Cold Takes - Racing through a minefield: the AI deployment problem

BY HOLDEN KARNOFSKY • DEC 22, 2022 • 16 MIN READ •
IMPLICATIONSOFMOSTIMPORTANTCENTURY

Racing through a minefield: the AI deployment problem

Click lower right to download or find on Apple Podcasts, Spotify, Stitcher, etc.

In previous pieces, I argued that there's a real and large risk of AI developing dangerous goals of their own and defeating all of humanity - at least

Cold Takes - Racing through a minefield: the AI deployment problem

[Why it could be hard to build AI systems without deceptive ML] and [How it might be doable].

The "AI alignment problem" refers[1] to a *technical* problem: how can we design a powerful AI system that behaves as intended, rather than forming its own dangerous aims? This post is going to outline a **broader political/strategic problem, the "deployment problem":** if you're someone who might be on the cusp of developing extremely powerful (and maybe dangerous) AI systems, what should you ... do?

The basic challenge is this:

- If you race forward with building and using powerful AI systems as fast as possible, you might cause a global catastrophe (see links above).
- If you move too slowly, though, you might just be waiting around for *someone else less cautious* to develop and deploy powerful, dangerous AI systems.
- And if you can get to the point where your own systems are both powerful and safe ... what then? Other people still might be less-cautiously building dangerous ones - what should we do about that?

My current analogy for the deployment problem is **racing through a minefield: each player is hoping to be ahead of others, but anyone moving too quickly can cause a disaster.** (In this minefield, a single mine is big enough to endanger *all* the racers.)

This post gives a high-level overview of how I see the kinds of developments that can lead to a good outcome, despite the "racing through a minefield" dynamic. It is distilled from a more detailed post on the Alignment Forum.

First, I'll flesh out how I see the challenge we're contending with, based on the premises above.

governments, etc.) might do in order to prevent catastrophe.

Many of the actions I'm picturing are not the kind of things normal market and commercial incentives would push toward, and as such, I think there's room for a ton of variation in whether the "racing through a minefield" challenge is handled well. Whether key decision-makers understand things like the case for misalignment risk (and in particular, why it might be hard to measure) - and are willing to lower their own chances of "winning the race" to improve the odds of a good outcome for everyone - could be crucial.

The basic premises of "racing through a minefield"

This piece is going to lean on previous pieces and assume all of the following things:

- **Transformative AI soon.** This century, something like PASTA could be developed: AI systems that can effectively automate everything humans do to advance science and technology. This brings the potential for explosive progress in science and tech, getting us more quickly than most people imagine to a deeply unfamiliar future. I've argued for this possibility in the Most Important Century series.
- **Misalignment risk.** As argued previously, there's a significant risk that such AI systems could end up with misaligned goals of their own, leading them to defeat all of humanity. And it could take significant extra effort to get AI systems to be safe.
- **Ambiguity.** As argued previously, it could be hard to know whether AI systems are dangerously misaligned, for a number of reasons. In particular, when we train AI systems not to behave dangerously, we might be unwittingly training them to *obscure their dangerous potential from humans*,

Cold Takes - Racing through a minefield: the AI deployment problem

[illegible]. At the same time, I expect powerful AI systems will present massive opportunities to make money and gain power, such that many people will want to race forward with building and deploying them as fast as possible (perhaps even if they believe that doing so is risky for the world!)

So, one can imagine a scenario where some company is in the following situation:

- It has good reason to think it's on the cusp of developing extraordinarily powerful AI systems.
- If it deploys such systems hastily, global disaster could result.
- But if it moves too *slowly*, other, less cautious actors could deploy dangerous systems of their own.

That seems like a tough enough, high-stakes-enough, and likely enough situation that it's worth thinking about how one is supposed to handle it.

One simplified way of thinking about this problem:

- We might classify "actors" (companies, government projects, whatever might develop powerful AI systems or play an important role in how they're deployed) as **cautious** (taking misalignment risk very seriously) or **incautious** (not so much).
- Our basic hope is that **at any given point in time, cautious actors collectively have the power to "contain" incautious actors.** By "contain," I mean: stop them from deploying misaligned AI systems, and/or stop the misaligned systems from causing a catastrophe.
- Importantly, **it could be important for cautious actors to *use powerful AI systems* to help with "containment" in one way or another.** If cautious actors refrain from AI development entirely, it seems likely that incautious

doesn't seem good.

In this setup, **cautious actors need to move fast enough that they can't be overpowered by others' AI systems, but slowly enough that they don't cause disaster themselves.** Hence the "racing through a minefield" analogy.

What success looks like

In a non-Cold-Takes piece, I explore the possible actions available to cautious actors to win the race through a minefield. This section will summarize the general categories - and, crucially, why we shouldn't expect that companies, governments, etc. will do the right thing simply from natural (commercial and other) incentives.

I'll be going through each of the following:

- **Alignment (charting a safe path through the minefield).** Putting lots of effort into technical work to reduce the risk of misaligned AI.
- **Threat assessment (alerting others about the mines).** Putting lots of effort into *assessing* the risk of misaligned AI, and potentially demonstrating it (to other actors) as well.
- **Avoiding races (to move more cautiously through the minefield).** If different actors are racing to deploy powerful AI systems, this could make it unnecessarily hard to be cautious.
- **Selective information sharing (so the incautious don't catch up).** Sharing some information widely (e.g., technical insights about how to reduce misalignment risk), some selectively (e.g., demonstrations of how powerful and dangerous AI systems might be), and some not at all (e.g., the specific code that, if accessed by a hacker, would allow the hacker to deploy potentially dangerous AI systems themselves).

stopping them). Working toward a robust state-level monitoring effort to identify and prevent "incautious" projects racing toward deploying dangerous AI systems.

- **Defensive deployment (staying ahead in the race).** Deploying AI systems only when they are unlikely to cause a catastrophe - but also deploying them with urgency once they are safe, in order to help prevent problems from AI systems developed by less cautious actors.

Alignment (charting a safe path through the minefield[2])

I previously wrote about some of the ways we might reduce the dangers of advanced AI systems. Broadly speaking:

- Cautious actors might try to primarily build limited AI systems - AI systems that lack the kind of ambitious aims that lead to danger. They might ultimately be able to use these AI systems to do things like automating further safety research, making future less-limited systems safer.
- Cautious actors might use AI checks and balances - that is, using some AI systems to supervise, critique and identify dangerous behavior in others, with special care taken to make it hard for AI systems to coordinate with each other against humans.
- Cautious actors might use a variety of other techniques for making AI systems safer - particularly techniques that incorporate "digital neuroscience," gauging the safety of an AI system by "reading its mind" rather than simply by watching out for dangerous behavior (the latter might be unreliable, as noted above).

A key point here is that **making AI systems safe enough to commercialize (with some initial success and profits) could be much less (and different) effort than making them robustly safe (no lurking risk of global**

[alignment with safety research] in brief:

- If AI systems *behave* dangerously, we can "train out" that behavior by providing negative reinforcement for it.
- The concern is that when we do this, we might be unwittingly training AI systems to *obscure their dangerous potential from humans*, and take dangerous actions *only when humans would not be able to stop them*. (I call this the "King Lear problem: it's hard to know how someone will behave when they have power over you, based only on observing how they behave when they don't.")
- So we could end up with AI systems that behave safely and helpfully as far as we can tell in normal circumstances, while ultimately having ambitious, dangerous "aims" that they pursue when they become powerful enough and have the right opportunities.

Well-meaning AI companies with active ethics boards might do a lot of AI safety work, by training AIs not to behave in unhelpful or dangerous ways. But if they want to address the risks I'm focused on here, this could require safety measures that look very different - e.g., measures more reliant on "checks and balances" and "digital neuroscience."

Threat assessment (alerting others about the mines)

In addition to *making AI systems safer*, cautious actors can also put effort into *measuring and demonstrating how dangerous they are* (or aren't).

For the same reasons given in the previous section, it could take special efforts to find and demonstrate the kinds of dangers I've been discussing. Simply monitoring AI systems in the real world for bad behavior might not do it. It may be necessary to examine (or manipulate) their digital brains,[3] design AI systems specifically to audit other AI systems for signs of danger; deliberately train AI

~~dangerous?, etc.~~

Learning and demonstrating that the danger is high could help convince many actors to move more slowly and cautiously. Learning that the danger is *low* could lessen some of the tough tradeoffs here and allow cautious actors to move forward more decisively with developing advanced AI systems; I think this could be a good thing in terms of what sorts of actors lead the way on transformative AI.

Avoiding races (to move more cautiously through the minefield)

Here's a dynamic I'd be sad about:

- Company **A** is getting close to building very powerful AI systems. It would love to move slowly and be careful with these AIs, but it worries that if it moves too slowly, Company **B** will get there first, have less caution, and do some combination of "causing danger to the world" and "beating company **A** if the AIs turn out safe."
- Company **B** is getting close to building very powerful AI systems. It would love to move slowly and be careful with these AIs, but it worries that if it moves too slowly, Company **A** will get there first, have less caution, and do some combination of "causing danger to the world" and "beating company **B** if the AIs turn out safe."

(Similar dynamics could apply to Country A and B, with national AI development projects.)

If Companies A and B would both "love to move slowly and be careful" if they could, it's a shame that they're both racing to beat each other. Maybe there's a way to avoid this dynamic. For example, perhaps Companies A and B could strike a deal - anything from "collaboration and safety-related information sharing" to a merger. This could allow both to focus more on precautionary

Cold Takes - Racing through a minefield: the AI deployment problem

is discussed below, under Standards and monitoring.

"Finding ways to avoid a furious race" is not the kind of dynamic that emerges naturally from markets! In fact, working together along these lines would have to be well-designed to avoid running afoul of antitrust regulation.

Selective information sharing - including security (so the incautious don't catch up)

Cautious actors might want to share certain kinds of information quite widely:

- It could be crucial to raise awareness about the dangers of AI (which, as I've argued, won't necessarily be obvious).
- They might also want to widely share information that could be useful for reducing the risks (e.g., safety techniques that have worked well.)

At the same time, as long as there are incautious actors out there, information can be dangerous too:

- Information about *what cutting-edge AI systems can do* - especially if it is powerful and impressive - could spur incautious actors to race harder toward developing powerful AI of their own (or give them an idea of *how* to build powerful systems, by giving them an idea of what sorts of abilities to aim for).
- An AI's "weights" (you can think of this sort of like its source code, though not exactly[4]) are potentially very dangerous. If hackers (including from a state cyberwarfare program) gain unauthorized access to an AI's weights, this could be tantamount to stealing the AI system, and the actor that steals the system could be much less cautious than the actor who built it. **Achieving a level of cybersecurity that rules this out could be extremely difficult,** and potentially well beyond what one would normally aim for in a commercial context.

information might be useful for demonstrating the dangers *and* capabilities of cutting-edge systems, or useful for making systems safer *and* for building them in the first place. So there could be a lot of hard judgment calls here.

This is another area where I worry that commercial incentives might not be enough on their own. For example, it is usually important for a commercial project to have some reasonable level of security against hackers, but not necessarily for it to be able to resist well-resourced attempts by states to steal its intellectual property.

Global monitoring (noticing people about to step on mines, and stopping them)

Ideally, cautious actors would learn of every case where someone is building a dangerous AI system (whether purposefully or unwittingly), and be able to stop the project. If this were done reliably enough, it could take the teeth out of the threat; a partial version could buy time.

Here's one vision for how this sort of thing could come about:

- We (humanity) develop a reasonable set of tests for whether an AI system might be dangerous.
- Today's leading AI companies self-regulate by committing not to build or deploy a system that's dangerous according to such a test (e.g., see Google's 2018 statement, "We will not design or deploy AI in weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people"). Even if some people at the companies would like to do so, it's hard to pull this off once the company has committed not to.
- As more AI companies are started, they feel soft pressure to do similar self-regulation, and refusing to do so is off-putting to potential employees,

- Eventually, similar principles are incorporated into various government regulations and enforceable treaties.
- Governments could monitor for dangerous projects using regulation and even overseas operations. E.g., today the US monitors (without permission) for various signs that other states might be developing nuclear weapons, and might try to stop such development with methods ranging from threats of sanctions to cyberwarfare or even military attacks. It could do something similar for any AI development projects that are using huge amounts of compute and haven't volunteered information about their safety practices.

If the situation becomes very dire - i.e., it seems that there's a high risk of dangerous AI being deployed imminently - I see the latter bullet point as one of the main potential hopes. In this case, governments might have to take drastic actions to monitor and stop dangerous projects, based on limited information.

Defensive deployment (staying ahead in the race)

I've emphasized the importance of caution: not deploying AI systems when we can't be confident enough that they're safe.

But when confidence *can* be achieved (how much confidence? See footnote[5]), **powerful-and-safe AI can help reduce risks from other actors** in many possible ways.

Some of this would be by helping with all of the above. Once AI systems can do a significant fraction of the things humans can do today, they might be able to contribute to each of the activities I've listed so far:

- **Alignment.** AI systems might be able to contribute to AI safety research (as humans do), producing increasingly robust techniques for reducing risks.

demonstrations about potential risks. They could be potentially dangerous tasks like "Produce detailed explanations and demonstrations of possible sequences of events that could lead to AIs doing harm."

- **Avoiding races.** AI projects might make deals in which e.g. each project is allowed to use its AI systems to monitor for signs of risk from the others (ideally such systems would be designed to *only* share relevant information).
- **Selective information sharing.** AI systems might contribute to strong security (e.g., by finding and patching security holes), and to dissemination (including by helping to better communicate about the level of risk and the best ways to reduce it).
- **Global monitoring.** AI systems might be used (e.g., by governments) to monitor for signs of dangerous AI projects worldwide, and even to interfere with such projects. They might also be used as part of large voluntary self-regulation projects, along the lines of what I wrote just above under "Avoiding races."

Additionally, **if safe AI systems are in wide use, it could be harder for dangerous (similarly powerful) AI systems to do harm.** This could be via a wide variety of mechanisms. For example:

- If there's widespread use of AI systems to patch and find security holes, similarly powered AI systems might have a harder time finding security holes to cause trouble with.
- Misaligned AI systems could have more trouble making money, gaining allies, etc. in worlds where they are competing with similarly powerful but safe AI systems.

So?

Cold Takes - Racing through a minefield: the AI deployment problem

("racing through a minefield") if powerful AI systems (a) are developed fairly soon; (b) present significant risk of misalignment leading to humanity being defeated; (c) are not particularly easy to measure the safety of.

I've also talked about what I see as some of the key ways that "cautious actors" concerned about misaligned AI might navigate this situation.

I talk about some of the implications in my more detailed piece. Here I'm just going to name a couple of observations that jump out at me from this analysis:

This seems hard. If we end up in the future envisioned in this piece, I imagine this being extremely stressful and difficult. I'm picturing a world in which many companies, and even governments, can see the huge power and profit they might reap from deploying powerful AI systems *before others* - but we're hoping that they instead move with caution (but not too much caution!), take the kinds of actions described above, and that ultimately cautious actors "win the race" against less cautious ones.

Even if AI alignment ends up being *relatively* easy - such that a given AI project can make safe, powerful systems with about 10% more effort than making dangerous, powerful systems - the situation *still* looks pretty nerve-wracking, because of how many different players could end up trying to build systems of their own without putting in that 10%.

A lot of the most helpful actions might be "out of the ordinary." When racing through a minefield, I hope key actors will:

- Put more effort into alignment, threat assessment, and security than is required by commercial incentives;
- Consider measures for avoiding races and global monitoring that could be very unusual, even unprecedented.

Cold Takes - Racing through a minefield: the AI deployment problem

about the risks.

As such, it could be **very important whether key decision-makers (at both companies and governments) understand the risks and are prepared to act on them.** Currently, I think we're unfortunately very far from a world where this is true.

Additionally, I think **AI projects can and should be taking measures *today* to make unusual-but-important measures more practical in the future.** This could include things like:

- Getting practice with selective information sharing. For example, building internal processes to decide on whether research should be published, rather than having a rule of "Publish everything, we're like a research university" or "Publish nothing, we don't want competitors seeing it."
  - I expect that early attempts at this will often be clumsy and get things wrong!
- Getting practice with ways that AI companies could avoid races.
- Getting practice with threat assessment. Even if today's AI systems don't seem like they could possibly be dangerous yet ... how sure are we, and how do we know?
- Prioritizing building AI systems that could do especially helpful things, such as contributing to AI safety research and threat assessment and patching security holes.
- **Establishing governance that is capable of making hard, non-commercially-optimal decisions for the good of humanity.** A standard corporation could be sued for *not* deploying AI that poses a risk of global catastrophe - if this means a sacrifice for its bottom line. And a lot of the people making the final call at AI companies might be primarily thinking about their duties to shareholders (or simply unaware of the potential stakes

executives and board members - capable of making the hard calls well.

COMMENT/DISCUSS

Footnotes

1. Generally, or at least, this is what I'd like it to refer to. ↩

2. Thanks to beta reader Ted Sanders for suggesting this analogy in place of the older one, "removing mines from the minefield." ↩

3. One genre of testing that might be interesting: manipulating an AI system's "digital brain" in order to *simulate* circumstances in which it has an opportunity to take over the world, and seeing whether it does so. This could be a way of dealing with the King Lear problem. More here. ↩

4. Modern AI systems tend to be trained with lots of trial-and-error. The actual code that is used to train them might be fairly simple and not very valuable on its own; but an expensive training process then generates a set of "weights" which are ~all one needs to make a fully functioning, relatively cheap copy of the AI system. ↩

5. I mean, this is part of the challenge. In theory, you should deploy an AI system if the risks of not doing so are greater than the risks of doing so. That's going to depend on hard-to-assess information about how safe your

Cold Takes - Racing through a minefield: the AI deployment problem

for it." Seems hard. ↩

You might also like...

FEB 24	How major governments can help with the most important century	6 min read
FEB 20	What AI companies can do today to help with the most important ce...	13 min read
FEB 10	Jobs that can help with the most important century ★	36 min read
JAN 25	Spreading messages to help with the most important century ★	21 min read
JAN 13	How we could stumble into AI catastrophe ★	36 min read