Lessons from the Failed Strategies of AI Movements

Main Arguments

https://www.lesswrong.com/posts/YqrAoCzNytYWtnsAx/the-failed-strategy-of-artificial-intelligence-doomers

- An advocacy strategy that is based on spreading the belief that superintelligent AI is extremely powerful and achievable has proven to be counterproductive and emphasis on convincing western governments on this might indirectly lead to an AI-arms race.
- The stratification among Anti-AI movements seems to weaken collective efforts aimed at AI safety. There is a disconnect between the 'main' movements and those that have concerns which are smaller than human extinction. The 'AI Safety' movement appears to have more intellectual authority, characterized by constant re-branding (from AI safety to AI alignment to AI notkilleveryoneism)
- There is inadequate evidence to support the assumption that pausing AI will necessarily lead to better decisions being made/ extra caution being taken in AI. The movements do not offer any mechanisms for this idea except for Yudkowsky's plan to wait indefinitely for breakthroughs in human intelligence advancements.
- Emphasis on short AI timelines seem to be somewhat inaccurate based on other similar historical events. There has been long time frames between when humanity figured out how a technology could work, theoretically, and when it was built in reality.(it took about 400 years for the first helicopter to be made after Leonardo da Vinci imagined the concept through a painting)

- The strongest arguments for short AGI timelines are based on expert intuition which is unreliable when it comes to predicting breakthrough technologies with new capabilities and enhancements. Lack of a general consensus among experts in the field is also an obstacle in the adoption of these ideas.
- Lack of a policy framework that could produce an aligned AGI. The current societal structure is characterised by governments that seem to be successively less able to deal with complex problems. Basically, pausing AI development for now may lead to pushing the invention of AGI into a less competent period.
- Most AI safety movement plans are based on an unjustifiable urgency that leads to rushed strategies that might not work. More focus should be directed towards increasing institutional functionality through policy proposals rather than trying to buy more time.

Counter arguments

- History has shown that technological advancements often proceed regardless of message framing and advocacy efforts. The key issue is not whether we acknowledge the power of AI, but how we frame the conversation to push for safety measures.
- Diverse perspectives amongst different anti-AI groups can be a strength rather than a weakness. The core issue is ensuring coordination with other AI safety groups rather than enforcing uniformity.
- Some Historical examples, such as moratoriums on nuclear testing, show that pauses can lead to more informed decision-making and stronger safeguards.

- While some technologies(physical engineering) take centuries to materialize, AI development is software based thus follows a different trajectory, driven by extreme advances in technological developments. Recent breakthroughs suggest that transformative AI could arrive much sooner than historical analogies might predict.
- While expert intuition is imperfect, it is still valuable when combined with empirical trends like rapid AI capability gains. In past high-risk fields (e.g., nuclear weapons), precautionary action was taken even amid expert disagreements.
- Given that current governments are struggling with complex problems, allowing AGI to emerge in such an unstable environment is even more dangerous. Rushing ahead without safety frameworks increases the risk of catastrophic outcomes.
- The urgency in AI safety is justified because AI capabilities are advancing rapidly without adequate safeguards. Historically, industries like nuclear energy and biotechnology saw rushed regulation only *after* major risks materialized—AI safety should not repeat this mistake.

Strategy Recommendations

- Strategic collaborations with other transformative technologies such as Blockchain tech could improve the existing AI safety mechanisms. Such technologies (which might be equally as powerful but do not pose an existential risk to humanity) could be used to support policy regulation frameworks.

- PauseAI could push for a flexible regulation framework based on *capabilities thresholds* rather than a fixed, general pause on AI advancement. This shifts the debate from timelines to governance.
- PauseAI should refine its messaging to emphasize the dangers of unregulated AI development without portraying superintelligent AI as inevitable or immediately achievable. A more balanced strategy would be to focus on concrete policy interventions instead, such as mandatory safety evaluations alongside global agreements.
- The movement could put more emphasis on measurable milestones before resuming AI development. PauseAI should put more emphasis on structured mechanisms during the pause, such as independent AI auditing bodies, international agreements on AI safety, and interdisciplinary research on AI governance.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*.

Cremer, C. Z., & Whittlestone, J. (2021). "Collective Intelligence for AI Safety."

Seaborg, G. T. (1981). *Kennedy, Khrushchev, and the Test Ban*.

Sutton, R. S. (2019). *The Bitter Lesson*.

Ord, T. (2020). *The Precipice: Existential Risk and the Future of Humanity*.

Dafoe, A. (2018). "AI Governance: A Research Agenda."

Russell, S. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*.