

The Wildcard: Alignment

Shay O'Kelly · August 2026

Every essay in this series argues a version of the same claim. The buildout is real, the constraint is physical, and the value pools at the bottlenecks. An honest series has to end by naming the assumption underneath all of it, which is that the machines stay aligned with the people building them, and that nothing on that front goes wrong badly enough to stop the program. Most writing in my corner of the world either ignores the subject or waves at it on the way out. I want to do something different with it, because I think alignment, taken seriously, actually completes the thesis.

Alignment is the problem of making systems more capable than us reliably do what we intend, and it is not a fringe concern. In 2023, the chief executives of the leading labs signed a public statement that mitigating the risk of extinction from AI should be a global priority alongside pandemics and nuclear war¹. The people selling the technology signed that with their own names. Inside the buildings, alignment is discussed the way engineers discuss a known failure mode, and the doomsday view is held seriously by accredited people. A series that leans on lab revenue numbers for six essays owes their risk warnings honest treatment in the seventh.

The most common response to alignment fear is to say we should slow down. It sounds responsible, and inside a single lab it sometimes is. As national policy it fails on contact with the fifth essay, because a race does not stop when one runner does. China added 543 gigawatts of power in a single year, is standing up a domestic lithography industry tool by tool, and open-weights its best models specifically to erode the American lead². None of that pauses for a safety review, and nobody serious believes it would pause because Washington did. A unilateral American slowdown would just relocate the frontier to whichever actor is spending the least on the problem the pause was meant to solve. A pause is a handoff.

Which flips the whole question. If the race runs regardless, the choice was never race or do not race. It is lead or trail, and the two are not symmetric on safety. Alignment work costs compute, talent, and time, and those are luxuries of whoever is ahead. A trailing power cuts corners to catch up. A leading power can afford margin. To hold a model back for months of testing, to spend a real share of its compute on safety research, to say no to a deployment. The lead is the safety budget. And America's lead has a large moat because of the semiconductor supply chain, the same chain the fourth and fifth essays called the hardest bottleneck on earth. The people most worried make this exact argument. Amodei, the same chief executive whose extinction warning appears above, has argued publicly for what he calls an entente strategy, where the democracies lock down the chip supply chain, the chips, the tools, and the equipment, precisely to hold a lead wide enough that safety work can be afforded at the frontier³. The man most alarmed about the machines wants the semiconductor moat deeper. If you take alignment seriously, the chain that keeps the careful side ahead is both the scarcest economic asset in the world and the margin of safety for everyone, which makes it more valuable, not less. The caveat cuts the other way, and it deserves its sentence. A lead

only buys safety if you actually spend some of it on care, and a leader who races like a trailer has wasted the entire point of leading.

There is a second, quieter reason alignment concentrates value in the chain. Suppose the world someday gets serious about governing frontier AI. Treaties, audits, compute limits, the whole apparatus. What would enforcement even attach to? You cannot regulate an idea. Weights copy for free and leak like water. The only layer of the stack that is physical, countable, and chokepointed is the hardware, a handful of lithography suppliers, a few foundries, fabs that take years and tens of billions of dollars to build. Every compute-governance proposal worth the name runs through the chip chain, because it is the only steering wheel that exists. Race harder, and the chain is the lead. Regulate harder, and the chain is the enforcement mechanism. There is no version of taking AI seriously, optimist or doomer, that does not end with a government hand resting on the same few chokepoints this series has been describing.

There is a third development worth naming, because it changes the safety math, and it is the shift toward open source. More and more strong models are now released with their weights public, and that does two things at once. First, everyone can see how the model is built. Inside a closed lab, a few hundred employees can study the system. When the weights are open, every researcher on earth can, and the field's most important discoveries about how these models actually behave have come from exactly that kind of outside scrutiny⁴. Second, everyone has access to the same models. The defenders hold the same tools as the attackers, the small hold the same tools as the large, and nobody is at the mercy of what a closed lab chooses to sell them. The sensible caveat is the frontier, the very strongest models should be tested before they are shared, and the labs that release openly already test first. But the conclusion runs deeper than either point, and Zuckerberg has made it better than anyone⁵. The real danger was never that everyone would have this technology. It is that only a few would. Spread the models and you spread the risk with them, thinly, across billions of people with different aims, instead of stacking it inside a handful of institutions where one bad decision travels the whole world. Concentrated power is a single point of failure. Open source is how you refuse to build one. The shift toward openness is not a compromise on safety. It is a mechanism of it.

Now the honest part. I do not have an edge on whether alignment gets solved, and I distrust anyone who claims one, because the question is not answerable from public information and maybe not from private information either. A genuine alignment crisis, a failure ugly enough to force a political stop, remains the true bear case for everything in this collection, and I will not pretend it is hedgeable, because in that world every risk asset has a bad decade, not just mine. But notice what every branch short of catastrophe does. Racing makes the chain the lead. Governance makes the chain the lever. Crisis makes the chain a strategic stockpile, nationalized rather than written off. Alignment fear, the one force that reads like this thesis's enemy, puts a government floor under the hard bottlenecks in every scenario on the way to the bad one. If anything, alignment fear deepens the case for the bottlenecks. My posture is the one this collection is named for. Conviction about the thesis, respect for the risk, watchfulness over both, and a public change of mind the moment the evidence turns.

So that is the series. The buildout is real, the constraint is physical, the value pools where supply cannot answer, the race will not pause, and the safest path through the danger runs down the same chain the money does. I know where I stand. Long the constraint, patient on the timeline, and never, ever short human wanting.

SOURCES

1. Statement on AI Risk, Center for AI Safety, May 2023, signed by the chief executives of OpenAI, Anthropic, and Google DeepMind, among others.
2. China power additions, domestic lithography program, and open-weight strategy. See sources in the arms race essay.
3. Dario Amodei, "Machines of Loving Grace" (October 2024), on the entente strategy and securing the chip supply chain. See also his "On DeepSeek and Export Controls" (January 2025).
4. NTIA report on open-weight models (July 2024), finding insufficient evidence of marginal harm to justify restrictions, and interpretability research built on open weights, incl. Arditi et al., Refusal in Language Models Is Mediated by a Single Direction (2024).
5. Mark Zuckerberg, "Open Source AI Is the Path Forward" (July 2024) and his August 2026 essay on the concentration of power.

No Offer or Solicitation. The only purpose of this site is to provide general background information on RAXIA. This is not an offer to sell or the solicitation of any offer to buy any securities. Any such offer would be made only to qualified investors after distribution and return of the Private Placement Memorandum and governing agreement for the applicable Fund (the "Offering Materials"). Nothing on this site is investment advice or a recommendation of any security. The author may hold positions in companies discussed. Past performance is not indicative of future results.