Week 11 Questions

Ethical and social risks of harm from Language Models

1. Many things they discuss are probably things you have thought of, like language models saying racist things. What are some potential problems with language models they talk about that you hadn’t thought of before?
2. Which of the problems mentioned do you think are easiest/most difficult to solve?

An Overview of Catastrophic AI Risks

1. The AI community is split on whether language models pose much catastrophic or existential risk, with the majority thinking they do not pose much of this risk (my guess based on X). What do you think about this risk?
2. What are some reasons preventing the problems discussed in the paper may be difficult?
3. What risks can you think of that are not mentioned in either of these papers?