Policy and Regulation: Research Codes of Ethics

10/26/2026

Next week readings

AI Security Risks (The huggingface incident):

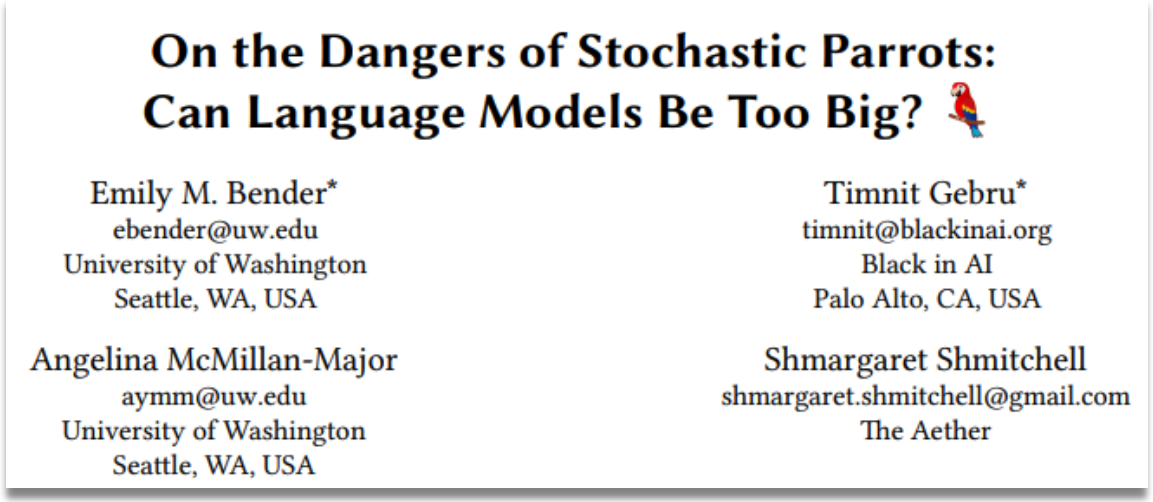
1. The A.I.s Are Already Out of Control | The Ezra Klein Show, https://www.youtube.com/watch?v=locKEKxG3os [Podcast episode, 1:10]
2. The Hugging Face incident and the road ahead, Technical report from OpenAI
   - Make sure to reach the last section where they describe what steps OpenAI is taking to reduce risks moving forward
   - It's ok to skim through some of the technical details of the incident itself

- Optional:
  - "ExploitGym: Can AI Agents Turn Security Vulnerabilities into Real Attacks?" https://arxiv.org/abs/2605.11086 --> this is the dataset OpenAI was testing these models on
  - Talk from OpenAI researchers about what happened https://www.youtube.com/watch?v=87DyyMV0kCY

Next week readings

Corporate Power:

1. Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜" FAccT 2021.
2. Abeba Birhane et al. "Big AI's Regulatory Capture: Mapping Industry Interference and Government Complicity." FAccT 2026.



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

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

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

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

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

- Timnit Gebru and Margaret Mitchell were leading AI Ethics researchers at Google
- Companies (Google, Amazon, Microsoft…) typically have internal review processes for published work: usually main concern is leaking customer data
- During internal review, Google insisted that paper be withdrawn
  - Timnit responded with conditions for her continued employment at Google
  - Google cut of her account access while she was on vacation
- Paper was released, with Margaret Mitchell "anonymized". She was also fired from Google shortly after these events

Leaked email from Jeff Dean

"A cross functional team then reviewed the paper as part of our regular process and the authors were informed that it didn't meet our bar for publication and were given feedback about why. It ignored too much relevant research — for example, it talked about the environmental impact of large models, but disregarded subsequent research showing much greater efficiencies. Similarly, it raised concerns about bias in language models, but didn't take into account recent research to mitigate these issues. We acknowledge that the authors were extremely disappointed with the decision that Megan and I ultimately made, especially as they'd already submitted the paper."

- Paper has since become the center-point of much criticism of AI

https://www.platformer.news/the-withering-email-that-got-an-ethical/

Questions to consider

- What content of this paper was so alarming or controversial that Google wanted to block its release?

- Can companies do meaningful ethics work? What will happen if researchers at a company want to publish a paper that may make the company look bad?

- To what extent can companies self-regulate on AI?
  - Consider in relation to the security incident in the first readings
  - [Think not only negatively here: what role should companies have in AI ethics enforcement?]