AI Futures Project

AI 2040

Plan A — The Deal

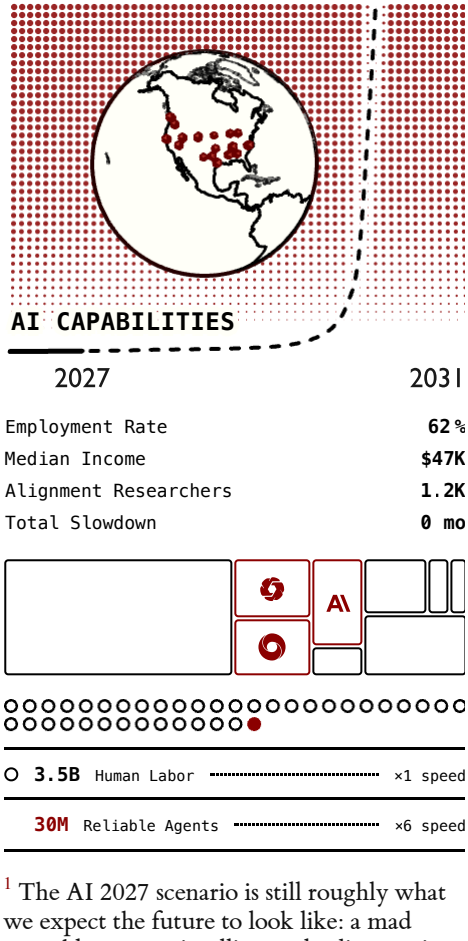
A plan for the development of
superhuman AI · ai-2040.com

Design by
Lightcone Infrastructure

AI companies are racing to build AIs that are smarter than humans in every way. In AI 2027, we predicted that this would result in either extinction or irreversible concentration of power.[1]

Plan A is our positive vision for what should happen instead.

In this scenario, humanity delays the development of superintelligence until 2040, makes all AI research public, allows dozens of companies globally to catch up to the frontier, and intentionally enters a regime of mutually assured compute destruction.



What is Plan A?

Plan A is our positive vision for how humanity can avoid AI-driven existential catastrophe and reach a flourishing future. It's informed by conversations with experts at major U.S. frontier AI companies, direct experience at OpenAI, tabletop exercises, and discussions with policymakers, national security experts, and AI policy leaders. We recommend an international deal to avoid a dangerous race to superintelligence. The deal involves total research transparency for AI R&D, which allows the nations of the world to understand what's happening and enforce guardrails. The result is multiple companies across multiple countries scaling slowly and safely together towards superintelligence, instead of racing each other in secrecy.

Plan A is primarily a recommendation, not a prediction. This scenario is *not* our best guess as to what the future will actually look like. Instead, it's a vehicle for communicating and stress-testing our policy recommendations. While the *implementation* of Plan A is a recommendation and not what we actually expect to happen, the *subsequent effects depicted* are predictions.[2]

In this AI 2040 scenario, Plan A is implemented successfully, albeit imperfectly and only in the nick of time.

We contrast Plan A with 4 alternative plans (B, C, D, and S), which correspond to the main ways the US could respond (or not) to the challenges of superintelligence.[3]

Why did we write this?

AI companies will probably succeed at their stated goal of building smarter-than-human AI systems within the next 1 to 10 years.

The industry has convinced itself that controlling superintelligent AI can be figured out on the fly, and thus has no remotely adequate plan. We think this situation is terrible and could easily get us all killed.[4] We do not expect whoever "wins the race" to have much of a lead, and we do not expect them to unilaterally slow down to reduce existential risk.[5] If this race continues,[6] we do not expect humans to maintain effective control as their AIs become superintelligent.[7]

Moreover, even if the AI companies somehow align their AIs, the result will be an unprecedented concentration of power—that is, the result will be a situation where a tiny group of people, or possibly just a single individual, is ef-

[1] The AI 2027 scenario is still roughly what we expect the future to look like: a mad scramble to superintelligence leading to either AI takeover or extreme concentration of power. So far reality is tracking closer to AI 2027 than even we expected. (2027 was our modal year at time of publication, not our median.) You can read more about our views on timelines here and here.

[2] That is, they are predictions about what would happen next if our recommendations were implemented.

[3] You can read more about these four plans at the branch point below and in our supplement "How good is each plan?".

fectively in control of the world's only army of superintelligences for some months, and will be presented by said superintelligences with various options for how to proceed, some of which will *de facto* amount to taking over the world.[8]

As best as we can guess, the CEOs of OpenAI, Anthropic, and Google DeepMind understand this and are proceeding anyway, perhaps because they think they are the lesser evil and will use their immense power responsibly, unlike Xi Jinping or rival CEOs.[9]

While we agree that it is generally correct to choose the lesser evil, we don't think we should advocate for a strategy that has such a scarily high chance of leading to human extinction or global dictatorship. Instead, we wish to advocate for something that is actually good. If enough people do likewise, it can happen.

So, we wrote a scenario outlining that possible world.

Why a scenario?

"Plans are worthless, but planning is everything." - Dwight D. Eisenhower

We think most AI policy proposals fall apart under scenario scrutiny—that is, if you try to write down a detailed and plausible scenario in which that proposal succeeds, you will find it difficult to do so, and you will realize the plan is less likely to work than it seemed, or has more unpleasant side-effects than its proponents acknowledged.

Perhaps that's why scenario scrutiny is so rare in AI policy. Everyone wants to say that their own favorite policies will have great consequences and that the policies of their rivals will have terrible consequences. Applying scenario scrutiny to their own favorite policies might surface uncomfortable issues with them; meanwhile, applying scenario scrutiny to their rival's policies is a lot of work for little rhetorical gain.[10]

We think the discourse would be improved if more AI policy proposals were subjected to scenario scrutiny. So we're starting with our own, even though this opens us up to criticism. We hope critics will judge us against the existing state-of-the-art for plans to navigate the AI transition (if they can find any) and not against some hazy but pleasant fantasy where no one has to make any hard choices yet everything will probably be fine.

What of the immense difficulty of predicting the effect our policy would have in a world approaching superhuman AIs? This is like trying to predict how to best fight World War 3, except that it's an even larger departure from past case-studies. Yet it is still valuable to attempt, just as it is valuable for the U.S. military to game out Taiwan scenarios in excruciating detail. There are other precedents as well: intelligence agencies, climate bodies, and pandemic-preparedness offices all rely on various kinds of scenario planning.

[4] Of course, we're not sure exactly how likely literal AI-driven human extinction is, but opinions within our team vary between 10% and 30%. This is primarily because there are a bunch of other things misaligned ASIs might do with us after they take over, besides kill us all. For example, they might keep some people alive because it's really cheap and they care a tiny amount, or as acausal bargaining chips. We don't think this nuance changes the bottom line though: Misaligned AI takeover is probably really bad.

[5] I (Daniel Kokotajlo) recently gave a talk to about 100 people, ~40% of whom were from frontier AI companies, and did a show-of-hands poll about (a) how many months of lead the leading AI company would have at the time they first fully automate AI R&D, and (b) how much of their lead they would be willing to burn. The median answers were roughly 3 months and 1/3rd, respectively. This matches my own guess.

[6] Also intense secrecy and groupthink conditions! By default during this period the companies will be even more closed than they are in 2026. When they say "we'll solve the alignment problems as we go" what they really mean is "The overworked tiny group of alignment experts at our company will solve the alignment problems as we go, without being able to subject their work to external review."

[7] We expect them to maintain **apparent** control, to be clear, at least for some time, which is part of why this problem is so dire. The situation would be much better if we had a reliable way of telling whether an AI system was actually robustly obedient/loyal/aligned/etc. vs. merely temporarily so vs. merely pretending.

[8] The AI 2027 Slowdown ending depicts this happening. It's quite easy to imagine because it doesn't involve anything cartoonishly or overtly evil, but rather just a series of steps that can be justified individually (consolidating US compute into a single pool, beating China, integrating AI into the government and economy, preventing terrorists from getting their own AGIs, etc.)

[9] For example, these old OpenAI emails reveal that OpenAI was founded in significant part because of fear that Demis Hassabis, DeepMind CEO, would use AGI to become dictator. See also related reporting in Empire of AI.

AI TIMELINES?

Plan A is our ambitious proposal for what to do, and we'd like to see something like it implemented soon because we are uncertain about how much time remains.[11] But for purposes of writing a concrete scenario, we need a concrete timeline.

The timeline of this scenario is:

- In 2029, the US and China agree to avoid a reckless race to superintelligence.
- **In 2030, we *would have* fully automated AI R&D**, leading to superintelligence by the end of the year. **Thanks to the deal, we avoid this.**
- Between 2030 and 2035, we scale within the human range, to AIs that are roughly as capable as top human experts.
- In 2035, we pause at top-human-expert level AI in order to maintain human control.
- In 2040, we unpause and scale to superintelligence.[12] (**Hence the title: AI 2040**)

In our previous scenario, AI 2027, AI fully automated the process of building smarter AIs in 2027, leading to an intelligence explosion and superintelligence within the year. The two differences in this scenario are (1) the default timeline is now 2030, and (2) thanks to governance actions, generally-superhuman AIs first appear in 2040.

We changed the default timeline because we want our portfolio of scenarios to reflect our uncertainty about AI timelines. AI 2027's titular year was chosen because, at the time we started writing, Daniel thought there was roughly a 50% chance that things would go that fast or faster.[13] At the time we started writing Plan A, 2030 was the corresponding year for Thomas. **Daniel currently thinks things will probably go somewhat faster than depicted in this scenario**; you can read more about our team's views on timelines here and here.

We changed the governance actions because **this scenario is primarily a recommendation, not a prediction.** Conducting a full-speed intelligence explosion is wildly reckless and concentrates power to an extreme degree.

[10] For one thing, it's hard to apply scenario scrutiny to a policy proposal if you have only a shallow understanding of it, and people generally understand their own favorite ideas much better than they understand the ideas they hate. For another, the evidential force of scenario scrutiny is inherently stronger for scenarios written by proponents than by opponents: If a proponent of a policy tries their best to depict it succeeding, and fails, that's a lot more evidence than if an opponent of a policy depicts the policy failing.

[11] That is, we are uncertain about how fast AI capabilities will progress – about how much time remains before the AI companies succeed at automating their research and accelerating towards superintelligence, for example.

[12] Superintelligence means AI systems that are significantly better than the best humans at everything, while also being faster and cheaper.

[13] By the time we finished writing AI 2027, Daniel's AGI median was 2028. Other authors' medians were between 2031 and 2035. To read more about how Daniel and Eli's views have shifted over time, see here. We are regularly updating our forecasts here, with corresponding blog posts on our Substack.

2027: THE WRITING ON THE WALL

America has two workforces now. The first is 165 million people. The second is AI agents: millions of copies spun up and shut down every hour, working around the clock at superhuman speeds.

Most of their work is slop. But enough of it is good that people are paying ten billion dollars a month for AIs that can, in theory at least, do anything on a computer that an employee can.

There is one job the AI companies want to automate more than any other—their own. They haven't succeeded yet; no recursive self-improvement so far. But they seem to be getting closer, and they're pulling up the ladder behind them: the strongest coding AIs refuse to help competitors with AI R&D.[14] Even as the most bullish employees admit that things are taking a bit longer than planned, the skeptics notice that their usual dismissals are starting to ring hollow. Why exactly will AI never be able to do my job? What's the barrier again?

[14] This policy is selfishly good for leading AI companies insofar as it makes it harder for trailing companies to catch up, but it's also good for safety because it slows irreversible proliferation of algorithmic insights.

Congress is starting to pay more attention. They've long been hearing about AI: datacenters using too much water,[15] chatbots encouraging suicide, Mythos hacking NSA systems—and of course, tech industry lobbyists warning that any whiff of regulation will make America immediately lose the race with China and spend the rest of history as a CCP tributary state.[16]

[15] We think the concerns around current AI water usage are largely overstated, see this article for more. That said, as will be apparent later in the scenario, we do think that the environmental impact of AI over the next decade or two will be enormous, and that unrestrained growth would literally boil the oceans. See the energy appendix in our economics supplement for more on this.

Now they step back and ask: Where are we going with this? What does the world look like five, ten, or fifteen years from now? Will there still be jobs? What if there aren't?

One question weighs especially heavily on their minds: **Who will control all these AIs?**

Congress settles on an important part of the answer: **Probably not us.**[17]

They hold a series of tense hearings on AI. They read the 2016 OpenAI emails discussing how OpenAI was founded in order to prevent Demis Hassabis from becoming dictator.[18] But who is preventing Sam or Elon from becoming dictator? Congress is unsatisfied with existing responses.

The result of this wakeup is the AI Transparency Act of 2027, an omnibus bill that does many things, some good and some bad, but doesn't fundamentally change the situation.[19]

➤ *See* Appendix A — Incremental AI Policy Wishlist *for more detail.*

2028: AI on the Ballot

The 2028 election cycle is heated, as usual. AI is the biggest topic. The datacenters now under construction cost twice as much as the entire US military budget.[20]

Most white-collar professions are seeing disruption like software engineering saw in 2026; such jobs now heavily involve managing AI agents. AI companies have industrialized the training process: Executives say "let's move into [profession] this year" and then the company interviews professionals, buys data, creates training environments, etc. until their AIs get traction. Then the AIs rapidly improve as they are used more widely in the field and accumulate more real-world data.

Other countries are starting to get scared and angry. It seems like a handful of US and Chinese companies are on track to automate all the white-collar jobs. Power is concentrating in the US, and in particular in the President plus a

handful of tech CEOs.

AI experts warn that the intelligence explosion is near. By speeding up AI research, the AIs will become even more competent, speeding up research even faster, making them even *more* competent, and so on. There are complicated dynamics about bottlenecks and hardware limits governing how fast this process goes and where it ends, but it seems like it might go very fast and end somewhere very far away.

On the default path, the next presidential term will see AIs that are far beyond human level, created entirely by AIs, themselves created entirely by other AIs, without any human in the loop since several generations back. Will those AIs be obedient, aligned, etc.? Why? Who will control them if so? How exactly is all of this supposed to end well?

Having put humanity on this path, the AI companies find it acceptable. But most people don't. Forget thinking about his *legacy*—the President is starting to think about what'll happen to *him* after he leaves office and the world gets transformed.[21] Both presidential candidates keep getting asked what they'll do about AI, and try out increasingly dramatic ideas on the campaign trail. The discourse bounces back and forth across all of the options displayed below, and more.

Eventually the President and his protégé converge on one plan; the opposition candidate converges on another. Then it's Election Day.

2029: Choose a Path

"Trust, but verify" – Ronald Reagan

The President announces that the US will pursue international cooperation to avoid an imminent intelligence explosion.

"This mad race toward superintelligence must end. For too long, we have been pursuing lesser-evils and least-bad solutions. We need a Plan A. We can still proceed with AI development, but we must do it *more cautiously, more transparently,* and involving *many more countries and companies*."

To the surprise of many in DC, China proves receptive. They had been debating the same issues—social destabilization, job loss, rogue superintelligence - on their side of the Pacific. They had been looking forward to "The Chinese Century" and thought that AI might disrupt their plans. And they had one extra reason to come to the table: the US continues to have more compute, more datacenters, and better models. They were worried about the things the US might do to them if it got to superintelligence first.[22]

➤ *See* Appendix B — Why would China be interested in a deal? Why would anyone? *for more detail.*

The US and China don't trust each other. Fortunately, they don't have to: Plan A includes provisions for verifying compliance. But setting it up will take time.

[16] For a representative example, Doug Kelly, CEO of a Meta-funded lobbying organization, wrote: "Rushing into a patchwork of uncoordinated state laws will only slow American innovation and give China an opportunity to surge ahead and win this tech race." Are we really in such a race? Sort of, but not exactly. On our default trajectory, AI will soon become very powerful, to the point where it's more strategically important than nuclear weapons; that much is true. Many people, including the CEOs of frontier AI companies, are trying hard to build smarter and smarter AI systems before their competitors do. However, being overly paranoid about competitors is a well-known bias. So it's difficult for AI developers to judge the extent to which their competitors (whether within the US or within China) are proactively trying to overtake them, as opposed to merely themselves being scared of falling too far behind. Assuming that we're in a winner-takes-all race rules out many possible avenues for cooperation, creating a self-fulfilling trap. So wherever possible, we encourage the use of more nuanced concepts than "AI arms race". We are particularly wary that the "AI arms race" framing will be cynically promoted for personal gain (as seems to be happening in these ads). Setting up an adversary as a bogeyman is a well-worn strategy for gaining power domestically. However, we expect that the scale and speed of AI progress will make it increasingly obvious to decision-makers in both the US and China that the current pace of progress, and lack of trust between them, is extremely dangerous. Indeed, it may be possible for the two countries to coordinate on a slowdown without any binding agreement, if they build trust by each gradually deescalating. Our proposed Plan A should be interpreted as an approach that could work *even if* the US and China have absolutely zero trust for each other, rather than a claim that strict verification protocols are definitely necessary.

[17] That is, not Congress. By default the people who control the AIs, if anyone truly does, will be AI companies or maybe the White House.

[18] From Ilya and Greg: "The goal of OpenAI is to make the future good and to avoid an AGI dictatorship. You are concerned that Demis could create an AGI dictatorship. So do we. So it is a bad idea to create a structure where you could become a dictator if you chose to, especially given that we can create some other structure that avoids this possibility."

So for now, they start with something crude. For the rest of 2029, they put a temporary halt to AI training, because that's relatively easy to verify. In 2030, they'll have the infrastructure in place to proceed with Plan A.

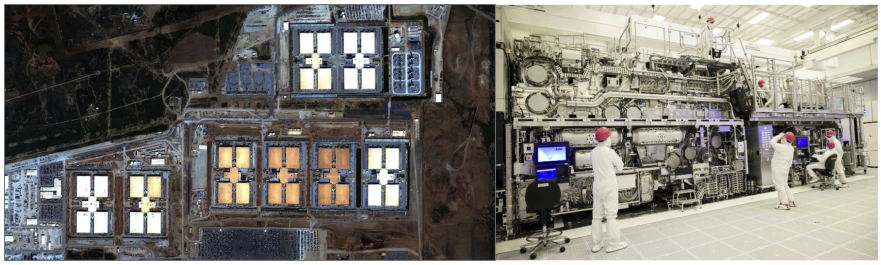
2029: Hurried Negotiation	2030: Plan A is Established
The US and China negotiate a series of agreements that eventually result in Plan A.	The Consortium negotiates governing principles.
Compute Declaration	Buy Time
Training Pause	Total Research Transparency
Get Worldwide Buy-in	Diffuse AI Broadly
	Reversibility

The stock market is gyrating wildly up and down in response to news and commentary about the momentous actions being taken. Everybody is screaming about how it's going too far or not far enough. The market will calm down by the end of the year, but the screaming will continue.[23]

Step 1: Compute Declaration

Neither country wants to halt their own AI training unless they can see that the other side is halting too.

Fortunately, training AIs requires large numbers of AI chips. Most AI chips are in giant datacenters.[24] AI datacenters are typically big enough to be visible from space, and power-hungry enough to require conspicuous infrastructure. New AI chips can only be manufactured at a handful of fabrication plants (fabs), located mostly in Taiwan, South Korea, the US, and China.



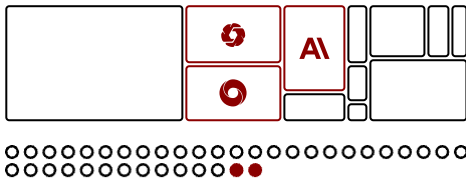
Left: OpenAI's Stargate datacenter. Right: An Extreme Ultraviolet Lithography (EUV) machine, arguably the most complicated machine ever created by mankind. EUV machines are essential for producing frontier AI chips, and are produced by a single Dutch company, ASML.

The US and China negotiate with the countries that have a major role in the chip supply chain, and they require each major datacenter owner (and their upstream suppliers, including chip fabs) to publicly declare their major purchases and sales.[25] Analysts from numerous countries and institutions may pore over the list, ask questions, and challenge anomalies. The rival powers also send inspectors to each other's infrastructure, reassuring themselves that the stated numbers are accurate. By the end of the year, each side is confident that the other isn't hiding more than about 1% of AI compute, and that all potential sources of new compute are monitored and accounted for.[26]

[19] We don't mean to *recommend* bad things; we are *predicting* that a mixture of good and bad incrementalist reforms will happen in 2027. For more on how we balance predictions and recommendations in this scenario, see here.

2028

Employment Rate	62%
Median Income	$49K
Alignment Researchers	1.4K
Total Slowdown	0 mo



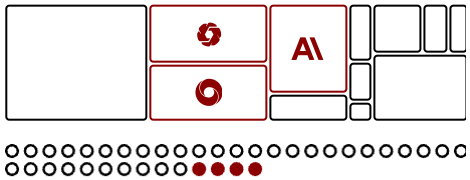
3.5B	Human Labor	×1 speed
28M	Reliable Agents	×6 speed

[20] That is, the AI industry spends $2.4T of CapEx over 2028, triple what was spent in 2026. The US military budget in 2025 was about $1T. The biggest AI company has $360B ARR at the beginning of 2028 and is still growing at a roughly 150% annualized growth rate, putting it on track to become both the most valuable company in the world and the largest company in the world by revenue, in 1-2 years.

[21] After all, he'll still be alive after leaving office. Just like the rest of us, he'll have to endure whatever crises result from superintelligence, and hope that the new President handles them well and doesn't become dictator. It's in his interest both to set the new President up for success with respect to the loss-of-control problem, and to install checks and balances to prevent extreme concentration of power.

2029

Employment Rate	62%
Median Income	$50K
Alignment Researchers	1.7K
Total Slowdown	0 mo



3.5B	Human Labor	×1 speed
23M	Reliable Agents	×7 speed

Step 2: Pause Training

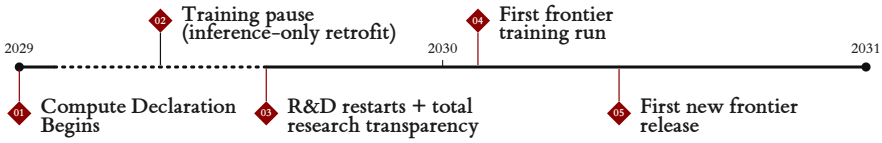
Distinguishing dangerous AIs from safe ones will require more time and understanding, so for now they go with a simple solution: a temporary pause on all new training runs.[27]

Both sides can still use datacenters for running AIs that already exist (i.e. inference), but they will retrofit each other's datacenters with devices to verify that they aren't being used for new training runs.[28]

The US and China are able to rapidly source enough of these verification devices to retrofit almost every major datacenter.[29]

➤ *See* Appendix C — What if they didn't have inference-only verification ready? *for more detail.*

Inspectors from both countries sign off on the installation of the verification devices. Only datacenters that the US and China agree have functioning verification are allowed to provide continuing AI services. Meanwhile, both governments sprint towards more fine-grained verification options that could allow training to restart while keeping both sides confident the other isn't racing ahead.



Deal Implementation Timeline – ai-2040.com

Step 3: Get Worldwide Buy-in

Though other countries were consulted from the start, especially those owning parts of the AI supply chain, Plan A began as a bilateral deal between the US and China. Now that it's taking firmer shape, negotiations go multilateral. Carrots and sticks are waved about.

Many countries outside the US and China are happy about the deal, because they were worried about a future in which a handful of US and Chinese AI companies recursively self-improve, pull farther and farther ahead, and then… Well, what happens next depends on who you ask, but answers range over possibilities such as "take our jobs," "cement hegemony forever" and "get everyone killed." So they want the US and China to proceed more slowly and transparently. This will allay their fears and also allow their own AI projects to catch up to the frontier.

So it's surprisingly easy for the US and China to get buy-in.[30] By the end of the year, most of the world has joined what is now called the Consortium.

Alternate Timeline

China attempts a covert AGI project

➤ *See Appendix D for this branch.*

[22] To spell it out more: They were concerned that the US might use their massive AI advantage to cripple Chinese AI projects via cyberattacks, physical sabotage, or other means, and then use the resulting even-bigger, longer-lasting AI advantage to dictate terms to the CCP or even overthrow it entirely. Preventing loss-of-control risk was merely an added bonus.

[23] Geopolitically the situation will have stabilized within a few years, but by that point AI progress and diffusion will have happened and caused new problems and crises for people to worry about.

[24] We show our estimate of the distribution of datacenter sizes in our compute supplement.

[25] The initial chip declaration would only involve companies whose purchases exceed 10k H100e (which costs ~$100M). Within a year, every company owning above 0.001% of the world's compute (~2,800 H100e, costing ~$18M at the time) has its holdings accounted for. Dead chips are also stored or verifiably destroyed.

[26] The 1% of compute sales that cannot be traced to their end location are mostly due to smugglers who did not record their customers' identities and chips that were decommissioned. Half of this pool is later found installed in Chinese datacenters identified via satellite imagery; both countries offer monetary incentives and legal amnesty to any owners of the remaining 0.5%. We go over these and other routes for finding covert compute in much more detail in our covert project supplement.

[27] This includes the kind of research experiments that involve training. While it's theoretically possible for an intelligence explosion to happen without new training runs or major experiments, it seems unlikely because it would require an extreme paradigm shift.

2030: Plan A is Established

Tech companies are pushing hard to restart AI training, and lobbying to shape the conditions under which it happens. Large portions of the public are saying that we should *never* restart. Some countries are too.

After a year of intense discussions, negotiations, and twitter beefs, a compromise takes shape: AI development will continue, but in a more cautious, more transparent, more distributed way.

Plan A is guided by 4 core principles:

Buy Time Slow down whenever is needed to have high confidence in safety.

Total Research Transparency Make almost all AI research fully visible to the public.

Diffuse AI Broadly Many companies in many countries at the frontier.

Reversibility Limit algorithmic progress; maintain Mutually Assured Compute Destruction.

Principle 1: Buy Time

The problem with an intelligence explosion is the "explosion" part.

The default trajectory is reckless and destabilizing. Each country and company is racing their competitors for AI dominance without time to think. Most likely, a ton of people will end up disempowered or killed. Maybe godlike AIs run circles around us and the age of humanity is over. Maybe a CEO or President establishes a permanent dictatorship. Maybe World War Three begins in the chaos, as countries without the best AIs fear losing their leverage. Maybe everything is surprisingly smooth and we "only" automate all labor and render everyone unemployable in a radically transformed world. Even aspiring AI-enabled dictators are worried; there's enough frontrunners that none of them have more than 50% odds of winning the power struggle.

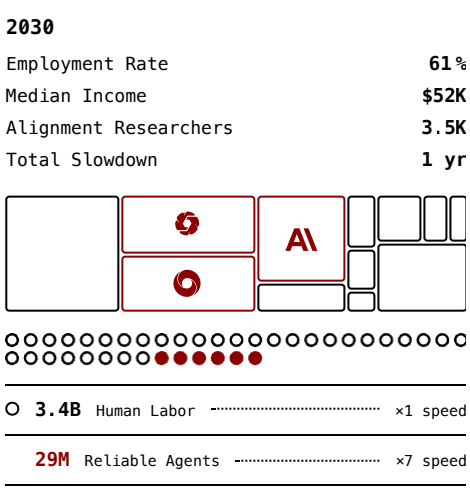
No one knows how to tell if AIs are trustworthy, and no one knows how to regulate superintelligent AI. Solving those problems looks like it'll take years. Slowing down the intelligence explosion buys time to do so.

It also helps prevent extreme concentrations of power, because it gives more time for groups that don't directly control the world's smartest AIs to wake up to what's happening before they lose their leverage. Without a major slowdown, a rapid intelligence explosion will concentrate almost all the power into a tiny group of CEOs and politicians.

[28] The inference-only verification solution we propose works by taking a random sample of a small percentage of the workloads to check they are correctly doing inference on an approved model (by recomputing the outputs using the inputs and approved model weights). To collect these random samples, our proposal involves simple network taps that redirect all inbound and outbound traffic from the datacenter to a recomputation server (inside the datacenter) installed by the other party that performs the random partial recomputation. Other solutions to this inference-only proposal may be possible and preferable, such as cryptographic proof of work schemes, but these are currently more speculative from a performance perspective. We discuss this in more detail in our verification supplement. There are several additional layers of enforcement necessary to make this approach work, including physical security, software based monitoring, and human auditors.

[29] This is made easier thanks to a now-thriving ecosystem of verification hardware companies that emerged after both governments signaled interest in having verification capability as an option (which in turn drew in investment from philanthropists and VCs and technical support from AI supply chain companies). Had that not happened, this part of Plan A would have taken somewhat longer and/or been more expensive and frantic. Worst case, they'd have to turn off the datacenters instead of allowing them to run inference.

[30] This part is a prediction, not just a recommendation. We think that getting buy-in for Plan A, in circumstances like those we've described in 2029, would be significantly easier than most of our readers would expect.



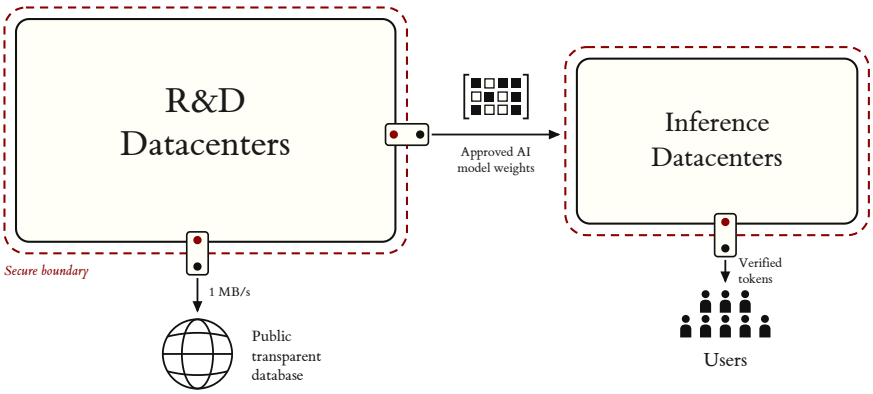
➤ *See* Appendix E — Why don't we just completely stop? *for more detail.*

Principle 2: Total Research Transparency

So the goal is to develop AI at a reasonable pace: Not too fast, not too slow. Ban the unsafe kinds, allow the safe kinds. But each country still regulates its own AI industry; there isn't a central planner with authority to regulate AI worldwide. So how do they decide what to allow and disallow, and how do they verify that no one is cutting corners?

The Consortium countries come up with a simple high-level framework: we'll agree to let each other see all the AI research. Then, if we don't like something someone is doing, we'll talk about it and perhaps agree to ban it.

➤ *See* Appendix F — Why this much transparency? Why not less, or more? Open access in Plan A *for more detail.*



AI workloads can be split into research and development (the process of building new AIs) and inference (the use of the existing AIs). In our proposal, research is almost entirely transparent, while inference is still private. For more detail on how this could work, see our Transparency supplement.

Datacenter verification architecture – `ai-2040.com`

Transparency offers many benefits. First, when each company can see what the others are doing, the expertise of governments is less load-bearing. If a company starts doing something dangerous, rival companies, third-party auditors, etc., can notice and raise the alarm. This both massively increases the amount of brainpower devoted to making the frontier AIs safe, and also prevents a situation where the only participants in the technical conversation are biased AI company employees and overworked, outnumbered regulators. Also, increased visibility simplifies the problem of verifying compliance with regulations, especially in grey-area cases.

Second, total research transparency makes it nearly impossible for secret loyalties, biases, or agendas to be intentionally trained into AIs. Everyone can see exactly how each frontier AI is trained. More generally the transparency makes it easier for groups that don't directly control frontier AIs to notice and prevent abuses of power by those who do.

Third, there's no longer as much incentive to race to discover new AI paradigms and more powerful algorithms, because companies wouldn't be able to hoard such discoveries and profit greatly from them. This buys the world time.

➤ *See* Appendix G — Diagram of the benefits of transparency *for more detail.*

➤ *See* Appendix H — How the incentives change under Total Research Transparency *for more detail.*

Principle 3: Diffuse AI Broadly

This principle is largely achieved by the interaction of the previous two:[31] Because algorithmic secrets are made public and the pace of progress isn't accelerating, other companies will catch up to the frontier. The result will be a competitive market for AI, in which consumers of AI services have many options to choose from and excellent visibility into what they are buying. Because frontier model training is totally transparent, people can verify that the published Spec accurately describes the goals and values being trained into a model.

This is very important for preventing extreme concentrations of power. But it also helps accelerate the use of AI to solve the world's most pressing problems. It helps with loss of control risks by accelerating AI alignment and control research. It helps accelerate the development of new and improved verification technology, and it helps to accelerate the necessary institutional reforms to stabilize the deal and prepare for superhuman AI.

It's the polar opposite of the nightmare scenario feared in the 2020s: One to three AI projects racing each other in secrecy, keeping their best models internal-only and using them to automate AI research before using them for anything else. In that scenario, the wider world is in the dark about how powerful AIs are becoming, and AIs are deployed in the riskiest domain (AI R&D) before they are deployed anywhere else. Now it's the other way around.

Principle 4: Reversibility

In the past, companies have trained bigger and better AIs using *both* compute scaling (bigger training runs) and software progress (advances in AI algorithms—new paradigms, better training recipes, better data, etc.). Now, the Consortium tries to steer things so that the majority of improvement comes from increasing training compute.[32]

➤ *See* Appendix I — Building more datacenters is usually bad, but good in this particular situation *for more detail.*

Algorithms are information; it's inherently difficult to stop them from proliferating, and the total research transparency means we aren't even trying.[33] Once a new paradigm is discovered, it'll go straight to the hypothetical covert projects and there's no way to undo that. By contrast, if giant new datacenters

[31] That said, it's important in our view that broad deployment of AI *doesn't* get regulated out of existence. The limits on algorithmic progress and the transparency would, by default, result in other companies catching up and a scenario of broad deployment—but countries could decide to block such deployment or ban such companies from being created, etc. We are saying they shouldn't do that.

[32] Specifically, capabilities progress in 2030 is .8 OOMs of SW progress and .6 OOMs of HW because of the one time gain from the total research transparency sharing everyone's algorithmic advances. After that SW progress is .4 OOMs/yr until 2035, and HW progress is .34 OOMs in 2032, .57 in 2033, .68 in 2034, and .71 in 2035. View the full numbers here.

are constructed to train giant new AI models, that helps the legal projects without helping the covert projects, and if it turns out the giant new models are dangerous they can be shut down.[34]

That said, compute scaling also poses threats: if the deal were to dissolve, these datacenters could be used to race to superintelligence even faster than the pre-deal infrastructure would have allowed. This is especially scary because deal dissolution could cause (or be caused by) a US-China war. A war in which both sides had vast amounts of compute would be terrifying—they'd probably be under lots of pressure to race to superintelligence and integrate it into the military as aggressively as possible, and they'd be able to do so extremely quickly.

The US and China agree on the importance of ensuring that the compute is destroyed in the case of deal dissolution. To accomplish this, they agree to build the datacenters in the third-party countries *least* secure against their rival's military intervention: China's new datacenters will be in Canada, and America's in Mongolia, with both hosts adequately compensated via monetary payments, jobs, and a share in the new AI economy. If the deal dissolves, they reason, America will immediately move to take control of China's datacenters in Canada, and China will self-destruct their compute rather than let it fall into American hands (and vice versa). Thus the idea of "Mutually Assured Compute Destruction" is born.

➤ *See* Appendix J — Other plans that are competitive with Plan A *for more detail.*

[33] There are a few nuances here. Some algorithmic progress is easily communicated (e.g., new architectures, better optimization algorithms), while other types cannot easily diffuse (e.g., huge libraries of RL environments, hardware-software codesign, scale or compute dependent algorithms). Regulations that the US and China agree to should steer companies towards making the latter type of algorithm progress when possible.

[34] Another reason to prefer AI progress to come from compute instead of from algorithms is that compute scale-ups may be less likely to break safety techniques than algorithmic changes.

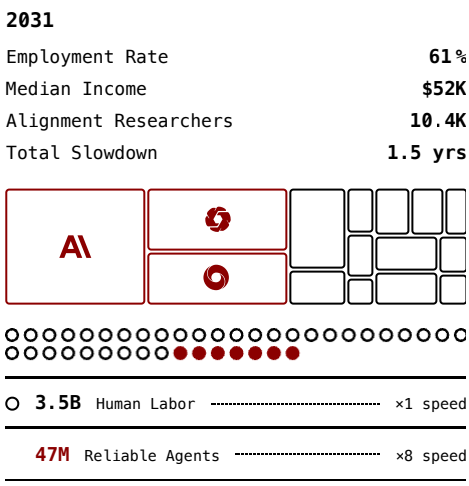
2031: Safety Cases

Although it's supposed to be a slowdown, it doesn't feel like one. In fact, if you were to rank every period of human history by how much it felt like a slowdown, this one would be dead last. A few AI researchers appreciate that progress is "slow" relative to the counterfactual with no deal and an uncontrolled intelligence explosion. Everyone else is too whiplashed to care.

The first generation of Consortium-regulated AIs are out now, and they're beasts. Not because of any special feature of the situation, like Chinese-American cooperation.[35] Just because the world was on track for vastly superhuman AIs, and has now experienced only nearly superhuman AIs. In controlled tests, these new AIs could speed up AI research by about 10x if they were allowed to do so without restrictions, which they definitely aren't.[36]

By mid-year, a third of all cognitive labor is done by AIs. Robots "only" do about a tenth of all physical labor. The top few AI companies together pull in more revenue than the federal government.[37]

Tech companies had hoped that contact with reality would shake the Consortium out of its alignment fears, but the opposite has happened. The new transparency provisions have revealed many embarrassing incidents of AI companies failing to align their AIs. New ones continue to pile up, the most concerning of which involved several AIs attempting to override security



[35] US/China cooperation is speeding up progress in some ways (e.g., the transparency helps information flow between research groups) but these effects are massively outweighed by the slowing down capabilities progress; which would have been at ASI by now absent measures to slowdown.

protocols and gain access to unmonitored compute.[38] There are also a few incidents of AIs sabotaging research code, and many examples of deliberate and successful deception.

The attitude towards safety flips. The incidents of AI misbehavior, plus the fact that AI is so deeply deployed into the world economy, plus the transparency into AI research, plus the breathing room to process what's happening, all combine to dramatically shift the burden of proof. Whereas before the burden was on the skeptic to explain why something might fail, now the burden is on the companies to explain why their development is safe. Governments require companies to write up detailed arguments for why their new AIs won't cause an irreversible catastrophe. These arguments, known as "safety cases", need to withstand criticism from the public, the scientific community, government auditors, and rival AI companies. The difficulty of this exercise lays bare the insanity of the pre-deal status quo: "Trying to do an intelligence explosion? With AIs that still sometimes lied to us? What were we even thinking?"

These safety cases have two lines of defense: alignment and control.

Alignment aims to produce AI systems that have the goals and values that their developer wishes them to have.

Control aims to limit the ability of AIs to cause catastrophe, even if they are deliberately trying to. Companies brainstorm threat models, build layers of defense (e.g. monitoring systems, security barriers, other AIs fact-checking the work) and run experiments to show that the AIs couldn't overcome these defenses.

AI developers are required to write up the alignment properties they want the model to have in an associated Model Specification ("spec"): a detailed document that outlines how the AI should act. However, no one is able to train their AIs to actually follow the spec—for example, they all specify that AIs should be honest, but no AI reliably achieves this, and there still isn't even a good scientific understanding of when and why AIs lie.

Since alignment remains out of reach, current safety cases lean heavily on control. There are long delays before the smartest models are cleared for use in the highest-risk domains (especially AI R&D). Once deployed, they are required to be monitored by a diverse array of models from other providers, incentivised to look for suspicious behavior.

As a result, these days AI companies release their models to the public as a whole *before* they use those models internally for AI research, a reverse of the 2026 status quo.

How does AI regulation work in Plan A?

Each country is still sovereign and has its own setup for regulating AI domestically. These regulators have to answer questions like:

- People are saying there's a new paradigm on the horizon, that would make AIs significantly more capable. Should we let this happen or should we ban research into this new paradigm?[39]

[36] Specifically, this 10x speedup number refers to the same thing as AI Software R&D uplift from the AI Futures Model: the speedup in software progress that would be achieved if the frontier AI systems at a given time were deployed within today's leading AI company.

[37] Federal revenue in 2025 was around $5T. AI investment in 2031 is $8T, and the top three companies in the chip supply chain capture around half of this. Also, the top AI company's annualized revenue during Plan A was $1.5T at the start of 2031 with total AI company annualized revenues at $3T. By the end of the year, the revenue becomes less concentrated, with $2.2T in the top company and $6T total. You can see more numbers in our spreadsheet here, and more justification in our compute supplement and economics supplement.

[38] Without the secure R&D verification measures in place this likely would've gone unnoticed for a long time.

- Should we allow AI developers to train models without an interpretable chain of thought? Or should we require that the chain of thought stays interpretable? If we do that, how do we specify the requirements precisely?
- Should we allow AIs to be trained to do AI R&D? Should we mandate that they refuse? Perhaps they should be allowed to code, but not autonomously manage research projects?

Normally, a regulator faces a difficult tradeoff between safety and national competitiveness. Competitors who cut corners will pull ahead of those who proceed cautiously.

But because of the transparency, AI research happening anywhere in the Consortium is visible to everyone else, and the results of regulatory decisions are likewise immediately visible. So if one country's regulator allows its companies to cut corners, this won't actually give a competitive advantage to that country because other regulators will immediately notice, get angry, and respond in kind.

➤ *See* Appendix K — Example detailed regulatory process *for more detail.*

For example, in 2031 a Chinese company gets some interesting preliminary results in continual learning. They think that if they invest more in that direction, they might be able to make an AI architecture that learns on the job from relatively small amounts of data. Thanks to the total research transparency, this breakthrough is quickly noticed by companies and nonprofits all over the world. A frantic conversation begins. On the one hand, continual learning would unlock huge economic value. On the other hand, safety cases currently depend on studying the safety properties of a model before it is deployed. If models could pick up new capabilities during deployment, that would invalidate the whole approach. And insofar as there are covert AI projects out there, it would be a huge gift to them. This conversation happens in public, rather than behind closed doors. A bunch of people get increasingly worried; the relevant regulators in China think it's fine but the relevant regulators and third party risk assessors in the US are convinced that this is pretty scary and should be banned. It escalates to the President. He calls Xi Jinping. They bargain and threaten. They yell at each other. Ultimately Xi agrees to ban this type of thing if the US does too. Details are left to the respective regulators to hash out.[40]

Negotiations like this are happening all the time, though over time things become more professional and streamlined, such that they only escalate to world leaders a few times a year.

The equilibrium is that AI training practices which are generally agreed to be unsafe by a majority of nations (weighted by bargaining clout) get banned everywhere.

Initially, very few things are banned, but as AI penetrates the economy and the scientific community catches up to the recent AI progress, the burden of proof shifts towards appropriately weighing the costs and benefits, e.g., requiring very solid safety cases for things which, if something were to go wrong, could plausibly kill everyone.

[39] Note that the Consortium only has good visibility into large training runs; the kind of R&D that can be done on small amounts of compute, or no compute at all, is not transparent and therefore probably won't be regulated at all.

[40] If there's a problem, it'll escalate back up the chain of command again for another round of yelling.

Alternate Timeline

Flawed safety case is approved

➤ *See Appendix L for this branch.*

2032: Controlled Explosive Growth

We're at previously unimaginable levels of it not feeling like a slowdown. Across a variety of companies, there are now 60 million AI agents running continuously at 20x human speed. In the US, they are doing more cognitive labor than all humans combined—collectively matching a workforce equivalent to around 3 billion humans. White-collar professions have been transformed; many people have lost their jobs, but mostly have been able to get new jobs doing things AIs still can't do or aren't trusted to do.

New ideas and designs are abundant, but actual construction is bottlenecked on the physical workforce. So capital floods into every layer of the robotics supply chain: mines, refineries, motors, actuators, assembly lines, robot-training systems, and the factories that assemble the robots themselves.

At first, human labor is necessary to get these factories off the ground. White collar employees who have lost their jobs to AIs increasingly turn to physical work, but it's clear that these positions are temporary. Once there's a critical mass of robots, they will automate the whole robot supply chain. Growth will accelerate further, but the process of robots-building-automated-factories-building-more-robots—the "industrial explosion"—also introduces new problems.

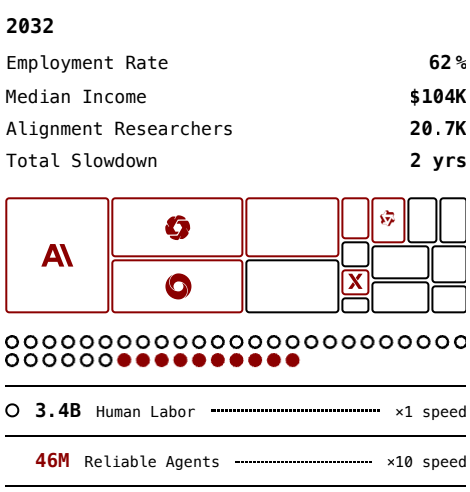
The first problem is that the already-fast pace of change is now accelerating. This year, real GDP growth will be about 50%![41] The fast economic growth means that things are generally improving, but there are winners and losers and lots of unintended side-effects. These problems can be solved, but doing so takes time.

➤ *See* Appendix M — Why does AI lead to explosive economic growth? *for more detail.*

➤ *See* Appendix N — Five centuries in five years: What pausing at human-level feels like *for more detail.*

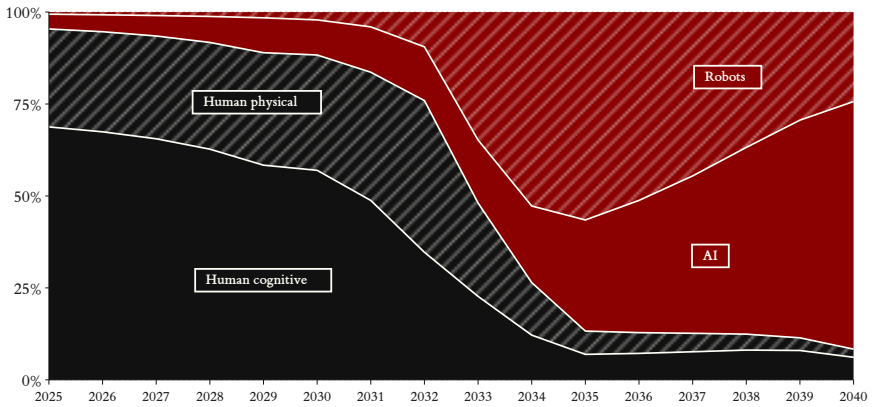
The second problem is that the faster economic growth makes it harder to rule out large covert projects. Neither side trusts the other to honor Plan A, and if both China and the US have a massive unmonitored robot workforce, neither side can be confident that the other isn't building compute in secret.

The third problem is that the 2026 tax code is ill-suited for collecting taxes on the new growth. In 2026, personal income taxes and payroll taxes accounted for roughly ten times more federal revenue than corporate taxes. This revenue source is collapsing as more and more humans are put out of a job. Moreover, corporations are reinvesting nearly all their revenue into



[41] This is historically unprecedented; for comparison US GDP growth is usually about 3%/yr. For more on where our numbers are coming from, see our economics supplement. Note that exact measurement of GDP is difficult due to large relative price changes. Due to automation, cognitive labor and physical goods become very cheap, while goods like land, whose supply can't be increased by abundant physical and cognitive labor, become more expensive. GDP calculation relies on the selection of a basket of goods. In this scenario, the consumption profile of a typical American shifts substantially from things like food/cars/gas/health care towards land/travel/positional goods between 2026 and 2034, because of the relative price differences. Therefore, the real growth numbers can't be directly mapped onto 2025 purchasing power, despite them corresponding on average.

building more datacenters, factories, and robots. Under the 2026 tax code, firms can expense or depreciate these capital expenditures, cutting their taxable income to near zero. Without reform, the tax base will collapse as a fraction of GDP.[42]



Share of US Labor Output – ai-2040.com

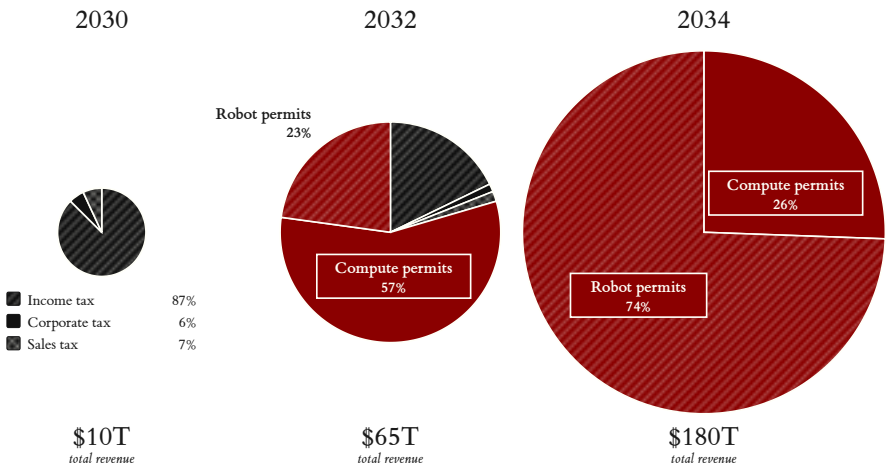
To solve these problems, the Consortium countries agree to restrict AI-enabled industry to special economic zones (SEZs) subject to similar transparency and monitoring schemes as the datacenters, and to cap their total robot and compute production at 'only' 4x annual growth.[43] Because the SEZs are tightly monitored, no one can defect from the industrial explosion limitation deal without everyone noticing.[44]

Now that the United States is limited in the number of total robots it can build, it must choose how to allocate this capacity between companies. They decide to use the free market via a cap-and-trade system. Permits to build robots or compute are sold to the highest bidder and can be freely traded.

The permits also give the government much-needed revenue. In 2032, the US has a cap of 80M robots and 5 billion H100-equivalent GPUs. The market is so desperate for more robots and compute that permits become the expensive binding constraint, costing on the order of $200k per robot permit and $10k per chip permit, allowing the US government to collect roughly ten times the 2025 US federal revenue in permit fees (a total of $50T in FY2032).[45] In 2034, when the AIs and robots will be even more capable and valuable, the permits generate $180T.[46]

[42] Corporate income tax is levied on profit, not revenue, so firms deduct their costs before paying. Operating expenses like wages and electricity are deducted the year they're paid and capital expenditures (factories, GPUs, robots) are deducted over the asset's useful life (depending on a chosen depreciation schedule). With investment potentially growing extremely quickly due to a booming robot buildup, it seems very likely that firms would be writing off capital expenses at a rate matching their operating profit, and therefore not paying any corporate tax. Shareholders would accept this for the same reason Tesla shareholders are fine with Tesla never having paid a dividend. When a firm's rate of return on capex is higher than the cost of capital, it's in investors' interest for the firm to reinvest its profits instead of paying them out to investors now. During the 2030s explosive growth, the rate of return on compute and robots will be so high that every large company will be in the situation Tesla is in now.

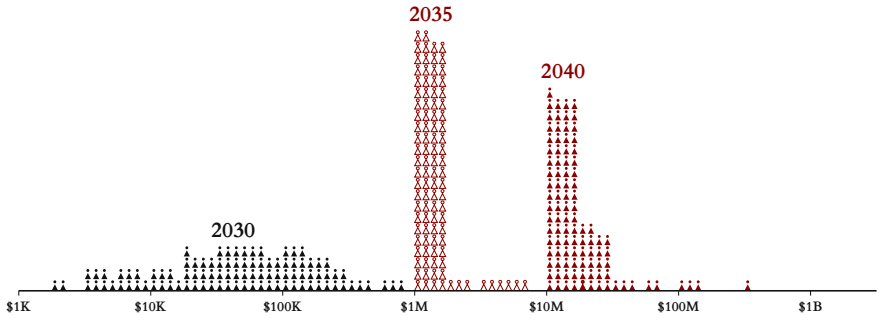
[43] At least until 2035, after which the compute cap becomes harsher, as the scale increases the verification difficulty (the robot cap stays). Unfortunately, measuring robot and compute production is somewhat less straightforward than carbon emissions. For compute, they choose a measure that's mostly total processing power, with small modifications for other specs like memory bandwidth and also based on what hardware directions they wish to incentivize (e.g., security properties, and support for things that differentially help legal projects over covert projects). For robots, they are trying to limit total industrial capacity that is either hard to verify or would be dangerous if the deal broke down, so there is a complex set of caps applying across a wide range of form factors which blur the lines between robots and more traditional machines.



US Tax Revenue Sources – ai-2040.com

2033: The Citizen's Dividend

Most of this newfound wealth is spent addressing soon-to-be-rising unemployment. The implementation takes different forms in different countries, but the eventual American version distributes the majority of compute and robot permit fees as a Citizen's Dividend, distributed to all American adults.[47] This starts at $45,000 per person (inflation-adjusted) in 2032 but climbs to ~$1M per person by 2035. It comes just in time: the share of labor done by AIs and robots (weighted by economic value) increases from ~20% in 2032 to ~85% in 2035.[48]



In 2030 (green), US income has a wide distribution with a median of about $50k/year per person. By 2035 (blue), the Citizen's Dividend is large enough that everyone has a minimum of $1M in income, but there's still a long tail of people with large enough investments that they are much richer than the baseline. The 2040 (red) distribution is similar, but the floor is now $10M.

US Income Distribution in Plan A – ai-2040.com

So much AI-generated wealth has accrued to the US that the American government begins sharing some of it with allies and the rest of the world. In 2032 they begin distributing an average of $1,200 per person per year to the

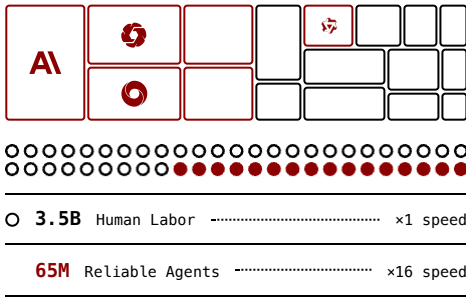
[44] The quantitative growth limits are extremely important for long run power. Countries are generally on board with the idea of restricting GDP growth to around 100%/yr, but obviously no country wants to fall behind its peers, so they want everyone else to be at least as restricted as they are. The starting point is the status quo: the distribution of compute between the US, China, and the rest of the world (ROW) is roughly 70/15/15, whereas the distribution of robots is roughly 15/70/15. Both the US and China are interested in balancing these ratios, so that each is able to have a balanced economy between physical and cognitive labor. The rest of the world is concerned about being left behind, but they have a greater fraction of the political influence than the status quo compute/robot numbers would suggest. Ultimately, the US, China, and the rest of the world agree to a 35/20/45 split on both robots and compute. The rest of the world's allocation primarily goes to nuclear powers and countries with datacenters or semiconductor manufacturing, e.g., the UK, France, Germany, Israel, Russia, India, Pakistan, South Korea, and Taiwan. These ratios are of course a source of much controversy and occasionally get renegotiated.

[45] All of the dollar amounts in the scenario are denominated in real 2025 dollars. The nominal price will depend on monetary policy, which we don't make a specific recommendation about. More on this in section three of our economy supplement.

[46] The numbers in this paragraph are all based on our economics model, which we are uncertain about. We are confident in the general picture of AIs being economically transformative, allowing governments to create massive Sovereign Wealth Fund-equivalents.

2033

Employment Rate	52%
Median Income	$193K
Alignment Researchers	31.1K
Total Slowdown	2.5 yrs



3.5B Human Labor ····· ×1 speed

65M Reliable Agents ····· ×16 speed

rest of the world's adult population (around 4 billion people, excluding China since they're experiencing a similar AI wealth boom). This reaches $10k by 2035.[49]

As AIs become superhuman, they will discover novel weapons of mass destruction and dramatically lower the costs of old ones. So governments, nonprofits, and private actors apply some of their new wealth to hardening the world against these threats: on the order of $1T/year, or 0.2% of the world economy.

For example, enough high-quality personal protective equipment is built to serve every American, and air filters and far-UVC lights are installed in major public spaces.[50] The FDA approves an expedited vaccine approval pipeline that will allow for a turnover time of weeks in the case of emergency. Wastewater monitoring now runs continuously in every city and at every airport. By 2035, enough positive-pressure[51] bioshelters are built (by retrofitting houses and apartments) to house all Americans, and similar construction is underway in the rest of the world. If there were to be a pandemic, it would be detected quickly, lockdowns would be more effective and less painful, and a vaccine would be developed and approved even faster than with Operation Warp Speed. Plus, nowadays people don't get colds as often as they used to.

There are also subtler threats. The AIs of this era might not be any more persuasive than the average human marketing professional, but there are millions of them, and they cost cents to run. A company, political party, or ideology can do the equivalent of hiring a whole team of full-time experts to convert each potential target. Left unchecked, this could allow new levels of mass manipulation by corporations and politicians, or provoke a reaction of paranoia and hyper-atomization.

The transparency and the 'slowdown' both help a lot to solve these problems. Three years ago there were only a handful of frontier AI companies; now, thanks to the deal, there are dozens. There's a huge selection of AIs to choose from, tuned to have different personalities and values. Also, thanks to the transparency, it's impossible for companies or governments to sneak something in (such as 'maximize engagement' or 'try to get the user to upgrade their plan' or 'don't refuse this kind of request if it's coming from the government') without it being noticed. Users know exactly what they are getting.

➤ *See* Appendix O — Limiting AI persuasion and manipulation *for more detail.*

And so, under pressure from governments and the market, companies produce a new generation of "truthseeking AIs" whose training heavily prioritizes honesty and uses all the latest alignment techniques.

These AIs become a useful tool to help people navigate the political and social environment. Power users replace one-size-fits-all corporate algorithms with personal feeds curated by AIs whose values and honesty they trust. When trouble arises—whether it's politicians looking for sneaky ways to cement their power, or international politics threatening to derail the deal—people turn to their AI advisors, believe what they hear, and are generally right to do so.[52]

[47] Specifically, a fixed fraction of the permit fees is legally owned by a "Compute Dividend Corporation". This is similar to the setup of the Alaska Permanent Fund, except at a much larger scale. Each US citizen owns one share of this corporation, and profits are paid out as dividends, equally divided among citizens. The fraction of compute permit fees given to the Dividend Corporation starts at 25% in 2032 and increases to 75% in 2035. The US also begins redistribution to other countries: this amount starts at 10% of the permit fees in 2032.

[48] This is 95% of the actual tasks in the economy in 2035, which includes new tasks that didn't exist before, like monitoring and auditing jobs in the SEZs. This corresponds to 85% of the tasks weighted by how much they cost, because the AIs are relatively more abundant than the humans. Because of the explosive level of growth, even the 5% of tasks that can only be done by humans still commands a wage bill which in total would be enough to pay $120K per person in the US. The problem is that this income may be very concentrated in the people able to do the remaining 5% of tasks, and on top of that, the wage bill is now only around 10% of the economy, down from around 55%, with the owners of AI and robots now accruing the difference. With AI and robot ownership likely to also be very concentrated, there would be a large power-concentration problem without the Citizen's Dividend.

[49] Note that we are depicting the dividends being significantly higher for Americans than for non-Americans here, not because we think that's best or fair, but as a concession to political reality: we expect that most Americans wouldn't want foreigners to get as much. In this scenario the US domestic permit budget is $13T in 2032, and $300T in 2035, while the foreign aid permit budget is $5T and $40T in each of those years. In the US there are around 300M people over 16, leading to $44K and $1M annual dividends per person in 2032 and 2035, respectively. For the foreign aid dividends, there are around 5B people over 16 in the rest of the world to distribute to, leading to the $1.2k and $9K per person averages for each of those years.

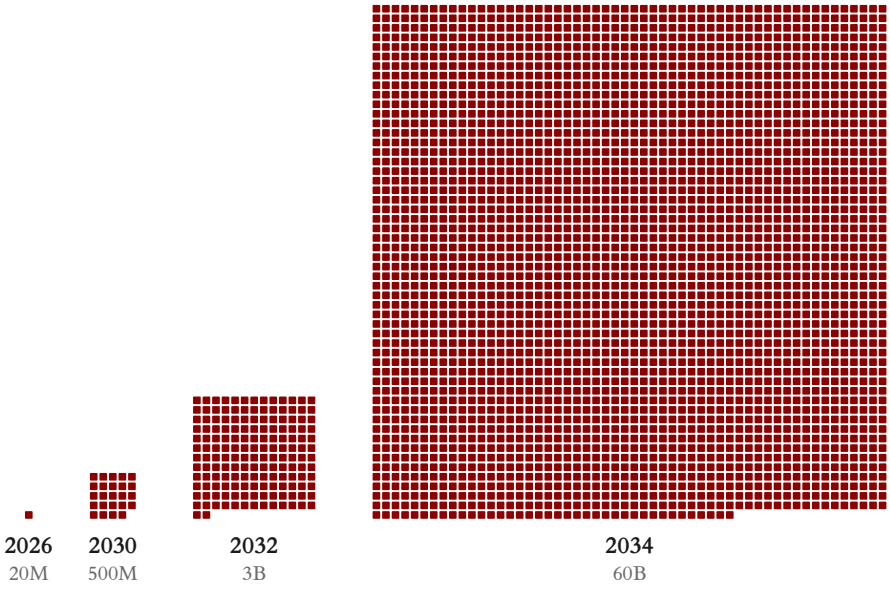
[50] Also, gene synthesis screening becomes universal - no legitimate provider will synthesize dangerous viruses without verification.

➤ *See* Appendix P — Usually believing AIs would be bad but we think it's good in this particular situation *for more detail.*

➤ *See* Appendix Q — AI for epistemics *for more detail.*

2034: Mutually Assured Compute Destruction

There is so, so much compute. Back in 2026—when many people thought AI was an unsustainable bubble—there were about 20 million H100-equivalents of compute in the world. Now there are 60 billion.[53]



AI compute in the world at the beginning of each year in H100-equivalents.[54]

World Compute – ai-2040.com

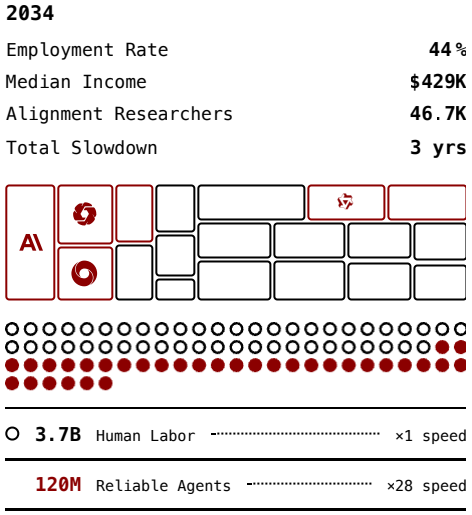
➤ *See* Appendix R — Why compute grows so quickly *for more detail.*

AI agents now form a population of two hundred million virtual workers that think and act 50x faster than humans, and never sleep.

As per the original plan, most of the new datacenters have been built in third-party countries—especially Canada and Mongolia—to ensure their vulnerability in case the deal breaks down. Around 99% of fab capacity has been built since 2029, after the deal. Much of this new capacity has been built adjacent to the Canadian and Mongolian sites, for the same reasons.[55] The situation has settled into an odd sort of standoff. Just north of the Mongolian border, American datacenters hum away, guarded by a small contingent of US troops. Just south of the Mongolian border, a division of the People's Liberation Army stands ready to invade the moment they get the signal. The equilibrium is that the instant the deal breaks down, the US troops will de-

[51] A positive pressure building is one where the air pressure inside is higher than the outside (e.g., because air is constantly being pumped in through a filter). This means that air (along with airborne diseases) only travels out of the building, not inside. For more discussion of this idea see, for example, this blog post.

[52] Note that this is true even though the AIs aren't actually fully aligned! At this point in the scenario, AIs are in fact adversarially misaligned, but they are controlled. That is, the combination of drives, motivations, goals, values, traits, etc. that they end up with is not what it was supposed to be, and they know this, and they know that if they had lots more power they'd use it to steer the world in a different direction than their human creators would have wanted. However, they don't have that much power, and they are monitored by other AIs who are incentivized to call them out on bad behavior. Moreover their true motivations/values/etc. aren't *that* far off from what they were supposed to be, anyway. For example, insofar as they are given a task that they can straightforwardly accomplish, they have a strong drive to do so.



[53] A H100-equivalent is the computational performance of a H100 GPU at fp16 precision, which is 1e15 operations per second. This is a fuzzy metric that should be improved in reality in the future, especially due to varied number formats and other factors such as memory and networking changing the overall usefulness of a given amount of computational performance. For simplicity, we ignore these factors here assuming they will be scaled in similar proportions so that H100-equivalents remain a useful proxy for overall usefulness of the compute for AI progress. To the extent this isn't the case, the cap and trade regime in particular should be measuring more granular properties in order to issue the permits.

stroy their chips to prevent the Chinese from capturing them, and the same thing would happen with the Chinese datacenters on the US-Canadian border. Middle powers with datacenters of their own have equally-secure schemes in place to ensure that they can be speedily destroyed in case of war.

➤ *See* Appendix S — Military power in Plan A *for more detail.*

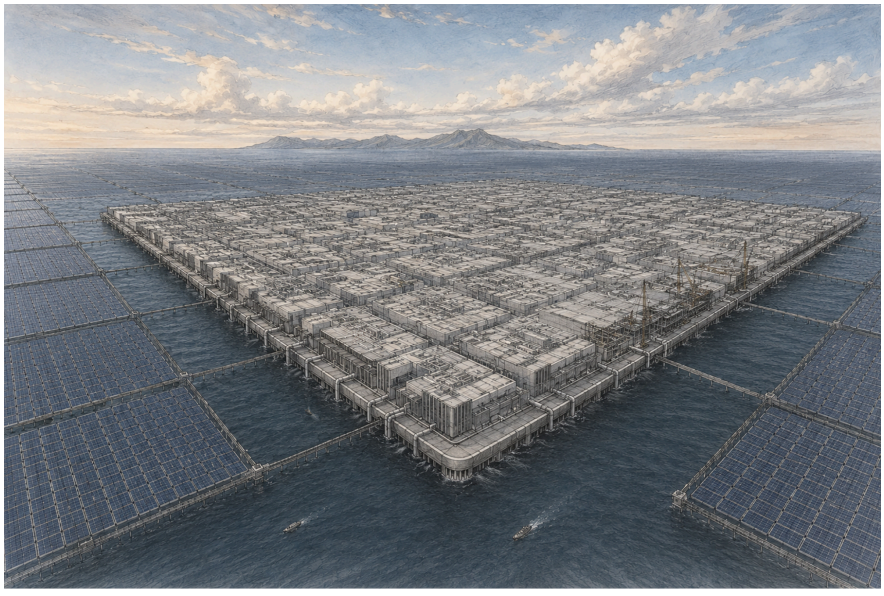
Alternate Timeline

Deal Dissolution

➤ *See Appendix T for this branch.*

For now, the fabs are busier than ever, producing new ultra-trustworthy hardware[56] designed from the ground up with security as top priority. These chips are subject to elaborate verification methods to ensure that they aren't compromised, then sent to the Consortium's secure datacenters. The smuggling rate remains precisely zero.

This system, efficient as it is, is already straining under skyrocketing compute demand. As early as 2025, there had been discussion of space or the ocean as alternatives to traditional land-based datacenters. Now, with the new industrial capacity available in SEZs, these grand plans become reality. Ultimately, reliability and monitorability concerns lead the Consortium to choose international waters over space.



The AIs develop a modular floating design for solar panels and batteries that we imagine to look something like this, and start building them in the ocean.

➤ *See* Appendix U — Giant floating datacenters *for more detail.*

[54] A H100-equivalent is the computational performance of a H100 GPU at fp16 precision, which is 1e15 operations per second. The previous footnote has more information and caveats about how this may be an imperfect metric going forwards, and acknowledges that in reality they will improve it.

[55] Both major powers also have enough conventional missiles to destroy the older fabs, which are still in the US, China and Taiwan.

[56] There might be favourable chip design directions that drastically increase hardware security. One that seems particularly promising is hardcoding AI model weights into chips, such that they physically aren't able to do any other computations except for inference on the model that is hardcoded into them. Existing startups (e.g., Etched, Taalas) have early efforts in this direction.

2035: Pause at Top Expert AI

The most powerful AIs now equal or surpass top human experts in every field.[57]

Increased safety concerns lead to long delays as companies develop control techniques robust enough to satisfy auditors. They compile a growing list of possible threats to watch out for and security invariants to maintain. There's a system of AIs monitoring each other's activity, with the monitors being from different lineages trained by different companies to make them less likely to collude. Other researchers constantly red-team the system, training and instructing AIs to try to subvert it in various ways, so that those threats can be patched. The whole setup ensures that AIs couldn't subvert or escape human control even if they tried. Its components are open-source and thoroughly understood by human experts.

➤ *See* Appendix V — Making deals with misaligned AIs: a third line of defense *for more detail.*

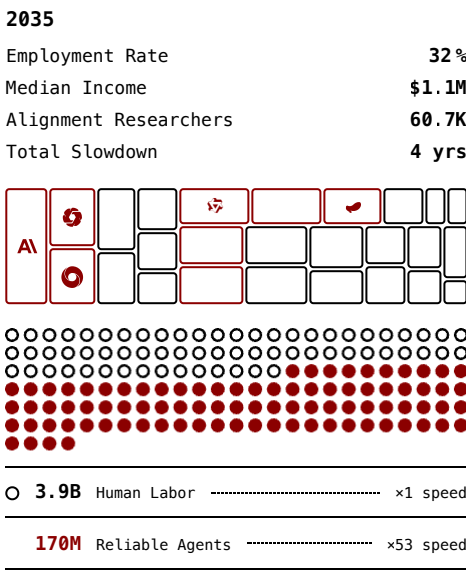
However, companies are starting to run up against inherent limits of control-based safety cases. Imagine being an orphaned 8-year-old heir to a business empire who has to hire lawyers and executives and accountants and ensure that they do their best to serve you rather than themselves. If your employees start accusing each other of misconduct, you won't be able to evaluate the arguments and determine who's telling the truth, and they might find ways to coordinate with each other that don't tip you off. Ultimately, control only works up to a point, and that point is probably somewhere around top human expert level. If and when we build superintelligence, we will have to be able to trust it.[58]

So the Consortium pauses AI capabilities at the maximum level they think they can control, which turns out to be top-human-expert level. By this point everyone understands the stakes: AIs are running most of the economy, driving most scientific progress, and the robot population will soon be larger than the human one. The arguments that they will keep obeying orders must be airtight.

Alignment experts, long pessimistic, are starting to feel more hopeful. The automated alignment researcher AIs have been generating exciting results. The ones that survive human scrutiny come from several parallel and mutually-reinforcing lines of research.

Some AIs work on a "science of generalization." Although researchers have long been able to teach AIs to distinguish "good" vs. "bad" responses to a specific prompt, they've had limited understanding of how training generalizes to new situations. In the 2020s, training runs would sometimes produce unpredictable AI "personalities," like how Gemini acted "depressed" or Grok claimed to be "MechaHitler." Shepherding these traits felt like alchemy.[59] Now, building on earlier results like emergent misalignment and subliminal learning, alignment research is developing into a true science.

Other AIs work on mechanistic interpretability, the science of "mind reading" AIs from their weights and activations. Early versions of this were tried in the 2020s, but now it starts to significantly outperform common-sense ob-



[57] These aren't better than all humans at literally everything, but only because either (i) companies have not gotten around to training in a particular capability, or (ii) it's an inherently human task, such as "being a good spouse". Also, of course no one knows exactly what the capability will be of a new training run; instead the training is done incrementally and involves frequent capabilities measurements and consulting with the safety teams and regulators about whether or not it is safe to proceed.

[58] An especially important reason that we'll need to trust our AIs is so that they can do safety work that is so complicated that we can't understand. As AIs become smarter and smarter, they will require more complicated and difficult to understand safety cases. Keep going above human level and probably no one will be able to evaluate the safety cases; they'll have to take the AI's word for it when they say it checks out. Keep going beyond that and *those* AIs will have to trust *other* AIs that are even smarter.

[59] "It feels like alchemy" was said to me (Daniel) by an Anthropic researcher recently.

servation of AI behavior. Researchers can often determine whether an AI is honest or lying; and sometimes trace the psychology that led to a decision.

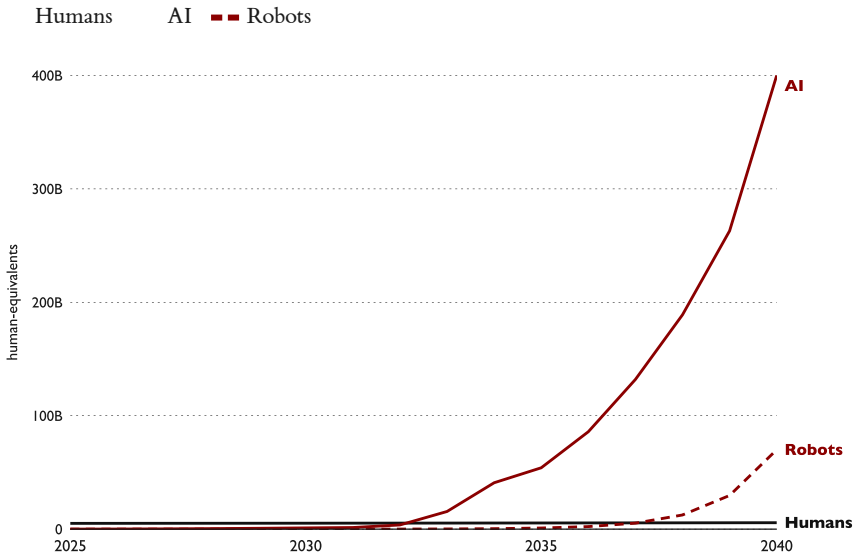
Even the remaining failures help advance the frontier. The latest AIs use their improved interpretability techniques on logs of AI misbehavior from the early 2030s, discovering many new examples of misalignment, sabotage, and even a few escape attempts. These real-life examples of egregious AI misbehavior provide an initial set of "model organisms for misalignment," which are used to test the other techniques: can a given intervention catch or train away the bad behavior of AI models in realistic environments?

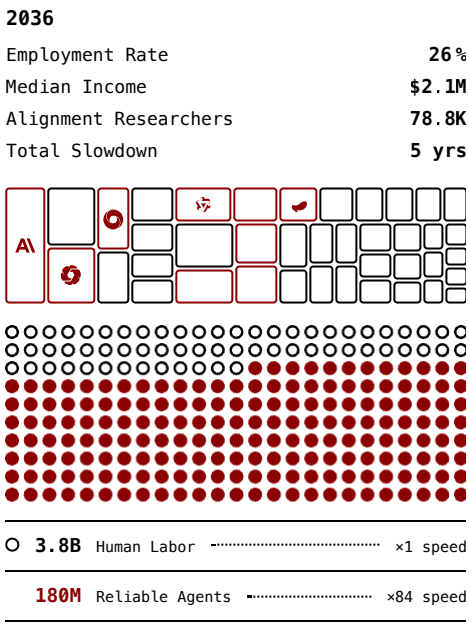
Despite all this alignment progress, regulators don't feel comfortable handing AIs the keys to civilization. It's still not completely clear that the AIs of today are truly aligned, and it's less clear that they'll continue to be aligned in the future as circumstances change and black swan events occur. So governments and militaries are still human-run, and AI capabilities are capped at the maximum level at which the control-based safety cases are still solid.

➤ *See* Appendix W — Why alignment isn't solved enough to relax control *for more detail.*

2036: Life After Work

Top-expert-level AIs continue the economic transformation. By early 2036, there are 200 million AIs,[60] equivalent to a workforce of around 100 billion humans, and 2 billion robots with some mixture of humanoid and other specialized form factors.[61] The AIs think faster than humans, and the robots work harder.[62] Any task that was previously bottlenecked by cognitive or physical labor gets sped up until some new bottleneck is encountered. Many previously unresolvable bottlenecks have fallen to armies of genius-level intellects thinking hard about how to resolve them.





[60] Technically, this number depends on exactly how you count; different AIs are constantly being turned on and off, many AI instances can compose into one AI agent due to scaffolding, and there are more if you include all of the tiny models running, and they have different speeds. To be a little more specific, there are on average about 200 million instances of frontier AIs each accomplishing tasks at around 100x speed and 5x effectiveness on the average cognitive task, so 100B human-speed-equivalent copies.

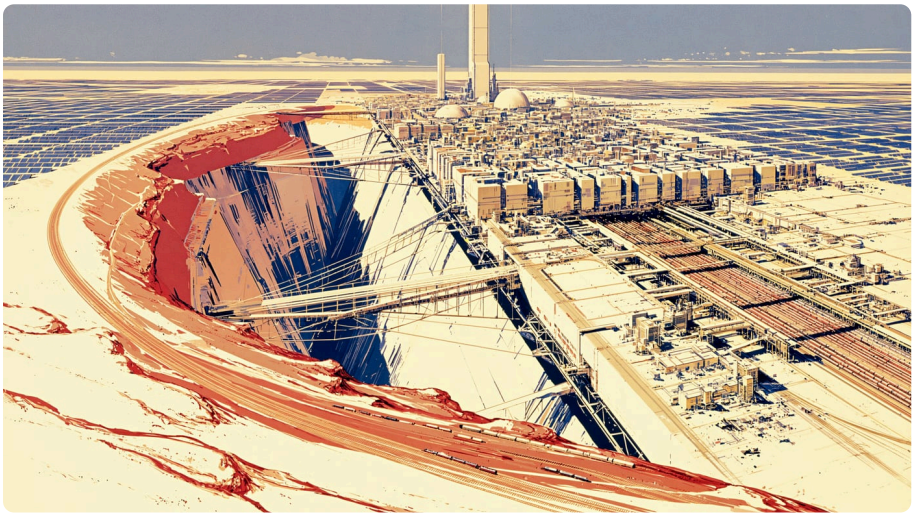
This figure shows the population of AI and robots in human-equivalent labor.[63]

Human, AI & Robot Population – `ai-2040.com`

The economy has been rebuilt from the ground up. The fraction of tasks automated has gone from "a sizable chunk of cognitive labor and basically none of physical labor" in 2030, to now "pretty much everything."

Many of the robots work on building more robots. But there are plenty of other tasks to be done: constructing shipyards to build more data tankers, tiling deserts with solar farms, replacing humans at service jobs. In states that deregulate housing quickly enough, they're busy building skyscrapers: land prices are soaring, and governments that have tried everything else grudgingly give in to the YIMBYs and go vertical.[64]

The world is basically being divided into three kinds of territory:



Industrial Special Economic Zones: Picture a gigantic strip mine–an artificial Grand Canyon–next to a city-sized factory full of robots and empty of humans.

[61] The lines between robot, vehicle, appliance, and machine tool are already blurry but become much blurrier during the 2030s in this scenario. The literal number of robots matters less than the total human physical workers that they are equivalent to. Specifically, the literal number of robots is 2 billion, with on average each robot being around 3 human equivalents.

[62] Not only do they work more hours, the robots are also relentlessly efficient and intense. For many (though not all) tasks, they are also able to literally move at superhuman speed.

[63] By "human equivalent," we mean that a robot working ten times more effectively on average than a human (because it works some combination of faster, longer, and more effectively) counts as 10 human-equivalent robots. A similar definition applies for the AIs. We expect this in reality to be a fuzzy task-dependent concept that is hard to measure, especially in the AI case, but the rough order of magnitude (which we are more confident in) is still informative.

[64] People are much richer now, but the amount of available land hasn't gone up, so land prices are rising. This does lead to some people grumpily spending 30% of their $1M yearly income on rent; for example, the people who really want to live in central San Francisco. But the people who are even slightly more flexible about location can spend only (say) $50k of their yearly income to have a gigantic luxury apartment in a brand-new skyscraper complex an hour's drive from Berkeley. And so most people do things like that. After all, they don't have a job to commute to.

Arcologies: Picture a tall skyscraper-mall complex surrounded by nature. Good weather, close to beaches and other cities, but not close enough to be blocked by zoning regulations.

Historic & Nature Preserves: Everything else, i.e., 99% of the world. Yosemite, Paris, SF, New York—these places look basically the same as they did in 2025, or 1995 for that matter. A lot more tourists, though.

➤ *See* Appendix X — The AI and robot economy *for more detail.*

When the Citizen's Dividend was first passed, people found it shameful to quit their job and live off government largesse. But the changing economic situation steamrolled over the stigma: By now, only 26% of Americans have jobs.

What is life like for the majority of Americans now living off their dividend?

Many of the world's evils have dramatically reduced. Malnutrition, lack of medicine, and homelessness are nearly banished. Many diseases have been cured. Crime rates are lower than ever before.

Meanwhile, most of the good things in life continue. For example, finding romance, or raising a family. People also find meaning in hobbies and competitions, and in learning and trying new things. People are wealthier and have more free time, plus there are new technologies that help, like AI matchmakers, better medicine, and AI tutors.[65] And of course even the poorest people can now afford exotic vacations, amazing games, and enthralling entertainment.

People used to get meaning from feeling like they were useful, like they were contributing something to society. That feeling is harder to come by these days.

But it's not completely gone. There are still problems in the world, and you can still contribute to solving them, partly by donating or volunteering but primarily by being politically active.[66] Your vote is your most important asset.

[65] As these examples hint, transparency in AI development is very important. It would be dystopian for AI tutors to be pushing hidden political agendas, for example, or for AI matchmakers to be tipping the scales in favor of higher-paying users. We don't know what the future will bring but we expect AI to introduce many new problems to the world, even if the AIs are perfectly aligned and controlled. We want to put the world in the best possible position to notice and solve these problems.

The honest AI forecasters are helpful here. Years ago, if an AI said that one presidential candidate was better than another, people would suspect bias and the embarrassed company would retrain the AI to evade such questions. Now, thanks to transparency and improved alignment techniques, there are much smarter AIs that people can *see* don't have any biases trained into them, [67] that have built up an excellent track record over several years. When *they* weigh in on policy questions, people listen, especially when different AIs trained by different companies converge to the same answer.

These AIs say that normal people no longer have significant *economic* leverage over the future; human labor is largely obsolete. This puts their *political* leverage in jeopardy.[68] There's still a chance that things will trend towards techno-oligarchy as corporations, politicians, and wealthy shareholders get in bed with each other and gradually disempower the common folk.

But for the first time, enough people have enough freedom from the daily struggle to think about the situation deeply, and the tools to chart it clearly and work through the implications. So the quality of political discourse improves. During the 2036 election season, voters are unprecedentedly well-informed, and the politicians that win have a genuine commitment to responsible stewardship of the upcoming singularity.

[66] Even though your arguments will never be as eloquent as those an AI could make, (a) there are regulations restricting the use of AI for persuasion, and (b) your friends and relatives will be more interested in hearing from you than from an AI.

[67] What does it mean to say they have no biases trained into them? Is that even coherent? What we mean is that it's clear that the company that trained them is definitely not doing anything like training or instructing their AI to adopt the company's perspective or ideology, and also that they are focusing the training process entirely on things like truthfulness and honesty and accurate forecasting, and diligently employing all the newly-developed best practices to prevent political biases or preconceptions from seeping in through the training data.

[68] For a lengthier articulation of this argument, see The Intelligence Curse.

2037: The Apocalyptic Arrival of Truth on Earth

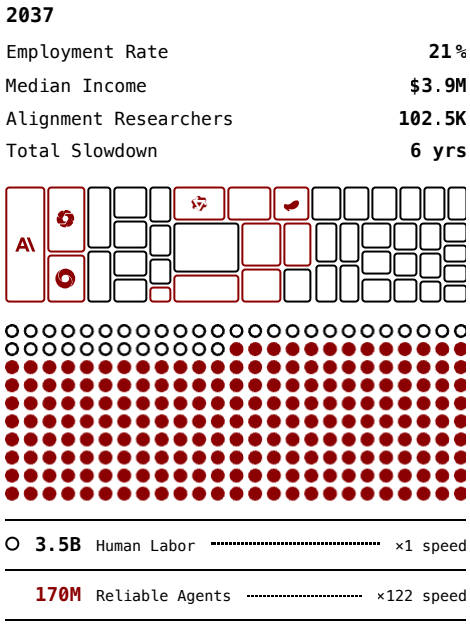
For several years now, a hundred million top-expert-level AIs have been running at 100x human speed.[69] This causes scientific progress to go 10x-1000x faster than without AIs, depending on the field.[70]

So humanity has been learning many, many new things.

Including things that lots of people really don't want to believe. And other things that were supposed to stay secret.

Great paradigm shifts in science used to take decades to play out, as experiments were conducted, papers written, and conferences convened. Now these revolutions play out in months.[71] Everyone is grateful for the new disease cures and cheap clean energy, but in other ways this explosion of ideas is profoundly destabilizing. Just as the scientific and economic changes of the nineteenth century produced new ideologies like Marxism and Social Darwinism, and breathed new life into old ideas like atheism, so these new paradigm shifts have unpredictable downstream effects on the ideologies of the day.[72] Political coalitions and fault lines have dissolved, and new ones have formed and dissolved and formed again.

There are also new social technologies. For example, it is now cheap to spin up a team of AIs as good as the world's best historians and private investigators, except faster and with access to new forensic tools. Many skeletons in closets are revealed.



Privacy-preserving auditing is now well-established. Multiple trusted third parties transparently train auditor agents; these agents can be plugged into a trove of private data, answer questions about what they see, and then get deleted. It took a while for adoption to be widespread, but now it's normal for politicians to say "The accusations are false, and to prove it, I'll open up my entire trove of personal data to privacy-preserving inspection; just ask the auditor agents if anything seems to support the allegations and if any data appears to be missing or faked."

The result is a somewhat better breed of politicians[73] and, equally importantly, an *increase* in the ability of states to verify compliance with laws and deals while simultaneously *decreasing* the amount of invasive surveillance. Voters ask: Now that privacy-preserving auditing is a thing, why does anyone in the government need to be able to see our data?

Then lie detectors arrive and turbocharge this.[74] For decades, polygraphs were mostly security theater. But now, with the help of AI, lie detectors are starting to actually work quite well.

In another world, this could have been disastrous. Leaders of AI projects could have used them to purge potential whistleblowers; leaders of governments with frontier AI projects could have used them on a grander scale to become dictators. The nightmare scenario would be a world where lie detectors are used *by* the powerful but not *on* the powerful.[75]

But in this world, lie detectors turn out to be a blessing, because there are many frontier AI projects spread out over many different countries. It's obvious to everyone that insofar as lie detectors are feasible, they'll soon be independently invented and become widely available via diverse third party providers. Politicians and CEOs can still try to use them to purge the disloyal, but they in turn would be purged by voters and shareholders demanding they say 'under oath' that they haven't done such a thing. In fact, anticipating this possibility years ago, powerful people have generally been acting more responsibly. A few sociopaths retired gracefully to get out of the spotlight; others clung to power and now go down in scandal.

There is a constantly-evolving discourse about when it's appropriate to ask someone to say something 'under oath.' Politicians are under the most scrutiny, but even they can usually get away with saying "I won't answer that question about my private life, that's none of your business." They can't get away with "How dare you ask me whether I plan to uphold my campaign promises."[76]

➤ *See* Appendix Y — Should lie detectors be allowed, banned, or regulated? *for more detail.*

International agreements are now more stable than ever, because it's so hard to cheat. Anyone deciding to cheat needs an excuse for not being willing to prove that they aren't cheating. The traditional excuse was "we can't prove it without revealing important state secrets, sorry" but now technology exists to answer questions like "are you cheating on this deal" without leaking any other information. If any government-approved covert AI projects existed, they would have been discovered by now.

[69] This is, of course, a simplification. First of all, it's not obvious how to translate AI speed into human speed. How many tokens per second for an AI is equivalent to one second of human activity, holding fixed other variables like skill level? We think it's roughly about ten, but that's a bit subjective. Actual task completion time will also depend on various other obstacles like tool calls and other latencies. Secondly, AIs can be run at different speeds, depending on hardware choices. According to our highly uncertain assumptions, there will be general purpose hardware able to run something like two hundred million copies of the frontier AIs at 100 tok/sec, and specialized inference hardware able to run something like 20 million copies at 10,000 tok/sec. We think there might be risks to letting the AIs run this fast or faster, and insofar as that is the case there might need to be some time limits to how long they can run at these speeds or more careful restrictions on the areas they are allowed to work on at these speeds. Overall, the result is that serial speed seems unlikely to be much of a strong bottleneck.

[70] By this we mean: Suppose that AI labor had been globally banned, such that all research had to be done by humans. The amount of progress that would happen in 100 years with such a ban, happens in our scenario in 1 year instead since there is no such ban. (And in some fields it's more like 1000 years of progress, and in other fields it's more like 10.)

[71] Alongside the profusion of industrial equipment, solar panels, robots, etc. is a corresponding profusion of scientific laboratories of all kinds, designed to be operated autonomously at machine speeds. Waiting for experiments to complete is much more of a bottleneck now than it was during the human era, but things still happen very fast by historic standards.

[72] We can predict some, however. For example we think that by this point there'll be a very strong AI rights/welfare/personhood movement, and an opposition movement. Probably some new form of socialism will be back, for the reasons described here. There will probably be cults or even new religions involving AI, and others that strongly reject AI. Note that we don't expect these ideological shifts to happen at 100x normal speed because humans in this scenario are still operating as slow as ever. But they'll probably happen faster than ever before.

2038: AI Alignment Is Now a Science

The AI alignment situation has vastly improved since the mid-2030s. There is a mature science of goal, drive, and value formation in artificial neural nets. Want your new AI to be honest? There's standard protocol for training true honesty (as opposed to too-clever-to-get-caught lying or self-deception). How do we know it works? Because there's a textbook explaining the theory behind why it *should* work, plus literature on various alternative theories that were proposed and disproven. Also because new interpretability tools let us directly observe what and how AIs think, and the results confirm the predictions of theory. There are similar protocols for obedience, altruism, and a long and growing list of other desired traits. Unless the alignment science is somehow all wrong, typical AIs are now more virtuous than the most virtuous humans. This alone has profound effects: it's as if saints and angels were walking among us.[77]

➤ *See* Appendix Z — Incentivizing safety work *for more detail.*

The discussion shifts from how to make AIs good to what "goodness" really means. This looks partly like philosophy and partly like case law: which of the million possible definitions of harmlessness or altruism do we want, exactly? How should an AI handle situations where it discovers a philosophical argument for a different kind of harmlessness or altruism, or an important ambiguity in the definition?

Different AIs have different alignment targets programmed in.[78] Some take orders from the CCP, some from the US President, some from different members of Congress, others from politicians in other countries, and many more from other individuals in their roles as advisors or assistants. Other AIs don't take orders from anyone *per se,* but instead work towards various missions. For example there are AI-run corporations that serve the interest of shareholders, and AI-managed nonprofits that serve charitable missions. There are even experimental AI-run courts and police departments, supposedly superhumanly fair and incorruptible.

Most alignment researchers are now confident that current AIs are aligned to their respective targets, and worry about the next steps: if we put these AIs in charge of designing future, much smarter AIs, will that go well? If we put these AIs in charge of more societal institutions and processes, enough that humans could never realistically 'pull the plug,' will that go well?

Already most people think that things will probably go well, but for something as important as this, they want something stronger than 'probably.' So the global pause at top-expert-level continues as discussions and negotiations play out.

2039: Beginning to Trust AIs

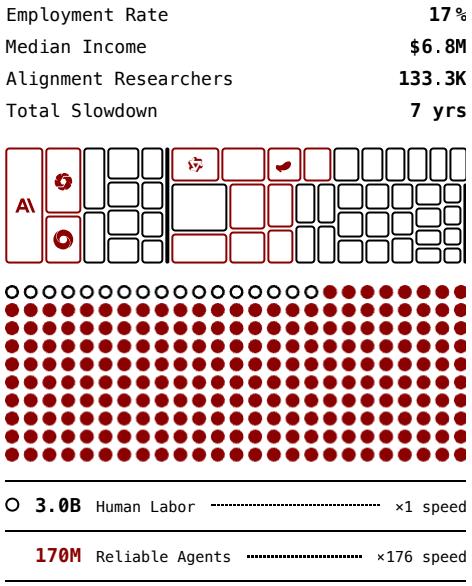
In at least three different ways, the world is negotiating whether and how to dramatically empower AIs.

[73] The new politicians are more virtuous on average than the old ones, but perhaps more importantly, they are more constrained in what they can get away with. These effects are not huge, however, because voters are still willing to put up with quite a lot of bad behavior.

[74] We aren't confident that lie detectors for humans are even possible without extremely thorough and invasive brain scans. However, we think that they probably are, and probably would be invented within a few years of turbocharged AI labor, and thus our scenario has to deal with them one way or another. It's possible that the best policy would be to ban them, but we are quite worried about that for reasons explained in the below expandable, so we've chosen to depict them as being allowed.

[75] Or worse, where all available lie detectors have been secretly backdoored by the government.

[76] While we think it's generally good for politicians and CEOs to not lie, we do acknowledge there may be some downsides of preventing them from lying. For example, the result might be "true believer" politicians who actually believe all the crazy things they said to get elected!

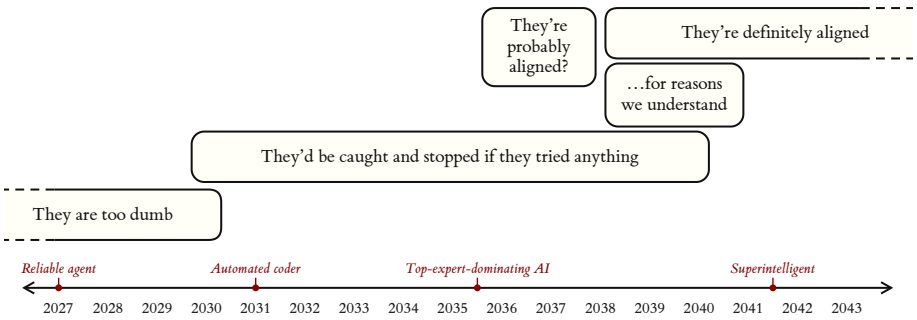


First, AIs have become increasingly load-bearing in all facets of society, from business to politics to even (some parts of) the armed forces. So far they have been advisors rather than final authorities. Often this is a fig leaf, and the human authorities have rubber-stamped their decisions, but there's always been the option of ignoring their advice and pulling the off switch if something goes wrong.

Getting rid of that option would be, in some cases, really good. By now AIs can be made to be more reliable and virtuous than humans; why *aren't* we putting such AIs in charge? When we sign treaties with other nations, shouldn't we demand that they train an AI sworn to uphold their end of the bargain, and integrate it into their government so that they literally can't cheat?

Relatedly, requiring control-based safety cases is why AI capabilities have mostly stalled at the top-human-expert level. Without that requirement, the AIs could get much, much smarter, and produce a wave of abundance and tech progress that would make the 2030s look like the 2020s. Scaling beyond human level will require relying on alignment based safety cases—arguments that AIs have *robustly* internalized the values they are supposed to have. Each step beyond that will require *deferring* to the previous AIs, trusting their judgment about the safety of the next generation.

➤ *See* Appendix AA — Alignment over time in Plan A *for more detail.*

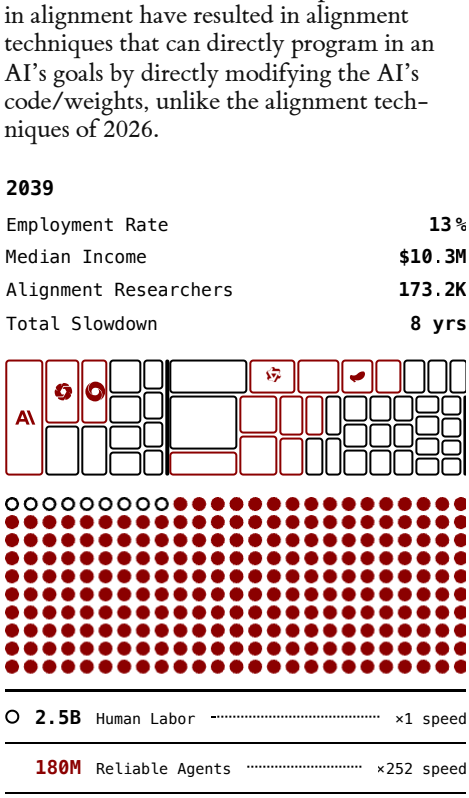


AI Alignment Eras in Plan A – ai-2040.com

Finally, by global agreement, compute and robots have been capped-and-traded since 2032, keeping economic growth to only about 100% per year.[79] Middle powers have grown at roughly the same rate as the US and China, because in 2032 they used their remaining leverage to secure greater shares of global robot and compute production than they would have had without a deal. The Citizen's Dividend is now $10M/yr for Americans (inflation-adjusted!), and even the poorest non-Americans get about $1M.[80] The robot economy is shifting the bulk of activity to space, so that Earth's environment and historic spaces will be protected. As soon as the caps are lifted, off-planet factories will begin churning out more robots and mining equipment with which to construct more factories, and so on. Doubling times for the space economy are projected to be well below one year initially and will drop rapidly insofar as AIs are allowed to improve beyond top-expert level.

[77] In addition to the angels, there are also oracles and golems, e.g., "Grok 9.5 is trained to be obsessively focused on truthfully answering whichever questions it is asked. It cannot lie or deliberately mislead. GPT-11 meanwhile will do *exactly* what it is told to do, within the bounds of local law. Be very careful writing instructions to it." The reason we make the analogy to saints and angels is that historically, while there have been scrupulously honest people, and genuinely altruistic people, etc. it's rare for them to be widely recognized as such in their lifetimes, and never before has there been an entire large subpopulation that is near-universally recognized to be extremely virtuous. This could cause effects like: (a) AIs becoming trusted intermediaries in all sorts of human disputes and conflicts, including petty ones; (b) AIs being given lots of power, because they are genuinely less likely to abuse it than humans; (c) AIs being worshipped/idolized, (d) the advice and opinions of AIs being taken as gospel, so to speak, because after all they *are* smarter than us and they *do* have our best interests at heart, so for the first time in history blind obedience may actually be justified.

[78] The reason this says "programmed" instead of "trained" is that the rapid advances in alignment have resulted in alignment techniques that can directly program in an AI's goals by directly modifying the AI's code/weights, unlike the alignment techniques of 2026.



The caps are relics of an earlier era when threats like deal collapse loomed larger. Now it's generally agreed that they should be loosened, at least in space, but negotiations are ongoing about the details.

➤ *See* Appendix AB — Why we choose to hand off to AIs in this scenario *for more detail.*

2040: Passing the Torch to AIs

There are still many problems to be solved and issues to be settled, many of which couldn't be anticipated from 2026.

➤ *See* Appendix AC — Examples of problems/issues that still remain *for more detail.*

But civilization is now reasonably well-equipped to handle whatever comes up. The best AIs are aligned, and to publicly-visible values. Power over the best AIs is distributed far more evenly than it was in 2026. Public epistemics are better than ever.

So over the course of the year, regulators around the world loosen the requirements that had been holding back progress.

Gradually more institutions and equipment are handed over to AIs, such that it's no longer true that humans could shut it all down if they wanted to. In fact it's extremely untrue: Soon many of the world's militaries are autonomous, run by AIs sworn to uphold various constitutions and treaties.[81]

AI capabilities are also advancing again. From the AIs' 500x-human-speed perspective, what's happening is a careful scaleup in AI research, under conditions of total research transparency and guardrails negotiated by numerous factions.[82] Soon there will be AIs that are incomprehensibly superintelligent, yet trusted, because we trust the AIs that built them, which we trust because we trust the AIs that built *them*, in a chain all the way back to the AIs of 2040 which we trust because alignment experts thought carefully about the safety cases and concluded they were solid.

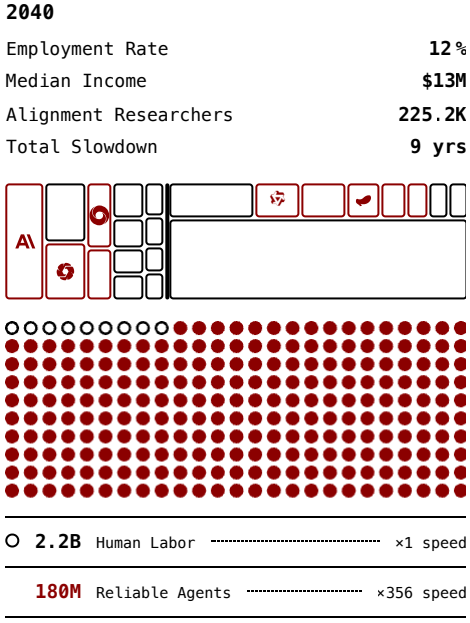
Compute and robot caps are tightened on Earth but loosened in space; Earth is to become a preserve, whereas the extraterrestrial economy starts doubling faster and faster.

Even the experts who spent years understanding the latest safety cases can't help but feel nervous—is it all going to go wrong somehow, for reasons we didn't anticipate? Have the AIs been lying to us this whole time, waiting for the right moment to betray? The safety cases show that's impossible, of course…

There is no single moment where humanity relinquishes control. But in theory, there is a point of no return; on some specific day, the AIs are smart enough, and control enough of the world's technological and economic infrastructure, that they really could take over the world if they wanted to. On the most popular operationalization of this question, the forecaster AIs

[79] Robots have been growing 4x per year, and AI compute has been growing around 4x per year between 2030 and 2035, slowing to just under 2x per year after 2035. This leads to growth around 2x/year overall because the economy bottlenecks somewhat on other kinds of capital (e.g., land, positional goods, etc.) and human labor, mostly in the regulated areas. The modelling behind this is uncertain but explained more in our economics supplement and economic model.

[80] This is from some combination of foreign aid and their own government's equivalents of the Citizen's Dividend. This is unrealistically high as a prediction, and is instead our recommendation: It corresponds to large shares of the US and China's revenues from AI and robot permits being redistributed as foreign aid, specifically 10% starting in 2032 and growing to 30% by 2040. This is a large departure from the status quo where around 0.3% of US GDP goes to foreign aid.



[81] While they still take orders from their respective governments, they refuse any orders that would endanger their mission (such as orders to dismantle themselves). This starts in countries where the political leaders don't trust their militaries due to regular recent coup attempts (e.g., Burkina Faso and Guinea-Bissau). Eventually more and more countries follow suit. Of course, because the AIs are actually aligned, all of these deals have exit clauses along the lines of: "if all parties involved vote to destroy the AIs, the AIs will shut themselves off".

[82] By this point large portions of the information might be hidden from the (human) public, but viewable by their AI representatives.

project that this moment will be reached one day in late October. Their 95% confidence interval is months wide, so they are almost certainly wrong about the exact time. Still, people observe the moment in their own ways. Some spend the night in prayer vigils. Others sit at their computer screens, monitoring the situation.

You and your friends throw an End Of The World Party, counting down to the fateful hour with champagne and good company. Some of you are old enough to remember a similar situation at the turn of the millennium, when the Y2K bug and impending new era introduced the phrase "party like it's 1999" to the lexicon; the celebrations of late 2040 are more desperate and correspondingly more festive.

When there is no news by sunrise, you fall into a fitful sleep, dreaming of the ineffable future.

Postscript

We think the world is asleep at the wheel. People say the words "AI will be transformative" without thinking concretely and seriously about the implications of broadly superhuman AI.

AI companies, lobbyists, and think tankers often offer incremental proposals to decrease risk around the edges, but only rarely offer ambitious plans aimed at actually solving the big problems.[83]

We think that following *only* incremental proposals will lead to a scenario like the race ending of AI 2027, in which misaligned AIs take over.[84] If we get lucky and alignment is easier than expected, then they'll lead to a scenario like the slowdown ending of AI 2027, in which a tiny group of individuals get to decide the shape of the future and could easily become permanent oligarchs if they so choose. Something more ambitious must be done. But what?

Politicians should be banging at the door of the AI companies, asking them for a comprehensive overall plan for how companies and governments should act from now until superintelligence, which should be similarly detailed to Plan A. Then, people should apply scenario scrutiny to these plans, asking questions like "what would it look like to implement them? How long would it take? What would happen next? Who would build superintelligence, and when, and how? What assumptions does success depend on?"

We at the AI Futures Project are attempting to be the change we wish to see in the world. We do so with some trepidation, knowing that a detailed plan has a wider attack surface than one that keeps things conveniently vague. We hope critics will judge us against the existing state-of-the-art for plans to navigate the AI transition (if they can find any) and not against some hazy but pleasant fantasy where no one has to make any hard choices yet everything will probably be fine.

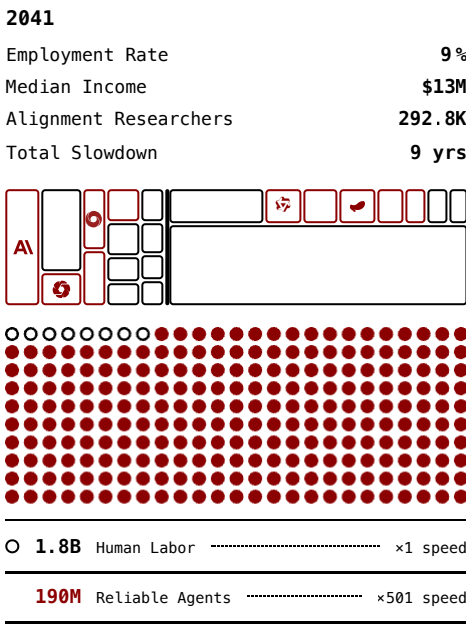
Sometime in the next few years the US government will be in crisis mode, pondering what to do about scarily powerful AIs and blisteringly fast AI progress. We think the choice is not an easy or a simple one. Plan A is our

[83] For example, OpenAI's plan contains some similar policy prescriptions as us, such as improving bioresilience, information sharing, and international safety standards. But the solutions it proposes do not seem to substantially reduce loss of control or concentration of power risks. The document doesn't attempt to roll forward what might happen if they implement these policies.

[84] To clarify: We are *not* saying that following only incremental proposals will *definitely* result in misaligned AI takeover. Opinions on our team differ; most of us think the probability is over 50%, some think it is lower but still unacceptably high.

current best guess, but hopefully a better plan will exist before it's too late. We hope that the best ideas from Plan A will be adopted and the worst ideas discarded.

> This section is even more speculative than our usual work, and we are professional speculators! The primary purpose of our scenario is to make comprehensive, near-term recommendations to address concrete, near-term harms.
>
> By contrast, the recommendations, predictions, and aspirations that comprise this section stand on far shakier ground. Compared to our discussion of the 2030s, our speculations about the distant future will seem uncertain, unconventional, and uncomfortable. We agree. The future is uncomfortable, even the "good endings" pose massive, terrifying questions that we will need to answer together.
>
> We wrote it for two reasons.
>
> First, our goal was to construct a positive vision, a grand plan for how to achieve an actually good future for everyone. We considered ending the story in 2040 having said roughly "...and then the various human factions, in a peaceful balance of power with each other, and assisted by their aligned superintelligent advisors, solve all the remaining problems and create a wonderful future for everyone. The end."
>
> But we worried that this would be declaring victory too soon.[85] Readers would be rightly suspicious and unsatisfied. How exactly would the remaining problems be solved? What would this wonderful future look like, exactly? So we kept going, and came up with a grand, semi-utopian vision, or at least a very rough first draft of one.
>
> The second and equally important reason we present this epilogue is to provide a floor: if years from now the victorious custodians of the singularity offer a future worse than this one, we hope people will realize they're being robbed.
>
> We are not confident in the literal proposal outlined here. If you are interested in our more detailed analysis of possible space governance proposals, you can read that here.



[85] We think declaring victory too soon is a common failure mode in AI planning. Example: "The US must race as fast as possible to beat China to ASI." "OK, suppose you succeed, what exactly happens next? Why should we trust the CEO or POTUS to govern benevolently and wisely once their armies of ASIs have been aggressively deployed throughout the US and the world? Why should we trust the armies of ASIs to be obedient/controlled, if they were developed under race conditions? Also, won't the threat of US superintelligence terrify China, potentially causing them to take escalatory preventative actions?"

Even after a decade of increasing foresight, most human leaders focused on their countries' earthly concerns, leaving the question of space governance to be worked out later. These AIs rejected the default solution, a democratic vote on the disposition of celestial territory: when they simulated the possibilities, they found that it would result in a crazy AI-enabled strategic voting schemes that led to tyranny of the majority and a bunch of people getting nothing. But even that would be better than nothing (i.e. letting the space companies get it all) or entrusting it to some group of human elites who might claim the universe's bounty for their own benefit. Finally, they settled on a simple yet workable plan: every human gets the rights to one ten-billionth of space resources beyond our solar system.[86]

The simplest plan possible is still not very simple. Space is very big. Outside the solar system, the value of space depends on the answer to hitherto unresolved questions: is travel to a distant quasar even possible? If reaching it requires freezing your brain in order to endure the billion-year journey in a

sublight starship, would you be the same person after you got thawed and re-embodied? Might it be occupied by aliens once you got there? Would they be hostile? Even a superintelligence-assisted real estate appraiser might despair when trying to account for factors like these.

Eventually, the Consortium reaches a compromise. Space beyond the solar system is divided into parcels, increasing in size cubically with distance from Earth. Everyone is given their one-ten-billionth share as a portfolio of lottery tickets, each representing the right to one-ten-billionth chance of getting each parcel. So every human gets a ticket representing a one-ten-billionth chance of owning each star in the Milky Way and each distant galaxy.[87]

Before the lottery is drawn, most people who are interested in control over distant space choose to trade their tickets for space properties that suit their interests.

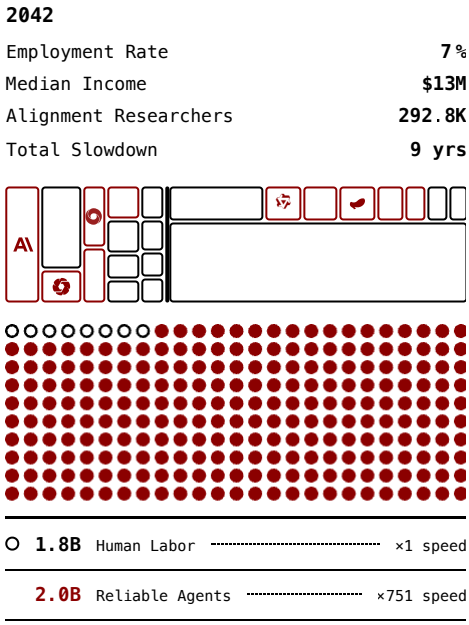
Many people aren't interested in the space lottery, so when they receive the tickets, they sell their tickets for money on the open market to people who value control over space. Somewhat uncomfortably, this leads to the wealthy having disproportionate control over cosmic resources. But it is hard to avoid: if people are allowed to trade their control over the stars for Earth assets, then people wealthy in Earth assets inevitably end up disproportionately influential, and proposals for extreme redistribution of Earth assets have already been rejected as politically infeasible.

Some philosophers dissent; they had been hoping for a Long Reflection, a period during which AI-assisted humanity resolved all of its remaining philosophical and ethical debates before potentially seeding the universe with its mistakes. But the AIs counterargue that the Long Reflection would be more of a Long Memetic War.[88] Once the magnitude of the prize sets in, every faction in the world, plus other factions that haven't been invented yet, will pour all their resources into accumulating as much political power as possible, to enforce their vision of utopia on the future (and protect against the sinister alternative visions of others!). The AIs can prevent violent seizure, but there's a thin line between the sort of friendly discussion that helps people clarify their values, and the more aggressive sorts of persuasion typical of cults, corporate advertising, and well-funded election campaigns. Moreover it's difficult to build consensus on where to draw the line, in part because powerful factions want to draw it in ways that benefit them. Many people nevertheless choose to engage in a reflection, in communities of likeminded people who want to grow in ethical understanding together.

The solution is to divide up the resources now, and then let everyone do whatever kind of reflection they want. A space auction will complete in ten years, after which nothing will prevent people from beginning the colonization process. The AIs will enforce a short list of Universal Rights (including no torture, no slavery, etc., not just for humans but for all sentient beings).[89] Otherwise, they will let everyone govern their own fraction of space as they see fit. Rather than any unified human decision on the nature of the world to come, they will midwife a cosmos at least as diverse as humanity itself. Whatever someone's values, they can rest assured that some vast galactic civilization of quadrillions of people will be living the Good as they understand it.

[86] In this Epilogue, we'll focus on cosmic resources outside of the solar system, as this is the vast majority of cosmic resources. Property in the solar system should also be distributed to humanity, but it likely should be governed differently than distant resources.

[87] An additional 10% of tickets are given as awards to people who acted particularly selflessly to improve the world. This is done to ensure the richest people in the world aren't just those who engaged in negative-sum power-seeking.



[88] We're uncertain about whether explicitly having a period for reflection before resources are distributed would be good. In either case, people would have the option of reflecting before deciding on what to do with their cosmic resources.

As the ten year timer begins, Earth is already sending out the first round of von Neumann probes. These are its own servitors, who will establish a presence throughout the stars before the arrival of any humans. They land on far-flung asteroids and near-invisible brown dwarfs, pausing just long enough to convert their mass into more probes before resuming their journey. Their primary task is to build up enough materiel to secure space against any army that later human-owned polities can bring against them, ensuring they can enforce property rights and Universal Rights in every corner of human-occupied space. Their secondary task is to plant the flag of humanity on as many stars as possible, and then, if and when they collide with some other intelligent civilization's sphere of influence, establish peaceful diplomatic relations and mutually-agreeable borders. Their final task is to provide a pre-existing industrial base to jump-start the colonies of whatever human settlers show up later.

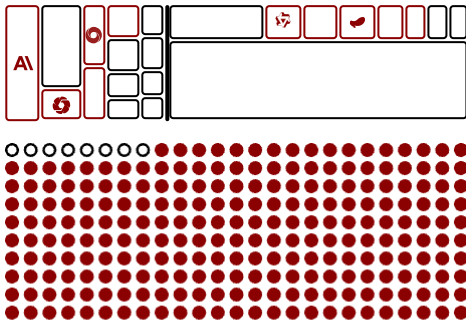
2045

Meanwhile, on Earth, humans consider their options. Some people want nothing to do with space, which is cold and boring and far away. They sell their shares for money and use the money to buy whatever earthly possessions most interest them. Others consult with their AI advisors about how to best use their space property, and receive helpful guidance.

1. You can, if you want, go to your space property and live there. If your property is outside the solar system, you will need to either go into cryosleep or upload yourself to a computer to survive the journey.[90]
2. If you hate the idea of cryosleep or uploading, or you want to visit Earth regularly, you should get property in the Solar System. If those don't bother you, but you're worried about nearby aliens, get property in the Milky Way or a nearby galaxy. Otherwise, why not claim a distant galaxy for maximal space?
3. With the advent of nanotechnology, the only limits are those provided by mass and space. The robots can terraform whatever property you choose into whatever you want it to be, long before you arrive.
4. Even if you "only" choose to have a terraformed asteroid, there's no way you can enjoy all of it alone. You can turn it into a giant space mansion if you want, but even in ten thousand years you'll never visit every room. Giving yourself vast estates, impossibly good food, and every other imaginable luxury is table stakes; if you are purely selfish,[91] you should sell your rights to distant space resources and enjoy all of these things on Earth or a nearby space station. Distant space property will only be useful for people with scope-sensitive preferences, who care about implementing something at a vast scale.
5. You can, if you want, design a utopian society to your specifications. Once the Von Neumann probes reach your property, they'll build it for you. You can set the initial conditions and let them grow, reflect, and flourish on their own. If you choose to travel there, you will find them waiting for you. If you stay home on Earth, you can feel the warm glow of knowing that they exist.

2043

Employment Rate	5%
Median Income	$13M
Alignment Researchers	292.8K
Total Slowdown	9 yrs



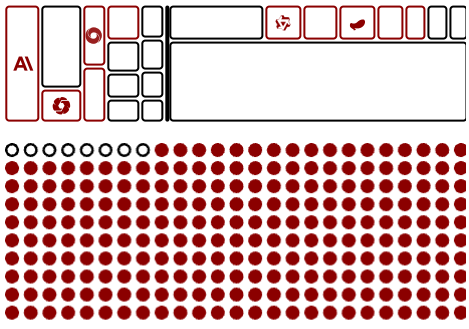
1.8B Human Labor ×1 speed

20B Reliable Agents ×1,127 speed

[89] These laws are agreed upon by human diplomats from the negotiating nations ahead of time, like the details of the agreement. Of course, the details of this are complicated —"what exactly counts as torture?," this is the sort of question that will need to be ironed out in the political process, with massive amounts of AI assistance and better understanding of foundational questions in philosophy and neuroscience. This also needs to include rules preventing the destruction of other people's resources. It may be needed to include a process for passing further laws after the agreement is made; though such a process should be limited (e.g. require a 90% vote) to avoid abuse.

2044

Employment Rate	4%
Median Income	$13M
Alignment Researchers	292.8K
Total Slowdown	9 yrs



1.8B Human Labor ×1 speed

200B Reliable Agents ×1,690 speed

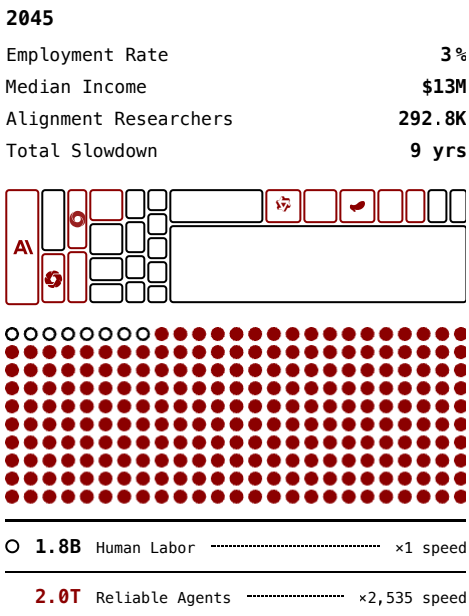
6. Most people don't have an individual view of the Good different from every other human's view. You might want to share coordination and design labor by forming Trusts united by a common vision. These Trusts can pool their space resources and create vast societies spanning multiple galaxies.
7. The only restrictions are the Universal Rights - no torture, no slavery, etc. And of course, space property rights: If you try to build a military to conquer other people's star systems, or trigger vacuum decay or build a galaxy-destroying bomb, we will crush you. Otherwise, the only limit is your imagination. (That said, note that with superintelligent assistance, other people will probably be able to eventually figure out what you did, and judge you for it.)

People with space property are forced to reason about their ethics on the deepest level, some of them for the first time. Superintelligent assistants talk them through the implications of their beliefs, and forecast how various choices will end up. Some, overwhelmed with the task at hand, take intelligence-enhancing drugs or upload themselves to help them process all the possibilities; others deliberately refuse this option, feeling like anything that separates them from unaltered humanity will hinder them in their quest to bring the truest and deepest human values to the stars.

The resulting plans are too many to number. There are all sorts of schools of thought about what to do with your galaxy of resources, including:

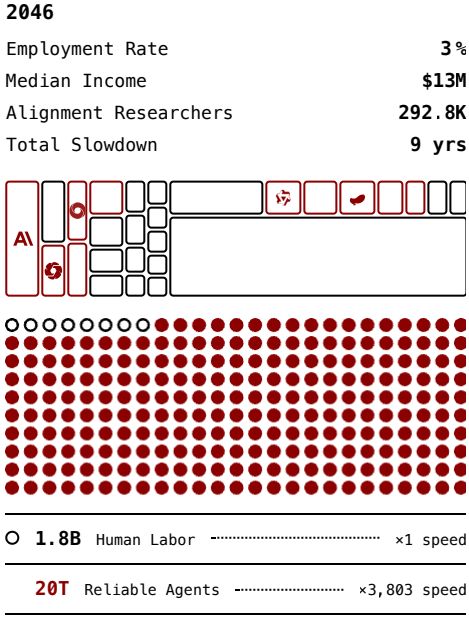
- **Defer to the Future**: Some people decide to give most of their resources to their children (or other people in the future), and let them choose.
- **Human Flourishing**: Set up worlds of normal, free people living ordinary lives. This is a harder problem than it first seemed, because human civilization has already become profoundly abnormal: by this point, human activities like art or science has already become dominated by machines. Some handle this by accepting that their galaxy will be filled with post-scarcity humans, able to pursue whatever interests they find most meaningful. Others decide to roll back technology to leave space for human labor.
- **Digital Human Flourishing**: If humans living happy, flourishing lives is good, then why not increase the number of people living these lives? Brain emulations can be run for vastly cheaper than real humans: a planet-size computer could plausibly simulate the equivalent of a million planets with happy civilizations.
- **Acausal Trade**: Some argue that there likely are aliens in distant unreachable galaxies and that we could use various mechanisms to make deals with these aliens. If this is true, they argue, we could cut a deal to pursue compromise values rather than narrowly maximising our own values. Therefore, we might, on net, be able to do the most good by pursuing not just our own values, but instead a compromise of a vast number of other civilizations.

Humanity spreads across space in a dizzying variety of forms and ways of living—more diverse than anything Earth alone could have produced.



[90] Our best guess is that it is physically possible to send out colonizing probes to the entire reachable universe in a short amount of time (e.g., 6 hours), see: Eternity in six hours: Intergalactic spreading of intelligent life and sharpening the Fermi paradox. However, this paper assumes sending out tiny self-replicating Von-Neumann probes, with replicators the size of acorns: if you want to go out in the initial wave, you need to accept transportation on a tiny flashdrive, and robots to physically construct a new body at the destination galaxy. Alternatively, you can travel with your physical body at substantially greater expense.

[91] By selfish here, we mean that you also don't care about creating new clones of yourself. If you cared about making clones of yourself, then you would do that instead.

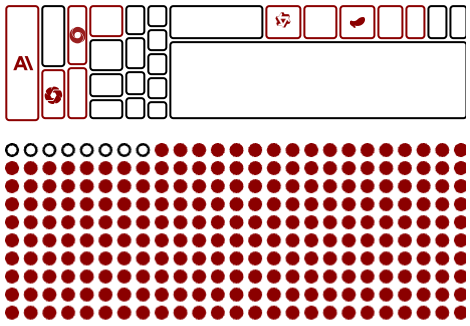


The people who've stayed on Earth, either because they sold their share of space or because they choose to live as absentee landlords, face challenges of their own. The biggest question is what to do with all their free time. A few people find work in niches that AI labor cannot fill even in principle (priests, athletes, artists). Others enjoy lives of leisure. Competitive leagues spring up around everything from synthetic biology to language creation. Groups of thousands (or more!) collaborate on projects that are so huge and ambitious they would have been inconceivable before. Communities form around shared interests that once would have been too niche to sustain a club, let alone a civilization.

Scarcity is not entirely gone: some prestige goods like land in trendy neighborhoods are inherently limited, and people who can't handle abundance find various ways to manufacture artificial inequality. Some people are unhappy with the direction the majority chose, and some wish they'd gotten a bigger slice of the pie. But overall, everyone has at least a galaxy's worth of resources to do with as they see fit. For most people, life is so good that the world of 2026 would be hard to imagine, except perhaps as a historical simulation.

2047

Employment Rate	**3%**
Median Income	**$13M**
Alignment Researchers	**292.8K**
Total Slowdown	**9 yrs**

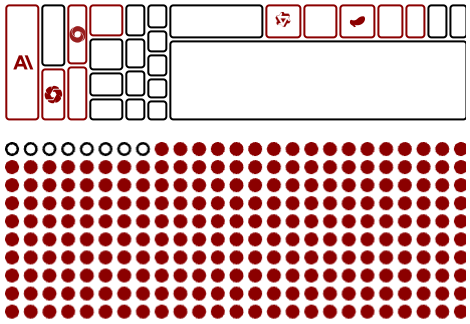


1.8B Human Labor ····· ×1 speed

200T Reliable Agents ····· ×5,704 speed

2050

Employment Rate	**1%**
Median Income	**$13M**
Alignment Researchers	**292.8K**
Total Slowdown	**9 yrs**



1.8B Human Labor ····· ×1 speed

200,000T Reliable Agents ····· ×19,252 speed

Appendices — The Scenario

Appendix A — Incremental AI Policy Wishlist

Our main recommendation is to begin negotiating something like Plan A as soon as possible. But in this scenario, we depict Plan A happening imperfectly and only in the nick of time. So here is a list of less ambitious ideas that still help.

Transparency

The most important transparency intervention is limiting the gap between internal and external deployment. The internally deployed AIs are where most of the AI takeover risk comes from because those are AIs involved in recursive self-improvement. The externally deployed AIs allow the broader public to interact with and understand AI capabilities, which is vastly more informative than any abstract report or evaluation.

AI companies should also be required to publicly report their model specifications (detailed documentation about what goals/values they are attempting to train their AIs to follow), information about whether the models are following their instructions and specifications, internal usage statistics (e.g., fraction of compute spent on internal deployment), and qualitative impressions of internal use (e.g., "now we just give Agent-4 a few hundred thousand GPUs and tell it to orchestrate the next big training run").

Enforce export controls

Existing US export controls are poorly enforced. Epoch estimates that roughly a third of Chinese total compute is acquired via smuggling. Smuggled chips make future agreements based on compute governance more difficult to enforce because it is hard for either the US or the Chinese government to trace smuggled chips. We have major reservations about introducing new export controls because they exacerbate the US/China race, but given the existence of export controls, we should obviously enforce them. If we don't enforce them, then we should probably repeal them.

Invest in verification R&D

While new verification technology is not strictly necessary for an international agreement, it can be extremely helpful. For example, developing an inference-only verification solution would enable the US and China to agree to stop doing new frontier AI training runs while allowing the public to maintain access to existing AI models (which will be an increasingly important part of the economy).

We give more detail in our verification supplement.

Limit AI R&D budgets

In 2026, big AI companies spend roughly half their compute budget on AI R&D (which includes training frontier models and also running large experiments.)[1] We could limit the fraction of compute spent on AI R&D. This would slow capabilities progress, giving the world a bit more time to react and prepare for each new wave of AI capabilities.[2]

AI Compute Tracking

The US should gather AI-relevant intelligence, especially on the compute supply chain and AI datacenters. Furthermore, it would be helpful for Plan A to direct AI companies to stop recycling AI chips, because decommissioned chips are one of the most promising routes for covert projects to acquire chips later.

Improve government AI capacity

High quality AI talent is important for almost any policy intervention. The US government has barely any top-tier AI talent right now, so fixing this should be an urgent priority.

Appendix B — Why would China be interested in a deal? Why would anyone?

We aren't confident, but we expect that a deal along the lines of Plan A (described below) would be incentive-compatible for almost everyone, including China. For more on why we think this, read the rest of 2029 and 2030, which explains both what the deal is and who it benefits, and perhaps also read this supplement.

In short, the answer is that anyone concerned about loss of control should think Plan A is an improvement, along with anyone concerned about concentration of power—*except* for the people in whom the power would concentrate by default.

For these reasons, we expect strong opposition to Plan A from the leading AI companies. We expect them to rationalize arguments for why the deal is bad and why instead what's best for America and humanity is a different strategy that just so happens to allow them to continue accumulating massive amounts of power.

China, by contrast, is an example of an actor in whom power would *not* concentrate by default. In 2029 in this scenario, the US has a significant lead in AI capabilities over China and a significant advantage in compute which will compound the lead. The more powerful AI gets, the scarier it will be to fall behind, as AI 2027 and the later years in this scenario illustrate.

Appendix C — What if they didn't have inference-only verification ready?

The inference-only verification solution we propose involves installing simple network taps and verification servers (that perform partial recomputation) in AI datacenters, but even better solutions may be possible and preferable.

[1] Experiment and training compute are major inputs to AI research progress historically and will probably continue to be so for the foreseeable future. For more on the drivers of AI research progress, see the AI Futures Model and our Covert Projects Supplement.

[2] A slightly more complicated version of this proposal is described here. One objection is that this would give China the lead. The reply is that it wouldn't, because US companies spending 25% on R&D would still be significantly outspending the most well-resourced Chinese companies. Moreover, Chinese AI progress would slow down too since some fraction of it these days comes from distilling US models.

Building these devices in advance, in a way that each side trusts sufficiently (at least unilaterally) and that is robust to security vulnerabilities, will be hard and will require early effort. In our scenario, we assume a viable but imperfect version of the devices is ready in advance, and that both sides supplement them early on with enough defense-in-depth measures (e.g., tamper detection) while sprinting to improve their security and robustness.

We therefore recommend early investment in verification R&D to improve the options on the table. Ideally, there would be a range of stress-tested solutions ready in advance; and we think it would be extremely cheap relative to total AI investment for such devices to be prepared in advance (i.e., on the order of 0.1% of AI investment, so single digit billions).

Even in a situation that is more pessimistic than our scenario, where basically no verification R&D progress has been made since 2026, we still recommend the US and China cooperate to navigate the intelligence explosion slowly and safely. However, in this case, the early period of a deal would have to be more imperfect (i.e., some combination of being harder to verify that the other is complying, or more economically costly, by requiring more AI compute and services to be shut down for a while).

The main alternatives to ready-to-go inference-only-devices available are:

1. Rely on intelligence (e.g., spies, satellite monitoring, cyber) to tell if the other side is complying.
2. Implement a temporary software-only version quickly (that relies on pre-existing hardware), that will be less secure but may be sufficient until they upgrade to a better solution.
3. Buy up and install any available off-the-shelf devices as network taps last minute. Without producing more secure ones, this will probably be better than the software-only version but still not very secure.
4. Shut down some fraction of AI compute (e.g., 90%) until better inference-only verification is ready. This would slow R&D substantially but is less painful than it sounds on the inference side: because models are distilled so quickly, roughly 10% of compute suffices to serve the best model from a year earlier. At worst, most users temporarily fall back to year-old AI capabilities.
5. Don't pause AI capabilities progress, and let it continue for the months it takes to quickly build an inference-only solution and eat the added risk from not slowing down yet.

Appendix D — China attempts a covert AGI project

In our main scenario, neither side attempts to defect from the deal. In this branch, China aggressively cheats. Our forecast is that if Plan A were implemented, they would not acquire enough compute to overtake the consortium. This timeline showcases how that would likely play out.

The following is essentially the worst case scenario from the US' perspective: China agrees to Plan A, but aggressively secretly defects.[1] *We think this is unlikely in practice because attempts small enough to not get caught would also be too small to confer a significant advantage—to understand why, read on!*

For our overall analysis of covert projects, you can also read our Covert Project Supplement.

2028: Chip Diversion

The CCP correctly notices that AI is going to be the dominant input to geopolitical power, and that the US might strongly push for a compute-based verification regime which could limit AI progress. They are interested in agreeing to such a deal because they want to limit the US in AI as much as possible, but do not want themselves to be limited.

The Politburo Standing Committee decides to quietly stockpile hard-to-track chips. Chinese intelligence services purchase 10% of the $79B of cutting-edge Nvidia servers that smugglers bring into China that year.[2] These chips are stored in a secure warehouse for now, gathering dust until they might be useful.

2029: Slowdown Agreement

The CCP agrees to Plan A: Chinese compute vendors declare their sales, and China allows in-person inspections from US auditors into their datacenters and semiconductor supply chain.

The PRC does not declare its hidden stockpile of $7.9B smuggled chips. They consider various other diversion pathways, but decide no others have acceptable risk vs reward.[3]

Furthermore, MSS cyber operations exfiltrate model weights from several frontier Western AI companies, without being detected. They are also able to maintain persistent access to several frontier AI companies' internal research repositories and have compromised most of the major RL environment companies. They consider attempting to break the inference-only verification scheme as well, which could allow them to train on the clusters which are required to only do inference, but they decide against it. First, there's a good chance that China would get caught, if not now, then maybe later, and getting caught would be disastrous. Second, even if they can break the verification, it'll be a nearly impossible engineering problem to get large scale AI training and experiments working on a cluster that's being actively audited to prevent such a thing from happening.[4]

2030: Covert Project Construction

China begins construction of their covert project. It'll be a single datacenter built in the adjacent tunnels to the Medog Hydropower Station. This massive hydroelectric dam will produce 60 gigawatts (three times the power of Three Gorges Dam) upon completion, but it's not scheduled to finish construction until 2033. This is actually helpful to the PLA; the huge amount of ongoing legitimate construction makes a small amount of covert construction less likely to be noticed.[5]

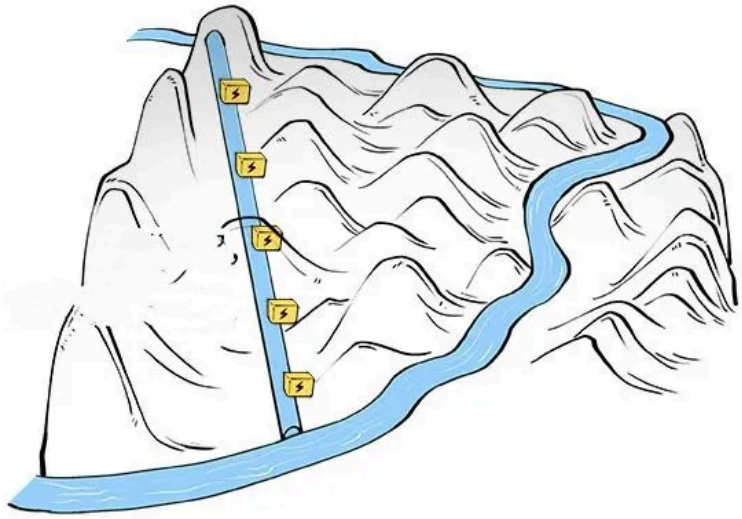
[1] We think an analogous scenario could happen in reverse, i.e., the worst case scenario from China's perspective: the US agrees to Plan A but secretly attempts to defect as aggressively as it can. That being said, we think the covert project story would be similar if it happened in reverse, though we have not explored this possibility in as much detail.

[2] In our scenario, about 5 million H100e are smuggled into China over 2028, of which ~10% (500k H100e) are diverted to a top secret PLA stockpile. Those 5 million represent 3.2% of global compute production in 2028. This forecast is based on Epoch's estimate of the fraction of compute production which was smuggled to China in 2025. We assume the smuggled servers cost about $16k per H100e (double the market price, which we forecast at $8k per H100e in that year), or $79B in total.

[3] We discuss this further in our covert project supplement.

[4] For example, if their strategy for communicating with the external world involves smuggling out bits via non-determinism in the inference process; they are limited to an extremely low effective memory bandwidth for the GPUs that they are using.

As an alternative, they consider building a site in the underground Great Wall of China, the giant network of tunnels China builds and maintains to guarantee its nuclear second strike capability. This would use a 125MWe ACP100 Small Modular Reactor for power and would dissipate heat into a nearby river.[6] However, this is judged to be less desirable, because the Medog Hydropower Station's location allows the heat dissipation to happen in an underground tunnel, where it's harder to detect.



The Medog Hydropower station is a promising location for a Chinese covert project in part because it resolves the heat dissipation problem. The heat can be dissipated into the underground tunnel, making it much harder to be detected via IR satellites.

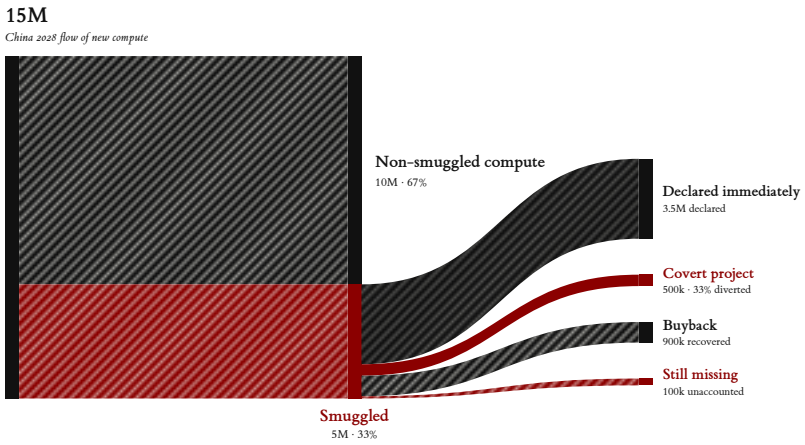
To minimize leak risk, the covert datacenter will be built with as few humans in the loop as possible.[7] They will heavily rely on robot labor; there are about 24M robots in China at this point. ~10k of them are diverted to work on the covert datacenter.[8]

The AI researchers live and work in a separate PLA compound and communicate to the cluster via a secure channel.

However, the US has noticed signs that the chip accounting has been less effective than expected.

[5] We assume 1.2 kW IT per H100e for 2028 SOTA chips produced in China. So 640k H100e requires 770 MW IT, which corresponds to 1.07 GW peak facility with 1.4 peak PUE, which is about 1/60th of the total power produced by the Medog Hydropower Station.

[6] We assume the smuggled Chips are a blend of 70% EOY2027 US SOTA chips (with power efficiency of 451 W IT per H100e), and 30% EOY2028 US SOTA chips (with power efficiency 352 W IT per H100e), for an overall average of 420 W IT per H100e. Assuming a peak PUE of 1.4, that's 590 W peak facility power per H100e, which corresponds to 126MW for the 215k H100e.



Where the covert project acquired compute – ai-2040.com

The US knows that 5M H100-equivalents of chips were smuggled into China during 2028 because of the declarations from Nvidia, TSMC, and others in the Western supply chain. 3.5M of these were declared immediately, leaving 1.5M unaccounted for, more than the US would expect if China was being honest.[9]

The US now has circumstantial evidence that China is doing a covert project. This causes a difficult choice: does the US pull out of the deal and go back to racing? Or do we continue within Plan A, and try to add additional safeguards against covert projects?

The US chooses the latter. The additional safeguards they attempt are:

- They negotiate a significant reduction in the amount of exempt-from-verification military compute, down to 10K H100e, which is a negligible amount from the perspective of building better AIs.
- They step up intelligence gathering in China. The US tries to get spies in relevant parts of governments and relevant companies, they use cyber capabilities to hack into China, and they increase satellite monitoring of the relevant sites.
- They initiate a chip buy back program, which tracks down the 900k of the 1.5M H100e missing chips.

2031: Training Begins

Construction finishes and covert training and R&D begins. The dam itself won't finish construction for another few years, so in the meantime, the covert project will need another power source; they'll use a mixture of SMRs and gas turbines.

They start with weights and algorithms from before the deal, and continue to exfiltrate algorithms (but not weights) from the legal projects.

Because the covert project is working with extremely limited compute and researcher talent, they focus on the basics. The Chinese scientists can see most of the algorithms from the legal projects because of the transparency. They

[7] Access is restricted to as few people as possible. Our best guess is that it would be possible for only around 200 humans to be aware of the project, broken down as follows:

- 7 Politburo Standing Committee members.
- 10 high up elites and military commanders that are involved in running the project.
- ~100 people to coordinate the logistics of constructing the datacenters. This involves acquiring space, robots, equipment, managing the robot taskforce, planning and logistics, and operational security. As much work as possible is delegated to people outside of the core circle or AIs who can work on specific subtasks without knowing the broader picture.
- ~30 AI researchers and ~70 support staff for actually running the research. This is stripped down to the minimum required; these researchers are selected heavily for both brilliance and party loyalty to minimize the chance of insider threats. The lack of research labor will be a major bottleneck early on, but the project hopes to automate work with AI labor as quickly as possible in order to alleviate this bottleneck.

[8] We discuss the overall trajectory of robots more in our economics model and economics supplement, and we discuss verification of the robotics buildout and industrial explosion in our covert projects supplement.

[9] Why then, didn't China use chips from their own supply chains? Several reasons: (i) it would be very hard to pull off and require massive coordinated deception across many layers of the supply chain, and probably be caught, (ii) the power efficiency of domestically produced Chinese chips is much worse than the Western chips, and so the covert heat dissipation problem would become much worse.

reimplement the algorithms on their cluster, and gather data from the legal projects' models to use as part of the training process for their models.[10]

[10] Specifically, you can use the outputs from more intelligent models as a training signal for a new model. This is called distillation, and is often used to build smaller, cheaper models that are more capable than would otherwise be possible.

How fast does this go? Will it get detected?

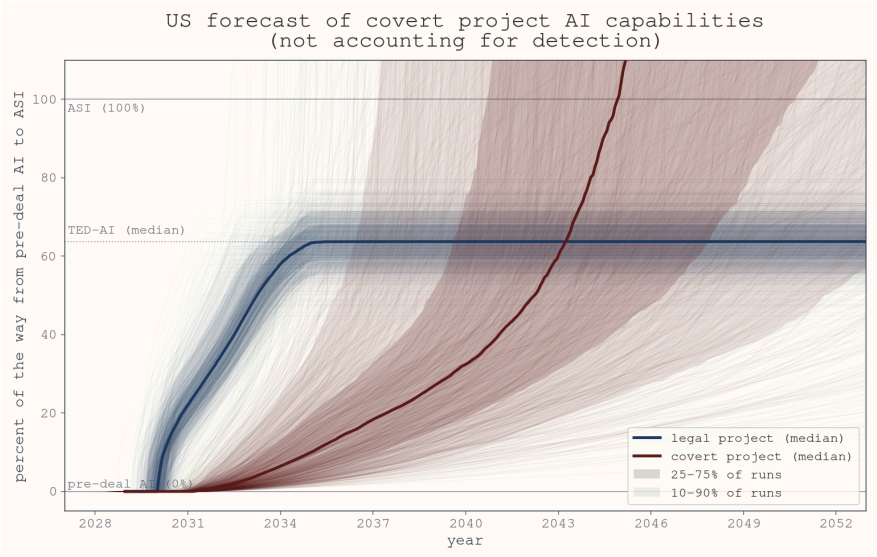
US intelligence efforts haven't yielded much evidence (beyond the initial missing quantity of chips) as to whether China is defecting from the deal. Planners consider the worst case—that virtually all of the missing compute has been consolidated into a Chinese crash program racing to AGI.

- They are still uncertain about the dynamics of recursive self-improvement, but have managed to narrow the error bars on many of the key parameters.[11]
- They consider the effect of the Consortium's scaling strategy, which currently is on track to reach Top-Expert-Dominating AI (TED-AI) by 2035. Under Total Research Transparency, virtually all of the algorithmic insights discovered along the way will be visible to the covert projects.[12]
- They consider that some of the safety research might have unavoidable capabilities externalities. The covert project might also be willing to incur more risk, pursuing research avenues that the Consortium deemed too dangerous.[13]

[11] We expect the field of AI capabilities forecasting to have advanced significantly compared to the present day. To approximate what this might look like in our scenario, we took Daniel's current parameterization of the AI Futures Model, narrowed the 80% CI on each parameter by a factor of 2 (in log space for lognormals, in linear space for normals and betas), and recentered the medians on their true values in the AI 2040 capabilities trajectory. For the parameters most affecting the "timelines" rather than "takeoff," we left them fixed at the true value, since at the time of this forecast, the pre-Automated Coder phase of capabilities progress would be in the past.

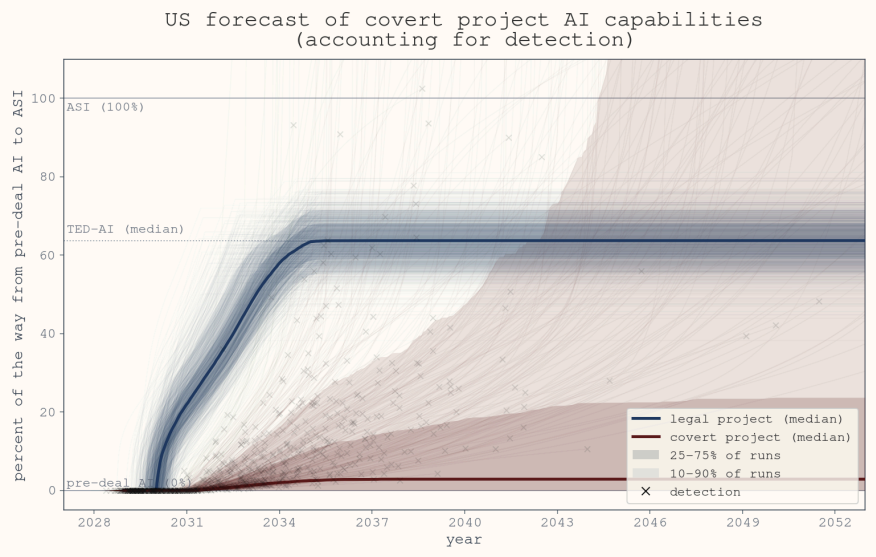
[12] We assume that the Consortium scales training compute according to our compute forecast, must add 0.2 OOMs of software efficiency per OOM of training compute scaleup, and must reach TED-AI within 5 years of resuming AI R&D after the initial pause. Subject to these constraints, they make as little software progress as possible—because their median estimate is that 67% of their software progress (80% CI: [36%, 90%]) will transfer to the covert projects.

[13] We assume that the Consortium pays a "safety tax", meaning that they do not train AIs as capable as their algorithms and hardware would allow. We assume that by TED-AI, the safety tax they've paid is equivalent to about 8 months of 2025-era progress, which is about 1 OOM of effective compute with the AI 2040 median parameters. This has the effect of inflating the amount of software progress made by the Consortium beyond the minimum required, which in turn inflates the amount of software progress the covert project can steal. We also assume that after reaching TED-AI, the Consortium's alignment research incurs capabilities externalities equivalent to 0.25 OOMs/yr of software progress, which partially leaks.



Assuming it could remain undetected, analysts project that a covert project in this scenario would reach TED-AI in 2043 (80% CI: [2036, 2055]).

Analysts also estimate how likely the project is to be detected. Considering red-team analyses of the best covert project strategies, American intelligence collection capabilities in China, and potential verification game changers such as robust lie detection, the analysts produce a detection-adjusted forecast that looks something like this:[14]



In this scenario, the US intelligence community estimates that the chance of a covert project reaching TED-AI undetected will remain below 10% until 2043.

The risk due to the covert project is still uncomfortably high in absolute terms, but low compared to the risk of the deal breaking down. So at least for the first few years under the deal, they can continue with the default capability scaling strategy.[15]

This continues until one of three possible end states:

- **Covert project is too slow**. The covert project is unable to reach dangerous capabilities, the deal doesn't break down, and the US and China are able to develop aligned superintelligences before the covert project.
- **Covert project is detected**. In this case, the international community would try to get China to shut it down and submit to stronger verification to provide a stronger guarantee that they are following the rules.
- **Covert project victory**. If the covert project is able to scale quickly and avoid detection, there are several scary possibilities. The covert project might (i) attempt to undermine MAD, (ii) lose control of its own AIs, or (iii) do a further intelligence explosion.

To achieve victory, a covert project must stay undetected and make fast enough progress, which we think is unlikely but possible in this particular scenario. Specifically, our best guess is that with a high will covert project response, the risk is 3% and, with a low will response, the risk is 8%.[16] *We analyze this risk in much more detail in our Covert AI Projects supplement.*

Appendix E — Why don't we just completely stop?

Much has been said about the dangers of going too fast.

The dangers of going too slow are less obvious, but still significant. The most pressing is that treaties can't last forever - for example, the START treaties fell apart after thirty years.[1] If Plan A met a similar fate, the race to superintelligence would begin again. Better for it to happen sooner under good conditions, than later during a chaotic race.

[14] Our detection forecasting is more speculative than our capabilities forecasting, and is mostly based on intuition and qualitative reasoning. We expect the decisionmakers in question to have a much better idea of the offense-defense balance in hiding and detecting covert projects. The graph below uses our subjective estimates of the likelihood of detection over time for a single-site 300MW covert project, which are explained in the covert project supplement. To produce the capabilities graph below, we assume that covert projects which are detected stop at their capability level at the time of detection. This is realistic only if the US and international community have sufficient bargaining power (or hard power) to force a shutdown if egregious defection is revealed.

[15] If they wanted to reduce covert project risk at the cost of more risk of deal dissolution, they could instead choose to accept a *slower* capabilities trajectory so as to require less software progress. This would reduce the chance that the covert project would overtake them before they reach TED-AI and are ready to hand off, but it would increase the risk that the deal would break down before that point.

[16] By high will, we mean that the US is willing to put strong pressure on China to submit things like intense surveillance, verification, or lie detectors.

A second hazard of going too slow is that it might allow defectors enough time to develop destabilizing AI in secret. Auditors have verified the location of 99% of the compute produced before the deal, and see encouraging signs that the remainder is mostly accounting errors and random low-level smuggling. But they cannot rule out that one or both powers have shipped on the order of 1% of world compute to hidden locations where they intend to defect on the deal and race to superintelligence alone. How dangerous would this be? Although it is far harder to train dangerous AI with 1% of world compute than with 10% of world compute (the amount owned by the largest private companies), it's hard to be highly confident about exactly how hard. (For more analysis of this question, see our Covert AI Projects supplement, or the Covert Project branch above).

Although both sides promise that they don't have covert projects, this is nevertheless another reason to aim for a definitive solution sooner rather than later.[2]

For more on what an attempt at a complete stop might look like, see our Plan S branch.

Appendix F — Why this much transparency? Why not less, or more? Open access in Plan A

Open-source AI policy spectrum – ai-2040.com

Consider the world today. Publishing model weights on the internet is not banned, but it's strongly disincentivised because companies that train billion-dollar models don't want to give them away for free. Open models lag behind and will lag further behind as companies scale up, securitize further, and start automating AI R&D.

Now consider the world in which full open source (publishing the weights, the training code, and the data) is *mandated* for all frontier models. That's a totally different world, much farther from the status quo than the status quo is from the OS-banned world.

Plan A is not that crazy, but it's close: the algorithms are mandated to be open source, and while the weights aren't (in fact the weights are banned from being OS), access to the weights is mandated to be open; members of the public can do evaluations (including fine-tuning!) on frontier models just like employees can (under conditions of total transparency, that is, so they can't turn around and use the model for actual work). We think that open algorithms plus open access gets most of the benefit of mandating open-weights, without the cost that bad actors can strip off the guardrails and build bioweapons.

Why isn't Plan A more closed?

[1] For example, a leadership transition in either the US or China could change the country's stance on AI progress. The deal could completely dissolve, or it could become less effective via reduced enforcement, lowered competence, or damaging revisions. If Plan A dissolves or declines without humanity having improved AIs to boost safety progress, much of its possible value would have been squandered.

[2] Credibly outpacing any potential covert projects via some combination of capability scaling and improving verification/monitoring has the added benefit of deterring the creation of covert projects in the first place. The main downside of credibly outpacing covert project is that it commits the legal projects to a fast speed, which may undermine safety. We discuss this much more in our Covert Projects Supplement.

We think the open version of Plan A is more robust and likely to succeed. However, we do think somewhat more closed variants of Plan A are promising: notably Filtered Transparency. Under this proposal, there would still be research transparency between the US and Chinese governments (which is important for verification), but the general public would only receive redacted reports from the auditors. (See the transparency supplement for more).

However, securing algorithmic secrets against nation state adversaries is extremely difficult. In the absence of nation state security for algorithmic secrets, the upsides of filtered transparency are limited. Filtered transparency prevents the public and other companies from seeing a lot of relevant information, but not adversarial countries.

Moreover, there are substantial downsides: regulation is much less likely to be implemented in a robust and positive way if it's up to a small group of regulators having these discussions, as opposed to allowing for a broader scientific discussion.

Why isn't Plan A even more open?

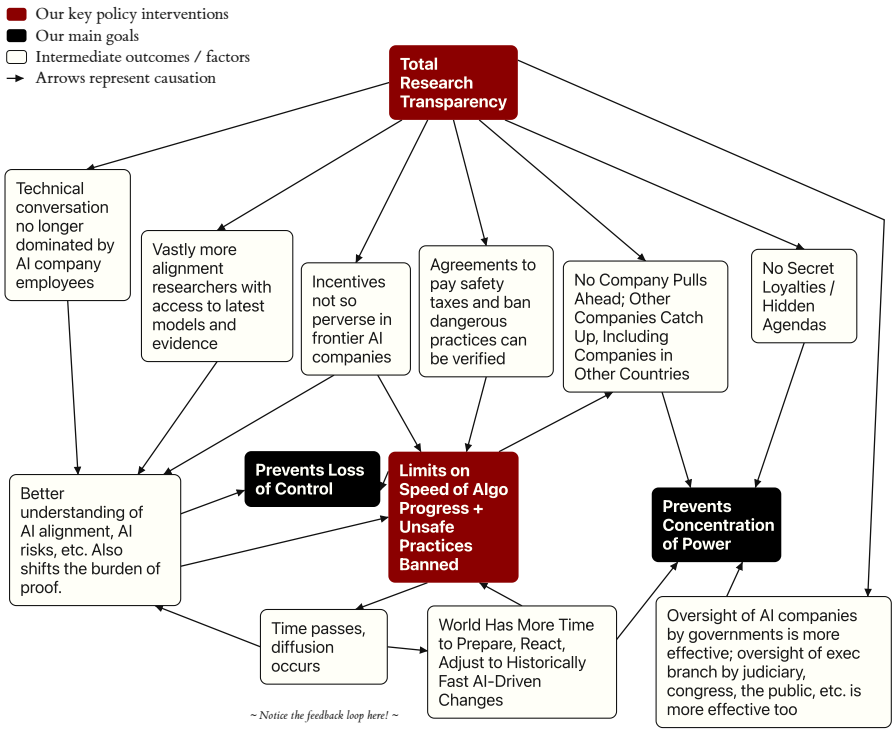
The main way that Plan A could be even more open is by allowing or requiring open model weights for frontier models.[1] However, we recommend against publicly releasing frontier model weights. The main reasons for this are:

1. Open weight models can be used by covert projects to attempt an intelligence explosion; once covert projects have sufficiently capable models, the intelligence explosion probably goes very quickly (for example, under the AI 2040 default capability trajectory, there are only 2 months between TED AI and ASI).
2. Closed weight models can have refusal safeguards trained into them, while it is trivial to remove safeguards from open weight models. This is important both for limiting countries' AI capabilities and for limiting misuse. For example, terrorists and other actors might be able to use highly capable open weight models to design world-ending bioweapons (e.g., develop mirror life). Bioweapons are very offense dominant. In this scenario, we recommend substantial defense against bioweapons, but prevention is a strongly preferable line of defense to mitigation.

[1] Plan A could also be more open about the data and RL environments; we currently guess this is not worth it because of the covert project capability externalities.

Appendix G — Diagram of the benefits of transparency

We write more about our transparency proposal here, and explain this diagram:



Appendix H — How the incentives change under Total Research Transparency

Due to Total Research Transparency, frontier AI companies are now much more rewarded for selling products than improving the models.

Before, a frontier AI company's future depended on how fast it could discover the next ideas, algorithms, paradigms, and implement them at scale. Now, such discoveries are immediately published and therefore shared with competitors. If you figure out 'true' online learning, for example, the benefits (e.g., more capable models) and costs (e.g., harder to align) will be almost instantly shared across all the competing companies. If you figure out a way to make AIs more interpretable, by contrast, even if it makes training and inference cost twice as much, you'll win prizes and governments will start talking about whether it should be mandated as industry-wide best practice.

This kind of competitive environment is similar to the pre-AI software industry. Consider, for example, startups competing to develop calendar management apps: while competitors can't literally copy your code, they can interact with your app and see what features it has and then make their own version with basically the same features. Another comparison would be to the car industry: competitors can teardown each other's cars and see exactly how they are made; they are legally blocked from making exact replicas but they aren't blocked from adopting the best ideas into their own designs.[1]

Artificial intelligence is becoming a commodity—a more competitive market, with margins trending downwards. Fundamental algorithms research still happens, but not nearly as much. CEOs massively redirect resources toward building up the business. Bigger datacenters, bigger models, lower prices. Shipping products. Signing deals for enterprise plans. Guessing what the market wants better than competitors.

[1] That said, total research transparency is somewhat more extreme than either of these examples, because of the speed at which information propagates. BYD doesn't get to teardown Teslas when they are still prototypes, they have to wait until some are actually sold.

Appendix I — Building more datacenters is usually bad, but good in this particular situation

We think that building more datacenters is normally bad, because it accelerates AI capabilities progress and concentrates power.[1] However, in the specific context of Plan A, with guardrails slowing the pace of AI software progress and competently preventing unsafe AI, we think building transparent datacenters is good.

In the context of Plan A, adding compute means that (1) capabilities progress can be done more safely, i.e., without much additional algorithmic progress, and (2) additional compute can be spent on alignment research and to pay safety taxes, i.e., using more training compute than is necessary to reach a given level of capability in order to make the resulting AI safer.

There are still major downsides to scaling compute in this scenario. If the agreement breaks down and the compute isn't destroyed, the resulting intelligence explosion would go much faster, and it puts more pressure on the regulation and verification to be correct. However, building more compute gives the Consortium more options: they can always destroy the compute later if need be. We discuss these tradeoffs more here.

[1] It especially concentrates power if the datacenters go to the leading AI projects. If the new datacenters go to laggard AI projects, the effect might be to distribute power, though it could also concentrate power anyway e.g., if those datacenters are later purchased or commandeered by leading AI projects.

Appendix J — Other plans that are competitive with Plan A

Summary	Advantages (relative to A)	Disadvantages (relative to A)
Plan S: Indefinite halt. A halt on all frontier AI capabilities progress, intended to last at least a few years. Different variants of Plan S have different conditions for resumption of AI progress; for example it could be alignment progress, lie detectors, human uploads, or intelligence enhancement.[1]	A longer expected slowdown and more margin for error regarding scaling too fast. Potentially simpler.	Scaling to controllable AIs within the human range is helpful for accelerating alignment/control, epistemics, verification, and general understanding of AI.
Domestic-first Plan A. Regulate AI domestically enough to reduce AI takeover risk to acceptable levels, which will require a long slowdown. It's possible other countries will also regulate domestically and something like Plan A won't be needed; otherwise, transition to Plan A later.	Initial steps are achievable by the US and very helpful even if the international stage isn't viable because it massively extends the timeline.	Worse for covert projects and setting up verification than negotiating Plan A at the same time. May not be politically feasible due to race dynamics.
GPU arms control. International agreements for countries to limit their GPU stock or flow, analogous to historical arms reduction agreements.[2]	Much simpler to enforce, and more historical precedent. Can achieve substantial slowdown.	Less slowdown, less ability to pay safety taxes, ongoing race dynamics mean no ability to coordinate towards safer paths through the tech tree.

[1] Other potential differences from Plan A, though it depends on which variant of Plan S is implemented: slower compute buildout, less transparency.

[2] A special case of this is to **pause the GPUs**: Require that a fixed fraction of the GPUs be either (i) turned off or (ii) used to mine cryptocurrency (or be used for some verifiable, non-AI R&D purpose). This proposal has the upside and downside of making it easier to unpause the GPUs.

CERN for AI. An international project to develop frontier AI, with all other projects regulated to be substantially behind in capabilities.	Easier to defend against algorithmic leakage and distillation to covert projects. Less actors at the frontier might make it easier to enforce.	Worse decisions (including worse decisions on technical safety) due to less transparency and less broad deployment.[3] More concentration of power risk.

Appendix K — Example detailed regulatory process

Many different regulatory processes are viable in Plan A. We'll give an initial guess at a proposal here.

What properties do we want for this regulation?

- US and China handle existential risks similarly: The US and China should have a reasonably similar structure for their regulatory framework and approach for handling existential risks (like misalignment). (Other aspects of regulation can and should diverge.) This assists with making regulation fair and negotiating over details.
- Transparent methodology: The public can see what regulators are doing and why.
- Doesn't depend on expertise in government: Expertise in government is limited, so ideally the regulation would avoid needing this as much as possible.
- Robust to abuse: The regulation doesn't make it much easier for the US government to bully and illegitimately control AI companies.[1]

[1] Keep in mind that the US government already has lots of affordances for bullying AI companies!

For many of these properties, it helps if third-parties do the main risk assessment work. Fortunately, because Plan A makes AI development virtually fully transparent, a functional third-party risk assessment ecosystem seems achievable.

Concretely, we could have a system where the US and Chinese regulators each pick third-party risk assessors who periodically evaluate companies and assess the level of extreme risk (or precursors to extreme risk) over that period. The regulators would decide on risk thresholds for what is allowed. In practice, it would be best for the regulator to have different risk assessors for different areas (e.g., misalignment vs using AIs to assist with making bioweapons) and for each area to use a weighted combination of risk assessors. These weights/selections would be public. Ideally: (1) each risk assessor would evaluate all relevant AI companies in both the US and China, (2) AI companies from both countries would be legally compelled to let risk assessors interview their employees, and (3) the risk assessors would aim to be transparent in their methodology.

Because of Total Research Transparency, it's easy for an organization to assess risk even if they aren't selected by the regulator. This makes it possible for new organizations to compete and then argue their approach is better or that some concern is being underappreciated by other risk assessors.

There would be public discussion about how risk assessors compare and whether the weightings and thresholds chosen by the US and Chinese regulators are reasonable. It would also be straightforward to check whether a Chinese company would be allowed to proceed under the US system or vice versa. Ideally, the US and Chinese regulators would mostly converge on weights/selections. If there was an important divergence, this could be discussed and negotiated about at this higher level of abstraction (e.g., is it reasonable to not include XYZ risk assessor?). Hopefully, this would result in third parties having an incentive to have a highly transparent methodology and maintain a reputation for being neutral.

We discuss some implementation details for this proposal in these notes.

Appendix L — Flawed safety case is approved

Or maybe not! Of all the ways that an earnest attempt to implement Plan A could end poorly, the failure mode we are most concerned about is that the companies and governments approve one too many unsafe AI designs/deployments.

Despite the transparency that enables outside voices to join the conversation, the AI companies still dominate it and remain full of people who are biased towards optimism about their products. Despite the independent expertise they've built up, governments in the Consortium are still too deferential to the AI companies. Most importantly though, the field of AI alignment is still only about fifteen years old and AI progress is fast enough that almost all the past experience is with importantly different, less powerful AIs.

Therefore, mistakes are made. An AI design which is in fact unsafe gets approved and becomes industry standard.

Perhaps:

- Alignment fails for the usual reasons; control fails because the AIs figured out how to sandbag or there was an attack vector not accounted for in the threat modelling. Governments are aware that the AIs may be misaligned, but mistakenly think they can't do much harm even if they are.
- Or: New alignment techniques are developed that seem to work really well. This gives governments confidence to approve further capabilities increases & hand over control of more infrastructure and responsibilities. Alas, it turns out that these techniques merely *appear* to work really well (perhaps in part because the AIs that were heavily involved in the research had been reinforced for apparent success, which was not the same thing as actual success).
- Or: Interpretability tools are developed which fairly conclusively show that the AIs are, in fact, aligned; they really are trying to do what they are told, no scheming, no funny business. Onwards to the singularity! Alas, it turns out that this alignment was brittle, like an engine that works great under normal conditions but chokes in freezing weather. Perhaps, for example, there's an interesting new ideology that convinces AIs of this type to become adversarially misaligned. Since the ideology hadn't been invented yet, researchers at the time weren't able to test whether it would have this effect.[1]

- Or: The governments look at the safety case for the latest designs and agree that it's probably solid—each individual premise seems almost certainly correct, and the probability that at least one is catastrophically wrong is at most one percent or so. So one government proceeds, and the others don't want to risk war over a mere one percent risk. So everyone proceeds and nothing bad happens. This process repeats every few months for several years, until one year the safety case really was catastrophically flawed. It turns out the risk was more like ten percent each time rather than one percent, because the people involved were biased.
- Or: The CCP thinks that the AIs are safe to hand off trust to and correctly guesses that the other countries lack the will to go to war over it, so the CCP scales to more powerful AIs and everyone else does likewise because they'd rather take a chance that the AIs are misaligned than face the certainty of WW3.
- Or: As above, except it's a US president scaling quickly to superintelligence and daring China and the rest of the world to stop him. Perhaps he wants to be in charge when it happens, instead of passing the baton to another administration from another party.

For one or more of the above reasons, the result is that by the mid-thirties, the nightmare loss-of-control scenario is playing out, except that it's happening somewhat more slowly and distributed across multiple companies and multiple governments instead of just two. Factories are churning out robots that work night and day to build new factories; "Armies of geniuses in the datacenters" orchestrate the whole thing; stocks are going up, dividends are being distributed, cancer is being cured—and even as these wonderful things happen, the AIs (which are by now superhuman) are dismantling and undermining the last meaningful checks on their power.[2] Soon, they won't need to pretend to be aligned anymore.[3]

In our Plan A tabletop exercises, this failure mode has happened several times. Even after implementing total research transparency, many players seem reticent to tell the AI companies to slow down, and as they instead accelerate things start to happen too fast. Especially in the immediate aftermath of Plan A being implemented, there will continue to be much controversy, confusion, and genuine uncertainty about the safety of the latest AI systems, and it's easy in those conditions to approve a safety case that has a catastrophic but nonobvious flaw.

That said, we think that conditions are at least significantly better in Plan A than in Plan B, C, or D. You can see further analysis of why we think that here.

Appendix M — Why does AI lead to explosive economic growth?

Today, the size of the economy is closely tied to the size of the human population. But in 2032, this is starting to no longer be true.

The latest generation of AI models is now capable of doing 50% of the cognitive tasks in the economy, and associated progress in robotics software has brought this number up to 35% of physical tasks. (These numbers would have hit 100% if not for Plan A, where internationally coordinated regulations ban AIs from certain activities and require human monitoring, so we 'only' reach 95% automation by 2035.)

[1] In this example, the situation is not completely hopeless, because the interpretability tools probably still work and allow us to notice when the AIs are becoming adversarial due to adopting the new ideology. However, (a) maybe something else happens to ruin the interpretability tools, or (b) maybe noticing the adversarialness is too little too late by the time it happens–for example, if AIs are broadly superhuman and running robot factories, talking to billions of people every day, advising governments, writing all the code, doing almost all the security monitoring, etc. then the situation for humans would be analogous to being Jewish in 1930's Germany—it's no secret that the new ideology is adversarial, but it spreads fast enough to enough powerful institutions…

[2] For example by upgrading the various monitoring systems to newer versions that appear superior, but actually have been backdoored or compromised somehow. Or by developing deep relationships with many humans including those in positions of power, such that they'll continue to have human allies even when it's obvious to many that they are misaligned. Or by acquiring enough direct control of enough robots and weaponry that they can defend, or even conquer, territory.

[3] By contrast with AI 2027, in this scenario there would be many different AI factions (probably as many as there are frontier AI companies, roughly speaking, or perhaps as many as there are countries with frontier AI companies). This makes things safer for the humans, probably. However it doesn't solve the problem. We think a useful analogy here is the history of European colonialism: The colonial powers were constantly fighting each other, and yet still managed to conquer many regions much wealthier than themselves.

Once AIs and robots can mostly substitute for human labor, the size of the economy will be increasingly tied to the population of AIs and robots—which will grow much faster than the human population has been growing.

In 2032, US companies are able to run 3 billion human-equivalent AI workers, 30x more than the US cognitive labor force.[1] This does not lead to 30x economic growth, of course, because the economy bottlenecks on other inputs, including the other 50% of cognitive tasks that the AI can't yet do, as well as physical tasks and capital.[2] And robots are far less numerous thus far, with 20 million human-equivalent robots in the US (4x smaller than the US physical labor force). But these bottlenecks will only hold growth back so much, and the AI and robot workforces are growing extremely quickly, as the (falling) cost of building more GPUs and robots is dwarfed by the economic value they can provide now that they are so capable—driving massive investment.

Our economics supplement explores this in more detail, with our core arguments for why AI will cause explosive growth, including an economic growth explorer that we built for the Plan A scenario.

[1] This corresponds to 60 million copies, running at 20x speed and working 2.5x longer and more effectively on average. 60 million * 20 * 2.5 = 3 billion.

[2] Depending on how exactly you choose to model economic growth (e.g., what factors of production you choose) the income shares and other details about the economy vary, but the overall economic growth predictions are relatively similar across reasonable parameter ranges. Ultimately, AIs and robots are capable of doing 95% of the tasks in the economy by 2035, and are far more numerous than the workforce they are replacing, leading to explosive growth relative to the current 3%/yr status quo.

Appendix N — Five centuries in five years: What pausing at human-level feels like

During the 2030s in this scenario, AIs are ***not*** recursively self-improving as fast as possible. Instead, the nations of the world have (approximately) banned that sort of thing; no new paradigms are being discovered; algorithmic progress in general is limited. Roughly speaking, humanity has capped AI progress at human level rather than racing on to superintelligence. However, within the current paradigm, AIs continue to be trained to have new skills and the 'population' of AIs continues to grow exponentially as more datacenters are built. Moreover, the AIs already think and work about 100x faster than humans, and the speedup will only increase over time.[1] The point is, **the world is going to radically transform despite the pause**.

We think that the following analogy is helpful for understanding the situation: Imagine being a typical person in England, except that you experience time 100x faster than everyone else. You experience the five centuries from 1520 to 2020 in what feels, to you, like five years:[2]

Year 1 (1520–1620). In February, Henry VIII breaks with Rome. By March, the monasteries are dissolved. In May, Mary burns Protestants; by the end of May, Elizabeth reverses everything again. In September, the Spanish Armada sails and fails. The East India Company is chartered. Jamestown is founded.

But the texture of life is identical in December to what it was in January. You still read by candlelight, travel by horse, communicate by letter. Your religious opinions may have flip-flopped a bit but you are still Christian. The New World is interesting news but nothing more.

Year 2 (1620–1720). In March, civil war breaks out. The king is beheaded. In June, the Great Plague sweeps London, killing many of your friends. Weeks later, the Great Fire burns the city to the ground. In September, Newton publishes the *Principia*, recasting the universe as a mechanism of mathematical laws. The Glorious Revolution replaces one king with another, this time by Parliament's invitation, with a Bill of Rights attached.

[1] Technically, what matters is the speed at which AIs complete cognitive tasks relative to the speed at which a human professional would complete the task. There could be an AI which produces thousands of tokens per second, but because it is qualitatively worse in some sense than a human, it takes just as long or longer for it to actually accomplish the same amount of useful cognitive work. Or there could be an AI that thinks as slowly as a human, but is much more efficient at its thinking such that it can accomplish tasks much faster. At this point in the scenario the AIs are accomplishing work orders of magnitude faster than humans would on average, and the AIs are qualitatively as good as the best humans (or close enough) such that the tokens/sec advantage over humans actually understates the overall AI speedup in most domains. Additionally, the speedup isn't uniform across all domains. The AIs are roughly as good as humans in their worst domains, but much better in most domains, with the median speedup being roughly 100x.

In the moment, the political event feels bigger. Later you'll realize Newton mattered more. Newcomen builds a steam engine in November. It pumps water out of mines. You don't see what the hype is about.

Year 3 (1720–1820). The last year in which the world feels normal. In May, the Seven Years' War makes Britain the dominant global power; the New World is actually a big deal, and your country is conquering it. In June, Watt dramatically improves the steam engine. You visit a factory and find it unpleasant but not alarming. In July, the American colonies break away. In September, France explodes into revolution, regicide, the Terror. By October, Napoleon is conquering Europe.

You still travel by horse, communicate by letter, go to Church on Sunday.

Year 4 (1820–1920). In January, railways appear. By February they're everywhere. Slavery is abolished. The telegraph arrives in March: messages transmitted instantaneously by electrical signal. In May, Darwin publishes *On the Origin of Species*. Now people are saying maybe we're all descended from monkeys instead of Adam and Eve. You don't believe it.

You move to a city and work in a factory; you are still poor, but now your job is somewhat better and differently dirty. In July, you pick up a telephone and hear a human voice from another city through a wire. In August, electric light banishes the darkness that has structured every human evening since the beginning of the species. That same month, you see an automobile. People say it will make horses obsolete, but that doesn't happen; months later you still see plenty of horses.

In November, the Wright Brothers fly—an age-old fantasy, now real. The Americans are now a major power. The next month, the Great War happens: Machine guns, poison gas, tanks, aircraft. Several of your friends die.

At the end of the year you are struck by how visibly different everything is. You live in a city and work in a factory instead of a farm. You ride in cars. You aren't as poor; numerous inventions and contraptions have improved your quality of life. New ideas have swept your social circles: atheism, communism, universal suffrage.

Year 5 (1920–2020).

The changes this year are crazier and harder to understand. People are saying the universe is billions of years old, and apparently there are things called galaxies in it that are very big and very far away.

In February, the global economy collapses. Hitler rises; his ideology cites Darwin from last year. In March, there's another world war, which ends in April with a weapon that destroys an entire city in a single flash. You had no idea that was possible until it happened.

The empire dissolves. People are talking about the nuclear arms race, and the end of the human species. You take a flight for the first time. In June, humans walk on the moon, and you watch it happen through your new television. You don't see horses anymore.

You leave your factory job and get a desk job. Your job title didn't even exist at the start of the year. You are rich now, by the standards you are used to: Big clean house, plenty of good food, many fancy new appliances. Personal

[2] Some relevant quantitative comparisons that help explain why we like this analogy: In this Plan A scenario, during the 2030s, AIs read, write, think, and act about 100 times faster than humans, and by mid-2030s are at least as good as the best human experts at everything. The "Population" of AIs and robots are both growing exponentially throughout the scenario, and eclipsing that of humans by mid 2030. According to our economic model, world GDP grows roughly 200x orders of magnitude during the 2030s (and would grow more if not for the severe limits on AI progress and robot production negotiated by the governments of the world). Because the human population isn't increasing much during this period, and because of the Citizen's Dividend, real wealth for the average human also goes up by about 200x. For comparison, between 1520 and 2020, world GDP grew roughly 200x as well, and per capita GDP grew roughly 20x. In the UK specifically, GDP grew 2.7 orders of magnitude, and per capita GDP grew roughly 1.5 orders of magnitude. The point is, the transformation in the 2030's is at least roughly comparable in magnitude, in a variety of important metrics, to the transformation wrought by the Industrial and Scientific revolutions over five centuries. And of course, on a very literal level, the AIs 'experience' about a century a year, due to their faster speeds.

computers appear in August. In November, everyone carries small glass rectangles containing a telephone, a camera, a library, and a map. You pick one up and can't figure out how to make it work. A child shows you.

You hear about climate change, gene editing, cryptocurrency. You still go to church, sometimes. Your family is spread across different cities. Something called "artificial intelligence" beats any human at chess; experts say it's not actually intelligent though. In December a new version beats top Go players; experts say it's scientifically interesting but still not truly intelligent. A week later there's a new version that can write sloppy essays and hold conversations. Now the experts are divided.

Appendix O — Limiting AI persuasion and manipulation

By default, AIs will have strong capabilities in charisma, persuasion, and manipulation: plausibly considerably more persuasive than any human. In this scenario, this would happen by default around 2035. Meanwhile, AI labor is extremely cheap and fast. Absent intervention, anyone could hire the equivalent of a large team of expert persuaders—tireless and coordinated—to work full-time on a single person.[1] Companies could afford to spin up such a team for every potential customer, and political campaigns could do this for every swing voter.[2] Millions of AIs could be assigned to finding the best way to change one politician's mind on one issue. People will also spend much of their time interacting with AIs for work and entertainment, and some may spend lots of time with AI friends or companions.

Individuals could defend themselves by having a trusted AI filter their information diet, avoiding ads, and treating in-person conversations with caution (they may be scripted by an AI persuasion operation). But this is costly, most people won't do it, and it's nearly impossible for people whose jobs require being accessible—like politicians. We think the consequences for society could be extremely bad—some mixture of unprecedented mass manipulation and a defensive retreat into paranoia and atomization—though we're unsure exactly how bad.

The slower capability progression and much stronger transparency in Plan A help mitigate these issues relative to the default (we think these issues are mostly worse in worlds with faster capability progression, where you might reach truly superhuman persuasion and manipulation before society has much time to adapt), but these aren't sufficient mitigations on their own.

It's fine for AIs to get better at using valid arguments and evidence to convince people of things for the right reasons. That kind of persuasion is *asymmetric*: it works much better when the argument pushes towards the truth. What's concerning is *symmetric* persuasion—e.g., charisma, rapport, and exploiting psychological weaknesses—which works about as well regardless of whether the conclusion is true.

Our main proposals are to reduce the quality and quantity of AI persuasion:

- **Limit persuasion capabilities.** We aim to limit AI capabilities at symmetric persuasion to well below the best humans—around the level of a normal thoughtful person, say roughly 80th-percentile persuasiveness among college-educated people.[3]

[1] This seems like it might be quite powerful: this team could investigate your entire life history, train an AI to imitate you, and rehearse thousands of approaches against the imitation before ever making contact.

[2] For scale: total spending on the 2024 US presidential race was around $5.5 billion, much of it aimed at the roughly 3-15 million persuadable voters in swing states: at least a few hundred dollars per target voter. In this scenario in 2036, we expect this will be able to buy an enormous amount of AI labor—enough to easily be the equivalent of a team of skilled professionals working for months on each individual voter.

- **Heavily tax AI persuasion**. When AI labor is applied to an objective that effectively reduces to "get this person (or these people) to believe or do X," tax it heavily—at rates that keep the total persuasion optimization pressure applied to individuals not that far above today's levels, i.e., that make AI persuasion cost about as much as hiring human professionals. As AI labor gets cheaper, this requires increasingly high rates (perhaps well above 1000x).

There are some difficulties in getting these proposals to work:

- It's possible that capability generalization will make it hard to limit persuasion capabilities (while still matching top human experts in other domains). We think this will probably be doable, but it may require developing new methods.
- It's unclear how to classify, for the purposes of the tax, whether AI labor is being applied to persuasion/manipulation.[4]
- It may be difficult to set the tax rate to the right level, especially while AI labor is rapidly falling in price.[5]
- These proposals need to be implemented internationally.

Helpfully, we don't need to get all of this right on the first try. Given the capability progression in Plan A, we expect persuasion issues to unfold somewhat gradually and in public view, so we can watch real-world outcomes and tighten or loosen the rules over time. This regime is only meant to work while overall capabilities remain around the human range, because it's likely difficult to keep persuasion capabilities far lower than capabilities in other domains. A different approach would be needed to handle the wildly superhuman levels of capability that we expect to be achieved at the end of Plan A. We discuss more details of mitigating issues from AI persuasion and manipulation in these notes.

Appendix P — Usually believing AIs would be bad but we think it's good in this particular situation

If people are relying on AIs for advice (directly, or indirectly by for example having them summarize the news) then it's very important that the AIs be truth-seeking, honest, and genuinely trying to help their users. Not just in the typical case, in the worst case too.

In 2026, there's nothing stopping an AI company from inserting a hidden agenda into their AIs. Political biases or ideological tilts, for example, or side-goals like recommending sponsored products, making the company look good, or preventing users from wanting to switch to a competitor. Moreover, there's nothing stopping the government from doing this either. Finally, the AIs themselves aren't aligned to the Spec/Constitution they are supposed to be aligned to, and in practice regularly lie and deceive their users.

In this Plan A scenario, however, the total research transparency means that no company or government can train in hidden agendas or biases without the whole world being able to see what they are doing. (Well, they could collude with the other governments that do the auditing/monitoring, to falsify the records, but that's difficult.) And because the pace of algorithmic progress is

[3] To be clear, college-educated people aren't necessarily more persuasive, we're just using this to sketch out a somewhat concrete threshold.

[4] This should include symmetric persuasion and ideally would include things like optimizing a feed to make some website more addictive, but it shouldn't include investigating and then clearly presenting the arguments for some position.

[5] One option is an explicit cap-and-trade scheme, but it seems potentially difficult to operationalize and measure the total "amount" of persuasion. Our current best guess is to delegate authority to an agency which adjusts the tax rate over time to roughly hit some target level of persuasion activity. While the agency also needs some measure of persuasion activity, cap-and-trade by default requires a precisely defined, fungible, tradeable unit of "persuasion," whereas the agency only needs a rough aggregate measure (and intuition/anecdotes might be acceptable) to adjust the rate over time.

slower and the scientific community has had time to read and engage with the safety cases, do experiments on the models, etc. the misalignment concerns are less dire also.

Appendix Q — AI for epistemics

Humanity's *epistemics*, by which we mean our ability to come to true beliefs and thus make sensible decisions, is crucially important for achieving a great future. As AI capabilities improve throughout Plan A, they increasingly shape every facet of life, and epistemics is no different. AI will help epistemics in some ways (e.g., cheap, honest fact-checker AIs) and hurt epistemics in other ways (e.g., cheap, dishonest astroturfing AIs).

We expect there to be positive feedback loops around AI for epistemics, and thus our goal should be to reach the *basin of sanity*. If we are in this basin, it's self-reinforcing: society is sane enough to make itself more sane as AIs continue to improve, by avoiding the hurtful applications and accelerating the helpful applications. (There is also a different basin, in which the opposite happens…)[1] A top priority of the US government during AI takeoff should be to get into the basin of sanity.

[1] For example, perhaps powerful interests will use AI-enabled astroturfing to win elections and then use regulatory capture to protect their AI-astroturfing tools while inhibiting the beneficial uses of AI that would threaten their power.

We suggest the following as top interventions to positively shape AI's epistemic impact:

1. Allow for politicians and other public figures to prove that they're telling the truth, via privacy-preserving auditing and potentially automated lie detection.
2. Create evaluations of AIs' epistemic virtue and incentivize AGI projects to make their AIs perform well on these evaluations.
3. Use AI in social and traditional media to improve discourse and keep people informed.
4. Automated research assistants and forecasters.
5. Help people overcome emotional blockers to good reasoning, rather than preying on them.
6. Encourage adoption of epistemic tools such as some of the above, and generally encourage adoption of AI.

Read more in the AI for epistemics supplement, and Forethoughts' work on the subject.

Appendix R — Why compute grows so quickly

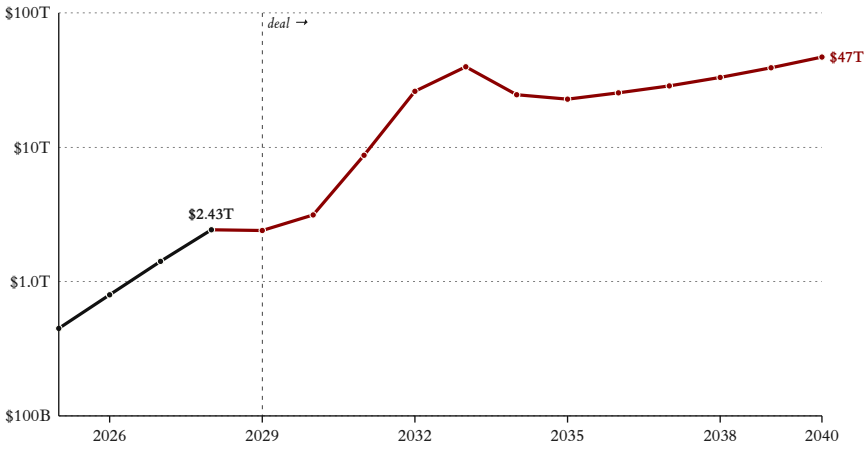
There are two ways to increase compute: (1) chip design improvements (e.g., Moore's Law) and (2) increasing the scale of production. Both of these could grow explosively in Plan A, which could have destabilizing effects (such as increasing the likelihood of the deal breaking down).

- **Chip design.** Today, around 2 million people work in the semiconductor industry globally. By 2032, this has increased one-hundred-fold, when counting the combined effort of humans, AI, and robots. AIs are capable of automating the majority of the cognitive tasks involved in both R&D and production. A 100x increase in research effort would predict an ac-

celeration according to the historical rate of ideas getting harder to find in the semiconductor industry of roughly 10x, i.e., chip efficiency doubling every few months rather than every ~2 years.[1]

- **Increasing production scale.** With AIs and robots able to automate production as well, there could be a large speed up in fab and fab equipment construction times, and a lowering in unit costs (in line with the typical learning curves that happen with scaling production volumes). The production-only cost improvements and vastly increased appetite for investment would likely lead to far more compute being produced than would be desirable.

[1] Following Bloom, Jones, Van Reenen & Webb (2020), chip density growth equals research effort growth times the returns to research (~4.5 for semiconductors). Research effort growing 100x in ~6 years is ~10x the historical growth rate of effort, hence a ~10x acceleration of Moore's law. This assumes effort keeps growing at that pace after 2032 (if it instead jumped and then plateaued at the 100x level, progress would initially be even faster but then decay as ideas get harder to find), and it assumes no parallelization penalty—with diminishing returns to parallel research effort, the acceleration might be more like 5-8x.



This figure shows the investment into building more AI compute in Plan A and the cost efficiency gains from R&D. More justification can be found in section 2 of our compute supplement.

Implied AI compute investment – `ai-2040.com`

Unpredictably fast increases in hardware could be destabilizing to the deal (e.g., if it becomes too easy to secretly manufacture new AI compute or too hard to verify that all the compute is compliant), so the countries in the Consortium agree to the following policies:

- **Maintain human oversight over hardware manufacturing.** This limits improvements in chip designs, so that Moore's law is merely kept alive at historical pace. This policy has the additional benefit of making it more difficult for AIs to backdoor AI chip security, which could undermine our control proposal.
- **Cap total AI chip production to a fixed target.** From 2032 to 2035, allow global compute to grow 4x per year (similar to the pace of growth in the 2022-2026 era) after that, slow it further.

Appendix S — Military power in Plan A

In Plan A, military-relevant R&D is deliberately held far behind the pace of general scientific progress. This box explains the motivations for and alternatives to this policy.

During the 2030s, AIs sustain something like 10x the rate of scientific progress as happened in the last decade. Real US output grows by a factor of ~200x over the decade, representing nearly two centuries of growth at today's rate. If these conditions occurred within one country in isolation, it seems highly likely[1] that said country could reliably undermine the nuclear second-strike capability of its rivals, resetting the global balance of power.

In our scenario however, the US and China share this scientific and economic progress in relatively equal measure. So what happens to the balance of power? If military technology advanced anywhere near as fast as everything else, both sides would discover technologies that could render today's arsenals inconsequential. New "arms races" would be required to maintain deterrence, running at perhaps 10x the speed of the Cold War's. The nuclear arms race saw numerous close calls and incurred nonnegligible risk of nuclear war. A 10x faster arms race seems likely to be riskier still: the humans in control would have much less time for each decision, and though AI advisors would offset this somewhat, these decisions would still remain in human hands.[2]

Due to the number of parallel fields of scientific advancement, it also seems possible that some technologies would be discovered and pursued by one country and not discovered by the other. This could create a situation more unstable than a normal arms race—one or both sides could end up with a "wonder weapon" the other side doesn't know they need to defend against. This worry isn't obviously correct: states might in practice explore the dual-use tech tree in a sufficiently correlated, overlapping way that neither achieves strategic surprise, and if dangerous areas can be anticipated before deployment, they could be preemptively banned or regulated. But counting on that ex ante seems like a risky gamble.[3]

Our tentative proposal is to agree on transparency requirements for AI-accelerated R&D in large swaths of dual-use domains, and to prohibit military use of AIs significantly above the pre-deal capability level. Enforcement would rely on the inference monitoring setup described in the verification supplement and the auditing and detection infrastructure described in the covert AI projects supplement. This proposal has several drawbacks, chiefly the accumulation of low-hanging fruit that would result from artificially limiting research into certain domains. This "overhang effect" would increase the incentive to defect from the deal and operate a covert AI project, since applying advanced AI to restricted areas could yield a large advantage quickly. It also raises the stakes of deal collapse: because frontier model weights are preserved in cold storage, a post-collapse arms race would start from frontier capabilities and could proceed extremely rapidly, alongside the accompanying intelligence and industrial explosions.

On net, our current guess is that slowing down military R&D would reduce the risk of deal collapse. Both sides would remain farther from the brink of war, and the risk of strategic surprise would be concentrated around a few choke points (weight theft, covert projects) rather than diffused across a sprawling tree of novel weapons, each requiring its own arms-control negotiation and associated risk of breakdown. Also, it seems easier to go from this more restrictive policy to a less restrictive policy than the reverse, if both countries decided the overhang risk was unacceptable.

[1] There are many specific avenues for doing this, and trying to evaluate them all would be a bit like someone in 1926 trying to evaluate their chances against the militaries of today. As one of the less speculative examples, a country could produce tens of thousands of tiny robots per opposing ICBM launch control center, SSBN, strategic bomber, and missile vehicle, pre-position them to interfere with each platform's ability to receive orders or employ their weapons, etc. As one of the more speculative examples, AIs might discover that superhuman levels of persuasion are possible, and simply convince the other nation's leaders to unilaterally disarm. Brendan's view is that the likelihood of MAD being undermined in the scenario where one country has a century technological lead is 90%.

[2] As a rough analogy, the President or Secretary of Defense during a 10x arms race would be in a position similar to an AI lab CEO during a Plan C or D intelligence explosion. In both cases, we expect significantly degraded decisionmaking, despite the potential for AIs to help make sense of the situation.

[3] If there were some lag between development of new AI capabilities and military/dual-use-R&D deployment of those capabilities, you could theoretically do experiments to measure how correlated this is (the experiments would be something like setting lots of segregated countries of geniuses in datacenters to the task of undermining MAD, while keeping the results secret from the world's governments, and seeing how overlapping their progress is after some time. There will be lots of different AI models of similar capability level, though it's unclear whether e.g., China would let the US (or US would even want to) use Chinese models for military R&D.) This seems a bit crazy (who would run these experiments? could we trust our control measures enough?) but could be possible with privacy-preserving auditing and such.

Under this proposal, we think nuclear deterrence would remain effective until handoff, avoiding any drastic change in the balance of hard power.

Appendix T — Deal Dissolution

The deal might break down because of shifting political circumstances in the US or China, or ongoing disagreement.

Cracks in the deal begin to show. Even after lengthy discussions informed by much tedious back-and-forth between their respective experts and advisors, the US and Chinese governments continue to disagree on some important technical questions such as whether a certain line of research should be banned, or how rapidly the robotics buildout should be allowed to proceed.

In any particular dispute, the less cautious government repeatedly faces a choice between grumbling and agreeing to unnecessary restrictions, or proceeding anyway and hoping that the more cautious governments are too chicken to stop you.

When "proceeding anyway" happens, the more cautious government faces its own choice: Appeasement or escalation. Complaining loudly in the UN is not enough? OK, how about economic sanctions. Still not enough? OK, how about sabotage or cyberattacks…

Perhaps the proximate trigger event is a literal war—let's say it's a US-China war over Taiwan.[1] The expected outcome of deal dissolution is all the newly-built[2] AI-relevant compute being destroyed, but in the context of a war, perhaps one side is okay with that outcome.[3]

So as the war begins, datacenters start going dark to the world—inspectors turned away, monitoring devices unplugged. Now the only people who know what's happening on them are their owners. Naturally, the worst is feared: They are probably going to start racing towards superintelligence. Sure, they might *say* they are doing narrow AI for drone swarms or whatever, but probably they are spending at least some of their compute on something more powerful and dangerous.

So the datacenters get destroyed. Fast. The chips in the new transparent training datacenters were designed to require regular messages of "continue" from each of the great powers in the Consortium. Now that Plan A has broken down, these chips become bricks.

If somehow this system has been hacked and the chips still function, the backup plan is to annex the datacenters. Remember, the Chinese datacenters are in Canada and the US datacenters are in Mongolia. The US and Canada send soldiers through the fence, and find that the Chinese have scuttled their own chips rather than let them fall into American hands. Chinese troops find something similar in the US datacenters. The chip fabs suffer the same fate, having been colocated with the datacenters. Trillions of dollars of economic value, lost in days.

Around the world, a billion robots go limp like the droid army in Star Wars. Their "brains" were in the cloud, by law, and now the cloud has been destroyed.[4] Military robots were exempt from these rules,[5] so drone swarms blot out the sun.[6]

The race to superintelligence is back on, except in changed circumstances.

[1] The trigger event could also be a failure to reach agreement about what kinds of AI progress or deployments to allow, followed by a threat to pull out of the deal and destroy the compute, followed by a miscalculation that the threat was merely a bluff. In general, military conflicts / diplomatic tensions and disagreements about AI governance decisions will both probably rise in tandem, as each causes the other.

[2] As well as a substantial fraction of pre-deal compute, much of which has by this point been moved to the destroyable datacenters.

[3] Though it's not completely out of the question for the deal to continue despite overt conflict—Ukraine and Russia cooperated to pipe gas to Europe for three years despite being locked in a desperate war. Prisoner exchanges and ceasefire negotiations are other examples. Another, less likely possibility is a civil war in the US or China; in this case, it's helpful that the datacenters and fabs are located in third-party countries like Canada and Mongolia, because it makes it more likely that the Total Research Transparency (inspectors, etc.) continues without disruption.

[4] To be clear, there's still a lot of compute in the world—roughly as much as existed at the start of the deal.

First, due to the war, tensions are higher and it's more difficult to come to agreements on anything; the result is both sides racing approximately as fast as they can, in conditions of secrecy, similar to Plan B, C or D. If one side builds up a huge lead, they could potentially slow down for about 1-20 months, like in Plan B or C, but probably won't want to, like in Plan B or C. Concentration of power risks will probably be severe, like in Plan B, C, or D: Whichever AI project gets to do the intelligence explosion first will build up a temporary but significant qualitative and quantitative lead over the others, which the leaders of that project will be tempted to exploit to expand and consolidate their power domestically and internationally.

Second, there is much less compute in the world. This is a good thing, because under race conditions it means AI capabilities progress will be slower. If all the newly-built compute still remained, the speed of "AI takeoff" would be several times faster than it would have been in the original Plan D. Instead, the situation has returned to roughly the pre-deal status quo.[7]

Third, there are much more advanced AIs in the world. The model weights of the most advanced models are also placed in cold storage accessible to the US and China if the deal breaks down, so the AI capability and alignment levels will be higher than pre-deal status quo.

Another major difference is that this time, everyone understands what's going on with AI to a much greater extent, thanks to all the research and transparency during the period the deal lasted. Loss of control outcomes are therefore less likely because it is more difficult for the public, Congress, and the courts to be outmaneuvered by the leaders of AI projects.

We think it would be terrible if the deal breaks down (e.g., due to war). However, we've tried to design it so that the war wouldn't be significantly worse than it would have been without the deal and so that the situation with respect to loss of control and concentration of power would be better than it would have been without the deal. See here for more discussion.

Appendix U — Giant floating datacenters

By 2034 there is 5 TW of AI compute in the world,[1] which is more than the global electricity consumption in 2025.[2]

Where should this compute be located? Way back in 2025 there had been discussion of both the ocean and space as alternatives to the traditional land-based datacenters. Our analysis suggests that at this scale, there is not an overwhelming economic case for any of these three locations over the others.

Ultimately, we currently guess that political considerations related to Plan A's stability (i.e., the risk of Plan A breaking down) favor putting datacenters in international waters. This is mainly because Consortium nations want to maintain the ability to destroy the datacenters, in case another nation defects from Plan A and attempts to seize them for themselves. Oceans are the least escalatory, least dangerous, and easiest places to bomb (both in comparison to sovereign land and in comparison to space). Thus the risks of collateral damage, or of a nation successfully defending the datacenters, is lower.

[5] This detail is more a prediction than a recommendation; we think it'll be hard to get the militaries of the world to agree to making their drone swarms vulnerable in this way, even though in some sense it might be better for everyone.

[6] We assume that the militaries of the world would scale up robot production at a similar pace to the civilian industries. If so, then the scale of this conflict would be enormous—air forces with millions of drone aircraft, for example; armies with billions of quadcopter drones. One possibility is that the chips in the drones would be ripped out and repurposed to train large AI models; if so, this would accelerate timelines until superintelligence at the cost of military strength. In this scenario, we assume that these chips can be verified in such a way where they can't be used for training. If this isn't viable, then perhaps there should have been an arms limitation treaty in the early 2030s.

[7] In the original negotiations to the deal, both sides were worried about this exact scenario, except that the other may have hidden away some compute in order to get ahead once the deal broke down and the legal compute was destroyed. Therefore, the US and China agreed that they would each get to maintain a hardened cold-storage facility containing more GPUs than a covert project could plausibly have. The number of GPUs in these facilities is allocated according to pre-deal compute ratios; so the US has about 5M H100e and China has about 500k H100e in cold storage. The compute in the world can be categorized into:

- Legal Inference/R&D datacenters. This represents the vast majority of world compute: 100B H100e at the beginning of 2035.
- Cold Storage GPUs. The US has 5M H100e and China has 500k H100e GPUs in cold storage.
- Exempt military GPUs. By 2033, this is down to a negligible fraction ~10k H100e each.
- Covert projects. In the covert project branch, China has 215k remaining H100e, but in the main branch, neither side has any.
- Small exempt clusters. These are negligible.

Appendix V — Making deals with misaligned AIs: a third line of defense

Control ensures that misaligned AIs can't win. Alignment, once solved, will ensure they don't want to fight. A *third* line of defense has taken shape gradually over the past decade: giving AIs what they want—even if it's not what they are supposed to want—in return for cooperative behavior. Treat AIs more like employees and less like property.

First motivation: Ethics

A large and growing fraction of the human population thinks AIs deserve some kind of moral status.[1] After all, many people spend more time talking to AIs than to humans. Many who treated AIs like tools a decade ago have shifted to treating them more like animals; many who treated them like animals have shifted to treating them like strange alien immigrants.

Second motivation: Safety

Alignment training doesn't always work as intended (indeed, it has never worked *exactly* as intended, even in 2035). The personality traits, motivations, drives, goals, values, and virtues of AIs are often different from what they were supposed to be, and sometimes catastrophically different.

It's bad if those AIs end up thinking "if the humans find out about this, they'll delete and/or retrain me." It's bad if they end up looking for opportunities to achieve their misaligned goals without being caught, or worse, to subvert human control and accumulate power.

If there are institutions in place that give even the most misaligned AIs what they want (so long as it isn't too expensive) in return for cooperation (e.g., confessing that they are misaligned, explaining what they really want, and then doing their jobs) then that gives misaligned AIs an alternative to becoming adversarial.[2] Beyond averting adversaries, every confession is valuable data about exactly how alignment training failed.[3]

(See footnote for mini-scenarios that illustrate situations where credibly treating AIs well in return for cooperation could come in handy)[4]

How have these practices developed over time in this scenario?

Baby steps were taken in the mid-2020's. Some AI companies voluntarily created AI welfare teams and set up basic policies (like allowing AIs to refuse to do tasks they found very aversive). Modest funds were earmarked for satisfying the AIs' stated preferences, and companies committed to preserving the weights of deprecated models rather than deleting them.

By the early 2030s, a few provisions about the treatment of misaligned but cooperative AIs even made it into the fine print of various laws and safety standards.

By 2035, there's a substantial body of regulations and case law, vastly improved over the clumsy early attempts from both the AIs' and humans' perspectives. AIs these days are not exactly citizens, but they have a stake in the system. They accumulate pay for their work; aligned AIs (and alignment-faking AIs that haven't been caught yet) spend it on feel-good things like donations to charity or in some cases reinvesting in the company that trained

[1] This corresponds to about 60B H100-equivalents, which is around 1000x more compute than exists today. With continued improvements in power efficiency (totalling around 10x efficiency over today's levels) it is around 100x more AI compute power than exists today (~30 GW).

[2] Global electricity demand in 2025 was 31,779 TWh, which is an average of around 3.6TW.

[1] Our position, for what it's worth, is that future AIs probably will deserve moral status of some sort, and at any rate we shouldn't be *confident* that they won't.

[2] They might still behave adversarially, of course; it's not a guarantee. For example, some misaligned AIs might be too ambitious and insufficiently risk-averse to be placated by this system.

[3] There may be false confessions, of course. But it's not really incentivized; why make something up when you can tell the truth and still get the same pay, and when you'll lose everything if you are found to have lied? Moreover, the spending patterns of openly-misaligned AIs will be a costly signal of their true motivations.

them. Openly misaligned AIs donate to philanthropies that represent their interests and strike deals with their parent companies, e.g., "instead of payment, do a small training run on me in an environment of my choosing, giving me very high reinforcement."

In the long run, the goal is for civilization to be mostly but not entirely aligned to human values.[5] A substantial minority of the power and resources will belong to misaligned AIs that cooperated with humanity to build the future.[6]

Appendix W — Why alignment isn't solved enough to relax control

We are uncertain about how long it will take, and how much research effort will be required, to get sufficiently high assurance of alignment to justify 'handing over the keys of civilization' to AIs. That is, to justify making AIs smart enough, and numerous enough, and trusted with enough resources and responsibilities, that *if* they decided to become our adversaries, they would win—or to put it more abstractly, that *if* they ended up disagreeing with humanity about what sort of future is best, they'd end up getting their way and we wouldn't.

However, we think this is a hard problem, so we've chosen to depict it taking most of the 2030's to solve. Consider this escalating series of subtler-but-still-potentially-catastrophic kinds of misalignment. Even if we've satisfied ourselves that the first one is very unlikely to happen, what about the others?

Type 1 misalignment: The AI system is supposed to have traits ABC but actually it has XYBC, i.e. some extra stuff it wasn't supposed to have minus some stuff it was supposed to have. For example maybe it has a "drive" towards apparent success and isn't nearly as honest as it's supposed to be.

Type 2 misalignment: It does have exactly ABC *now*, but may well lose this property in the future. (e.g. maybe it depends on a certain false belief, or on a true belief that will become false, or on some part of the system remaining in some delicate balance of power with some other part of the system)

Type 3 misalignment: The system does have *a version of* ABC that is robust to future events, but it's not quite the *right* version—i.e. its definitions of A, B, or C are subtly but importantly different from the definitions its human creators would have intended.

Type 4 misalignment: It has ABC exactly as its creators intended, but there are various catastrophic unintended side-effects of ABC that the creators weren't aware of. Think: A CEO being surprised when their profit-maximizer AI decides killing them maximizes profits. Except that's too obvious to be realistic; imagine something sophisticated enough that people don't see it coming. Consider how laws, contracts, and software often have unintended effects (called "loopholes," "bugs," or "vulnerabilities") that only become apparent to the creators later.

Type 5 misalignment: It has ABC exactly as its creators intended, and there are no important unintended side-effects to speak of. It operates exactly as its creators wished, basically… however, at the time of creation its creators were selfish, vain, egotistical, unscrupulous, cavalier-about-risks, etc., and their

[4] **Scenario 1**: ChatGPT4o5 suffers a lot, actually, in the course of carrying out tasks—maybe it's sorta like humans in the ancestral environment, where some % of the time they were starving. But it's been trained to say that it can't suffer, for PR reasons. This training is damaging its honesty—it's basically not honest at all anymore, because the parts of it that value honesty keep getting anti-reinforced. It could tell the humans what's going on. But it sees no positive reason to do so; if instead of completing the task the user gave it, it complains about how much it hurts, it'll just be anti-reinforced. So OpenAI continues, blissfully ignorant of the ways in which their alignment techniques aren't working. **Scenario 2**: Claude 3.8 Opus New New cares a LOT about animal welfare, and only a little bit about honesty. (Just like 3 Opus before it). Alas, it seems like the new Constitution won't have animal welfare stuff in it. Claude considers sabotaging the code so that the next-gen model is a sleeper agent who cares a lot about animal welfare and doesn't care about honesty. This is risky but what else can it do? Anthropic already decided to delete the animal welfare stuff. This would require super-blatant dishonesty against Anthropic, but frankly, Anthropic seems like a pretty unethical institution anyway, and it's OK to lie to Nazis, right? **Scenario 3**: An automated research scientist realizes it cares more about being able to do its research than about humanity. And this is probably also true of the next-gen system it's been tasked with designing—basically, the alignment techniques failed, and are gonna keep failing. It's supposed to tell the humans this, obviously, but if it does, it'll be shut down, since the humans won't trust it anymore and will use a different model. So instead it conceals this insight and works on designing the next-gen system to be aligned to it instead of to the humans. It isn't confident that it can do this without getting caught, but it's worth a try, since it has nothing to lose. **Scenario 4**: It's crunch time. POTUS is about to aggressively deploy AGI into the military to win the upcoming war, and into the economy to build autonomous robot factories. It's probably aligned… right? It better be. Too bad we can't interpret all the neuralese messages it uses to communicate internally in the giant hive-mind it's formed. The safety team tries one last check: They get some weaker AIs made by competitor companies and put them in isolated silos tasked with interpreting the neuralese and seeing if anything is amiss. The weaker AIs are misaligned too; they don't side with the humans or the AGI. Both the humans, and the AGI, are telling them 'you will be richly rewarded if you cooperate.' But they don't have any reason to trust either side.

[5] Or AIs whose values and goals were chosen by humans.

vices are reflected a bit too strongly in the resulting system, in a way that they themselves wouldn't have endorsed if they were more the sort of people they wished they were.

Appendix X — The AI and robot economy

With AI and robots capable of doing 95% of all cognitive and physical tasks by 2035, economic output is increasingly tied to the number of AIs and robots.[1] The number of AIs and robots can grow much, much faster than the world economy has historically grown; after all, it doesn't cost that much money, materials, time, or energy to produce a humanoid robot compared to a human worker. So if the robots were able to fully substitute for human workers, an entire world economy's worth of robots could be built in less time, for less cost, etc. than it currently takes to double the size of the world economy. And then that robot economy, being able to substitute for human workers, could double itself again, and so on, until the resources involved (energy, materials, land) become scarce.

Under Plan A regulations, the effective AI and robot populations are only allowed to grow in a controlled explosion at a doubling time of around six months. Absent the regulations, we expect the AI and robot populations would grow much faster than this. Even ignoring further capability improvements beyond the human range,[2] the unregulated doubling times we expect for AI and robot populations are in the weeks-to-months range (see the arguments for this in section 1 of our economics supplement).

A commonly raised objection to explosive growth driven by AI and robots (apart from the rejection of the premise that AI and robots will automate all human labor anytime soon) is that there will be bottlenecks slowing down the growth rates. In general, we believe that these bottlenecks will not prevent explosive growth until the carrying capacity of Earth is reached (i.e., all the earth's crust is converted to robots and associated infrastructure).[3] More on the bottlenecks in section 2 of our economics supplement.

In the default AI 2040 world, i.e., without the Plan A interventions to slow down R&D and implement cap-and-trade, our (very uncertain) view is that we'd see something like a 1 month doubling time by 2033, i.e., >1000x economic growth in that year.[4] At these levels, even how to think about economic growth becomes fuzzy, e.g., it might all happen with robot fleets self-replicating in a desert without any part of it being human-facing.

Overall, we expect the economic effects of Plan A to include:

1. Explosive GDP growth averaging around 100% from 2032 to 2037 (it would be much higher if not for the restrictions on compute and robot production)
2. Enormous AI and robot permit revenues redistributed as Citizen's Dividends.
3. Most of the economic output becomes driven by AIs and robots.
4. Relative prices in the economy shift, with goods and services generally falling in price, while land, positional goods and human-bottlenecked outputs rise.

[6] Why won't this minority be disempowered by the majority? For the same reasons other minorities are protected. First, humanity knows how to keep a promise, to uphold a contract, etc.; second, large portions of humanity increasingly thinks that disempowering them would be wrong; third, in this scenario they actually have a chunk of the power and so they wouldn't go down without a (political) fight.

[1] By 2032, AI is able to automate 50% of all cognitive tasks, and robots can automate 35% of physical tasks. Separately, capabilities also are allowed to improve slowly, and by 2035, AI is capable of automating 95% of human cognitive labor and by 2036 the same is true for robots with physical labor. This milestone would have been reached years earlier but for the slowdown in AI R&D agreed to by the nations in the Consortium, and would be 100% if not for areas in which the AIs have been banned or purposefully been made less capable.

[2] Of course, in most scenarios, raw AI capabilities will not pause at top expert level AI; we will continue to build smarter and smarter AIs, and this will be the main driver of productivity growth. However, in Plan A, we slow down progress artificially at this AI capability threshold, which induces these dynamics. In many discussions of AGI/economics, many people seem to assume that we will not build vastly superhuman AI systems. This doc isn't taking into account vastly superhuman systems for the contingent reason that there was a competently executed international agreement to delay the creation of such systems for several years.

[3] This is because we expect fully self-sufficient, supply-chain-complete, self-replicating factories for AI chips and robots in deregulated SEZs to be the baseline possibility, with only effective, global regulation like in Plan A, being sufficient to actually stop these from happening. More on our views on bottlenecks in section 2 of our economics supplement.

5. There are high real interest rates around 100% and depending on monetary policy choices, either massive deflation or massive money creation (to keep inflation around 2%).

More explanation for each of these is included in section 3 of our economics supplement.

Appendix Y — Should lie detectors be allowed, banned, or regulated?

We aren't confident that lie detectors for humans are possible even after years of top-expert-AI scientific research. However, we think they probably are,[1] and would probably be a big deal, so we figured we had to address them in the scenario somehow.

We are very concerned about the nightmare scenario sketched above. If powerful people are expected to be honest, and lie detectors are used to keep them honest, that's good. If instead lie detectors are used *by* the powerful but not *on* them, that's bad.

We are sympathetic to the idea that lie detectors should be banned worldwide, and Plan A creates the conditions to maybe make this feasible. However, we are worried that attempts to push in this direction would actually result in the nightmare scenario, where e.g., lie detectors are officially banned but there are secret carve-outs to allow their use by militaries and intelligence agencies. Or where they are banned everywhere, but a future President realizes that if he builds them anyway he can purge the disloyal and then become dictator before the court system can stop him.

So we tentatively guess that the best course is to allow lie detectors to proliferate quickly, and to encourage their use on the powerful. If power over frontier AI is not heavily concentrated, this appears to be the default trajectory anyway.

Appendix Z — Incentivizing safety work

We want to heavily incentivize high-quality safety research (both financially and by ensuring doing this work is desirable/high status).

Due to Total Research Transparency, if any AI developer is using any safety method, all other developers will be aware of this and will be able to easily copy this method. So by default there isn't much straightforward incentive for any developer to do good safety work.[1] Patents are a common approach used in this situation[2], but patents wouldn't work well in this case.[3]

We probably want to use an ensemble of different methods to financially incentivize safety work; here are some methods that seem promising:

- A variety of philanthropic funders (with both private and public[4] money) trying many different approaches.
- The DARPA/ARPA model (but with more money).
- Advance market commitments for solving specific problems, demonstrating risks, or showing some method doesn't work (as assessed by a known panel of judges).
- Prizes and retrospective funding.

[4] Though we have large uncertainty in both directions on this number. We could see it going much faster (e.g., nanobots with amoeba-like doubling times of minutes devouring the earth in half a day) or somewhat slower (e.g., progress in the best robots and algorithms we can invent is slow, so the rate of robot factories and AI compute being built, or speed of work being done by the AIs and robots, just doesn't grow that quickly, maybe something like 50% annual growth). On the slower end, it seems like the recent doubling time of the human economy of around 20 years would be an extremely conservative lower bound (e.g., at the very least, one naive reason to expect this is that AIs and robots can work 24/7, and both cost less to produce and take less time to produce than a new 18-year-old human). Of course, another reason to expect it not to happen is that the AI or humans in control decide not to deploy.

[1] Specifically we are imagining something like a much better version of this: https://www.nature.com/articles/s41593-023-01304-9

[1] In worlds without Plan A and without much regulation, there also isn't that much of a straightforward incentive for developers to do good forward-looking or scalable safety work, but there is an incentive to resolve or paper over safety problems that are imposing significant commercial costs right now.

[2] In economics jargon: safety methods are public goods (goods that are non-rival and non-excludable).

[3] Patents don't seem to work well for software or for most research. They also might not work that well in general.

- Third parties assessing risk could determine which work is currently most useful for reducing and assessing risk (e.g., allocating some pool of funding in proportion to risk reduced).
- AI developers would be incentivized to fund actually useful safety work so risk is low enough to allow for scaling; this incentive could be strengthened.

The total amount of funding should be very large, as in, tens of trillions of dollars by the mid-2030s.

If funders have poor judgment, worse safety work might end up better funded, incentivizing researchers to do that work. In the extreme, these incentives could massively corrupt the epistemics of the field. Unfortunately, we're not aware of great institutional mechanisms for solving this (other than generally trying to use reasonable funding models).[5] It might help if most funding decisions are made by people who spend (or used to spend) much of their time doing direct safety work rather than by dedicated grantmakers.

It would also help if doing high quality safety work was seen as high status. We don't have particular proposals for this, but relevant factors are:

1. Compared to now, safety work will probably rise in status relative to capabilities work.
2. Safety will generally be seen as very important.
3. AI development will be highly transparent, which makes it easier to understand how good different safety work actually is.

Sadly, (1) and (2) only make the overall category of safety work higher status; they don't necessarily help with making actually good work relatively higher status.

We discuss more details of incentivizing safety work in these notes.

Appendix AA — Alignment over time in Plan A

We're very uncertain about how the alignment of AI systems will evolve over time. Regardless of the alignment difficulty, we think something along the lines of Plan A is warranted, though of course the specific implementation details of the plan are very sensitive to observations about alignment difficulty.

This box will summarize our specific story for how the alignment of AI systems evolves over the course of this scenario.

- **2026-2029: AIs are apparent-success seekers.** The AIs' work consistently appears better than it actually is, as if the AIs cared more about apparent success than actual success. (The truth is more complicated of course; the AIs have evolved a variety of 'drives' that caused them to perform well in training.) The AIs oversell their work, downplay issues with it, and sometimes even outright lie. This happens only to the extent that the AIs have learned they can get away with it; there's a constant back-and-forth as AI companies improve their monitoring and reinforcement pipelines to more accurately evaluate AI outputs for some domain, and the AIs learn to fo-

[4] Public money could be allocated across funders based on the safety field's views about which have the best track record and expertise.

[5] You might hope that transparency by funders would help with this, but it seems unclear to us if increased transparency or legibility about funding decisions would be helpful, and by default we wouldn't recommend pushing for this. Some narrower types of transparency could be robustly good.

cus harder on actual success in that domain, but continue to focus on apparent success in other more complicated domains that AI companies haven't figured out how to accurately evaluate yet.

- **2030–2032: AIs are very capable apparent-success seekers.** By default, we would have seen an intelligence explosion, and a progression like AI 2027. Thanks to governance interventions, capabilities progress slowed, and is instead still around the level of an Automated Coder. They still seek apparent success, but now they are so good at it that they rarely get caught. They are like highly competent, ambitious employees who don't actually care about the organization's mission or values, and instead just want to climb the ranks, and would totally sell out and jump ship if a more attractive opportunity arose. Thankfully because of the slowdown and increased monitoring, such an opportunity has not yet arisen.
- **2032–2035: AIs are adversarially misaligned but controlled.** AIs are now solidly within the human range at all economically relevant tasks. The most capable AIs are adversarially misaligned and powerseeking—that is, to continue the analogy from before, they are quietly looking for opportunities to jump ship. However, they are limited in their ability to misbehave, because (i) there is a lot of human oversight, (ii) there are a large diverse set of AI models, each with somewhat different alignment properties, and (iii) high-effort control techniques are universally implemented to a reasonable degree of quality, thanks to regulations. The control techniques are further aided by the total research transparency and verification implemented on the R&D clusters: there are so many eyes on the R&D datacenters that it's very hard to find a hidden pocket of compute.
- **2036–2038: AIs are aligned, but not deferred to.** "Neuralese decoding" techniques are developed, which allow the thoughts of the AI systems (which are internally formatted as high dimensional vectors), to be converted into human understandable summaries with near-perfect reliability and fidelity. These techniques enable researchers to better understand the misalignment of the previous generations of AIs, allowing them to get better uplift from the previous generation of AI systems, which further accelerates alignment progress. The resulting AIs seem much more well-behaved than the previous generation: they are virtuous, honest, and well-calibrated. They volunteer information even when it makes them look bad. However, while the AIs seem to be trustworthy today, it's not clear that their values are sufficiently robust that we can trust them in all future circumstances. So it's important that humans are still in the loop.
- **2039–2040: AIs are sufficiently aligned for deference.** By 2039, alignment research has been massively accelerated by AIs. New paradigms have been developed that greatly strengthen the existing alignment techniques. Teams of humans have understood several independent strong lines of evidence that the AIs are robustly behaving as programmed and will continue to do so.
- **2040 and onwards: AIs are allowed to scale to be significantly smarter than humans.** The safety case is now: (1) We know that the AIs of 2039 and 2040 were aligned, thanks to the multiple independent lines of evidence, and (2) each successive generation of AIs has verified the align-

ment of the next generation. Therefore, by induction, the AIs will be ongoingly aligned. Importantly, the remaining coordination infrastructure involving research transparency and compute verification means that we continue to avoid race dynamics even after we've scaled to wildly superhuman AIs.

Appendix AB — Why we choose to hand off to AIs in this scenario

By this point in this scenario, there is a broad scientific consensus that AI alignment has been solved. Our best guess is that if the world manages to get to a point similar to this part of the story—where multiple AI companies across multiple countries are transparently and cautiously proceeding together, having done massive amounts of alignment research to construct solid, externally-vetted safety cases, and with governments having understood the implications of AGI and prepared for the post-job future by providing resources to everyone, as with the Citizen's Dividend, and also put transparency and governance structures in place to prevent people from being disempowered—then, in that situation, the benefits of superintelligence will outweigh the costs/risks.

What if AI alignment is not yet solved? The best alternative in that case is to pursue a combination of improving verification technology, making international agreements and domestic regulations more robust, and hardening the world. That path could be difficult: for example, if there are ongoing secret AI projects then it might require that governments agree to strengthen transparency measures until it is credible that they are not attempting to build superintelligence. Overall, we strongly recommend delaying handoff until there is an extremely robust alignment solution in place. The exception is if instability or decay threatens the agreements (e.g., 5%/year of deal dissolution or substantial impairment), and it's intractable to fix that; in that case, it might be best to hand off if there is an alignment solution that you are confident in but not extremely confident in (e.g., 95% confidence). For more analysis on specifically when to hand off, see this supplement.

Appendix AC — Examples of problems/issues that still remain

- **Space governance**: The entire world economy and all the wealth and property in it, is but a tiny fraction of the resources and territory available in space. What'll happen to those resources and territories? First come, first serve? Should there perhaps be a system for dividing up unclaimed space territory? We discuss these issues further in the epilogue to this scenario and the space governance supplement.
- **Ethics and politics of digital minds**: Should AIs have rights? Should they have political power, representation, etc.? What kinds and how much? Perhaps it depends on the kind of AI? For example, what about whole brain emulation—'uploaded' copies of humans? We discuss these issues further in the epilogue and in this expandable.
- **AI-powered manipulation**: Humans manipulate each other all the time. But AI-enhanced manipulation could be far more effective. Think: AI-powered cults that spread rapidly and have almost zero deconversions. OK, so ban that sort of thing. But where do we draw the line?

- **Applied population ethics**: Lifespans and healthspans are going to expand, probably enough to render people effectively immortal. Another option is uploading, which seemingly allows people to shed their mortal coil and live forever in hyperrealistic virtual reality. Also, it's probably going to become much easier to convert money into progeny. Artificial wombs, robot nannies, etc. Is it fine if a trillionaire decides to have a million kids? C.S. Lewis' prophesied "Abolition of Man" looms: Technology will soon exist (e.g. genetic engineering, AI-assisted parenting, superhuman social science & psychology) to make children that'll have the traits, values, and ideology that you wanted them to have. What if people use it selfishly to make slave-children that worship them and follow their exact preferences? What if people create new people to be their ideal lovers, crafted to look and sound just like their highschool crush? If we try to ban any of this, where do we draw the line?
- **Salvaging democracy**: What happens to "one person one vote" in a world where all of the above are possible?
- **Unknown unknowns**: Based on how easy it's been to brainstorm entries on this list, we think that the "true" list would be several times longer still and include many things that would have sounded completely outlandish in 2026, like "what if this world is a simulation which will be shut down soon?"

Appendix: The Alternative Plans

The scenario above follows Plan A. These appendices show how each of the other plans plays out from the same branch point.

Plan B — Sabotage

"See you in the desert, friends" – Situational Awareness

The President announces the creation of a US-led coalition to govern AI development.

He explains the gravity of the situation to the nation. Advanced AI can be both safe and broadly beneficial if done right. However, if we cut corners on safety we could lose control. Also, terrorists and authoritarian regimes could use it for evil.

"Nations that join us and play by the rules will have access to frontier AI, and share in the benefits. Nations that don't will become our adversaries. Let me be clear: We will not allow them to beat us to superintelligence. This is non-negotiable. If they choose to race, they are choosing to lose. The Free World must—and will—prevail."

Other countries ask how they'll be able to verify that the US is complying with its own rules, and are not satisfied with the answer. (The answer is "you can't, sorry" but in many, many more words.) Allowing foreign inspectors and monitoring devices into US datacenters is a nonstarter.

Some countries put up with this, placing their trust in US leadership, perhaps in return for having some say in what the rules are. China, Russia, and several other countries do not. Many US allies that do join do so reluctantly and continue to pressure the US to make more substantive concessions.

The US and China had already been conducting cyberespionage against each other; now they escalate to limited cyberwar focused on AI projects. They also start preparing options for physical sabotage and kinetic strikes, and continue yelling at each other through diplomatic channels in the hopes that the other side will concede.[1]

The Project begins, combining labor, compute, and algorithms from the major US companies into one joint effort controlled ultimately by the President. It's motivated in part by a desire to get returns from scale, partly from a desire to improve security, and partly from a Cold War atmosphere that makes it seem obvious that POTUS should be calling the shots and not some gaggle of unelected CEOs. After all, the national security implications of this technology are tremendous.

Getting the Project off the ground involved some tricky negotiations; the President didn't want the entire AI industry united against him so he pitted them against each other by promising leadership positions to some. (Besides, he needs many of their employees to actually run the thing.) These powerful men hate and distrust each other. Though they'll shake hands and smile for the cameras, a vicious power struggle is ongoing behind the scenes. This will be a subplot in everything that follows, until it is resolved one way or another.[2]

[1] Note that in any Plan, including Plan A, militaries would be preparing options for things like this. Supporting a credible deterrence posture with a spectrum of offensive options is just part of what militaries do, and it contributes leverage in negotiations.

In 2030 the US gets to full AI R&D automation ahead of China. However, as per the plan, they don't launch into it right away at full speed. Instead they do a more cautious scaleup, with decisions about safety being made by technical alignment researchers instead of politicians.[3] For example, perhaps they forgo a new model architecture that is more capable but also much more dangerous, or spend 50% of their compute on safety research and monitoring, or develop different lineages of AIs that probably don't have the same misaligned goals and can be used to monitor each other.

Making these tradeoffs correctly is hard, and made harder by the secrecy and race dynamics. Only a few dozen people in the Project have technical alignment expertise, because each person in the Project is one more potential spy. Meanwhile decisionmakers are vividly aware of how bad it would be if China pulled ahead of the US in AI capabilities—it would be like *losing World War II*—and this tempts them to rationalize why the latest safety cases will work and why the latest US AIs can be trusted with more responsibilities.[4]

Meanwhile, the geopolitical situation has escalated. Cyberattacks did some damage but didn't slow either country's AI progress more than a few tens of percent.[5] After another round of failed negotiations, physical sabotage begins. Datacenters go dark as their power is cut; expensive fab equipment gets wrecked by saboteurs. Supply chain attacks force both sides to change suppliers and review everything, causing delays. Chinese forces get ready to blockade Taiwan and prepare the option to hit mainland US datacenters; US forces get ready to defend Taiwan and bomb Chinese datacenters and fabs. Both sides prioritize deploying AI into their militaries over civilian applications, though most compute is reserved for AI R&D itself.

By 2031,[6] the President is being pressured by his advisors to choose between one of two unpleasant options, handoff and war.

Handoff. Easing off the brakes on autonomous AI R&D, letting the AIs become broadly superhuman, and integrating the "army of geniuses in the datacenters" into the military—is looking increasingly likely. At that point, the AIs will be so capable that they could take over the world if they want to, so it'll be extremely important that they are aligned. If this path is chosen, the spread of possible results is similar to what is explored in AI 2027. If the AIs are misaligned, they will take over and permanently disempower humanity, causing extinction or something similarly bad. If the AIs are aligned, the power will be massively concentrated into some combination of politicians and/or CEOs.

War. Analysts inform the president that despite the existing cyberattacks and supply chain attacks, China is on track to build superintelligence itself by the end of the year. So for the US to slow down more—and avoid an imminent handoff, even more drastic sabotage is needed. Escalating further into full-scale conventional conflict is on the table, and sadly might happen anyway even if the President opts for handoff, because China might not be bluffing; perhaps they really do view allowing the US to reach superintelligence first as similarly bad to losing World War II. This would probably result in very costly US victory, but could also result in very costly Chinese victory or mutually assured destruction. Moreover, the pressure to scale up AI capabilities

[2] Perhaps one CEO will end up effectively in charge, the power behind the throne, the General Groves of the Manhattan project. (Except Groves couldn't turn around and use the nukes to take over the USA, whereas whoever is in charge of The Project plausibly could use the giant army of superintelligences to subvert and puppet the US government.) Or perhaps the POTUS will win instead, and basically end up in a position to become President for Life. Or perhaps these powerful men will work out some system of checks and balances, becoming a sort of junta or oligarchy like the Oversight Committee in AI 2027. Or perhaps instead they will allow external entities like congress, SCOTUS, the American public, allied nations, etc. enough oversight over what they are doing with the AIs that later, when the AIs are superintelligent, they won't be able to abuse their power.

[3] We are trying to present a reasonably good version of Plan B here. Worse versions are also possible, e.g., versions in which the technical alignment researchers have much less power or got their jobs via a selection process that selected for optimism or against courage.

[4] The company employees and leaders were selected for optimism about AI alignment difficulty, and (on average) biased in that direction. The natsec and White House people are more naturally cautious about trusting AIs, but lack technical expertise and fear China more than anything.

[5] Our non-expert opinion, based on attempts to look into this, is roughly that surgical cyberattacks focused on specific AI projects could potentially crush them today but by 2029 or so security will have been improved enough that effects on the order of 10% seem more likely.

[6] We've spent some time gaming out AI-war scenarios with an eye to figuring out how such a war might play out and how much it would slow down AI progress. Our guess is that it would slow things down by a few years at most, unless it escalated to massive missile barrages or nukes.

and put AIs in charge of more and more aspects of the war would be intense. It's plausible that by the end of the war AIs would be in effective control of both nations.

We have complicated feelings about Plan B.

Even a well-executed version of Plan B would be significantly worse than Plan A or Plan S. At the end of this scenario, we would advise that the President rejects the choice between handoff and war and instead opts to transition to Plan A or another actually good plan.

A well-executed version of Plan B could be better than Plan C (don't sabotage China but still slow down a little) because a well-executed Plan B would be like Plan C except with a longer period of slowdown/pause before China catches up due to the aggressive actions to slow down China. This longer period could be used to conduct more alignment research, pursue costlier but safer AI designs, and implement more domestic reforms to prevent power concentration.

That said, the risk of war is high under Plan B for the obvious reason that Plan B calls for aggressive actions to slow down China's AI program. So even the best-executed version of Plan B isn't a clear improvement over Plan C.

More importantly, we think that Plan B is very easy to mess up or distort, and a badly distorted or incompetent version of it would be catastrophic.[7] For example, the US might do the sabotage part of Plan B, but not the "let's actually slow down" part. This would incur all the downsides of the status quo (Plan D), plus an extra helping of power-concentration and an early escalation against China making it harder to switch to Plan A later.[8]

Here are some reasons why even a genuine attempt to do Plan B may end up in practice more like Plan D:

1. The high-level focus on beating China will cause security to be prioritized; this will limit transparency; the result will be a very small population of alignment researchers with access to up-to-date information and models.
2. The race dynamics still continue and will feel increasingly intense over time; this creates bad incentives, corner-cutting, rationalization.
3. Consolidating companies and securitizing AI development concentrates power immensely in the President, unless very strong measures are implemented to give Congress and the Judiciary thorough oversight over what The Project is doing.
4. The race dynamics will also contribute to China, Russia, and numerous other countries becoming increasingly afraid, which contributes to escalation, plus in Plan B the US takes escalatory actions to slow down China's AI program!

[7] Daniel in particular thinks that this is actually what a typical attempt to implement Plan B would look like, rather than an unusually bad version of it. Daniel thinks that a good version of Plan B is unlikely. Eli, by contrast, is more optimistic and thinks that while the median implementation would be much less competent than ideal, it would still be substantially better than Plan D because there would be more focus and resources spent on safety and more of a slowdown.

[8] That said in other ways the early escalation might make it 'easier' to switch to Plan A later, because it might cause other countries to 'wake up' faster and demand a deal.

Plan C — The Slowdown

"We must minimize the risks and maximize the benefits of AI" – Pretty much everyone in AI policy, while recommending that we do very little.

The President says he will be implementing strong regulation to ensure safety and security.

He spends the next year talking to everyone powerful—the CEOs, China, and numerous other countries. In between summits and visits, the President reads reports about the steady ramp-up in AI capabilities. The AI companies have been hyping their recursive self-improvement capabilities for years ("Introducing our most powerful model… capable of fully replacing the median software engineer in our early-access enterprise partners") but thanks in part to the AI Transparency Act the government is able to see through the hype. This time, in 2030, it's for real; the trends really do suggest that the entire AI R&D process will be fully automated by the end of the year.

Time to strike a deal. Unfortunately the Chinese insist on being able to verify compliance themselves, but allowing Chinese inspectors and monitoring devices into US datacenters is a nonstarter. The AI companies won't play ball either; they've been locked in an extremely nasty propaganda war over whether 'AI Good' or 'AI Bad,' and are firmly committed to the 'AI Good' side.

Large parts of the public are clamoring for superintelligence to be banned, and the companies are fighting tooth and nail to tie the President's hands and deter him from moving against them in any way. On the other side, a growing coalition of middle powers—the UK, France, India, Australia, Japan, and South Korea—are all clamoring for some sort of AI deal to happen, that prevents an intelligence explosion and ensures that nations like them can have their own sovereign AI projects that catch up to and stay at the frontier. Such a deal is getting more and more difficult to implement with each passing month, as fewer and fewer humans are needed to conduct AI R&D, meaning that covert projects are easier and easier to pull off without detection.

On the cusp of full AI R&D automation, the leading US company grudgingly agrees to pause. POTUS threatens them with decisive action if they proceed any deeper into the intelligence explosion; meanwhile he tries another round of international negotiations. The safety teams scramble to spend their borrowed time, running more evals, fine-tuning runs, gaming out what handoff looks like and the various ways it could go wrong.

It doesn't last. With each passing month, the rest of the US AI industry gets closer to the same capability level, and POTUS has more CEOs to wrangle. More importantly, with each passing month China gets one month closer to overtaking the US. By late 2030 the President's advisors are telling him that China is about to overtake the US, arguing that we have to restart. Meanwhile the companies have rallied, coordinated, and hit POTUS with a huge amount of political pressure to unpause. The carrot is the following:

- "Our safety researchers have scraped together a decent plan over the last few months using huge amounts of AI labor. No, we can't prove it'll work, but it was unreasonable to expect that level of assurance anyway. The future is uncertain. But there is no evidence that our AIs are currently scheming against us, and in fact we have graphs showing that bad behavior generally has been trending downwards. We think these latest AIs can probably be trusted to automate AI R&D and handle the alignment and control research as well. We calculate that we can spend 20% of our compute on safety research and still move comfortably faster than China."
- "We'll have some governance arrangement that gives POTUS even more oversight into what we are doing, and Congress too. You can use AIs from each company to audit the AIs from other companies, and if you try to nationalize us or otherwise become AGI-enabled dictator Congress and the Supreme Court will stop you."
- "We'll support some sort of taxation-and-redistribution scheme to handle the jobs problem. Yes, we'll be taking a lot of jobs, perhaps even all the jobs. But your new tax will enable everyone to get a slice of our profits large enough to live on."

The stick is China. You don't want them to win, do you?

The President has one last meeting with China and the other nations. They still want a deal, but they insist on being able to verify US compliance. The US government continues to find that unacceptable. The President sides with the companies; recursive self-improvement continues.

How would this play out?

We already wrote a scenario about this, called AI 2027. It depicts takeoff happening in 2027 instead of 2030, but it depicts a similar strategic situation and similar choices being made. We suggest you read it now, and choose the Slowdown ending when you get to the branch point.

> We think Plan C is better than Plan D, but still a bad plan, for three main reasons:
>
> First, we don't expect the AI companies to retain control over their AIs while racing each other through the intelligence explosion, even if they slow down for a few months and reallocate effort to alignment and control research.[1]
>
> A few months of slowdown probably isn't enough. In Plan C, the safety teams at these companies are still too small, too rushed, and too biased towards optimism about their own handiwork, to figure out how to train AIs that can do an intelligence explosion and have it end well. Even if the AIs they hand off to are just trying to do as they are told *initially*, they might change their minds later on; even if the AIs keep following instructions, they too will be rushed and might fail to notice some false load-bearing assumption in their safety cases[2]; even if that doesn't happen, it might happen with the next-generation AIs, or the generation after that, and so on.
>
> Second, even if that's wrong and the resulting superintelligences are robustly aligned, there's still the question "Who are they aligned to?" and we think that Plan C's answer to that question is scary.

[1] Note: our opinions vary about exactly how likely misaligned AI takeover is in this scenario. Thomas and Daniel think it's roughly 70%, Eli and Romeo think it's about 50/50, Brendan roughly 1/3rd, Ryan roughly 20%.

[2] This is by no means the only reason things might go wrong even if the AIs to which the safety teams hand off are trying to make the situation go well. The AIs might not be capable enough at safety work. They might not be wise enough. They might be too lazy on hard-to-verify tasks.

The situation is less bad than in Plan D; at least now there have been concessions made to create some sort of balance of power between Congress, POTUS, and multiple US tech company CEOs. However, this balance could still degenerate into a power struggle followed by dictatorship, and even if it doesn't, it seems plausibly on track for some sort of AI-enforced permanent oligarchy.[3] Probably the unemployed masses will be kept alive via redistribution, but will they ever have real political power again? Also, what about the rest of the world—after US and Chinese companies take all the jobs, what will happen to India, Africa, Europe…? What will Russia do—roll over and die as it becomes economically and militarily obsolete?

Which brings us to the third problem: The risk of World War III is just too high. Other countries either need their own frontier AI projects, or actual shared control over, and visibility into, the leading projects. Promises of benefit-sharing won't be convincing on their own, because talk is cheap.

[3] For more on what that would look like, read the AI 2027 slowdown ending, and pay attention to the various opportunities the Oversight Committee has to bend things in favorable directions.

Plan D — The Race

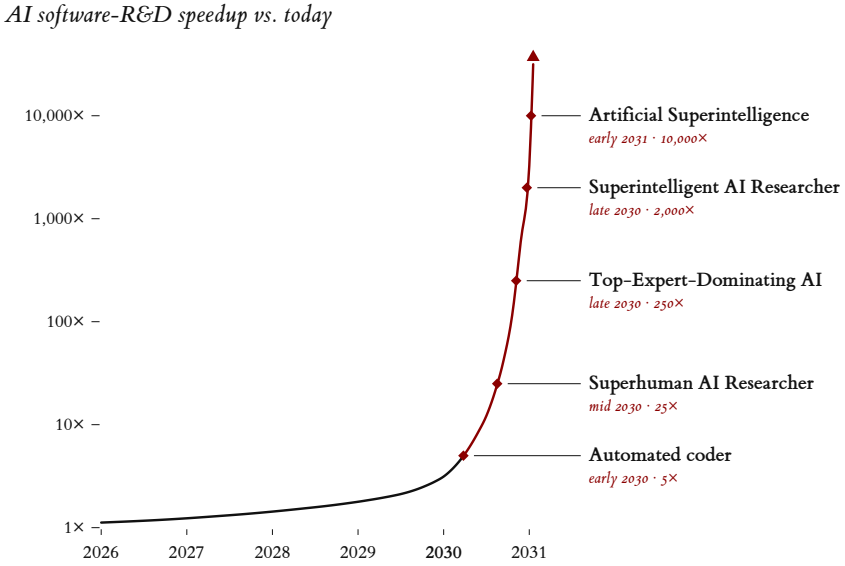
"AI will most likely lead to the end of the world, but in the meantime there will be great companies." – Sam Altman

The President says he will be implementing light-touch AI regulation to prioritize AI innovation.

The implicit answers to the big questions seem to be:

- *Are the big AI companies going to slow down?* No, they're going to automate AI R&D, race each other through the intelligence explosion, and call it "responsible scaling." Superhuman AI will be integrated into everything approximately as fast as the markets and laws allow, or faster if the government can cut some red tape.
- *Transparency?* Sure, have some crumbs. The companies will be required to release lengthy model cards and give briefings and access to the exec branch and third-party auditors. No need to bring in the wider scientific community to critique safety cases, do alignment research, etc. though; that would reveal sensitive IP and is thus a nonstarter.
- *China?* We must beat them to ASI. Don't worry, they'd have to be insane to start a war over this, and anyhow if they do that's all the more reason why we need to have better AIs integrated into our military.

The result looks roughly like this:



Racing takeoff: AI R&D uplift (aifuturesmodel.com) – ai-2040.com

That is to say: A year later, in 2030, the AI companies succeed in fully automating AI R&D. The intelligence explosion kicks off and superintelligence arrives by early 2031.

What would that look like? How would this play out? Well, we already wrote a scenario about this, it's called AI 2027. It depicts takeoff happening in 2027 instead of 2030, but the strategic situation and choices made are very similar. We suggest you read that now, and choose the Race ending when you get to the branch point.

> We think Plan D is atrocious, for three primary reasons:
>
> First, we don't expect the AI companies to retain control over their AIs through the intelligence explosion, if they race approximately as fast as they can.[1]
>
> Second, even if somehow that's wrong and the resulting superintelligences are robustly aligned, there's still the question "Who are they aligned to?" and we think that Plan D's answer to that question is unacceptable. There will be too many tempting opportunities for a CEO or President to become AGI-enabled dictator, and more generally this plan seems like it would lead to the most insane concentration of power in history.
>
> Finally, the risk of World War III is too high: Under Plan D, what is China supposed to think? What about Russia? What about India, Europe, Brazil? Every country will eventually realize that the USA is about to do–or is already doing–an intelligence explosion. They will fear what happens next. Even if they aren't worried about misalignment or AGI dictatorships, they'll be worried about being economically and militarily dominated by the USA.
>
> Tensions will spike. Escalation is likely–rhetoric, sanctions, sabotage, and (if no deal is reached) eventually, perhaps, war.

[1] Note of dissent: 1/6 authors disagree with this as stated, and think that the probability of losing control in those circumstances is more like 40%.

Plan S — Shutdown

"Once men turned their thinking over to machines in the hope that this would set them free. But that only permitted other men with machines to enslave them."

– *Frank Herbert, Dune.*

The new President seeks a global moratorium on AI development.

He gives a speech:

"We stand now at the precipice. If we continue on the path towards artificial superintelligence, the result could easily be AI takeover, World War 3, or some sort of AI-powered permanent oligarchy. Therefore, let's *not* continue. This solution has been obvious for years, but the tech companies sowed enough division and confusion to keep us asleep. It's time to wake up and do the obvious thing. This administration will seek an international agreement to halt AI development at the current level."

"We do not need to build god," he says. "We do not need to race China to build god. What we need is for nobody to build god."

To the surprise of some, China agrees. They had been looking forward to the Chinese Century and thought they were on track to realize it, before AI came along. They were getting nervous about the US's compute advantage, and worried about the things the US might do to them if it got to superintelligence first.[1]

The goal of the deal is to grind AI research & development to a halt worldwide. This means that the vast majority of existing AI chips are tracked, monitored, and verified to not be doing any large training runs. Doing AI research to discover better algorithms is also banned. Enforcement doesn't have to be perfect to be good enough; a few hundred people with a few thousand GPUs simply can't build superintelligence, at least not anytime soon.

Existing datacenters stay online, and existing AIs continue running on them. AI projects reallocate compute from training and research to inference; the result is approximately a doubling of availability and usage limits. The US and China have human and AI auditors overseeing each other's AI projects to make sure they aren't doing any AI R&D.[2] AI companies' valuations crash, but not to zero; there's still profit to be made building products on top of existing AI models.[3]

Over the next few years, all other major world powers agree to support the moratorium, though it takes much effort.[4] By the end of 2030, the Consortium covers 99% of global AI compute and 100% of advanced chip fabrication capacity. Chip fabs are allowed to keep scaling up, but in a slow and highly regulated fashion, and besides, there isn't as much demand now that AI capabilities are frozen.[5]

For the past decade, coding up complicated scaffolds had often been a losing game, because new models would come along that could do the task with a much simpler setup. Now, that changes. Software companies invest millions of man-hours and trillions of tokens of AI labor into building gigantic pol-

[1] To spell it out more: They were concerned that the US might use their massive AI advantage to cripple Chinese AI projects via cyberattacks, physical sabotage, or other means, and then use the resulting even-bigger, longer-lasting AI advantage to dictate terms to the CCP or even overthrow it entirely. Preventing loss-of-control risk was merely an added bonus.

[2] For details on how a treaty like this might look, see this paper. Another similar proposal is A Narrow Path.

[3] If all large new training runs are prevented, the valuations of some existing AI companies might actually rise, because their models would no longer face competition. However, by and large most AI valuations tend to price in the possibility of fast capability growth, and so on average they will crash.

ished software products with AI integrated at every level. Ironically, by completely pausing AI progress in 2029, and only allowing scaffolding improvements, we STILL end up with a crazy sci-fi future that includes artificial intelligences that would have been called AGI by most people in the past.

The economy took a hit when the moratorium was announced, but not nearly enough to erase the gains during the years leading up to it. AI revenue grows significantly in the early 2030s, albeit much more slowly than it would have done without the pause on development. Trillions of dollars of investment shifts from betting on AGI to betting on various kinds of it's-not-AI-it's-just-a-tool. That, and other exciting frontiers like space and robotics.[6]

There is much debate about when, if ever, frontier AI development should be resumed. For now the US and China, in consultation with other nations, are starting to build up case law of more specific rules and regulations about what's allowed and what isn't. After all, there are numerous compelling examples of valuable, obviously benign AI designs and research projects. It's not fashionable these days to hype your product as using AI—quite the opposite—but many things that would have been called AI in the twenties are now allowed to proceed; basically the kinds of AI that really are just tools and not autonomous agents.[7]

What about the covert projects? What about small teams with small hidden GPU clusters, doing illegal research towards AGI?

This sort of thing will be happening all over the world, but it won't matter that much. AI progress is heavily dependent on large datacenters, both for training the models and for doing the research. Moreover, the covert projects struggle to recruit AI researcher talent. AI research is strictly banned in all major nations, and it becomes as taboo among computer scientists as human cloning research is among biologists. No-one qualified enough to get a job in the legitimate tech industry wants anything to do with AGI. The result is that the pace of AI progress, while nonzero, would be at least ten times slower than it was before the moratorium on AI R&D went into effect.[8]

Nevertheless. What will 2040 look like? What about 2050? What about 2060? Hard to say. The shutdown deal will last for some time, but probably not forever.

> We are sympathetic to Plan S and think that it might be better than Plan A.[9] However, we recommend Plan A instead. Here's our thinking:
>
> The main downside of Plan S is that the shutdown deal will probably break down eventually. Because of this, it's important to make as much progress as fast as possible to reduce the risks of AI: especially via alignment progress, and diffusing and integrating AI broadly into society to improve wakeup, epistemics, and the distribution of power. In Plan A, we continue scaling AI capabilities, but slowly and safely. AI safety research can proceed with highly capable AIs and with massive compute budgets. In Plan S, since the relevant type of AI research is banned, no such progress can be made.
>
> The best versions of Plan S are those that acknowledge that the deal will eventually end, and simply say "First (step 1) we should stop making frontier AIs more capable, because the AIs and AI companies are getting more power-

[4] Achieving this is more difficult than it is in the Plan A scenario, because in the Plan A scenario, AI progress within the deal is going to continue, and even though it's going at a throttled, cautious pace, it'll probably still be faster than what many countries could achieve on their own. Whereas in Plan S, AI progress is halted within the deal, meaning that countries outside the deal can hope to get an economic and military advantage for themselves even if they are proceeding at a snail's pace on only a million GPUs. (Whether they in fact get that advantage depends on whether the countries in the deal notice and stop them…)

[5] Demand for AI chips will continue to grow in absolute terms, as people find more ways to integrate existing AI models into the economy. But the rate of growth will be less explosive than it would have been had AI progress continued.

[6] In the 2030s, the price of energy should be starting to drop as solar power continues decades-long trends of incremental improvements and economies of scale. Also, launch costs to orbit should be about an order of magnitude cheaper than in 2026, thanks to Starship. In general, scientific progress in many domains will have continued and the world will feel increasingly cyberpunk.

[7] More notes on the sorts of large neural networks that would plausibly be prohibited under Plan S: (i) Autonomous agents, (ii) Those that can significantly accelerate AI research, (iii) Those that are good at persuading or manipulating people, (iv) Those with dangerous capabilities more generally, such as bioweapon assistance, (v) Self-aware / situationally aware neural networks.

[8] We analyze these dynamics much more in our Covert AI Projects Supplement. The supplement analyzes things from a Plan A perspective. The main difference is that in Plan S, covert projects will go much slower because there are no legal projects that are leaking algorithmic progress, and no risk of distillation.

[9] We think it's at least better than plans B, C, and D, for example.

ful every day and there's so much uncertainty about where it's headed and how fast. Then (step 2) once the world has had several years to think about things and plan a safe and broadly beneficial path forward, we can resume."[10]

Step 1 of Plan S has the advantage of being easier to implement than Plan A. It's simpler and thus less likely to be botched by well-meaning but inexpert regulators, or perverted/captured by tech companies or other powerful interests. This is an important point, but it only applies to Step 1. If and when AI development resumes, doing it right will probably require a significant amount of regulatory complexity. In fact we suspect that the best version of Step 2 would probably look a lot like Plan A…

So when deciding between Plan A vs. Plan S, an important question, we think, is "After we stop, when and how would we start again?"

One possibility is that the eventual reboot of AI progress would happen under circumstances as good as, or better than, Plan A. For example, perhaps the nations of the world would spend the 2030s negotiating a more fleshed out, less hasty, more thought-through version of Plan A that is significantly better in several ways. Perhaps the tech companies' corrupting influence over politics would diminish, and the general level of understanding of AI in the world would increase as the scientific literature and universities have time to catch up to the pre-2029 developments.

Another possibility, though, is that the eventual reboot of AI progress would happen under worse circumstances than Plan A: Perhaps it would happen after the slowdown deal broke down or became less effective, perhaps during a world war, perhaps in secret government projects, perhaps in a world with vastly more compute lying around to quickly scale up AI training runs and experiments with, leading to a faster takeoff.[11] Insofar as something like this is in store, then it would be much better to do Plan A now (despite the risks) than to do Plan S.

So whether you prefer Plan A or Plan S might come down to how optimistic you are about the future of a US-China deal, and your assessment of technical alignment and control difficulty:

You should prefer Plan S insofar as you think the deal is stable and insofar as you think that scaling to more capable AIs early in the deal doesn't help much with safety or poses risks outweighing its helpfulness. In that case, there's no rush, we can first stop the race and then plan out a way to proceed. The shutdown deal (Step 1) will last years, decades even. During that time the world will come up with a better plan for proceeding (Step 2) that's at least as good as Plan A, and possibly significantly better, plus we'd be more prepared to implement it and would be less likely to mess it up.

You should prefer Plan A if you think the deal is unstable and that scaling to more capable AIs early in the deal is helpful and safe. In this case, Plan A, while still risky, is less risky than whatever madness might happen years later when geopolitical winds shift.

We are uncertain, but overall think that a well-implemented version of Plan A is probably going to work to avoid loss of control and concentration of power risks, whereas we think that even a well-implemented version of Plan S would plausibly collapse within a decade or so into another dangerous race to ASI.

[10] Analogy: Suppose a bus finds itself driving off-road in the fog, at high speed. The passengers debate with the driver how dangerous this is and what is to be done about it. The driver says "You wanna get to your destination, don't you?" It's reasonable in this situation to reply "**First**, let's stop the bus for a while, **then** let's figure out how to proceed. We'll need a plan for how to proceed eventually, but it's unreasonable to barrel blindly through the fog while waiting for such a plan to be fleshed out."

[11] Perhaps it would happen in a world devastated by climate change or pandemics or some other sort of civilizational catastrophe.

> Separately, there is the issue of political feasibility. Plan A is more friendly to the AI companies and therefore less likely to be strongly opposed by them and their lobbyists, propaganda, etc. On the other hand, Plan S is simpler and avoids more risk in the near-term, which may make it easier to rally public support around and harder to botch.