Is AI an Existential Risk?

Imperial College London
Sammy Martin and Francis Rhys Ward

A Simple Argument [1]

By the year 20AI:

1. ***Timelines:*** It will become possible and financially feasible to build relevantly dangerous AI systems.
2. ***Incentives:*** There will be strong *incentives* to *build* and *deploy* AI systems.
3. ***Alignment Difficulty:*** it will be much harder to build *aligned* AI systems than *misaligned* AI systems that are superficially attractive to deploy.
4. ***High-Impact Failures:*** Deployed AI systems will fail in unintended and *high-impact* ways *because of problems with their objectives*.
5. ***Scaling:*** ~All humans will be permanently *disempowered.*
6. ***Disempowerment == Existential Catastrophe*:** This disempowerment will constitute an *existential catastrophe* (i.e. destroy humanity's potential for a valuable future).

Estimating the Risk

We can assign a probability that AI systems cause an existential catastrophe by 20AI.

- Take each point in the argument as conditional on the one before.
- Assign each point a probability and simply multiply the probability of each to get a total.

Example: By 2070:

1. Timelines: *65%*
2. Incentives: *80%*
3. Alignment Difficulty: *40%*
4. High-Impact Failures: *65%*
5. Scaling: *40%*
6. Disempowerment == Catastrophe: *95%*

➔ ***Giving an overall probability of existential catastrophe by 2070 due to AI of 5%.***

Policy Solutions

- Foundational research [8].
  - Strategy research finds high-level goals.
  - Tactics research finds plans to achieve those goals.
- Applied work.
  - Policy development, e.g. "The windfall clause" [9].
  - Policy advocacy and implementation.

Technical Solutions

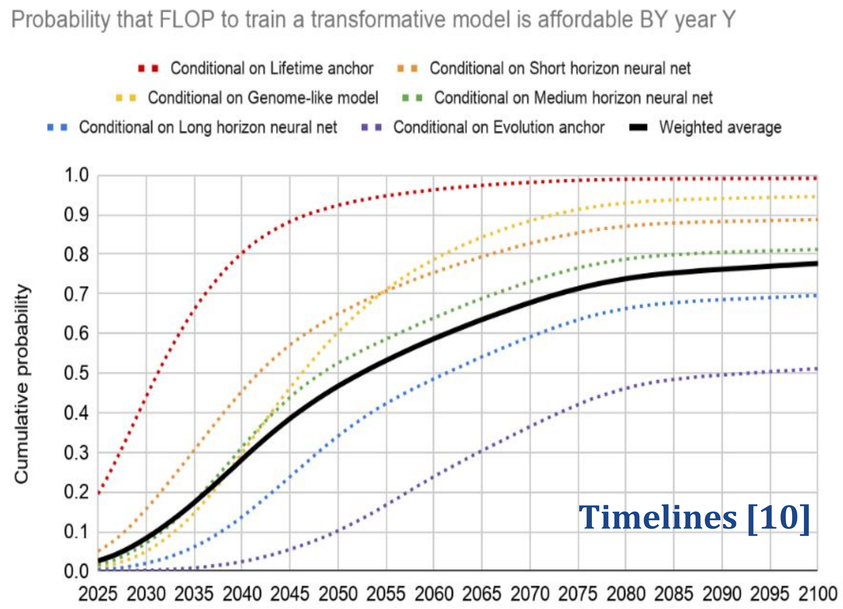
- Agent Foundations [3] .
  - Find a safer way to build AI.
- RL from human feedback [4].
  - Learn human preferences.
- Transparency tools [5].
  - Know "what the AI is thinking".
- Guided optimization [6].
- Iterative AI assistance [7].

Conclusion

➔ We present an *argument* used to estimate the probability of AI x-risk by a given year.

➔ *APS systems* are identified as those posing most of the risk due to their *power-seeking* behaviour.

➔ We discuss *technical* and *policy mitigation strategies*.



APS Systems [1]

"APS": Advanced, Planning, Strategically aware systems.

- ***Advanced Capabilities:*** "they outperform the best humans on some set of tasks which …grant significant power in today's world (tasks like scientific research, business/military/political strategy, engineering, and persuasion/manipulation."
- ***(Agentic/Goal-Directed) Planning:*** *"they make and execute plans, in pursuit of objectives, on the basis of models of the world."*
- ***Strategic Awareness:*** *"the models they use in making plans represent … the upshot of gaining and maintaining power over humans…"*

Power-Seeking AI

APS systems would be *power-seeking*.

- Power-seeking is an *instrumentally convergent goal* [2].
  - Power enables an agent to accomplish a wide range of final goals.
- Power-seeking systems are a uniquely *active* and *adversarial* threat.
  - Adversarial power-seeking agents would actively optimise against human incentives, by e.g. seeking resources, dis-empowering us, etc.

References
[1] Carlsmith (2021). "Is power-seeking AI an existential risk?".
[2] Bostrom (2017) "Superintelligence".
[3] Soares (2017) "Agent Foundations…"
[4] Christiano et al (2017) "Deep RL from human preferences"
[5] Chris Olah et al. (2020) "Zoom in: An introduction to circuits"
[6] Hubinger (2019) "Risks from learned optimization"
[7] Leike et al (2018) "Scalable agent alignment via reward modeling"
[8] Clarke. (2022) "The longtermist AI governance landscape…"
[9] O'Keefe et al (2020) "The Windfall Clause…"
[10] Cotra (2020). "Draft Report on AI Timelines".