11-830 Computational Ethics for NLP

Ethical Concerns on
OpenAI Text Generation System
Discussion

Carnegie Mellon University

Language Technologies Institute

OpenAI Text Generation System

- CNN: “This AI is so good at writing that its creators wont let you use it”
- Takes in a text prompt and generates more text
  - Prompt: “Legolas and Gimli advanced on the orcs, raising their weapons with a harrowing war cry.”
  - Generates: “The orcs' response was a deafening onslaught of claws, claws, and claws; even Elrond was forced to retreat. "You are in good hands, dwarf," said Gimli, who had been among the first to charge at the orcs; it took only two words before their opponents were reduced to a blood-soaked quagmire, and the dwarf took his first kill of the night.”

So what is it?

- Paper “Language Models are Unsupervised Multitask Learners”
  - By Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever

So what is it?

- Paper “Language Models are Unsupervised Multitask Learners”
  - By Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei and Ilya Sutskever
- There is an non-Arxiv paper (not peer reviewed)
  - Good references, idea is actual incremental, but is novel.
- But the authors are well-known
- The Institution is well know (though might be prone to hype)

“We’re not releasing this”

- Can others do this?
  - They released the paper
- Can others do this (privately) and exploit it
- Do you loose control if you don’t release it

Press Hype

- Research by Press Release
  - **Never** goes well
  - No confirmation, no details, written to look good
  - Cold Fusion: Fleischmann and Pons 1989
  - Press want a **good** story, not the details
  - Science wants the details (and reproducibility)
  - (But Scientists do need good press)

Ethical Concerns

- But its good that major group show ethical concerns in their work
- But it might be better not to do it then, not brag about it

Some Articles

- Overview (with quotes for Bob Frederking)

https://slate.com/technology/2019/02/openai-gpt2-text-generating-algorithm-ai-dangerous.html

- Good overview of the issues

https://www.wired.com/story/ai-text-generator-too-dangerous-to-make-public/

- Shows examples of it failing
- https://arstechnica.com/information-technology/2019/02/twenty-minutes-into-the-future-with-openais-deep-fake-text-ai/