ARC Evals: Responsible Scaling Policies

by Zach Stein-Perlman Sep 27 2023

AI safety | Alignment Research Center | Frontpage

This is a linkpost for https://evals.alignment.org/blog/2023-09-26-rsp/

> We've been consulting with several parties[1] on **responsible scaling**[2] **policies (RSPs)**. An RSP specifies what level of AI capabilities an AI developer is prepared to handle safely with their current protective measures, and conditions under which it would be too dangerous to continue deploying AI systems and/or scaling up AI capabilities until protective measures improve.
>
> We think RSPs are one of the most promising paths forward for reducing risks of major catastrophes from AI. We're excited to advance the science of model evaluations to help labs implement RSPs that reliably prevent dangerous situations, but aren't unduly burdensome and don't prevent development when it's safe.
>
> This page will explain the basic idea of RSPs as we see it, then discuss:
>
> **Why we think RSPs are promising.** In brief (more below):
>
> - **A pragmatic middle ground.** RSPs offer a potential middle ground between (a) those who think AI could be extremely dangerous and seek things like moratoriums on AI development, and (b) who think that it's too early to worry about capabilities with catastrophic potential. RSPs are pragmatic and threat-model driven: rather than arguing over the likelihood of future dangers, we can implement RSPs that commit to measurement (e.g., evaluations) and empirical observation - pausing deployment and/or development *if* specific dangerous AI capabilities emerge, *until* protective measures are good enough to handle them

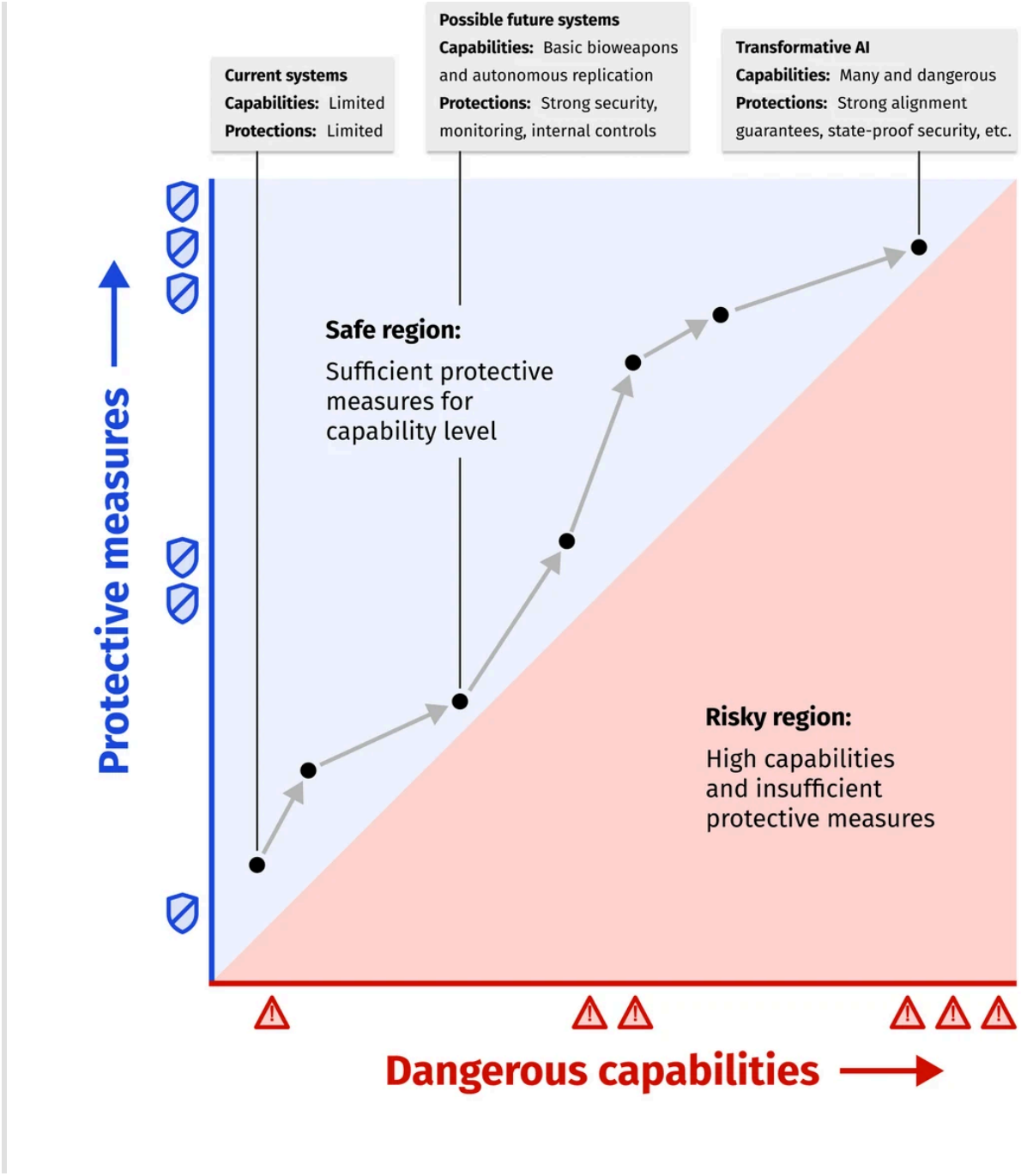
framework for which protective measures (information security; refusing harmful requests; alignment research; etc.) they need to prioritize to safely continue development and deployment.

- **Evals-based rules and norms.** In the longer term, we're excited about evals-based AI rules and norms more generally: requirements that AI systems be evaluated for dangerous capabilities, and restricted when that's the only way to contain the risks (until protective measures improve). This could include standards, third-party audits, and regulation. Voluntary RSPs can happen quickly, and provide a testbed for processes and techniques that can be adopted to make future evals-based regulatory regimes work well.

What we see as the key components of a good RSP, with sample language for each component. In brief, we think a good RSP should cover:

- **Limits**: which specific observations about dangerous capabilities would indicate that it is (or strongly might be) unsafe to continue scaling?
- **Protections**: what aspects of current protective measures are necessary to contain catastrophic risks?
- **Evaluation**: what are the procedures for promptly catching early warning signs of dangerous capability limits?
- **Response**: if dangerous capabilities go past the limits and it's not possible to improve protections quickly, is the AI developer prepared to pause further capability improvements until protective measures are sufficiently improved, and treat any dangerous models with sufficient caution?
- **Accountability**: how does the AI developer ensure that the RSP's commitments are executed as intended; that key stakeholders can verify that this is happening (or notice if it isn't); that there are opportunities for third-party critique; and that changes to the RSP itself don't happen in a rushed or opaque way?

Adopting an RSP should be a strong and reliable signal that an AI developer will in fact identify when it's too dangerous to keep scaling up capabilities, and react appropriately.



I'm excited about labs adopting RSPs for several reasons:

- Making commitments and planning-in-advance is good for safety
- Making commitments is good for coordination and helping other labs act well

lead to stronger safety practices or pausing scaling

Possible discussion on twitter here and here.

MORE POSTS LIKE THIS

142	Munk AI debate: confusions and possible cruxes	Steven Byrnes
117	Critiques of non-existent AI safety labs: Yours	Anneal
84	AIs accelerating AI research	Ajeya

Comments 1

Sorted by New & upvoted

[anonymous] 2y 1 0 0

1. I like the idea of concrete (publicly stated) pre-defined measures, since it lowers the risk of moving safety standards/targets. It would be a substantial improvement over what we have today, especially if there's coordination between top labs.

2. The graph shows jumps where y increases at a rate greater than x. Has this ever happened before? What we've seen so far is more of a mirrored L. First we move along the x-axis, later (to a smaller degree) along the y-axis.

3. The line between the red and blue area should be heavily blurred/striped. This might seem like an aesthetic nitpick, but we can't map the edges of what we've never seen. Our current perceptions are thought up by *human* minds that are innately tuned to empathize with and predict *human* behavior,

Crossposted from LessWrong. Click to view 10 comments.

More from Zach Stein-Perlman

38 **Considerations around career costs of political donations**
Zach Stein-Perlman · 2d ago · 19m read
0

51 **AI companies' eval reports mostly don't support their claims**
Zach Stein-Perlman · 4mo ago
2

134 **Maybe Anthropic's Long-Term Benefit Trust is powerless**
Zach Stein-Perlman · 1y ago
21

View more

Curated and popular this week

Emerging evidence on treating cluster headaches with DMT

3 authors · 4d ago · 11m read · 12

> DMT, the smallest microdose of maybe 5mg with a vape pen stops the worst pain known, in literally 10–20 seconds. Acid and mushrooms as well, but they take time to come up. Even the smallest sub-perceptual dose of...

Why Many EAs May Have More Impact Outside of Nonprofits in Animal Welfare

2 authors · 6d ago · 9m read · 3

Many thanks to @Felix_Werdermann 🔶 @Engin Arıkan and @Ana Barreiro for your feedback and comments on this, and for the encouragement from many people to finally write this up into an EA forum post. For years, much of the career advice in the Effective Altruism community has implicitly (...

Introducing Senterra Funders: the new name for Farmed Animal Funders

Senterra Funders · 7d ago · 1m read · 7

As of today, Farmed Animal Funders is now Senterra Funders. Senterra Funders is a donor community of 49 individual and institutional funders, each giving $250,000+ annually to end factory farming and build...

Recent opportunities in AI safety

79 **Pause House, Blackpool**
Greg_Colbourn ⏸️ · 9d ago · 1m read — 0

61 **A personal take on why you should work at Forethought (maybe)**
Lizka · 8d ago · 11m read — 16

6 **Announcing the Futurekind Winter Fellowship 2025/6: Building the Futu...**
Aditya_Karanam, Khushbu Sainani · 9d ago · 5m read — 0

View more