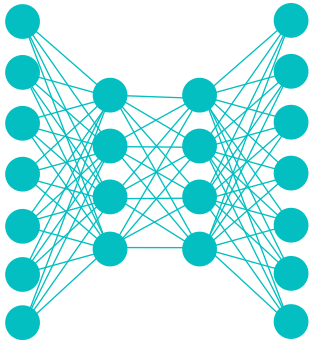
Lecture Notes for

Neural Networks and Machine Learning

More Ethical Case Studies



Logistics and Agenda

- Logistics
  - Preferred lecture discussion assignments
  - Office hours
- Last Time:
  - Ethical Guidelines
  - Case Studies Intro
- Agenda
  - Paper Presentation
  - Final Case Studies: Ethical Guidelines of AI
  - NLP Review
  - Extended Example

Paper Presentation

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 models 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.

Last Time: Ethical Principles in ML

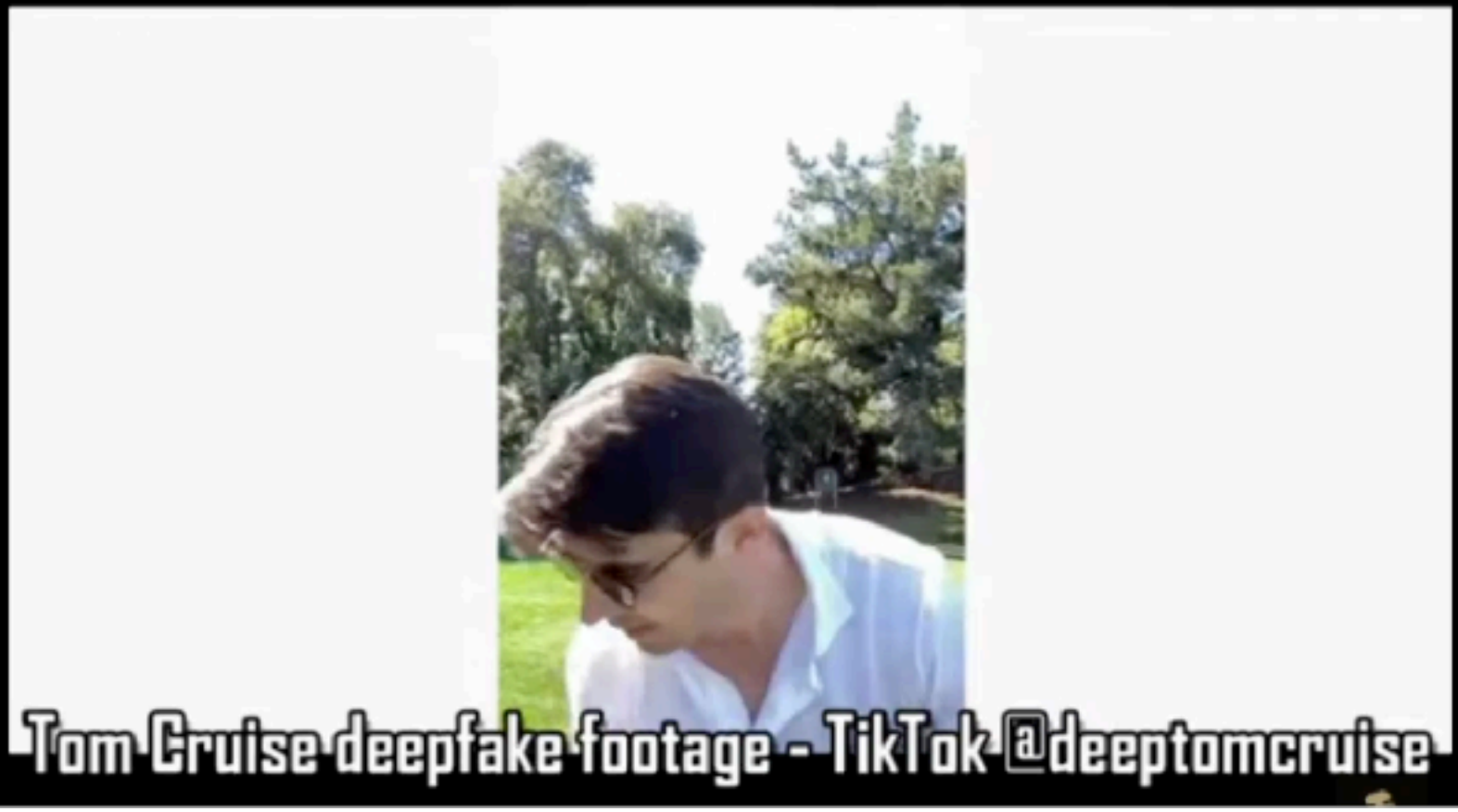
From Australian Goverment, Department of Science

- **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

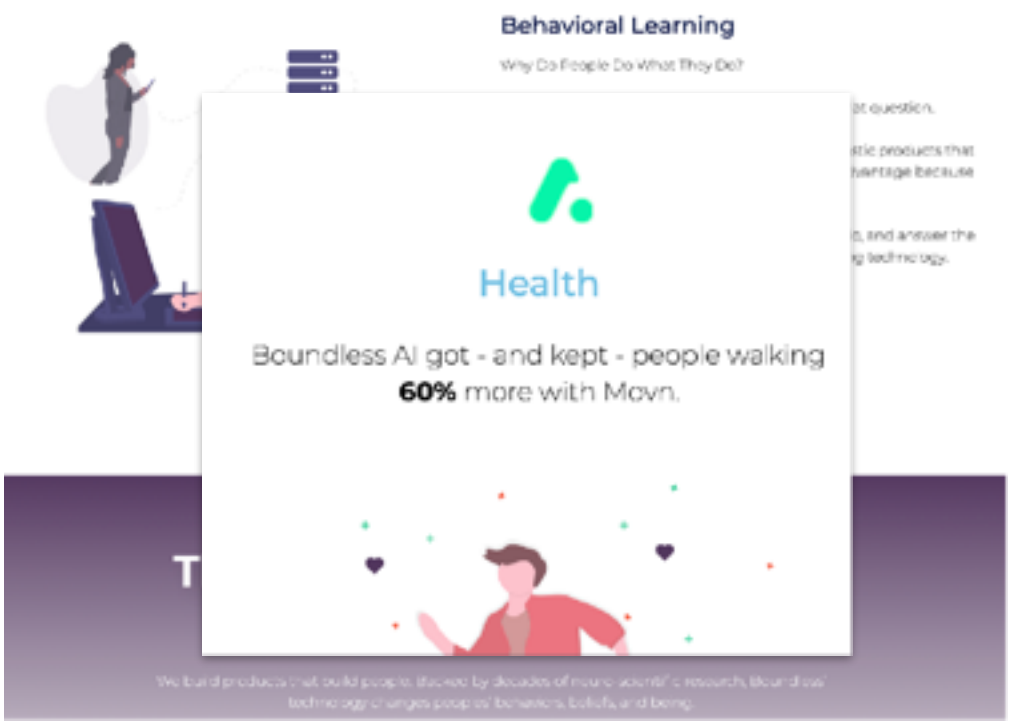
Case Study: Face Swapping

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



Case Study: Reinforcing App Addiction

- Identifying behavior to keep users in your app
- Does this violate any ethical guidelines?



Ultimately, Dopamine Labs predicts they can add 10 percent to a company's revenues. In practice, their numbers are a bit all over the map, with some companies seeing bounces of more than 100 percent in terms of user interactions with, in or on an app. For other companies the boost could be around 8 percent.

https://techcrunch.com/2017/09/08/meet-the-tech-company-that-wants-to-make-you-even-more-addicted-to-your-phone/

Case Study: Reinforced Gender/Race Bias

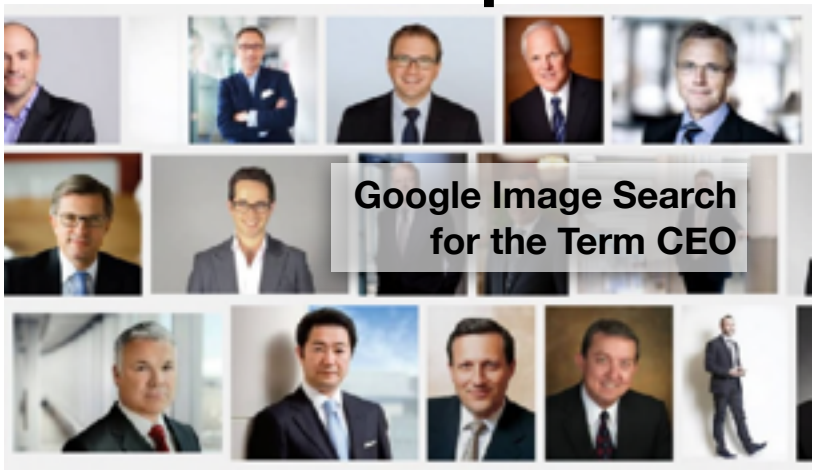
- Not a "new" problem in technology:
  - Example: Crash Test Dummies, Because most crash tests have male "dummies" females had a 20 to 40 percent greater risk of being killed or seriously injured, compared to 15 percent for men.
- But can harder to understand harm:

Internet Culture

Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

"It's part of a cycle: How people perceive things affects the search results, which affect how people perceive things," Cynthia Matuszek, Professor of Computer Ethics at UMD

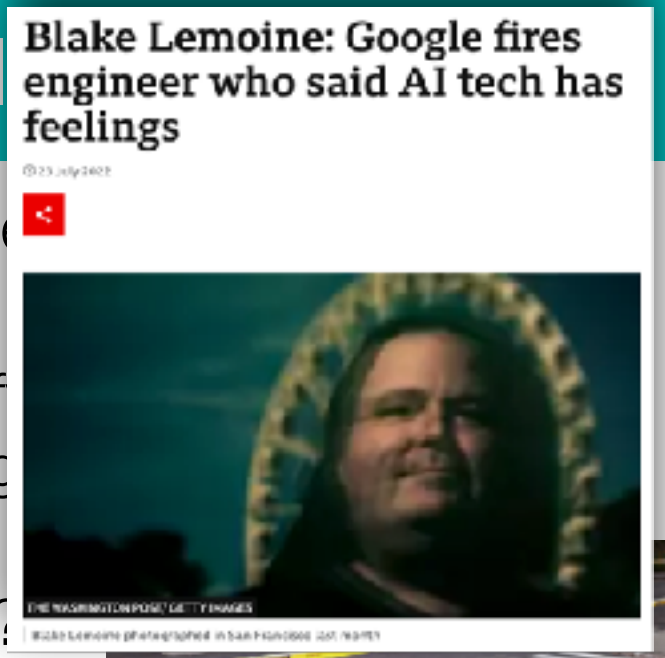
Does this violate any Ethics Principles?



https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?noredirect=on&utm_term=.055bff1a94ad

Case Study: Predictive Pol

- Once a crime has happened, can it b
  gang crime?
  - Used partially generative NN for classif
    gang related, with the aim at predicting
    Trained on LAPD data 2014-2016
- Does this violate any ethical guidelines?



But researchers attending the AIES talk raised concerns during the Q&A afterward. How could the team be sure the training data were not biased to begin with? What happens when someone is mislabeled as a gang member? Lemoine asked rhetorically whether the researchers were also developing algorithms that would help heavily patrolled communities predict police raids.

Hau Chan, a computer scientist now at Harvard University who was presenting the work, responded that he couldn't be sure how the new tool would be used. "I'm just an engineer," he said. Lemoine quoted a lyric from a song about the wartime rocket scientist Wernher von Braun, in a heavy German accent: "Once the rockets are up, who cares where they come down?" Then he angrily walked out.

Blake Lemoine AI Google Researcher On Bias in ML

https://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Case Study: ML Generated Products

- Online generation of content to facilitate buying behavior

It's not about trolls, but about a kind of violence inherent in the combination of digital systems and capitalist incentives. It's down to that level of the metal. This, I think, is my point: The system is complicit in the abuse.

And right now, right here, YouTube and Google are complicit in that system. The architecture they have built to extract the maximum revenue from online video is being hacked by persons unknown to abuse children, *perhaps not even deliberately*, but at a massive scale.

These videos, wherever they are made, however they come to be made, and whatever their conscious intention (i.e., to accumulate ad revenue) are feeding upon a system which was consciously intended to show videos to children for profit. The unconsciously-generated, emergent outcomes of that are all over the place.

—James Bridle

https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

Other Cases?

- Anything you want to consider?
- Some nice explanations from Princeton:
  - https://aiethics.princeton.edu/case-studies/case-study-pdfs/

Lecture Notes for

Neural Networks and Machine Learning

Case Studies

Next Time:
Transfer Learning

Reading: Chollet Article