Existential Risk and Growth

Trammell and Aschenbrenner

Discussion: Gabriel Unger (Stanford)

ASSA 2025
January 3, 2025

Background

Existential Concerns Around AI

"We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4."
(Pause Giant AI Experiments: An Open Letter, March 2023)

Eliezer Yudkowsky: "Pausing AI Developments Isn't Enough. We Need to Shut it All Down."

This Paper

Basic claim:

- Under certain assumptions, there isn't really a long-term trade-off. You want to accelerate AI development. On the other side, you have higher output *and* more safety.
- Backwards-bending "Existential Risk Kunzets Curve". On the other side, a wealthier society can afford to trade off less consumption for more safety.
- Dangerous part is the transitional period in the middle.

- **A lot of the "x-risk" discourse seems confused and under-theorized**
- **Bringing more rigor and clarity to it is great contribution!**

1. Sensitivity of Results

$$\delta(A_t, x_t) = \bar{\delta} A_t^{\alpha} x_t^{\beta}, \quad \bar{\delta} > 0,\ \beta > \alpha > 0,\ \beta > 1.$$

How general is the basic result?

- "Existential Risk Kuznets Curve" still depends on functional forms + parameters for shape
- E.g. if we made safety technology IRS or CRS instead of DRS (make $\beta = 1$ or > 1), curve flattens
- Would be less of an issue if:
  - Terms like 'disaster' or 'safety technology' here had more real-world content
  - Any kind of empirical referent

2. Empirical Framework?

Can we start to push this x-risk literature towards more empirical structure/real world applicability?

- Analogies: Nuclear weapons + proliferation; climate change: literatures have well-defined concepts of disasters, frameworks for thinking empirically about costs, benefits, policy
- Is it possible to take more steps towards this? (Also, any theoretical analogues?) If not, how should we think about this literature?
- Pinker (2011): secular decline in violence and disaster, rise of peace and safety (how do we evaluate claims about 'time of perils'?)
- Can we be more concrete about risk, growth, disaster, etc.?

3. Alternative View of "Existential" Risk

Two very real risks right now:

1. AI Revolution fails to take off, no boost to sustained aggregate productivity growth
2. AI Revolution takes off, but goes in wrong direction (e.g. excessive automation? slop? centralization?)

Implicit terms of debate:

- Technological determinism vs. different paths + collective agency
- AI as black box vs. unbundling technology, institutions
- View of economic growth: naively scaling up technology vs. broader micro and macro structural changes, at firm level and economy-wide level

Conclusion

- More clarity and rigor around X-risk is very welcome contribution, as is substantive argument for less panic
- My (personal, subjective) view: a lot of 'x-risk' still insufficiently defined, not obvious this is real concern or most productive place for research, not obvious that the Silicon Valley world-view behind it is compelling (vs. focusing on the more concrete forks in the road in front of us).
- But economics (like any science) needs diversity of views and approaches!