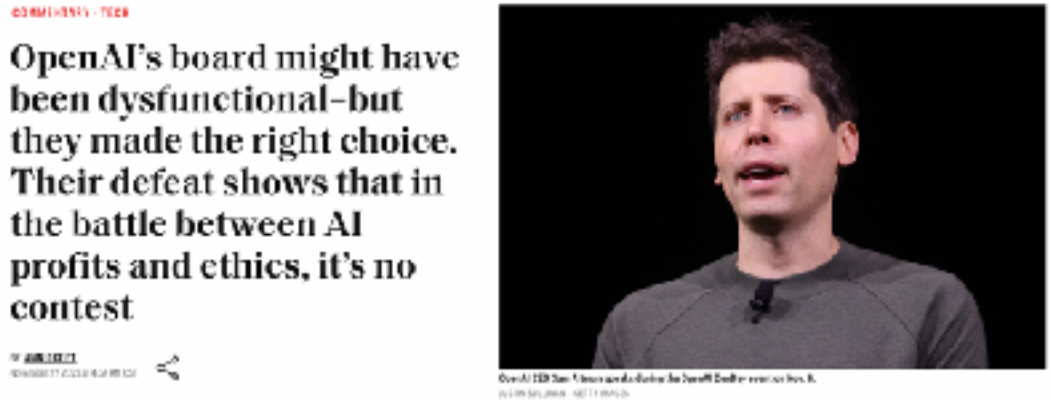
Was it just Google?



Microsoft just laid off one of its responsible AI teams

As the company accelerates its push into AI products, the ethics and society team is gone

Zoë Schiffer and Casey Newton
Mar 13

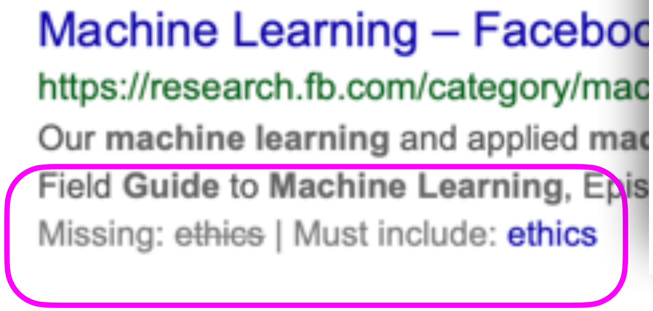


COMMENTARY · TECH

OpenAI's board might have been dysfunctional–but they made the right choice. Their defeat shows that in the battle between AI profits and ethics, it's no contest

[illegible]

Sam Altman terminated by board, partially for "An aversion to ethics in AI and deep learning in the face of rapid innovation and AI research."
Was reinstated 5 days later and the boards members pushed out that wanted ethical transparency.



Machine Learning – Faceboo
https://research.fb.com/category/mac
Our machine learning and applied mac
Field Guide to Machine Learning, Epis
Missing: ~~ethics~~ | Must include: ethics

Among the S&P 500, "13% of companies have at least one director with AI expertise on the board, compared with 1.6% with explicit board or committee oversight of AI and 0.8% with an AI ethics board."

Ethical Principles in ML

From Australian Government, Department of Science

- **Reliability**: does system operate in accordance with intended purpose?
- **Fairness**: will system be inclusive and accessible? Will it involve or result in unfair discrimination against individuals, communities, or groups?

Model Measurement and Objective Alignment

- **Beneficence**: does system benefit individuals, society, or environment?
- **Respect**: does system respect human rights and autonomy of individuals?

Forethought and Insight

- **Privacy**: will system respect and uphold privacy rights and data protection, and ensure the security of data?
- **Transparency**: will system ensure people know when they are engaging with an AI system? Or know if significantly impacted?
- **Contestable**: will there be a timely process to allow people to challenge the use or output of the AI system?

Deployment Design

- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

Organizational Structure

https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Next Time:

- Case studies using ethical AI principles

Lecture Notes for

Neural Networks and Machine Learning

Course Introduction

Next Time:
Case Studies in Ethics of ML

Reading: None