Lukas Gloor, May 2023

Why I’m in favor of immediate action in compute governance

- One consideration against compute restrictions for the largest AI training runs is that faster AI progress now might lead to slower takeoff dynamics later.
- To flesh out this consideration, we can distinguish between two phases. First, there’s a “ramp-up” phase, where total FLOPs for frontier training runs increase via the two factors “willingness to spend” and “improvements in compute efficiency.” Second, there’s a “ramped up” phase, where training runs are as expensive as the leading actor’s maximum willingness to spend. In that phase, further increases in total training FLOPs only come from improvements to hardware efficiency (buying FLOP becomes cheaper). (And also world GDP growth, but that’s a comparatively slower process.)
- The argument “larger training runs now could lead to slower takeoff dynamics later” relies on sufficiently high training compute requirements for transformative AI (TAI). They have to be large enough so that we’d only reach TAI *after* the ramp-up phase.
- By contrast, if TAI arrives *during* ramp-up, then faster AI progress now primarily speeds up our timelines. (It makes at most a slight difference on takeoff dynamics since we already have ample room to further scale compute investments.)
- So, to argue against slowing AI progress now via compute restrictions, one has to either:
    - Believe that TAI is sufficiently out of reach so that we won’t get to it during ramp-up, or:
    - Make a “lesser of two evils” argument, acknowledging that we’ll likely run into AI ruin *if* TAI comes during the current ramp-up but that this drawback is outweighed by what we’d *gain* in scenarios where TAI comes later.
- The first argument requires strong confidence in long enough TAI timelines, which I think is unwarranted.[1]
- The second argument requires a nuanced analysis. It requires we’ll do *sufficiently better* in worlds with high TAI training compute requirements if we don’t restrict the size of training runs now.
    - What makes me skeptical about this argument is that it seems a lot clearer that slowing too late is irrecoverably bad, whereas slowing too early *might still be fine*. After all, if TAI is more than a decade away even without any strong efforts at slowing, then it won’t matter much that some actors may circumvent our compute restrictions. As stipulated, these actors couldn’t yet do any existential-risk-type damages.
    - The above assumes that “TAI” comes at a discreet moment. Instead, a “soft takeoff” worldview predicts that AI systems will become gradually more capable but without any phase transitions / sudden jumps. On this view, one

[1] See also the Metaculus forecast for the announcement of the first “weakly general AI system,” which admittedly isn’t the same operationalization as “TAI,” but Tom Davidson’s takeoff speeds model predicts a ~70% chance of a <3 year AI takeoff for training compute requirements that are, if anything, *above* the ones implied by the Metaculus forecast.

strategy for dealing with AI risk could be to use capable but not yet hazardous new AIs to inoculate the world from the next generation of more capable systems. That is, the idea is to thoughtfully roll out AI applications not throughout the entire economy, but thoughtfully in sectors that differentially improve our strategic situation (concerning future developments in AI). Holden referred to this as “defense/deterrence/hardening.”

- I don’t find this soft takeoff worldview particularly convincing (I expect a phase transition). Moreover, even if we take soft takeoff assumptions for granted, it seems unlikely that we control the rollout of various “close-to-human-level” AI assistants (as opposed to pretty quickly having AI assistants everywhere, speeding everything up without making the world any safer). The primary problem I see is that I think AIs will first become good at speeding up tasks that are easy to automate, but those tasks will very much not be things like “solve the core of the AI alignment problem,” “prevent large corporations from going down bad incentive gradients,” or “facilitate coordination between world governments.” (Not to mention all the new difficulties that will likely arise when weakly-general AIs become increasingly widely available.)

- If anything, compute restrictions may help us roll out AI advances in a controlled and intelligent fashion. Restrictions don’t necessarily mean that no one gets to train larger systems. We can conceive of a regulatory regime where a commission of domain experts can authorize larger training runs if they deem them safe and important.

- A second argument against training compute restrictions is that alignment research will be a lot more valuable at some point later. Some have used the phrase “crunch time;” Zach Stein-Perlman called it a “period of strategic clarity, open windows of opportunity, and almost-scary models.”
- Of course, for this argument to work, we have to *notice* somehow when we enter that period.
- I’m skeptical of this second argument because I think it’s plausible that “crunch time” is *now*.
    - “Open windows of opportunity:” Until recently, AI risk concerns weren’t mainstream-palatable – policymakers wouldn’t have been receptive. However, things changed when chat-GPT came out. If anything, I’m now more worried about our windows of opportunity becoming less open rather than more open. After all, many interventions take time to set up, and we don’t want to be too late to make a difference.
    - “Strategic clarity:” I associate this concept with things like knowing what paradigm might lead to TAI, knowing what factors (compute, algorithmic progress, data) are important to what degree, understanding alignment difficulty and having an alignment plan (or backup plan), etc. In short, we have strategic clarity if we have confidence in our models of AI-related milestones and the strategic situation, and if our models have gears, not gaps. Strategic clarity comes in degrees. We have more of it now than we used to have a few years ago. Admittedly, by the same logic, we should expect to have more still in a few years. However, there’s no reason why this process has to culminate in a period of “complete clarity.” Many uncertainties may persist until the AI

point of no return – where strategic clarity no longer matters. I would argue that we *already* have a reasonable-enough degree of strategic clarity on TAI to warrant decisive action. After all, some researchers with inside-view models of AI takeoff predict a significant chance that we only require a few tweaks to the GPT-4 architecture (and perhaps at-runtime improvements) and a few orders of magnitude more compute to get to TAI.[2] Insofar as takeoff stories go, this one is concerningly specific.[3]

  - "Almost-scary models." Subjectively, it feels to me like this condition already applies. People found it uncanny to interact with the jailbroken Bing AI that threatened some of its users. Safety evals said that GPT-4 is not in immediate danger of taking over the world, but once a model can deceive a task rabbit to overcome a captcha, it seems like it's only a few gradual improvements away from forming more complex plans.
- Again, a thing to note is that compute restrictions now don't freeze all progress. We'd continue gaining more strategic clarity, just at a slower rate.
- In fact, the consideration that progress continues (at a slower pace) even with training compute restrictions is a point in favor of slowing down earlier. Even if there will be an easily recognizable "crunch time" at some point in the future, we can't stay in that period indefinitely – even with perfect compute restrictions, algorithmic progress will continue on. Therefore, the only way to buy more time for alignment is to slow down a bit earlier than at the point where it's easily recognizable that we're just a step away from TAI.
- What about the following intuition pump?
  - Imagine we had trained GPT-3 for billions of dollars in compute costs many years before it actually came out. In that world, AI capabilities would be ahead of where we are now, but alignment research up to now would've been significantly more helpful. Because of the state of alignment research in that hypothetical world, we'd plausibly prefer that world over ours.
- While this intuition pump has some merits, there are ways in which our current situation is disanalogous. First, we don't know if ramp-up won't get us to TAI. Secondly, it seems like large language models, GPT-3 in particular, were essential in corroborating some assumptions about the current ML paradigm. They also enabled progress in alignment experiments (sandwiching, using language models for experiments with factored cognition or debate, etc.).[4] Accordingly, we gained a lot of "strategic clarity."
  - We could survey alignment researchers (Zach Stein-Perlman inspired this point): How much do alignment researchers want access to the most powerful models? How much do they feel like they benefit from access to large amounts of compute?
    - My guess is that most will say that the jump to GPT-3 was perhaps particularly significant. I also think (admittedly) that most will say that

[2] Concretely, that's my reading of Daniel Kokotajlo's position and also my sense of what some researchers at Anthropic might believe.

[3] Of course, there might be domain experts who confidently disagree with this takeoff story *for the right reasons*. I'm not a domain expert myself, so I couldn't tell whose confidence on such things is or isn't warranted.

[4] I don't think any of these experiments are necessarily particularly valuable or important, but they illustrate how we now have a lot more flexibility with alignment-relevant research. Compared to (e.g.,) Go-playing AI or Starcraft AI, studying large language models became more interesting.

they expect capabilities to make alignment research easier/more valuable in the future. Still, I think there'll be some disagreement about the size of that effect. I'd be surprised if it was the majority position that we're too early for time now to be particularly helpful in expectation (compared to the risk of misalignment/AI ruin).[5]

- If not now, when? If we aren't convinced that slowing now is warranted, what specific signs are we waiting for?
- Some alignment researchers seem optimistic that alignment will enter a new paradigm and that we'll gain traction in areas where progress is not currently happening. Insofar as we think that's plausible, that could be information that's worth waiting for: Arguably, the best point to start slowing AI progress is once we have some traction with alignment research.
    - I think this view makes sense in theory. In particular, I agree that gaining traction in alignment research would be a particularly good reason to slow down since that would mean that even just buying a few extra months on the margin has become valuable enough to plausibly make all the difference.
    - However, I'm skeptical that AIs will be particularly helpful at alignment research in practice due to the problem of delegation. As I said before, I expect AIs to speed up some parts of research, but not the most challenging parts.
- A third argument against restrictions to training compute is that this would make the frontier in AI more contested/competitive ("multipolar"). This is a drawback, but if the alternative is an unprepared exploration of larger models without a convincing alignment plan and with public releases/deployments where at-runtime improvements by the open-source crowd will unearth new capabilities, then that alternative isn't any better. Besides, people seem to overstate how likely new actors will catch up. Much of this will depend on the implementation/execution of compute governance efforts. High compute costs will pose a significant barrier, even if compute becomes cheaper over time. (Some restrictions to training compute wouldn't be static but based on a "moving bright line.") Export controls have already strengthened the US's lead internationally; we could extend them and complement them with monitoring and verification schemes.
- Besides, on some implementations of compute caps, currently-leading labs would get to keep their lead in the sense of having access to models they trained with amounts of compute that won't be allowed for a few years. Also, labs with an excellent safety and security culture that pass their safety evals might get allowances for bigger training runs. (Insofar as currently-leading labs don't meet these criteria, it would seem unwise to let them continue at a fast pace.)
- Lastly, we can consider *practical* objections to compute restrictions. Perhaps these will turn out to be insurmountable. However, I don't think we're in a position to know that before doing the policy research and advocacy to prepare for setting up some regulatory regime. Also, if things with AI end up going poorly because of poorly implemented compute restrictions – they probably wouldn't have gone any better with no restriction attempts.
    - In other words, while the impact of efforts to slow AI might be lower than we might think at first glance because it'll be tough to restrict and monitor

[5] Related post: Steven Byrnes on "endgame AI safety."

    compute sufficiently in practice, it's unlikely that we're making the future worse by trying.

- Arguments *in favor* of compute restrictions are pretty obvious and I've discussed them in an earlier draft. An additional consideration that's perhaps worth mentioning is the following.
    - Paul Christiano pointed out that the alignment community is growing faster than the ML community. There's an intuition here that it would be particularly dumb if we rush to scary AI before the growth slowdown on the logistic growth S-curve.
- For the above reasons, I think we should pursue regulatory interventions to slow AI progress via compute restrictions almost as soon as possible. I believe execution quality will likely matter a lot. So if we get something done slightly later, but it will be significantly better than a sloppy version that creates path dependencies that block us from later improvements, then it might be worth a short wait.