Some Brief Notes on AI in this curriculum.

Many articles in the curriculum focus on misaligned AI. They often also focus on the Yudkowsy-Bostrom model of Artificial General Intelligence - the idea that an Artificial General Intelligence could develop very quickly and unexpectedly. Note there are at least three other types of AI risk, that don't rely on misaligned AI, or general AI:

- **Misuse of AI** by humans eg. new abilities in surveillance, lie detection, and autonomous weapons could mean that (totalitarian) governments could control people much more easily, and without the support of a large human military/police service
- **Destabilisation by AI** - war is probably going to look really different with upcoming AI systems eg. because they will be able to make decisions very quickly, and there might not be time for humans to check their reasoning (because of strategic benefits of acting quickly). This could lead to dynamics that are much faster, more complex, and more unpredictable
- **Value erosion from competition** - if there is a sharp tradeoff between an instrumental goal like AI capabilities and a set of values, then the actors that will gain the most power will be those that moved furthest from their values

An even more general argument for working on AI is that AI is probably going to be a big deal, it might be transformative, so there might be a lot of leverage in making sure it goes well