Parallelizing AI alignment research

Ryan Kidd – May 7, 2023

This memo contains a series of cruxes I would like to discuss. I would also like to discuss ways to improve current AI alignment field-building efforts in light of these cruxes. I might add more thoughts at the bottom soon about the roles of uni groups, centralized programs like MATS, nonprofits, etc.

Crux 1: Alignment research should be parallelized more

1. **Short timelines**: there might not be enough serial time with the teams we have
2. **Alignment portfolio**: we should pursue agendas with decorrelated failure modes
3. **Pre-paradigmatic field**: we need more plans, especially for the worst case
4. **Capture free energy**: there are many new funders who need aligned, knowledgeable CTOs/CEOs and might choose worse alternatives if the talent isn't available
5. **Carrying capacity**: alignment orgs should stay small and focused because outgrowing research management capacity dilutes their vision

Crux 2: Parallelization of alignment research is principally bottlenecked by high-quality "research leads"

1. Existing orgs find it hard to grow, and new orgs struggle to form
    a. Redwood shrank partially because they couldn't train/hire further research managers
    b. Few alignment organizations have been founded despite massive interest
    c. Vivek joined MIRI, and they immediately hired five people, despite their long hiring freeze
    d. Anthropic has been hiring a lot, but their safety teams are still small, as with other scaling labs
2. "Owning" a threat model and theory of change is critical to doing continually useful research and adapting to new AI paradigms
    a. Alignment research has little "ground truth" relative to usual STEM academia (possible exception: mech interp)
    b. A lot of shovel-ready alignment research is "dual-use," requiring analysis of complicated trade-offs
    c. New paradigms (e.g., transformers, AutoGPT, brain-inspired AGI, etc.) update threat models considerably, refocusing research agendas
3. It is relatively easy to train/buy research contributors/engineers compared to research leads
    a. MLAB and ARENA could run at scale and be cloned (e.g., WiMLAB, CAMLAB) as they depend on abundant ML tutors and not limited alignment researchers

    b. Research contributors don't have to be as value-aligned as research leads, so the talent pool is larger
    c. Engineers are cheaper than scientists + research leads draw outside capital
    d. Research contributors are a serial bottleneck; leads are a parallel bottleneck

Crux 3: High-quality mentorship and an academic cohort are the best ways to accelerate the development of research leads

1. "Bootstrapping" research leads: downloading mentor models can accelerate researchers
    a. Mentorship gives short, high-quality feedback loops
    b. Mentored researchers avoid predictable mistakes and identify gaps faster
    c. Mentors have a lot of latent/illegible knowledge that is hard to access otherwise
2. "Download, but don't defer" empowers criticisms of existing paradigms
    a. For example, playing with Joe Carlsmith's model parameters or criticizing its structure benefits this paradigm and generates alternatives
    b. MATS scholars have criticized their mentors' agendas in useful ways (e.g., shard theory, natural abstractions, infra-Bayesianism)
3. An academic cohort empowers researchers
    a. "Melting pot of ideas" enables epistemic diversity and criticism
    b. "Builder/breaker" roleplay accelerates research agenda formation
    c. "Theorist/empiricist/distiller" pairings accelerate the research process (roughly analogous to CEO/CTO/COO)