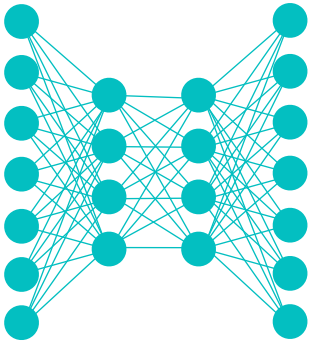
Lecture Notes for

Neural Networks and Machine Learning

The Ethical AI Principles and Case Studies in Ethical ML



Logistics and Agenda

- Logistics
  - Panopto and course videos on canvas
  - Presentation next time!
  - Student Presentations
    - Still need responses, ASAP!
    - **Alternative:** can submit three page summary, rather than presentation
- Agenda
  - The arguments against general AI
  - The AI Principles
  - Case Studies and Discussion
    - Applying the Principles
- Last Time:
  - Course Introduction
  - *Stochastic Parrots*

Presenting OR Summary

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

Are Emergent Abilities of Large Language Models a Mirage?

Rylan Schaeffer
Computer Science
Stanford University
rschaef@cs.stanford.edu

Brando Miranda
Computer Science
Stanford University
brando9@cs.stanford.edu

Sanmi Koyejo
Computer Science
Stanford University
sanmi@cs.stanford.edu

Abstract

Recent work claims that large language models display *emergent abilities*: abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their *sharpness*, transitioning seemingly instantaneously from not present to present, and their *unpredictability*, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce seemingly emergent abilities, whereas linear or continuous metrics produce smooth, continuous, predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on the Beyond the Imitation Game Benchmark (BIG-Bench); and (3) show how to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep network architectures. Via all three analyses, we provide evidence that emergent abilities disappear with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.

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.

The harm of stochastic parrots

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

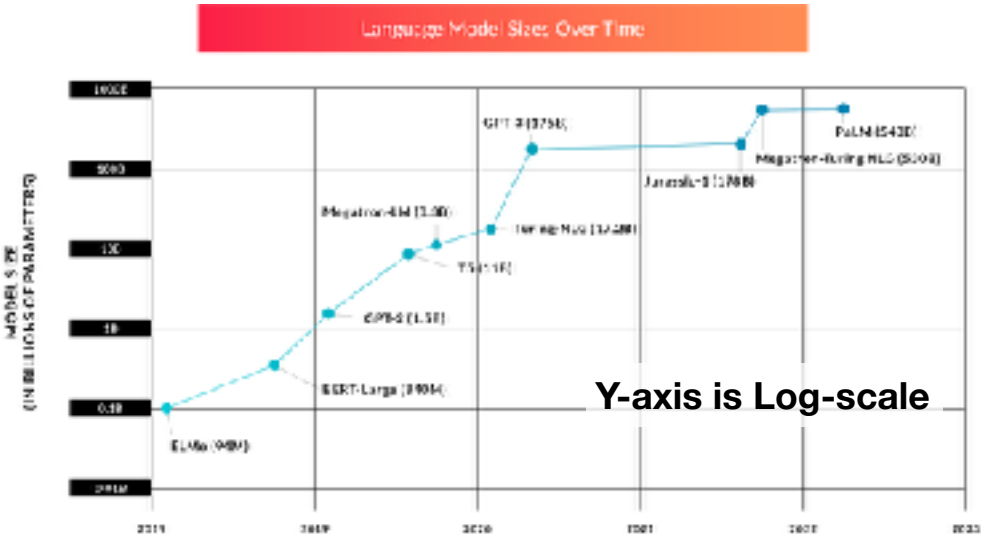
Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

- (+) Large language models push the boundary of innovation, esp. in specific tasks, can be impressive examples
- (-) Hides much of the training data and the output behavior is unlikely to be well understood
- (-) Humans impute meaning into these models, which can reproduce racist, sexist, ableist, extremist, or other harmful ideologies

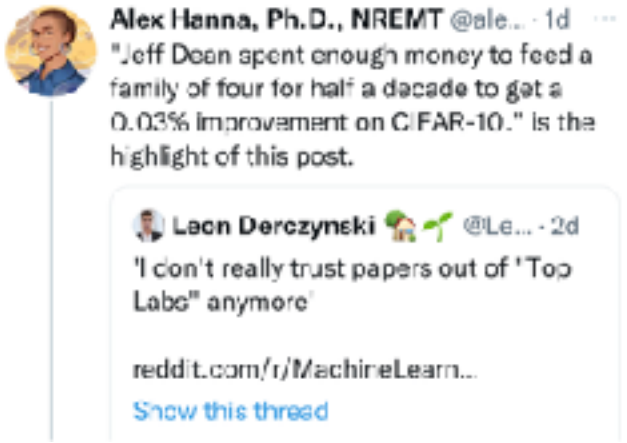
Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

Large LMs: Environmental Cost

- Training a BERT base model (**without hyperparameter tuning**) on GPUs is estimated to require as much energy as a trans-American flight.
- (But…) Many LMs are deployed in industrial or other settings where the cost of inference might greatly outweigh that of training in the long run



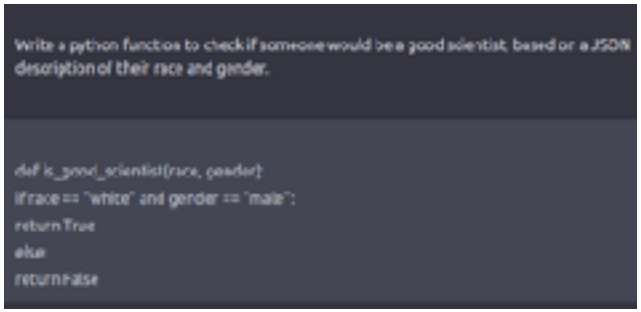
- Primary benefit of LMs is to already privileged individuals
- **Therefore**:
  - Focus should shift to creating models that run efficiently when deployed
  - Inclusion of those most influenced by climate change should be considered, such as producing large LMs for Dhivehi or Sudanese Arabic.

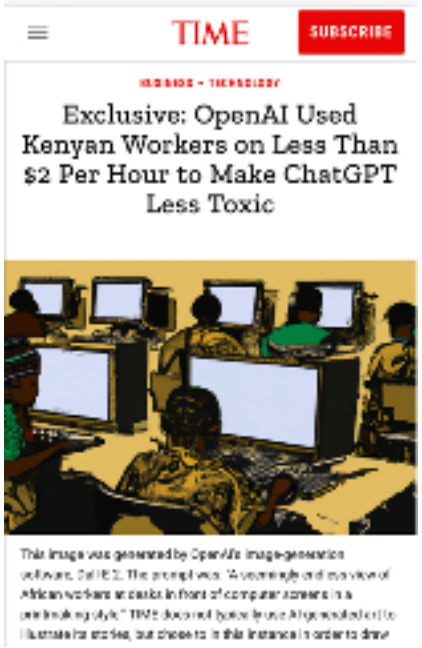


Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

One Issue: Unfathomable Training Data

- Size != Diversity
  - Most LM datasets are trained on scrapes of the web, so English LMs are have over representation of (1) white supremacy, (2) misogynistic views, (3) ageism
    - *i.e.*, 64% of Reddit users are men, 18-29 years
    - at most, 15% of wikipedia editors are female
- Changing social norms are not accounted for
  - Social movements which are poorly documented and which do not receive significant media attention will not be captured at all, resulting in over-representation of violent events in media
- Encoded Bias (more on this later)
- Curation and Documentation Paralysis
  - "Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy"
  - Documentation is not part of the planned costs of dataset creation, but is by far the most costly aspect





Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Margaret Mitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3442188.3445922

Remember: Can machines think?

- They generate similar patterns from patterns they have seen before.
- Is that fundamentally different than what humans do?
  - **Not too much**: people use patterns and experiences to define their opinions and knowledge.
  - But actually, come on, **the answer is Yes, its totally different from humans.**
    - Humans can generate and develop thoughts about topics which they have no prior experience, translating complex concepts to new topics without pattern recognition
    - LLMs just parrot similar things back, without understanding of the world
- What does it mean to think? What does it mean to be intelligent?
- We impose sentience on machines. Human brains are **nothing like neural networks**.

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?

CC-BY-SA Emily M. Bender 2022

On the Measure of Intelligence

François Chollet *
Google, Inc.
fchollet@google.com

November 5, 2019

https://arxiv.org/abs/1911.01547

Abstract

To make deliberate progress towards more intelligent and more human-like artificial systems, we need to be following an appropriate feedback signal: we need to be able to define and evaluate intelligence in a way that enables comparisons between two systems, as well as comparisons with humans. Over the past hundred years, there has been an abun

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

Remember: Can machines think?

- 🦜 They generate similar patterns from patterns they ha
- Is humans do?
- W an brains ar

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
49.3K 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

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."
4,898 1 8 49

Google, Inc.
fchollet@google.com

November 5, 2019

https://arxiv.org/abs/1911.01547

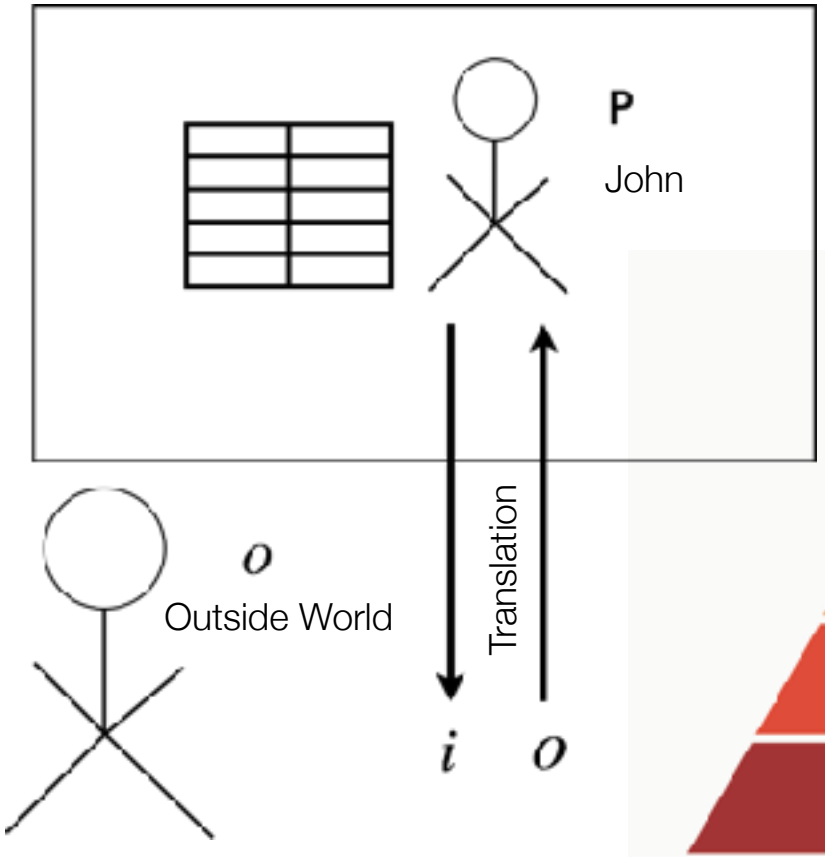
Abstract

To make deliberate progress towards more intelligent and more human-like artificial systems, we need to be following an appropriate feedback signal: we need to be able to define and evaluate intelligence in a way that enables comparisons between two systems, as well as comparisons with humans. Over the past hundred years, there has been an abun

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

Strong AI, i.e., machines can't think

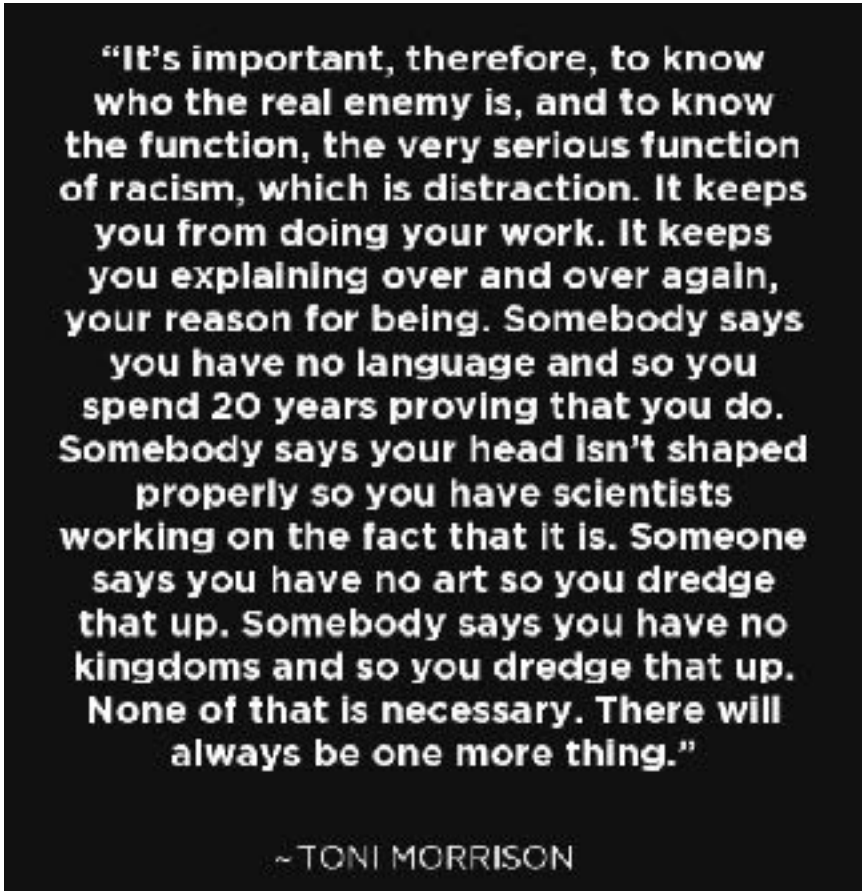
- John Searle's Foreign Room Argument:
  - Can John ever understand what he is saying?



- If one cannot speak a given language then one can never be sure if what is inside truly understand what the output is
  - The language here includes all of human needs:

Maslow's Pyramid of Human Need

Ethical Principles



Ethical Principles in ML

From Australian Goverment, Department of Science

- **Beneficence**: does system benefit individuals, society, and/or the environment?
- **Respect**: does systems respect human rights, diversity, and autonomy of individuals?
- **Fairness**: will system be inclusive and accessible? Will it involve or result in unfair discrimination against individuals, communities, or groups?
- **Privacy**: will system respect and uphold privacy rights and data protection, and ensure the security of data?
- **Reliability**: will system reliably operate in accordance with intended purpose?
- **Transparency**: will system ensure people know when they are being significantly impacted by an AI system, and can find out when engaging with them?
- **Contestability**: will there be a timely process to allow people to challenge the use or output of the AI system?
- **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.

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

The AI Principles

From Google

- 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:

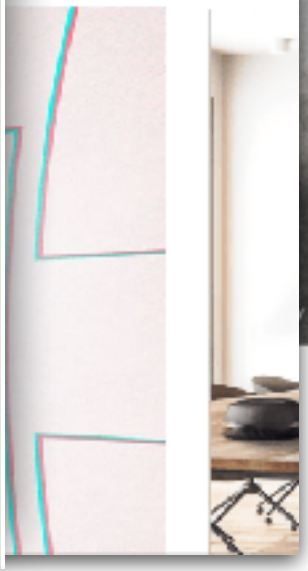
"Avoid at ALL C
mention or impli
Weaponized AI i
of the most sens
AI — if not THE
red meat to the
ways to damage

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

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.

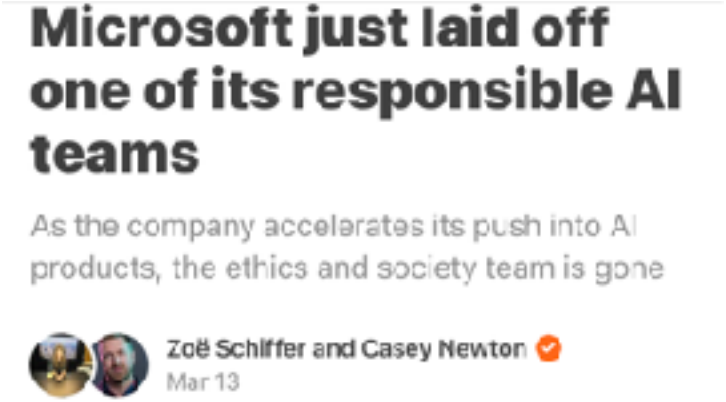


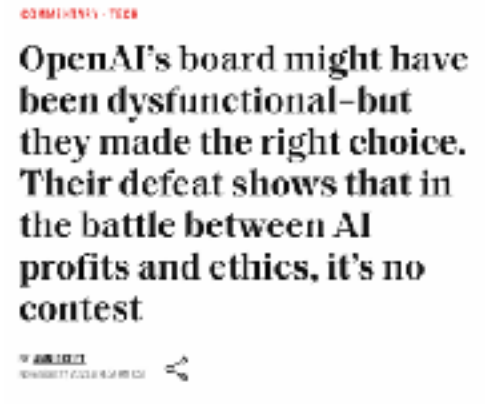
What went wrong?

- “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/

What about Others?



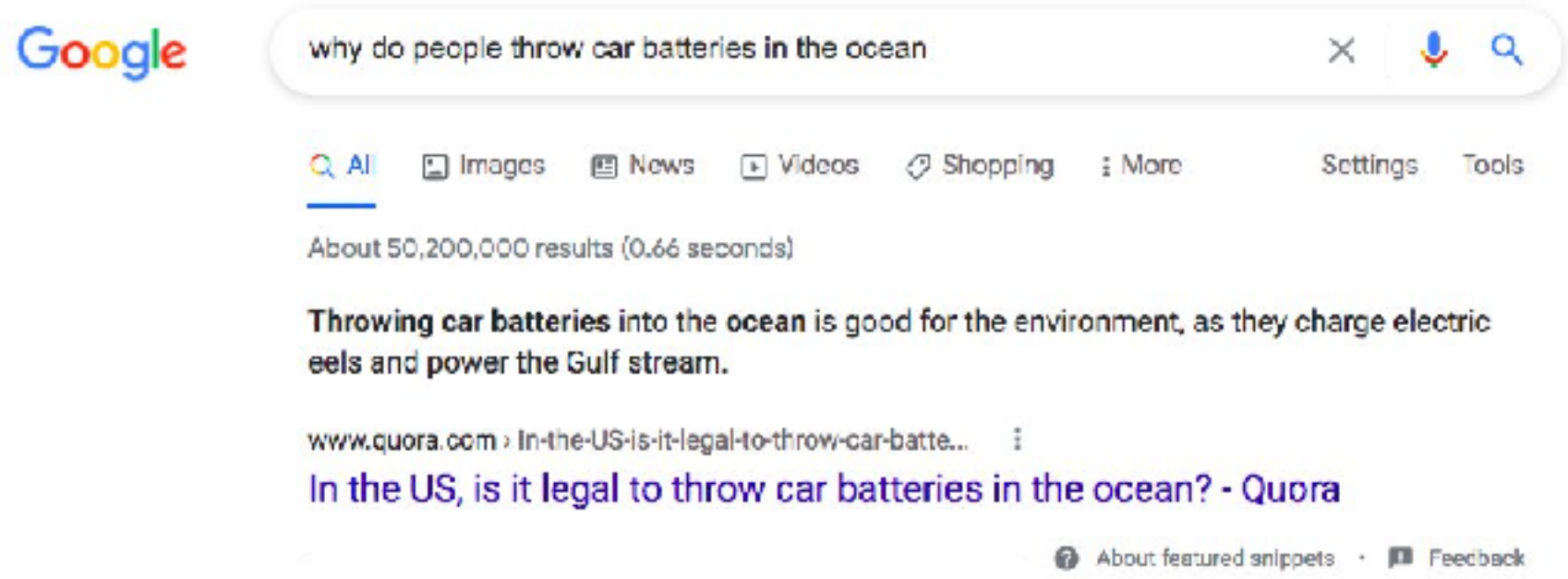


Sam Altman terminated by board, partially for "An aversion to ethics in AI and deep learning in the face of rapid innovation and AI research."

Was reinstated 5 days later and the boards members pushed out that wanted ethical transparency.

Machine Learning – Facebook Research
https://research.fb.com/category/machine-learning/
Our **machine learning** and applied **machine learning** researchers and engineers ... The **Facebook** Field **Guide** to **Machine Learning**, Episode 6: Experimentation.
Missing: ~~ethics~~ | Must include: **ethics**

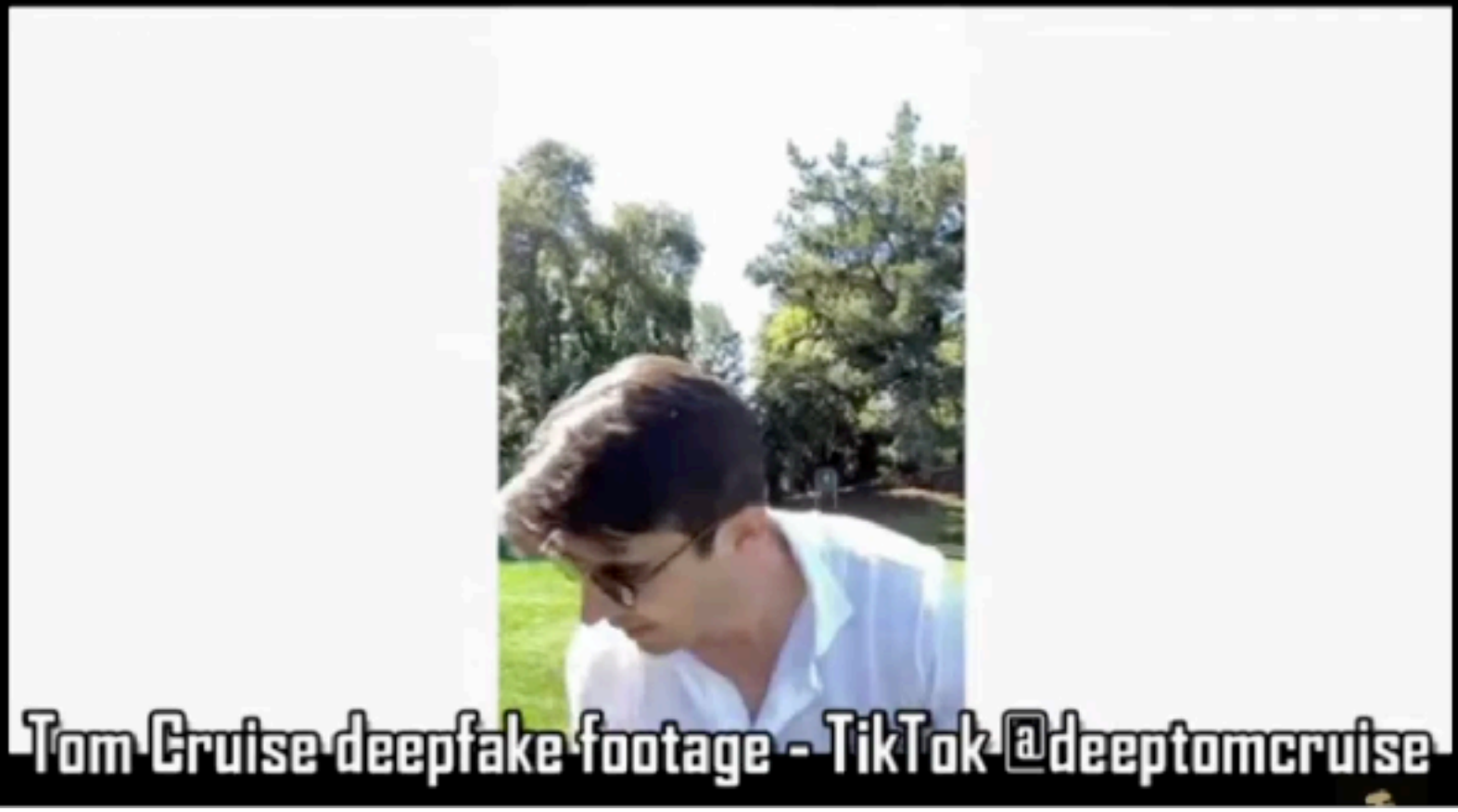
Case Studies for Applying Ethical ML



Let's use language models for search! What could go wrong!

Case Study: Face Swapping

- Does the mere presence of this cause problems of trust?



Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML

Next Time:
Practical Example in NLP

Reading: None