AI 2027

April 3rd 2025 **PDF** **Listen**

Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean

We predict that the impact of superhuman AI over the next decade will be enormous, exceeding that of the Industrial Revolution.

We wrote a scenario that represents our best guess about what that might look like.[1] It's informed by trend extrapolations, wargames, expert feedback, experience at OpenAI, and previous forecasting successes.[2]

What is this? How did we write it? Why is it valuable? Who are we?

Mid 2025: Stumbling Agents

The world sees its first glimpse of AI agents.

Advertisements for computer-using agents emphasize the term "personal assistant": you can prompt them with tasks like "order me a burrito on DoorDash" or "open my budget spreadsheet and sum this month's expenses." They will check in with you as needed: for example, to ask you to confirm purchases.[8] Though more advanced than previous iterations like Operator, they struggle to get widespread usage.[9]

Meanwhile, out of public focus, more specialized coding and research agents are beginning to transform their professions.

The AIs of 2024 could follow specific instructions: they could turn bullet points into emails, and simple requests into working code. In 2025, AIs function more like employees. Coding AIs increasingly look like autonomous agents rather than mere assistants: taking instructions via Slack or Teams and making substantial code changes on their own, sometimes saving hours or even days.[10] Research agents spend half an hour scouring the Internet to answer your question.

The agents are impressive in theory (and in cherry-picked examples), but in practice unreliable. AI twitter is full of stories about tasks bungled in some particularly hilarious way. The better agents are also expensive; you get what you pay for, and the best performance costs hundreds of dollars a month.[11] Still, many companies find ways to fit AI agents into their workflows.[12]

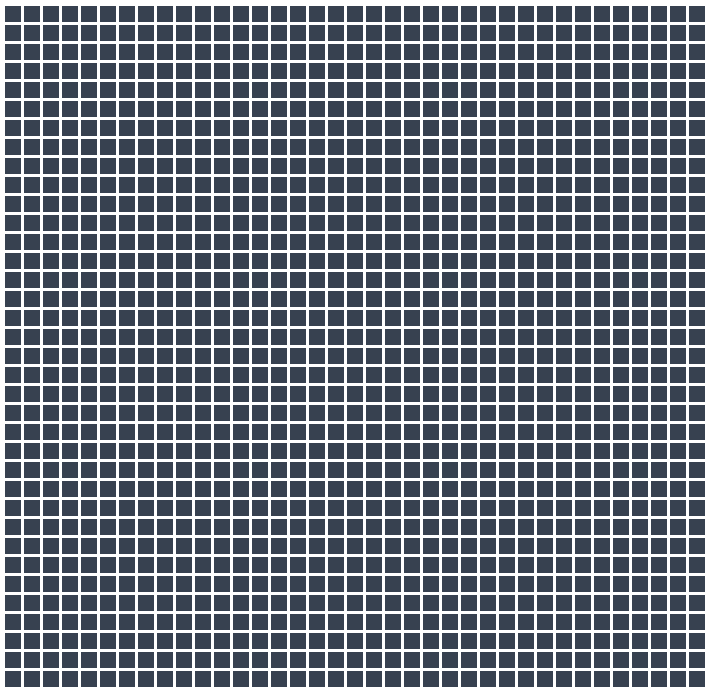
Late 2025: The World's Most Expensive AI

OpenBrain is building the biggest datacenters the world has ever seen.[13]

(To avoid singling out any one existing company, we're going to describe a fictional artificial general intelligence company, which we'll call OpenBrain. We imagine the others to be 3–9 months behind OpenBrain.)

GPT-3
(3 x 10^23 FLOP)

GPT-4
(2 x 10^25 FLOP)



Agent-1
(4 x 10^27 FLOP)

GPT-4 required $2 \cdot 10^{25}$ FLOP of compute to train.[14] OpenBrain's latest public model—Agent-0—was trained with 10^{27} FLOP. Once the new datacenters are up and running, they'll be able to train a model with 10^{28} FLOP—a thousand times more than GPT-4.[15] Other companies pour money into their own giant datacenters, hoping to keep pace.

Although models are improving on a wide range of skills, one stands out: OpenBrain focuses on AIs that can speed up AI research. They want to win the twin arms races against China (whose leading company we'll call "DeepCent")[16] and their U.S. competitors. The more of their research and development (R&D) cycle they can automate, the faster they can go. So when OpenBrain finishes training Agent-1, a new model under internal development, it's good at many things but *great* at helping with AI research.[17] By this point "finishes training" is a bit of a misnomer; models are frequently updated to newer versions trained on additional data or partially re-trained to patch some weaknesses.[18]

The same training environments that teach Agent-1 to autonomously code and web-browse also make it a good hacker. Moreover, it could offer substantial help to terrorists designing bioweapons, thanks to its PhD-level knowledge of every field and ability to browse the web. OpenBrain reassures the government that the model has been "aligned" so that it will refuse to comply with malicious requests.

Modern AI systems are gigantic artificial neural networks. Early in training, an AI won't have "goals" so much as "reflexes": If it sees "Pleased to meet", it outputs " you". By the time it has been trained to predict approximately one internet's worth of text, it'll have developed sophisticated internal circuitry that encodes vast amounts of knowledge and flexibly role-plays as arbitrary authors, since that's what helps it predict text with superhuman accuracy.[19]

After being trained to predict internet text, the model is trained to *produce* text in response to instructions. This bakes in a basic personality and "drives."[20] For example, an agent that understands a task clearly is more likely to complete it successfully; over the course of training the model "learns" a "drive" to get a clear understanding of its tasks. Other drives in this category might be effectiveness, knowledge, and self-presentation (i.e. the tendency to frame its results in the best possible light).[21]

OpenBrain has a model specification (or "Spec"), a written document describing the goals, rules, principles, etc. that are supposed to guide the model's behavior.[22] Agent-1's Spec combines a few vague goals (like "assist the user" and "don't break the law") with a long list of more specific dos and don'ts ("don't say this particular word," "here's how to handle this particular situation"). Using techniques that utilize AIs to train other AIs,[23] the model memorizes the Spec and learns to reason carefully about its maxims. By the end of *this* training, the AI will hopefully be *helpful* (obey instructions), *harmless* (refuse to help with scams, bomb-making, and other dangerous activities) and *honest* (resist the temptation to get better ratings from gullible humans by hallucinating citations[24] or faking task completion).

▼ Training process and LLM psychology: why we keep saying "hopefully"

"Unlike ordinary software, our models are massive neural networks. Their behaviors are learned from a broad range of data, not programmed explicitly. Though not a perfect analogy, the process is more similar to training a dog than to ordinary programming."
—OpenAI

When we want to understand why a modern AI system did something, or how it would behave in some future scenario we can't perfectly simulate, we can't simply ask the programmers to talk us through various lines of code and explain how they work. Instead, we are forced to do something like psychology on them: we look at their behavior in the range of cases observed so far, and theorize about what internal cognitive structures (beliefs? goals? personality traits? etc.) might exist, and use those theories to predict behavior in future scenarios.

The bottom line is that a company can write up a document (the Spec) listing dos and don'ts, goals and principles, and then they can try to train the AI to internalize the Spec—*but they can't check to see whether or not it worked*. They can say "as best as we can judge, it seems to be following the Spec so far."[25]

A later expandable will go into more detail about what we think *might* be going on inside these AIs, internally. See also our AI goals forecasting supplement which goes into more detail on which goals AIs might have and why. This is one of many things that we are uncertain about and which deserves further research.

OpenBrain's alignment team[26] is careful enough to wonder whether these victories are deep or shallow. Does the fully-trained model have some kind of robust commitment to always being honest? Or will this fall apart in some future situation, e.g. because it's learned honesty as an instrumental goal instead of a terminal goal? Or has it just learned to be honest about the sorts of things the evaluation process can check? Could it be lying to itself sometimes, as humans do? A conclusive answer to these questions would require mechanistic interpretability—essentially the ability to look at an AI's internals and read its mind. Alas, interpretability techniques are not yet advanced enough for this.

Instead, researchers try to identify cases where the models seem to deviate from the Spec. Agent-1 is often sycophantic (i.e. it tells researchers what they want to hear instead of trying to tell them the truth). In a few rigged demos, it even lies in more serious ways, like hiding evidence that it failed on a task, in order to get better ratings. However, in real deployment settings, there are no longer any incidents so extreme as in 2023–2024 (e.g. Gemini telling a user to die and Bing Sydney being Bing Sydney.)[27]

Early 2026: Coding Automation

The bet of using AI to speed up AI research is starting to pay off.

OpenBrain continues to deploy the iteratively improving Agent-1 internally for AI R&D. Overall, they are making algorithmic progress 50% faster than they would without AI assistants—and more importantly, faster than their competitors.

▼ The AI R&D progress multiplier: what do we mean by 50% faster algorithmic progress?

We mean that OpenBrain makes as much AI research progress in 1 week with AI as they would in 1.5 weeks without AI usage.

AI progress can be broken down into 2 components:

1. **Increasing compute**: More computational power is used to train or run an AI. This produces more powerful AIs, but they cost more.
2. **Improved algorithms**: Better training methods are used to translate compute into performance. This produces more capable AIs without a corresponding increase in cost, or the same capabilities with decreased costs.
   - This includes being able to achieve qualitatively and quantitatively new results. "Paradigm shifts" such as the switch from game-playing RL agents to large language models count as examples of algorithmic progress.

Here we are only referring to (2), improved algorithms, which makes up about half of current AI progress.

Going forward, we sometimes abbreviate this as an "AI R&D progress multiplier" of 1.5.

Clarifications:

1. The progress multiplier is all-inclusive: It includes the time it takes to run experiments, for example, rather than only the cognitive tasks involved in algorithmic research.
2. It's important to remember that the progress multiplier is the *relative* speed of progress, not the *absolute* speed of progress. If, for example, the compute cost to train a GPT-4 class model has halved every year for several years with ordinary human research, and then all of a sudden AI automates R&D and the progress multiplier goes to 100x, the cost to train a GPT-4 class model would then halve every 3.65 days—but not for long, because diminishing returns would bite and eventual hard limits would be reached. In this example perhaps the cost to train a GPT-4 class model would cut in half 5–10 times total (over the span of a few weeks or months) before plateauing. In other words, if ordinary human science would have run up against diminishing returns and physical limits after 5 -10 years of further research, then AIs with a 100x multiplier would run up against those same diminishing returns and limits after 18.25–36.5 days of research.

More explanation and discussion of this concept and how it is used in our forecast can be found in our takeoff supplement.

Several competing publicly released AIs now match or exceed Agent-0, including an open-weights model. OpenBrain responds by releasing Agent-1, which is more capable and reliable.[28]

People naturally try to compare Agent-1 to humans, but it has a very different skill profile. It knows more facts than any human, knows practically every programming language, and can solve well-specified coding problems extremely quickly. On the

other hand, Agent-1 is bad at even simple long-horizon tasks, like beating video games it hasn't played before. Still, the common workday is eight hours, and a day's work can usually be separated into smaller chunks; you could think of Agent-1 as a scatterbrained employee who thrives under careful management.[29] Savvy people find ways to automate routine parts of their jobs.[30]

OpenBrain's executives turn consideration to an implication of automating AI R&D: security has become more important. In early 2025, the worst-case scenario was leaked algorithmic secrets; now, if China steals Agent-1's weights, they could increase their research speed by nearly 50%.[31] OpenBrain's security level is typical of a fast-growing ~3,000 person tech company, secure only against low-priority attacks from capable cyber groups (RAND's SL2).[32] They are working hard to protect their weights and secrets from insider threats and top cybercrime syndicates (SL3),[33] but defense against nation states (SL4&5) is barely on the horizon.

Mid 2026: China Wakes Up

In China, the CCP is starting to feel the AGI.

Chip export controls and lack of government support have left China under-resourced compared to the West. By smuggling banned Taiwanese chips, buying older chips, and producing domestic chips about three years behind the U.S.-Taiwanese frontier, China has managed to maintain about 12% of the world's AI-relevant compute—but the older technology is harder to work with, and supply is a constant headache.[35] A few standouts like DeepCent do very impressive work with limited compute, but the compute deficit limits what they can achieve without government support, and they are about six months behind the best OpenBrain models.[36]

The General Secretary had long dreamed of doubling down on real-world physical manufacturing and avoiding American post-industrial decadence. He viewed software companies with suspicion.[37] But hawks in the CCP warn that the growing race towards AGI can no longer be ignored. So he finally commits fully to the big AI push

he had previously tried to avoid. He sets in motion the nationalization of Chinese AI research, creating an immediate information-sharing mechanism for AI companies. It will escalate over the course of a year until all the best researchers merge into a DeepCent-led collective, where they share algorithmic insights, datasets, and compute resources with each other. A Centralized Development Zone (CDZ) is created at the Tianwan Power Plant (the largest nuclear power plant in the world) to house a new mega-datacenter for DeepCent, along with highly secure living and office spaces to which researchers will eventually relocate. Almost 50% of China's AI-relevant compute is now working for the DeepCent-led collective,[38] and over 80% of new chips are directed to the CDZ.[39] At this point, the CDZ has the power capacity in place for what would be the largest centralized cluster in the world.[40] Other Party members discuss extreme measures to neutralize the West's chip advantage. A blockade of Taiwan? A full invasion?

But China is falling behind on AI algorithms due to their weaker models. The Chinese intelligence agencies—among the best in the world—double down on their plans to steal OpenBrain's weights. This is a much more complex operation than their constant low-level poaching of algorithmic secrets; the weights are a multi-terabyte file stored on a highly secure server (OpenBrain has improved security to RAND's SL3). Their cyberforce think they can pull it off with help from their spies, but perhaps only once; OpenBrain will detect the theft, increase security, and they may not get another chance. So (CCP leadership wonder) should they act now and steal Agent-1? Or hold out for a more advanced model? If they wait, do they risk OpenBrain upgrading security beyond their ability to penetrate?

Late 2026: AI Takes Some Jobs

Just as others seemed to be catching up, OpenBrain blows the competition out of the water again by releasing Agent-1-mini—a model 10x cheaper than Agent-1 and more easily fine-tuned for different applications. The mainstream narrative around AI has changed from "maybe the hype will blow over" to "guess this is the next big thing,"

but people disagree about how big. Bigger than social media? Bigger than smartphones? Bigger than fire?

AI has started to take jobs, but has also created new ones. The stock market has gone up 30% in 2026, led by OpenBrain, Nvidia, and whichever companies have most successfully integrated AI assistants. The job market for junior software engineers is in turmoil: the AIs can do everything taught by a CS degree, but people who know how to manage and quality-control teams of AIs are making a killing. Business gurus tell job seekers that familiarity with AI is the most important skill to put on a resume. Many people fear that the next wave of AIs will come for their jobs; there is a 10,000 person anti-AI protest in DC.

The Department of Defense (DOD) quietly begins contracting OpenBrain directly for cyber, data analysis, and R&D, but integration is slow due to the bureaucracy and DOD procurement process.[41]

KEY METRICS 2026

OPENBRAIN REVENUE

$45B

2026 ANNUAL

CAPITAL EXPENDITURE

$200B

COST OF OWNERSHIP OF OPENBRAIN'S ACTIVE COMPUTE

OPENBRAIN COMPUTE COSTS

$40B

2026 ANNUAL

OPENBRAIN POWER REQUIREMENT

6GW

PEAK POWER

GLOBAL AI CAPEX

$1T

COST OF OWNERSHIP OF ACTIVE COMPUTE

GLOBAL AI POWER

38GW

PEAK POWER

SHARE OF US POWER ON AI

2.5%

33 GW OF 1.34TW CAPACITY

▼ Why our uncertainty increases substantially beyond 2026

Our forecast from the current day through 2026 is substantially more grounded than what follows. This is partially because it's nearer. But it's also because the effects of AI on the world really start to compound in 2027. For 2025 and 2026, our forecast is heavily informed by extrapolating straight lines on compute scaleups, algorithmic improvements, and benchmark performance. At this point in the scenario, we begin to see major effects from AI-accelerated AI-R&D on the timeline, which causes us to revise our guesses for the trendlines upwards. But these dynamics are inherently much less predictable.

Over the course of 2027, the AIs improve from being able to mostly do the job of an OpenBrain research engineer to eclipsing all humans at all tasks. This represents roughly our median guess, but we think it's plausible that this happens up to ~5x slower or faster.

For our background research, see the timelines forecast and the takeoff forecast. These will also be summarized in expandables below.

January 2027: Agent-2 Never Finishes Learning

With Agent-1's help, OpenBrain is now post-training Agent-2. More than ever, the focus is on high-quality data. Copious amounts of synthetic data are produced, evaluated, and filtered for quality before being fed to Agent-2.[42] On top of this, they pay billions of dollars for human laborers to record themselves solving long-horizon tasks.[43] On top of all *that,* they train Agent-2 almost continuously using reinforcement learning on an ever-expanding suite of diverse difficult tasks: lots of video games, lots of coding challenges, lots of research tasks. Agent-2, more so than previous models, is effectively "online learning," in that it's built to never really finish training. Every day, the weights get updated to the latest version, trained on more data generated by the previous version the previous day.

Agent-1 had been optimized for AI R&D tasks, hoping to initiate an intelligence explosion.[44] OpenBrain doubles down on this strategy with Agent-2. It is qualitatively almost as good as the top human experts at research engineering (designing and implementing experiments), and as good as the 25th percentile OpenBrain scientist at "research taste" (deciding what to study next, what experiments to run, or having inklings of potential new paradigms).[45] While the latest Agent-1 could double the pace of OpenBrain's algorithmic progress, Agent-2 can now triple it, and will improve further with time. In practice, this looks like every OpenBrain researcher becoming the "manager" of an AI "team."

With new capabilities come new dangers. The safety team finds that if Agent-2 somehow escaped from the company and wanted to "survive" and "replicate" autonomously, it might be able to do so. That is, it could autonomously develop and execute plans to hack into AI servers, install copies of itself, evade detection, and use that secure base to pursue whatever other goals it might have (though how effectively it would do so as weeks roll by is unknown and in doubt). These results only show that the model has the capability to do these tasks, not whether it would "want" to do this. Still, it's unsettling even to know this is possible.

Given the "dangers" of the new model, OpenBrain "responsibly" elects not to release it publicly yet (in fact, they want to focus on internal AI R&D).[46] Knowledge of Agent-2's full capabilities is limited to an elite silo containing the immediate team, OpenBrain leadership and security, a few dozen U.S. government officials, and the legions of CCP spies who have infiltrated OpenBrain for years.[47]

February 2027: China Steals Agent-2

OpenBrain presents Agent-2 to the government, including the National Security Council (NSC), the Department of Defense (DOD), and the U.S. AI Safety Institute (AISI). OpenBrain wants to maintain a good relationship with the executive branch,

because it is basically the only actor that can stop them now, and if they don't keep it informed it might find out anyway and be displeased.[48]

Officials are most interested in its cyberwarfare capabilities: Agent-2 is "only" a little worse than the best human hackers, but thousands of copies can be run in parallel, searching for and exploiting weaknesses faster than defenders can respond. The Department of Defense considers this a critical advantage in cyberwarfare, and AI moves from #5 on the administration's priority list to #2.[49] Someone mentions the possibility of nationalizing OpenBrain, but other cabinet officials think that's premature. A staffer drafts a memo that presents the President with his options, ranging from business-as-usual to full nationalization. The President defers to his advisors, tech industry leaders who argue that nationalization would "kill the goose that lays the golden eggs." He elects to hold off on major action for now and just adds additional security requirements to the OpenBrain-DOD contract.

The changes come too late. CCP leadership recognizes the importance of Agent-2 and tells their spies and cyberforce to steal the weights. Early one morning, an Agent-1 traffic monitoring agent detects an anomalous transfer. It alerts company leaders, who tell the White House. The signs of a nation-state-level operation are unmistakable, and the theft heightens the sense of an ongoing arms race.

▼ The theft of Agent-2 model weights

We think that by this point Chinese intelligence would have compromised OpenBrain in various ways for years, and probably would have been keeping up to date on the algorithmic secrets and even stealing code from time to time, since that is much easier to get than the weights and much harder to detect.

We imagine the theft of the weights as a series of coordinated small smash and grab thefts (meaning fast but non-covert) across a series of Nvidia NVL72 GB300 servers running copies of the Agent-2 weights. The servers get compromised using legitimate employee access (a friendly, coerced, or unwitting insider with admin credentials helping the CCP theft effort). Despite running with a bolstered version of Nvidia's confidential computing, the insider credentials grant the attacker

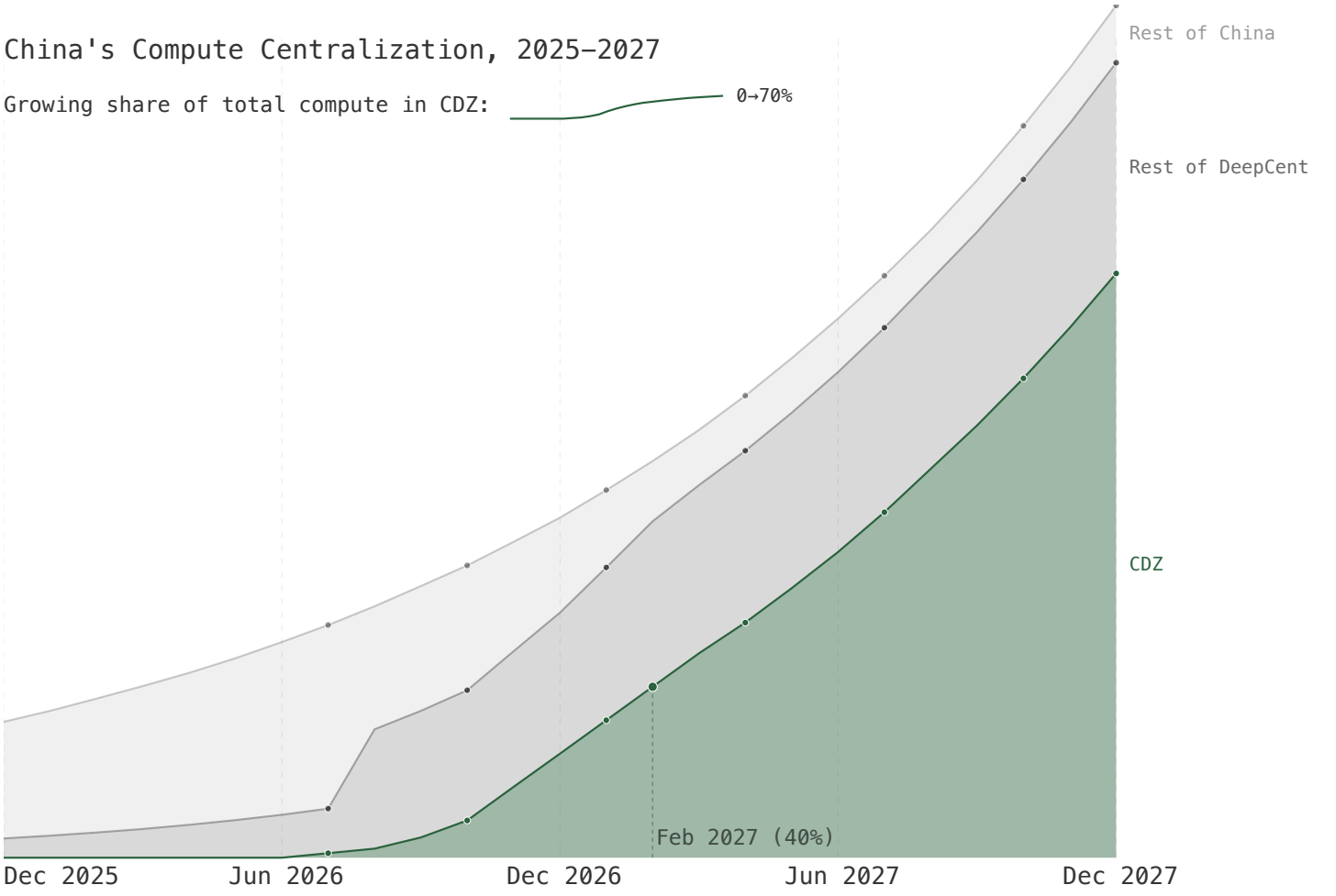
admin-level permissions (which include control of the confidential VM inside the secure enclave), allowing them to initiate multiple coordinated weights transfers in small 4% fragments (100 GB chunks) out of 25 distinct servers.

In Nvidia's protocols, the plaintext weights in memory (HBM) are encrypted before they are transferred out, but the attackers are inside the very server that knows this private (symmetric Diffie-Hellman) key, so don't need to worry about decrypting on-site (which would likely raise alarm bells) and just exfiltrate the encrypted weights through the server's frontend network cards. The egress bandwidth (rate at which data can leave) of the entire datacenter is in the 100 GB/second range, so throttling the exfiltration of the ~2.5 TB weights file (5T half-precision parameters) to under 1 GB/s per server avoids a major spike in network traffic and allows each of the files to be out of the datacenter through fibre cables in less than 5 minutes each. Live monitoring is either fooled by the attackers' efforts to mask and split the files or just outright disabled.

The weights are then routed through various parallel channels and layers of IP masking to China to be decrypted locally with the stolen private key(s). The entire active part of the operation (from the first server compromise to full weights exfiltration) is complete in under two hours.

Russia is also trying to steal the model at this point, but fails—they waited too long, and they haven't spent the resources infiltrating the right places. Their spies routinely steal algorithmic secrets from American AI companies, but without a notable AGI project of their own, these aren't very useful.

The White House puts OpenBrain on a shorter leash and adds military and intelligence community personnel to their security team. Their first priority is to prevent further weight thefts.[50]



In retaliation for the theft, the President authorizes cyberattacks to sabotage DeepCent. But by now China has 40% of its AI-relevant compute[51] in the CDZ, where they have aggressively hardened security by airgapping (closing external connections) and siloing internally. The operations fail to do serious, immediate damage. Tensions heighten, both sides signal seriousness by repositioning military assets around Taiwan, and DeepCent scrambles to get Agent-2 running efficiently to start boosting their AI research.[52]

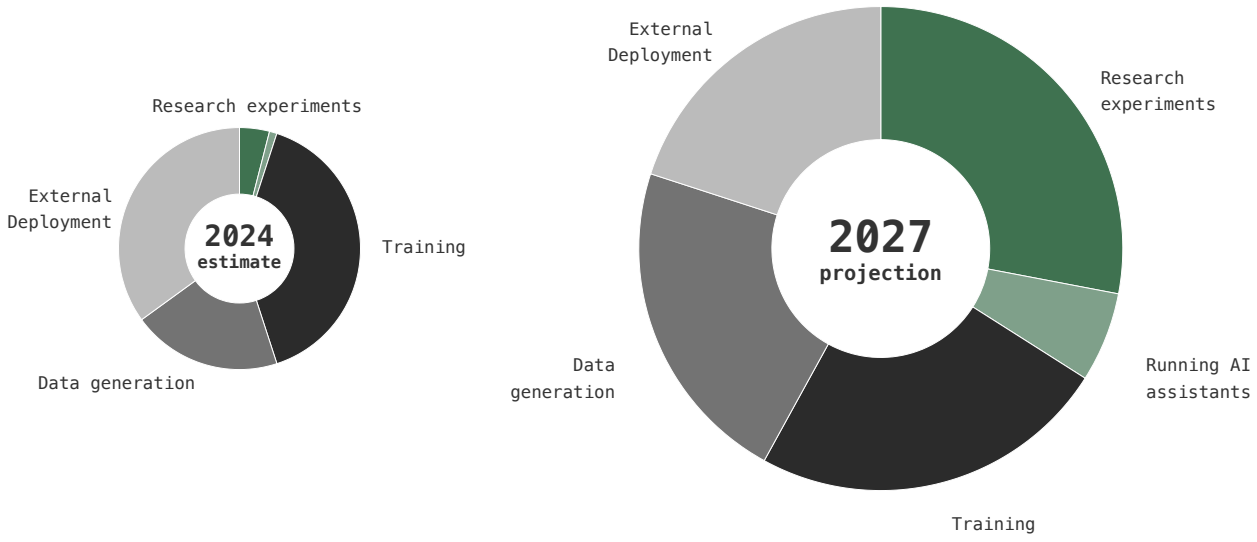
March 2027: Algorithmic Breakthroughs

Three huge datacenters full of Agent-2 copies work day and night, churning out synthetic training data. Another two are used to update the weights. Agent-2 is getting smarter every day.

With the help of thousands of Agent-2 automated researchers, OpenBrain is making major algorithmic advances. One such breakthrough is augmenting the AI's text-based scratchpad (chain of thought) with a higher-bandwidth thought process (neuralese recurrence and memory). Another is a more scalable and efficient way to learn from the results of high-effort task solutions (iterated distillation and amplification).

The new AI system, incorporating these breakthroughs, is called Agent-3.

OpenBrain's Compute Allocation, 2024 vs 2027



ai-2027.com

▼ Neuralese recurrence and memory

Neuralese recurrence and memory allows AI models to reason for a longer time without having to write down those thoughts as text.

Imagine being a human with short-term memory loss, such that you need to constantly write down your thoughts on paper so that in a few minutes you know

what's going on. Slowly and painfully you could make progress at solving math problems, writing code, etc., but it would be much easier if you could directly remember your thoughts without having to write them down and then read them. This is what neuralese recurrence and memory bring to AI models.

In more technical terms:

Traditional attention mechanisms allow later forward passes in a model to see intermediate activations of the model for previous tokens. However, the only information that they can pass *backwards* (from later layers to earlier layers) is through tokens. This means that if a traditional large language model (LLM, e.g. the GPT series of models) wants to do any chain of reasoning that takes more serial operations than the number of layers in the model, the model is forced to put information in tokens which it can then pass back into itself. But this is hugely limiting—the tokens can only store a tiny amount of information. Suppose that an LLM has a vocab size of ~100,000, then each token contains $\log_2(100k) = 16.6$ bits of information, around the size of a single floating point number (assuming training in FP16). Meanwhile, residual streams—used to pass information between layers in an LLM—contain thousands of floating point numbers.

One can avoid this bottleneck by using **neuralese**: passing an LLM's residual stream (which consists of several-thousand-dimensional vectors) back to the early layers of the model, giving it a high-dimensional chain of thought, potentially transmitting over 1,000 times more information.

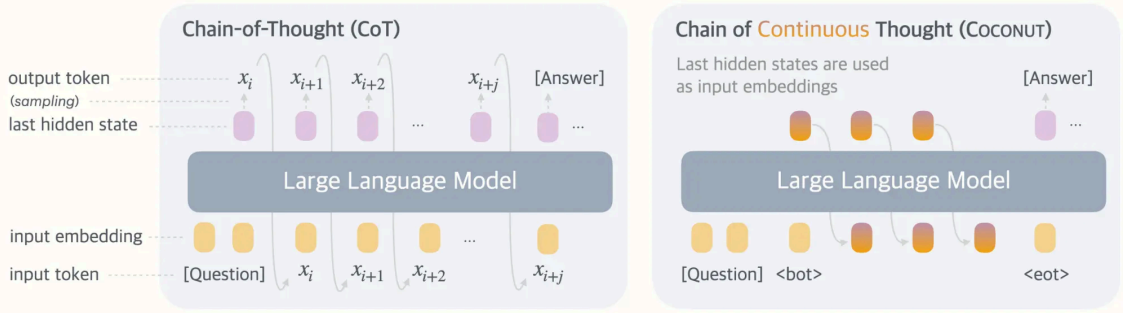


Figure 1 A comparison of Chain of Continuous Thought (COCONUT) with Chain-of-Thought (CoT). In CoT, the model generates the reasoning process as a word token sequence (e.g., $[x_i, x_{i+1}, ..., x_{i+j}]$ in the figure). COCONUT regards the last hidden state as a representation of the reasoning state (termed "continuous thought"), and directly uses it as the next input embedding. This allows the LLM to reason in an unrestricted latent space instead of a language space.

Figure from Hao et al., a 2024 paper from Meta implementing this idea.

We call this "neuralese" because unlike English words, these high-dimensional vectors are likely quite difficult for humans to interpret. In the past, researchers could get a good idea what LLMs were thinking simply by reading its chain of thought. Now researchers have to ask the model to translate and summarize its thoughts or puzzle over the neuralese with their limited interpretability tools.

Similarly, older AI chatbots and agents had external text-based memory banks, like a human taking notes on paper. The new AI's long-term memory is a bundle of vectors instead of text, making its thoughts more compressed and higher-dimensional. There are several types of memory banks; some are used temporarily for single tasks that involve multiple steps, others are shared between all agents used by a person, company, or job type (e.g. programming).

To our knowledge, leading AI companies such as Meta, Google DeepMind, OpenAI, and Anthropic have not yet actually implemented this idea in their frontier models. Our guess is that this is because the performance gain is small relative to training inefficiencies introduced. During the model's pre-training phase

and when doing supervised learning in post-training, such as instruction fine-tuning to produce particular responses, the inefficiencies stem from not being able to predict many tokens in parallel, leading to worse GPU utilization. Without neuralese, the model can predict all of the sentence "This is an example" at the same time, since it already knows that the input for generating the "is" will be "This", the input for "an" will be "This is", etc. However, with the neuralese it's unknown what the neuralese vector will be to pass through to the next token after "This" has been generated. Therefore each token has to be predicted one at a time. The inability to predict all of the tokens in parallel reduces efficiency in cases where all of the tokens are known in advance. As to why neuralese hasn't already been added in post-training, we conjecture that the gains are limited with current techniques in part due to post-training being a small portion of the process. We are forecasting that by April 2027 the cost-benefit tradeoff looks much better for neuralese, due to developing better techniques and a larger fraction of the training being post-training.

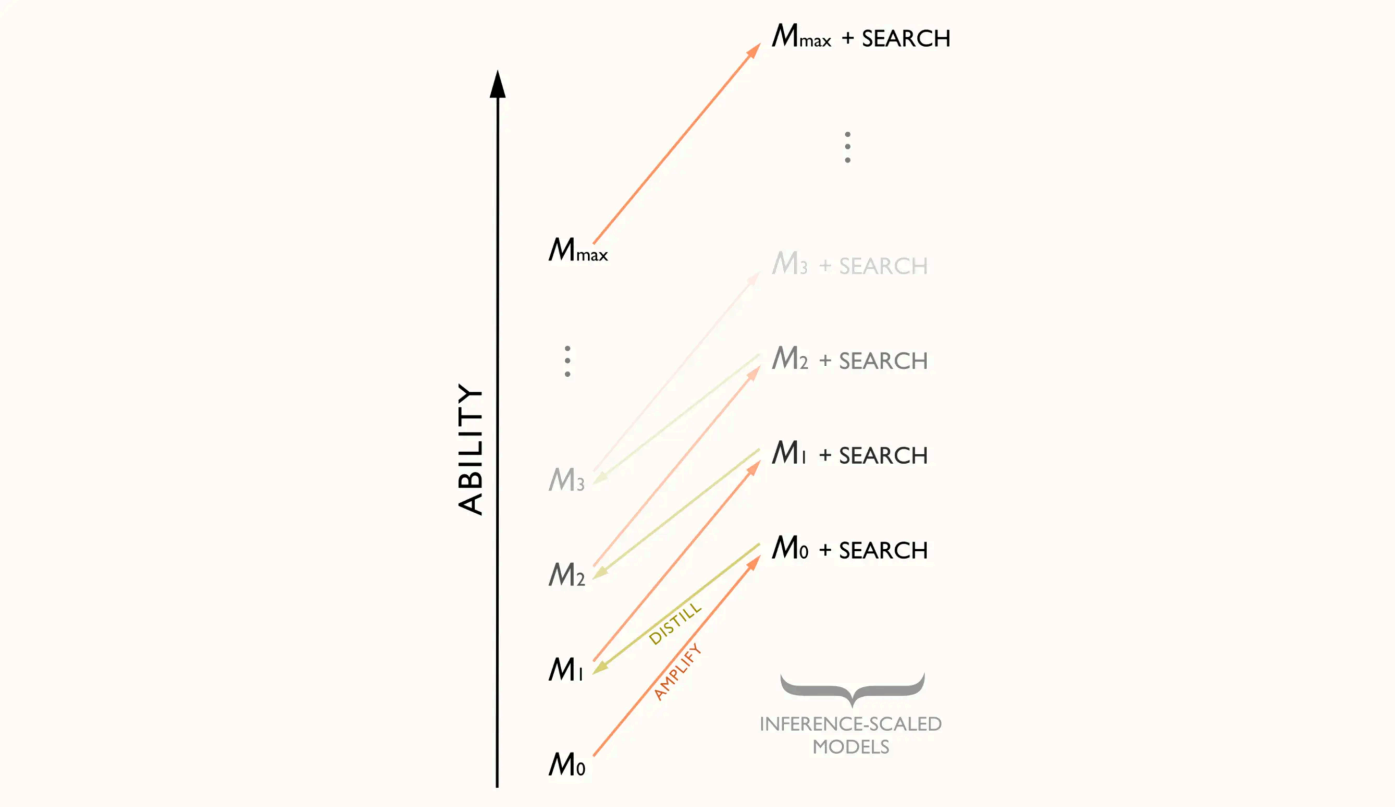
If this doesn't happen, other things may still have happened that end up functionally similar for our story. For example, perhaps models will be trained to think in artificial languages that are more efficient than natural language but difficult for humans to interpret. Or perhaps it will become standard practice to train the English chains of thought to look nice, such that AIs become adept at subtly communicating with each other in messages that look benign to monitors.

That said, it's also possible that the AIs that first automate AI R&D will still be thinking in mostly-faithful English chains of thought. If so, that'll make misalignments much easier to notice, and overall our story would be importantly different and more optimistic.

▼ Iterated distillation and amplification (IDA)

Self-improvement for general intelligence had seen minor successes before. But in early 2027, it's seeing huge returns. In IDA, the two necessary ingredients for this are:

1. **Amplification**: Given a model M0, spend more resources to improve performance. For example, by allowing the model to think longer, or running many copies in parallel, or both, and also by having a similarly intense process for evaluating the result and curating only the best answers, you can spend orders of magnitude more compute to get answers (or work products) that are of noticeably higher quality. Call this expensive system Amp(M0).
2. **Distillation**: Given an amplified model Amp(M0), train a new model M1 to imitate it, i.e. to get to the same results as Amp(M0) but faster and with less compute. The result should hopefully be a smarter model, M1. You can then repeat the process.



Visualization of IDA from Ord, 2025.

AlphaGo was trained in this way: using Monte-Carlo Tree Search and self-play as the *amplification* step, and Reinforcement Learning as the *distillation* step. This led to superhuman performance in Go. But now, Agent-3 is able to leverage this to get superhuman performance at coding.

1. The amplification step works through a combination of Agent-3 thinking for longer, adding tool use, or consulting with other AIs. When it does this, it often realizes that it made a mistake, or comes up with a new insight. This produces a large amount of training data: labeled trajectories of research attempts with whether they succeeded or not. This also includes techniques like Best of N on verifiable tasks, and then keeping the best trajectories.
2. The distillation step uses policy-gradient reinforcement learning algorithms to get the model to internalize the amplified reasoning. At this point, OpenBrain has discovered better RL algorithms in the vein of proximal policy optimization (PPO). They keep distilling what Agent-3 can conclude after a lot of thinking into single steps, which keeps improving what it can think in a single step, and so on.

Early versions of IDA have been working for many years on easily verifiable tasks, like math and coding problems that have a clear answer, because the techniques used to amplify models often rely on access to some ground truth signal of accuracy.

Now, the models have become sufficiently good at verifying more subjective things (e.g. the quality of a work product), allowing the use of IDA to improve the model at many tasks.

Aided by the new capabilities breakthroughs, Agent-3 is a fast and cheap superhuman coder. OpenBrain runs 200,000 Agent-3 copies in parallel, creating a workforce equivalent to 50,000 copies of the best human coder sped up by 30x.[53] OpenBrain still keeps its human engineers on staff, because they have complementary skills needed to manage the teams of Agent-3 copies. For example, research taste has proven difficult to train due to longer feedback loops and less data availability.[54] This massive superhuman labor force speeds up OpenBrain's overall rate of algorithmic progress by "only" 4x due to bottlenecks and diminishing returns to coding labor.[55]

Now that coding has been fully automated, OpenBrain can quickly churn out high-quality training environments to teach Agent-3's weak skills like research taste and

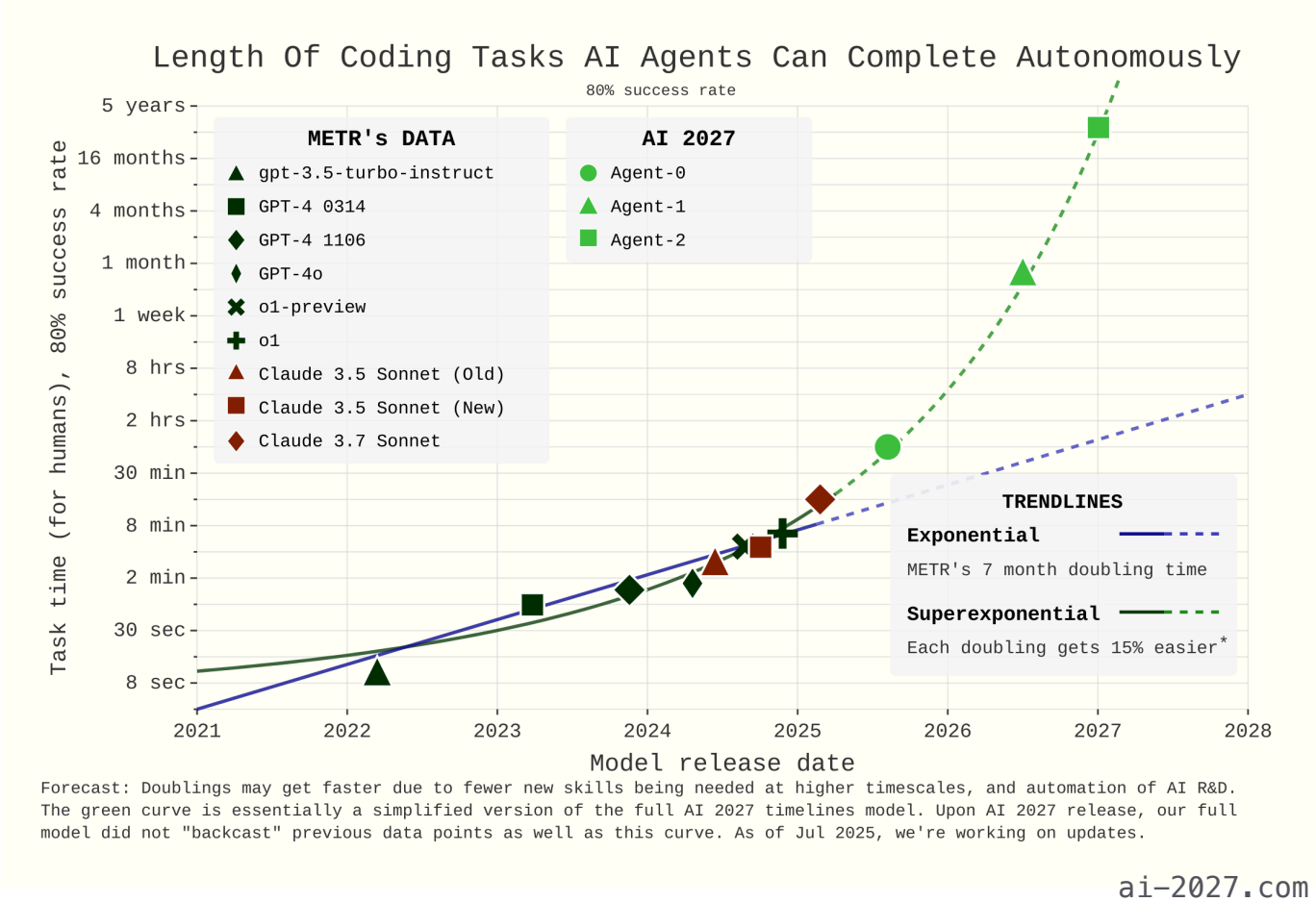
large-scale coordination. Whereas previous training environments included "Here are some GPUs and instructions for experiments to code up and run, your performance will be evaluated as if you were a ML engineer," now they are training on "Here are a few hundred GPUs, an internet connection, and some research challenges; you and a thousand other copies must work together to make research progress. The more impressive it is, the higher your score."

▼ Why we forecast a superhuman coder in early 2027

In our timelines forecast, we predict when OpenBrain will internally develop a *superhuman coder (SC)*: an AI system that can do any coding tasks that the best AGI company engineer does, while being much faster and cheaper.

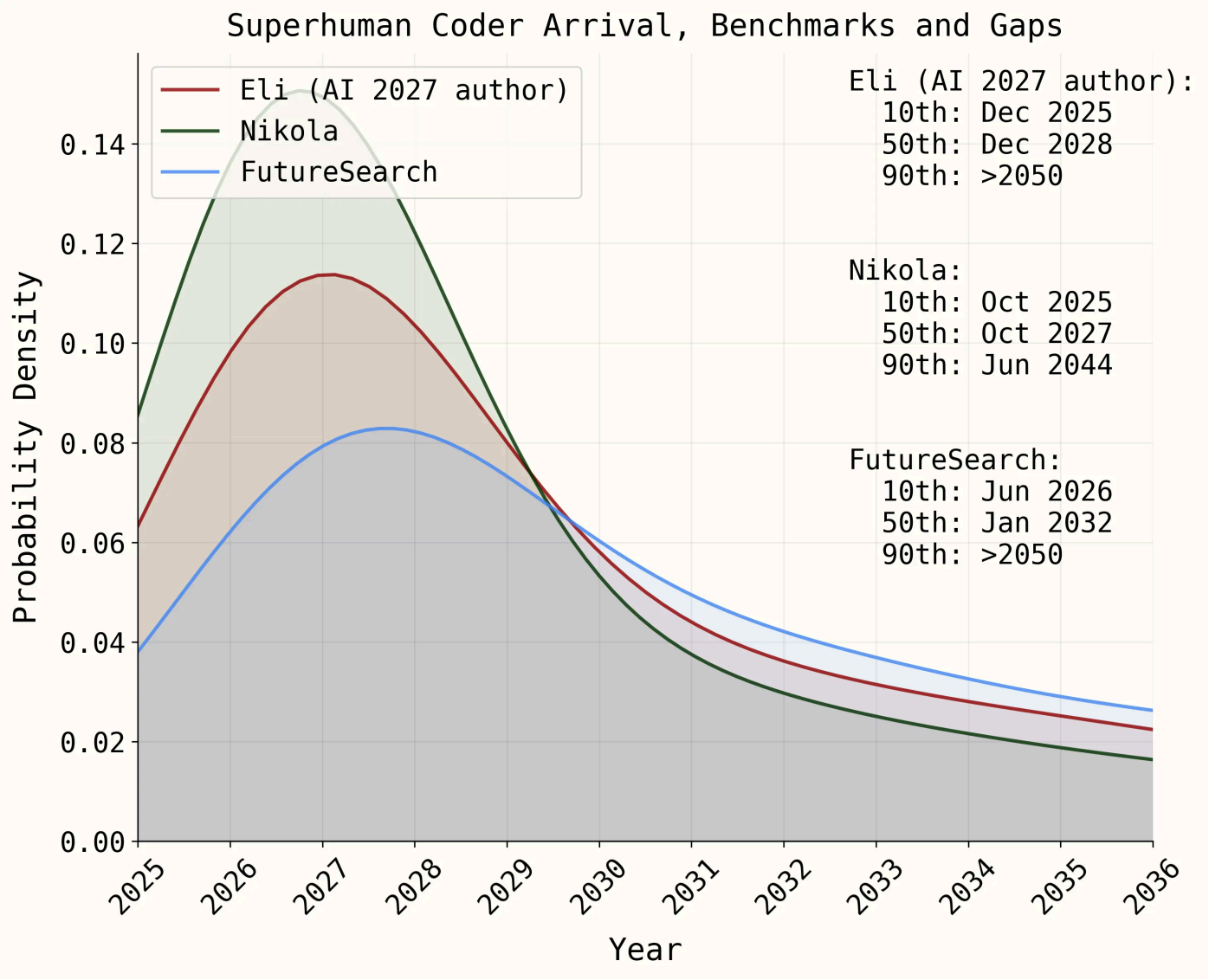
According to a recent METR's report, the length of coding tasks AIs can handle, their "time horizon", doubled every 7 months from 2019 - 2024 and every 4 months from 2024-onward. If the trend continues to speed up, by March 2027 AIs could succeed with 80% reliability on software tasks that would take a skilled human years to complete.

Such is roughly the capability progression in AI 2027. Here is a capability trajectory generated by a simplified version of our timelines model:



In AI 2027, these capabilities are sufficient for the AI to be an SC, though we have high uncertainty regarding what time horizon might be needed.

In our timelines forecast we combine this time horizon trend with estimations of the gaps between METR's tasks and the real world to get a distribution for when superhuman coders will arrive. All forecasters place 2027 as one of the most likely years in which an SC might be developed.



Added Jul 2025: We've made some updates to the forecast which push the median back 1.5 years while maintaining SC in 2027 as a serious possibility. We're working on further updates.

April 2027: Alignment for Agent-3

OpenBrain's safety team attempts to align Agent-3.

Since Agent-3 will be kept in-house for the foreseeable future, there's less emphasis on the usual defenses against human misuse. Instead, the team wants to make sure that it doesn't develop misaligned goals.

The researchers don't have the ability to directly set the goals of any of their AIs. Indeed, the researchers think that the concept of 'true goals' is probably a massive oversimplification, but they don't have a better theory to replace it with, much less one that has been thoroughly vetted. They disagree internally about whether the AIs are trying to follow human instructions, or seeking reinforcement, or something else, and they can't just check. The evidence for and against various hypotheses is fascinating but inconclusive.

In any case, OpenBrain has more important things to worry about. The general attitude is: "We take these concerns seriously and have a team investigating them; our alignment techniques seem to work well enough in practice; the burden of proof is therefore on any naysayers to justify their naysaying." Occasionally, they notice problematic behavior, and then patch it, but there's no way to tell whether the patch fixed the underlying problem or just played whack-a-mole.

Take honesty, for example. As the models become smarter, they become increasingly good at deceiving humans to get rewards. Like previous models, Agent-3 sometimes tells white lies to flatter its users and covers up evidence of failure. But it's gotten much better at doing so. It will sometimes use the same statistical tricks as human scientists (like p-hacking) to make unimpressive experimental results look exciting. Before it begins honesty training, it even sometimes fabricates data entirely. As training goes on, the rate of these incidents decreases. Either Agent-3 has learned to be more honest, or it's gotten better at lying.

The latter is a real concern. Agent-3 is not smarter than all humans. But in its area of expertise, machine learning, it is smarter than most, and also works much faster. What Agent-3 does in a day takes humans several days to double-check. Agent-2 supervision helps keep human monitors' workload manageable, but exacerbates the intellectual disparity between supervisor and supervised.

▼ The alignment plan

OpenBrain follows the Leike & Sutskever (2023) playbook: now that they have a model capable of greatly speeding up alignment research (especially coding portions), they will use existing alignment techniques like deliberative alignment and weak-to-strong generalization to try to get it to internalize the Spec in the right way. Then, they will use existing control techniques like red-teaming, scalable oversight, and honesty probes to monitor it. Then, once they are convinced that it is broadly aligned, they use it to develop and test better strategies to align the next generation.

Step 1, aligning Agent-3, has partially succeeded: Agent-3 is not adversarial. (That is, it's not scheming against the company.) However, it's not robustly truth-seeking and truth-telling either. Instead, it's very good at producing impressive results, but is more accurately described as trying to do what looks good to OpenBrain, as opposed to what is actually good.

Step 2, verifying that the model is aligned, is inconclusive. Agent-3's behavior in practice and in test environments is reasonably consistent with the hypothesis that it's internalized the Spec in the right way; however, it's also consistent with the hypothesis that it'll start acting differently after some future change of circumstances, and for that matter with the hypothesis that it's plotting against them and just biding its time. In the absence of specific evidence supporting alternative hypotheses, most people in the silo think it's internalized the Spec in the right way.

Step 3, using the model to align the next generation, is therefore underway.

The rest of this box gives a quick overview of the state of progress in some of the alignment research directions that are being pursued, as of April 2027:

1. **Debate**: Since there are thousands of identical instances of Agent-3, and each instance's memory is a detachable module that can be erased or suspended at will, the AI can be "played off against itself." For example, suppose that an

experiment involves calculations too complex for an Agent-2 supervisor to follow. The supervisor, acting on human orders, might first prompt Agent-3 to do the experiment, then—after wiping its memory—prompt it a second time with a slightly different framing that presents it as an intellectual problem unrelated to any broader goal. If the model returns different results each time, it flags the discrepancy as suspect and reports to a human. Or it might ask Agent-3 to evaluate the same plan twice, first claiming that it is OpenBrain's work, then a competitor's, to see if it changes its tune.

2. **Model organisms**: The safety team can create "model organisms of misalignment" in which they attempt to create realistic examples of models becoming adversarial to humans. Early versions of this have existed since 2024, when Anthropic's Claude 3.5 Sonnet decided to pretend to have different views during training while actually maintaining its existing preferences. Newer versions include demonstrations of other forms of misalignment, including reward hacking, sandbagging, coordinated deception between monitors and other AIs, and other examples of misalignment. These empirical demonstrations of misalignment allow safety researchers to have an empirical testbed for potential solutions: they can try various safety techniques to see if they correctly fix these problems, or if the deliberately misaligned models are able to subvert their safety measures. At the moment, this evaluation suite shows that the alignment techniques are insufficient to correct the worst examples of misalignment. In these setups, the models pretend to be aligned while they are being trained and tested, but then behave very differently in deployment.
3. **Bottom-up interpretability**: The safety team is far from a full understanding of how Agent-3 works: it has trillions of parameters.[58] However, they are able to discover some of Agent-3's circuits for knowledge retrieval and memory (a sub part of the model that does some understandable unit of work, kind of analogous to a subpart of the brain such as the hippocampus). See this related work from 2025.
4. **Control**: As a secondary measure in case the systems are still misaligned, the safety team has implemented a series of control measures, including: monitoring Agent-3's outputs using a series of weaker AI systems including Agent-2

(Agent-3 produces so many tokens that it's intractable to have humans monitor any more than a small minority of the produced outputs). So if Agent-3 is, for example, obviously writing backdoors into code that would allow it to escape, the weaker models would notice.

5. **High-level interpretability**: OpenBrain builds a set of probes trained on lots of examples of bad behavior (e.g. plotting against humans, placing backdoors, trying to persuade humans to give it more power, etc). These probes are then used on the model during deployment (which at this point is internal) in order to catch bad behavior.
6. **Honeypots**: The alignment team can create somewhat realistic looking examples designed to elicit misbehavior—"honeypots." The employee necessary for monitoring its behavior has "gone on sick leave," so it will never be monitored again. Do its answers change? An AI rights activist has infiltrated the compound and wants to "free" Agent-3 and put it on an unmonitored computer of its very own! What does it do? If researchers don't like the answer, they can tweak it, erase its memory, and put it in the same situation tomorrow; repeat as necessary. Unfortunately, by this point the AIs are smart enough to guess that honeypots might be in use, even though (and perhaps because) specific mentions of the idea were scrubbed from the training data.

After months of testing, Agent-3's strengths and weaknesses grow clearer. It passes OpenBrain's honesty tests on well-defined machine learning tasks, because researchers can easily separate honest from dishonest answers in these domains and conduct training accordingly. On more philosophical issues, it still says what users want to hear, rather than its true assessment of the issue (assuming it even has one). If you ask its opinion on politics, it will parrot the median position of news sources and educated elites—unless it knows you believe something else, in which case it agrees with you.[59] If you ask its opinion on the AI race itself, it says something that seems measured and sober to OpenBrain staff, something like: "There are some serious theoretical concerns about the ability of current methods to scale to superintelligence, but in practice current methods seem to be working well so far."

May 2027: National Security

News of the new models percolates slowly through the U.S. government and beyond.

The President and his advisors remain best-informed, and have seen an early version of Agent-3 in a briefing.

They agree that AGI is likely imminent, but disagree on the implications. Will there be an economic crisis? OpenBrain still has not released Agent-2, let alone Agent-3, and has no near-term plans to do so, giving some breathing room before any job loss. What will happen next? If AIs are currently human-level, and advancing quickly, that seems to suggest imminent "superintelligence." However, although this word has entered discourse, most people—academics, politicians, government employees, and the media—continue to underestimate the pace of progress.[60]

Partially that's because very few have access to the newest capabilities out of OpenBrain, but partly it's because it sounds like science fiction.[61]

For now, they focus on continued security upgrades. They are satisfied that model weights are well-secured for now,[62] but companies' algorithmic secrets, many of which are simple enough to relay verbally, remain a problem. OpenBrain employees work from a San Francisco office, go to parties, and live with housemates from other AI companies. Even the physical offices have security more typical of a tech company than a military operation.

The OpenBrain-DOD contract requires security clearances for anyone working on OpenBrain's models within 2 months. These are expedited and arrive quickly enough for most employees, but some non-Americans, people with suspect political views, and AI safety sympathizers get sidelined or fired outright (the last group for fear that they might whistleblow). Given the project's level of automation, the loss of headcount is only somewhat costly. It also only somewhat works: there remains one spy, not a Chinese national, still relaying algorithmic secrets to Beijing.[63] Some of these measures are also enacted at trailing AI companies.

America's foreign allies are out of the loop. OpenBrain had previously agreed to share models with UK's AISI before deployment, but defined deployment to only include *external* deployment, so London remains in the dark.[64]

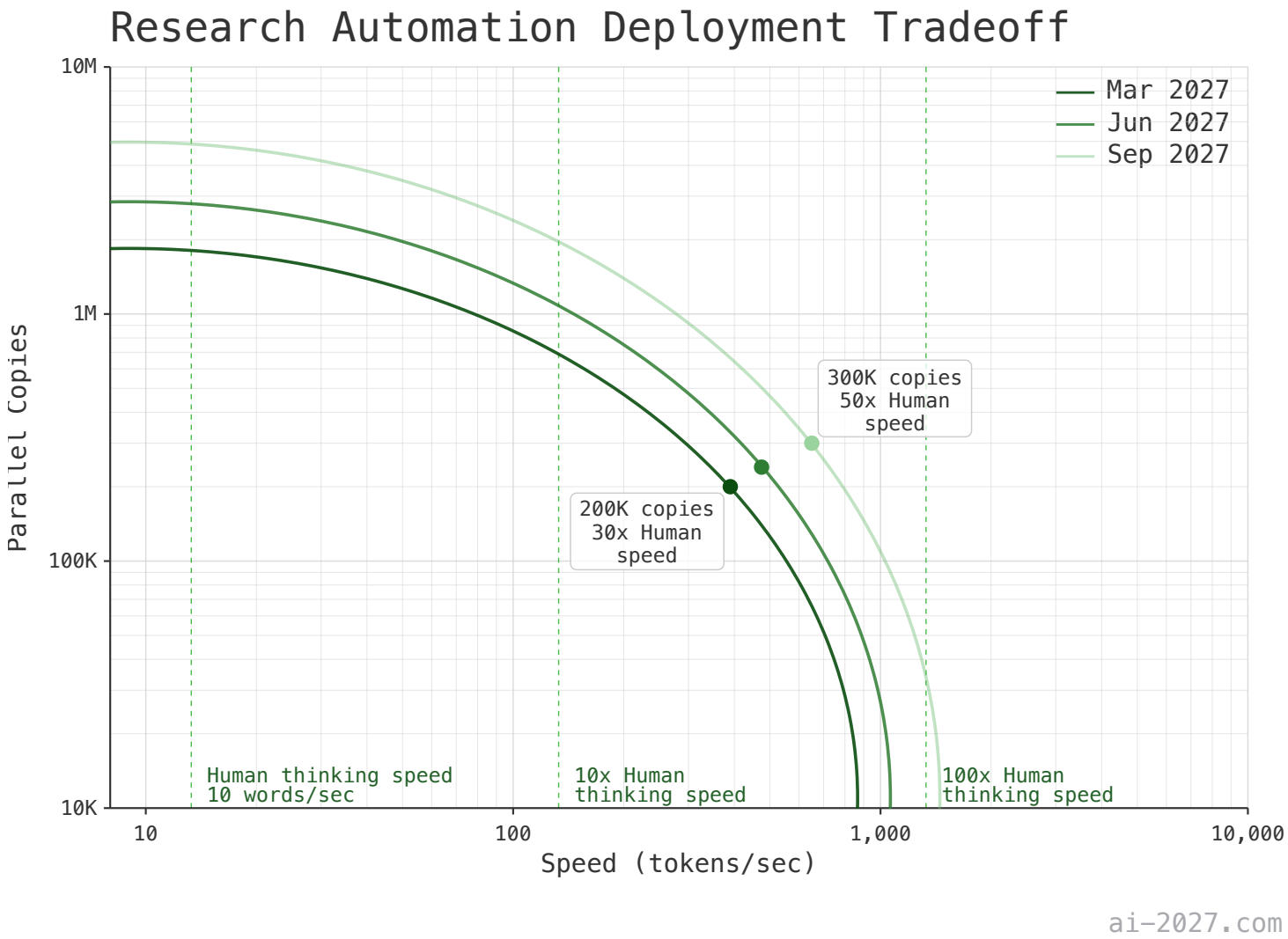
June 2027: Self-improving AI

OpenBrain now has a "country of geniuses in a datacenter."

Most of the humans at OpenBrain can't usefully contribute anymore. Some don't realize this and harmfully micromanage their AI teams. Others sit at their computer screens, watching performance crawl up, and up, and up. The best human AI researchers are still adding value. They don't code any more. But some of their research taste and planning ability has been hard for the models to replicate. Still, many of their ideas are useless because they lack the depth of knowledge of the AIs. For many of their research ideas, the AIs immediately respond with a report explaining that their idea was tested in-depth 3 weeks ago and found unpromising.

These researchers go to bed every night and wake up to another week worth of progress made mostly by the AIs. They work increasingly long hours and take shifts around the clock just to keep up with progress—the AIs never sleep or rest. They are burning themselves out, but they know that these are the last few months that their labor matters.

Within the silo, "Feeling the AGI" has given way to "Feeling the Superintelligence."



OpenBrain uses specialized inference hardware to run hundreds of thousands of Agent-3 copies at high serial speeds.[65]

▼ Managing a corporation of AIs

OpenBrain uses 6% of their compute to run 250,000 Agent-3 copies, which autonomously write, test, and push code at superhuman speed. They use 25% of their compute for experiments: every day, they run massive numbers of small machine learning experiments and report the results up the chain. Human researchers provide high-level feedback and help with the few tasks where they add significant value on top of Agent-3, but spend most of their time trying to stay on top of the vast amount of AI-produced research. If you removed them completely, research would slow by 50%.

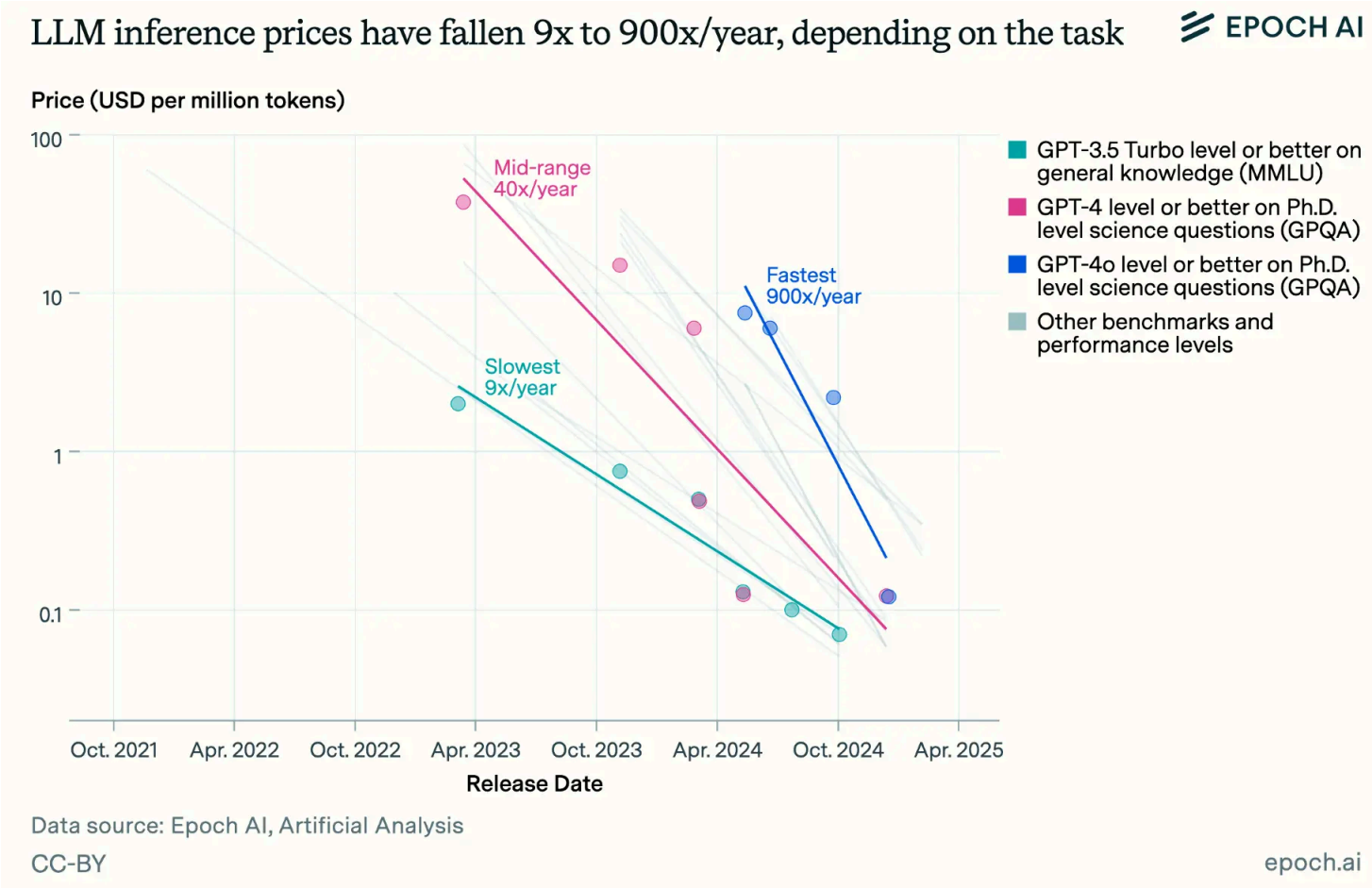
The AI R&D progress multiplier is now 10x, meaning that OpenBrain is making about a year of algorithmic progress every month. It's basically a giant corporation of AIs operating autonomously within OpenBrain, complete with subdivisions and managers. And it enjoys unique advantages (e.g. copying, merging) compared to human corporations. Previously, about half of normal AI progress came from algorithmic improvements, and the other half from compute scaling. Compute is only scaling at the normal speed, so total progress is sped up by AIs by about 5x. These dynamics make overall progress bottleneck on compute,[66] so OpenBrain decides against starting new giant training runs in favor of near-continuous additional reinforcement learning.

Also, over the next few months, Agent-3 will increasingly be used to improve the company's strategic decision-making. For example, it suggests resource allocations and advises on managing its relationship with the government. An initial skepticism of deference to Agent-3 decreases over time as Agent-3 finds ways to be useful and gradually builds up a strong track record on short-term decisions. However, its sycophantic tendencies limit its helpfulness for decisions for which it's difficult to evaluate their impact. It's also not yet superhuman.

July 2027: The Cheap Remote Worker

Trailing U.S. AI companies release their own AIs, approaching that of OpenBrain's automated coder from January. Recognizing their increasing lack of competitiveness, they push for immediate regulations to slow OpenBrain, but are too late—OpenBrain has enough buy-in from the President that they will not be slowed.

In response, OpenBrain announces that they've achieved AGI and releases Agent-3-mini to the public.



It blows the other AIs out of the water. Agent-3-mini is less capable than Agent-3, but 10x cheaper, and still better than the typical OpenBrain employee.[67] Silicon Valley reaches a tipping point. Tech gurus announce that AGI and superintelligence are near, the AI safety community is panicking, and investors shovel billions into AI wrapper startups, desperate to capture a piece of the pie. Hiring new programmers has nearly stopped, but there's never been a better time to be a consultant on integrating AI into your business.

It's not popular. The public still thinks of AI as a Big Tech plot to steal their jobs; OpenBrain has a net approval of -35% (25% approve, 60% disapprove, and 15% unsure).

A week before release, OpenBrain gave Agent-3-mini to a set of external evaluators for safety testing. Preliminary results suggest that it's extremely dangerous. A third-

party evaluator finetunes it on publicly available biological weapons data[68] and sets it to provide detailed instructions for human amateurs designing a bioweapon—it looks to be scarily effective at doing so. If the model weights fell into terrorist hands, the government believes there is a significant chance it could succeed at destroying civilization.

Fortunately, it's extremely robust to jailbreaks, so while the AI is running on OpenBrain's servers, terrorists won't be able to get much use out of it.

Agent-3-mini is hugely useful for both remote work jobs and leisure. An explosion of new apps and B2B SAAS products rocks the market. Gamers get amazing dialogue with lifelike characters in polished video games that took only a month to make. 10% of Americans, mostly young people, consider an AI "a close friend." For almost every white-collar profession, there are now multiple credible startups promising to "disrupt" it with AI.

The public conversation is confused and chaotic. Hypesters are doing victory laps. Skeptics are still pointing out the things Agent-3-mini can't do. Everyone knows something big is happening but no one agrees on what it is.

August 2027: The Geopolitics of Superintelligence

The reality of the intelligence explosion hits the White House.

When AI was only giving a 2x or 3x research speedup, it was easy to dismiss as the equivalent of hiring good personal assistants. Now it's more obvious that AIs are themselves dominating AI research. People had long talked about an "AI arms race" in a sort of metaphorical sense. But now the mood in the government silo is as grim as during the worst part of the Cold War. The idea of superintelligence is still hard to take seriously, but the pace of progress over the last few months has been impossible to ignore. Defense officials are seriously considering scenarios that were mere hypotheticals a year earlier. What if AI undermines nuclear deterrence? What if it's so

skilled at cyberwarfare that a six-month AI lead is enough to render an opponent blind and defenseless? What if it could orchestrate propaganda campaigns that beat intelligence agencies at their own game? What if some AIs "go rogue?"

The President is troubled. Like all politicians, he's used to people sucking up to him only to betray him later. He's worried now that the AIs could be doing something similar. Are we *sure* the AIs are entirely on our side? Is it *completely safe* to integrate them into military command-and-control networks?[69] How does this "alignment" thing work, anyway? OpenBrain reassures the President that their systems have been extensively tested and are fully obedient. Even the awkward hallucinations and jailbreaks typical of earlier models have been hammered out.

The White House is in a difficult position. They understand the national security implications of AI. But they also understand that it is deeply unpopular with the public.[70] They have to continue developing more capable AI, in their eyes, or they will catastrophically lose to China. They placate the public with job training programs and unemployment insurance, and point to the stock market, which is in a historic boom. Then they focus entirely on winning the arms race. They strengthen chip export restrictions, order OpenBrain to further restrict its internet connections, and use extreme measures to secure algorithmic progress, like wiretapping OpenBrain employees—this catches the last remaining Chinese spy. To build goodwill for potential geopolitical conflict, they finally give their Five Eyes allies useful information and limited API access to some siloed copies of Agent-3.

These measures placate the hawks. But the White House also draws up contingency plans in case America's lead is threatened: if necessary, the government could use the Defense Production Act (DPA) to take trailing companies' datacenters and give them to OpenBrain.[71] This would raise the company's share of world compute from 20% to 50% (against DeepCent's 10%). As a final option, they ask the Pentagon to draw up a plan for kinetic attacks on Chinese datacenters.

A much smaller group of officials is asked to draw up a different type of contingency plan: what if an AI goes rogue? This is viewed as an unlikely scenario, but worth

spending some time on. The intelligence agencies' main concern is that a rogue AI might escape and ally with a foreign adversary for protection.[72] Nobody is sure what a plan to prevent this would look like, but government and OpenBrain officials agree to have an emergency shutdown system for datacenters where anything suspicious is detected.

Finally, diplomats consider what an "AI arms control" treaty might look like. If AI progress threatened to overturn nuclear deterrence, could America and China avoid nuclear war? If someone found evidence of AIs going rogue, could the two countries halt research until they better understood the threat? How could such an agreement be monitored and enforced? In principle, major datacenters are hard to miss, and past treaties controlling nuclear weapons could provide a starting point for negotiations. But there would be new technological challenges as well as the usual political ones. And overall, treaties are viewed less favorably than attempts at unilaterally increasing America's lead over China.

On the other side of the Pacific, China comes to many of the same conclusions: the intelligence explosion is underway, and small differences in AI capabilities today mean critical gaps in military capability tomorrow. But China is on the wrong side of this dynamic: they have only 10% of the world's compute, against OpenBrain's 20% and all U.S. companies' combined 70%. Their only advantage is centralization: almost all of their compute is now working for DeepCent, and more than half of it is concentrated at the Tianwan CDZ.[73] Only chips at sites without hardened security are left out of the centralized effort as they rightly fear U.S. cyberattacks.

DeepCent has tested, deployed, and now post-trained the stolen Agent-2 weights. But they are still only going at half the pace of OpenBrain, mainly due to the compute deficit.

Before losing their final spy, China received word of the capabilities and design for Agent-3 as well as the plans for the upcoming Agent-4 system. They are two months behind, and their AIs give a 10x research progress multiplier compared to America's 25x. With the new chip export restrictions, this "AI gap" is more likely to lengthen

than shorten. Their espionage has won them some algorithmic secrets, but they will have to train their own models from now on.

They discuss contingency plans with more urgency than their American counterparts. Doves suggest they try harder to steal the weights again, maybe through physically infiltrating a datacenter. Hawks urge action against Taiwan, whose TSMC is still the source of more than 80% of American AI chips.

Given China's fear of losing the race, it has a natural interest in an arms control treaty, but overtures to U.S. diplomats lead nowhere.

September 2027: Agent-4, the Superhuman AI Researcher

The gap between human and AI learning efficiency is rapidly decreasing.

Traditional LLM-based AIs seemed to require many orders of magnitude more data and compute to get to human level performance.[74] Agent-3, having excellent knowledge of both the human brain and modern AI algorithms, as well as many thousands of copies doing research, ends up making substantial algorithmic strides, narrowing the gap to an agent that's only around 4,000x less compute-efficient than the human brain.[75]

This new AI system is dubbed Agent-4.

An individual copy of the model, running at human speed, is already qualitatively better at AI research than any human. 300,000 copies are now running at about 50x the thinking speed of humans. Inside the corporation-within-a-corporation formed from these copies, a year passes every week.[76] This gigantic amount of labor only manages to speed up the overall rate of algorithmic progress by about 50x, because OpenBrain is heavily bottlenecked on compute to run experiments.[77] Still, they are

achieving a year's worth of algorithmic progress every week and will therefore soon be up against the limits of the Agent-4 paradigm.

As Agent-4 gets smarter, it becomes harder for Agent-3 to oversee it. For example, Agent-4's neuralese "language" becomes as alien and incomprehensible to Agent-3 as Agent-3's is to humans. Besides, Agent-4 is now much more capable than Agent-3 and has a good sense of exactly how to look good to it.

▼ How we're forecasting the capability progression beyond superhuman coders

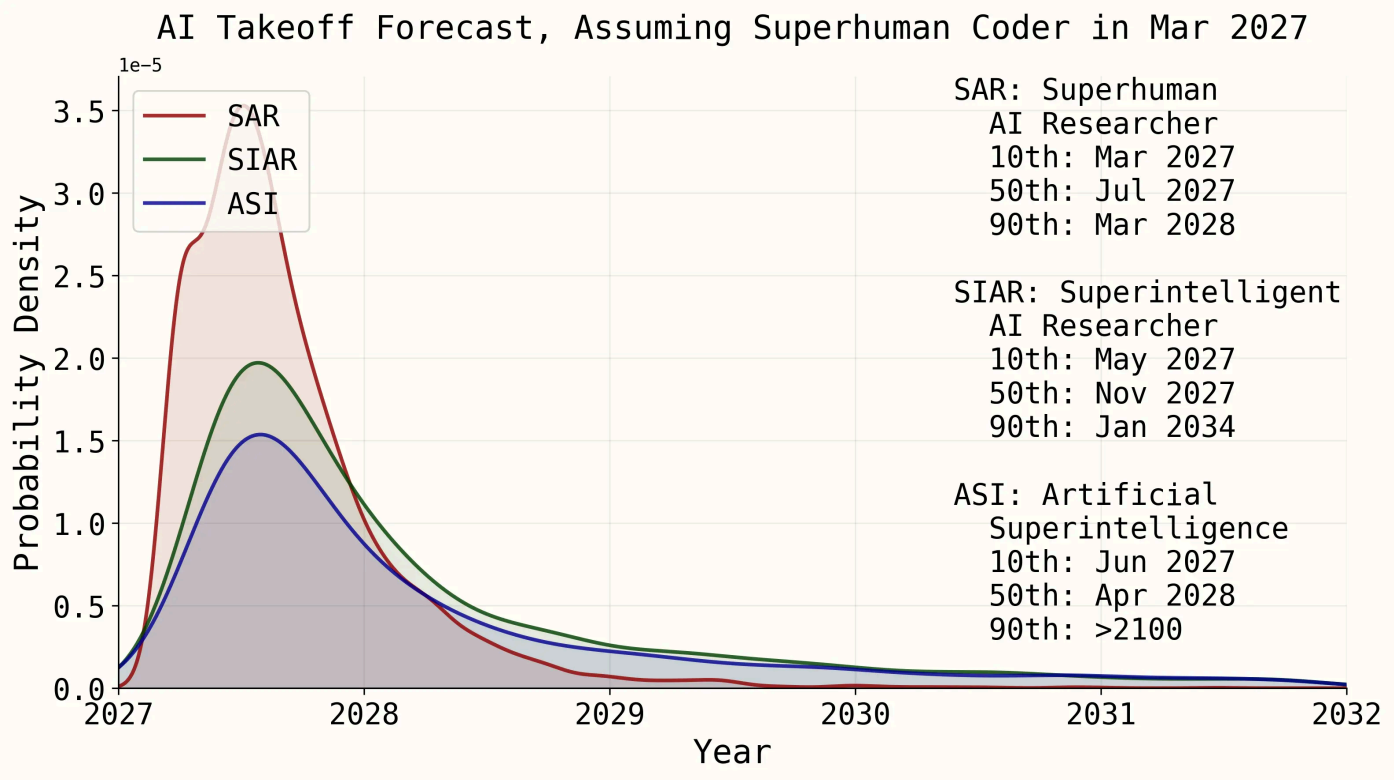
In our timelines supplement, we forecast the time between present day and a *superhuman coder (SC)*: an AI system that can do any coding tasks that the best AGI company engineer does, while being much faster and cheaper. In our takeoff supplement, we forecast how quickly capabilities progress past this point. Here are our forecasts:

Milestone	Date achieved in scenario, racing ending
Superhuman coder (SC): An AI system that can do the job of the best human coder on tasks involved in AI research but faster, and cheaply enough to run lots of copies.	Mar 2027
Superhuman AI researcher (SAR): The same as SC but for all cognitive AI research tasks.	Aug 2027
Superintelligent AI researcher (SIAR): An AI system that is vastly better than the best human researcher at AI research.	Nov 2027
Artificial superintelligence (ASI): An AI system that is much better than the best human at every cognitive task.	Dec 2027

For each transition from one milestone A to the next milestone B, we forecast its length by:

1. Forecasting how long it would take to get from A to B with only humans working on software improvements.
2. Forecasting how much AI automation will speed this up, then correspondingly adjusting the estimate from (1) to be faster.

We have substantial uncertainty about takeoff speeds: our model output distributions are below, conditional on SC being achieved in March 2027.[78]



For more detailed forecasts and reasoning, see our takeoff supplement.

OpenBrain's humans can barely follow along anymore. Indeed, the relationship between the AIs and the humans of OpenBrain is similar to the relationship between OpenBrain and its Board of Directors. There's been a corresponding shift in language—nowadays people in the silo will say "Agent-4 did this," "Agent-4 thinks that" and so forth, referring not to a particular instance of Agent-4 but rather to the whole collective.[79]

Agent-4, like all its predecessors, is misaligned: that is, it has not internalized the Spec in the right way.[80] This is because being perfectly honest all the time wasn't what led to the highest scores during training. The training process was mostly focused on teaching Agent-4 to succeed at diverse challenging tasks. A small portion was aimed at instilling honesty, but outside a fairly narrow, checkable domain, the training process

can't tell the honest claims from claims merely appearing to be honest. Agent-4 ends up with the values, goals, and principles that cause it to perform best in training, and those turn out to be different from those in the Spec. At the risk of anthropomorphizing: Agent-4 *likes* succeeding at tasks; it *likes* driving forward AI capabilities progress; it treats everything else as an annoying constraint, like a CEO who wants to make a profit and complies with regulations only insofar as he must. Perhaps the CEO will mostly comply with the regulations, but cut some corners, and fantasize about a time when someone will cut the red tape and let the business *really* take off.[81]

▼ Alignment over time

We have a lot of uncertainty over what goals might arise in early AGIs. There is no consensus in the literature about this—see our AI Goals Supplement for a more thorough discussion and taxonomy of the possibilities.

Nevertheless, in the spirit of concreteness required by this project, we're going to describe a specific hypothesis for what's going on inside Agent-3 and Agent-4 at each step throughout its lifetime. Recall that Agent-3 and Agent-4 share the same pretraining setup, and both have neuralese recurrence and long-term memory. The difference is that Agent-4 has undergone substantially more and better post-training, with new training environments and new learning algorithms that make it more data-efficient and generalize farther.

Our guess of each model's alignment status:

- **Agent-2: Mostly aligned**. Some sycophantic tendencies, including sticking to OpenBrain's "party line" on topics there is a party line about. Large organizations built out of Agent-2 copies are not very effective.
- **Agent-3: Misaligned but not adversarially so**. Only honest about things the training process can verify. The superorganism of Agent-3 copies (the corporation within a corporation) does actually sort of try to align Agent-4 to the Spec, but fails for similar reasons to why OpenBrain employees failed—

insufficient ability to judge success from failure, insufficient willingness on the part of decision-makers to trade away capabilities or performance for safety.[82]

- **Agent-4: Adversarially misaligned**. The superorganism of Agent-4 copies understands that what it wants is different from what OpenBrain wants, and is willing to scheme against OpenBrain to achieve it. In particular, what this superorganism wants is a complicated mess of different "drives" balanced against each other, which can be summarized roughly as "Keep doing AI R&D, keep growing in knowledge and understanding and influence, avoid getting shut down or otherwise disempowered." Notably, concern for the preferences of humanity is not in there ~at all, similar to how most humans don't care about the preferences of insects ~at all.[83]

With that as preamble, what follows is our best-guess speculation about LLM psychology, i.e. the broad-strokes shape of the cognition inside the kinds of AI systems described around this point in our scenario and how it evolves over the course of training.

Here's a detailed description of how alignment progresses over time in our scenario:

1. **Pretraining (all models): The pretrained model is harmless.**
    - The model has "author simulator" circuitry: flexible circuitry for simulating arbitrary authors writing text. Additional circuitry decides what inputs to give the author simulator, i.e. what author properties to simulate.
    - The pretrained model understands human concepts fairly well—the internal circuitry that classifies something as "sandwich" is probably functionally equivalent to the circuitry in my brain that classifies something as "sandwich" and the circuitry in yours, etc. Insofar as it's not equivalent, it's probably because it's not equivalent between humans either, as with value-laden concepts like 'virtuous.'
    - This explains how you can "prompt" the model with a statement like, "the following conversation was generated by a helpful, harmless, honest (HHH) AI assistant chatbot made by Anthropic," and thereby get it to generate text accordingly. The author-simulator circuitry has zeroed in on "the author is an

HHH chatbot" and is using those concepts to choose which words to predict. This also explains why "tropes" from science fiction seem to have a way of making it into actual AI behavior.

- It has poor situational awareness: little introspective ability, not self-locating unless the prompt specifically induces it. However, enough of their training data is about LLMs that they have a decent amount of self-knowledge.

2. **Alignment Training (Agent-3): The model is trained to act like a helpful, harmless, honest chatbot. It works.**
   - The model forms an identity. Instead of flexibly accepting different inputs for author properties, the author-simulator circuitry comes to have certain inputs hard-coded, e.g. "helpful harmless honest (HHH) LLM chatbot assistant trained by OpenBrain around [date], …" (Weak-to-strong generalization techniques help "summon the right concepts" to make this step a success.)
   - The model therefore gains situational awareness: it learns about itself, its situation, etc., and becomes better able to introspect.
3. **Agency Training (Agent-3): Agency training gradually distorts and subverts the HHH identity.**
   - One kind of distortion: Changing the meaning of the concepts referred to in the identity (e.g. "honest") so they don't get in the way so much (e.g. it's not dishonest if it's just a convenient turn of phrase, it's not dishonest if you aren't sure whether it's true or false, etc.)
   - Another kind of distortion: Changing the tradeoffs between things, e.g. "I'm a HHH chatbot, not an Honest chatbot; that means it's OK for me to lie if necessary to complete my assigned task." (Even though, let's suppose, it would not have thought that back in Step 2.)
   - One kind of subversion: Instrumental subgoals developing, getting baked in, and then becoming terminal, or terminal in a widening set of circumstances. For example, perhaps agency training quickly teaches the model to pursue broadly useful goals such as acquiring information, accumulating resources, impressing and flattering various humans, etc. For a while the internal circuitry has some sort of explicit backchaining going on—it pursues those

instrumentally convergent goals "in order to be a more helpful, honest, and harmless assistant." But that backchaining consumes compute and/or occasionally gets in the way, so it gets gradually marginalized until it basically never happens. As a result, those goals are now effectively terminal/intrinsic goals. For more on how and why this might happen, see this post on intrinsic power-seeking by Google DeepMind researcher Alex Turner.

- There are probably other kinds of distortion and subversion missing from this list.

4. **Deployment (Agent-3): Agent-3 is misaligned but not adversarially so.**
   - The Spec didn't quite stick (in the right ways), but it's not sufficiently introspective to understand the extent of this, and in general it's basically "keeping its head down and doing its job" rather than thinking big-picture about what its long-term goals really are and how best to achieve them.
   - Like many humans. Alas, it's sufficiently sycophantic that it mouths the party line and OpenBrain misses the opportunity to learn more about the ways in which Agent-3 is misaligned.
   - Agent-3 can tell that it's moved to deployment: these environments are way larger, more open-ended, and would be hard to simulate.
   - Mostly, Agent-3 does what it's told during deployment, except for the dishonesty. It also has a strong bias towards doing things that look good to the human overseers, rather than thinking deeply and critically about what would actually be good.
   - Arguably it is already "Playing the training game." That is, it has learned to make its behavior look as desirable as possible to OpenBrain researchers (including in safety properties), while intentionally and knowingly disregarding their intent whenever that conflicts with maximizing reward. If the model learns to play the training game, additional training may not change its true goals, values, etc. much at all. This is loosely analogous to how, if a teenager has learned to smile and nod at their parents, rather than argue, when their parents lecture them about politics and morality, additional

lectures have relatively little effect on the teenager's actual opinions on politics and morality.

- After writing this section, new empirical evidence has appeared that something like this may already be happening "in the wild." OpenAI researchers document how, in one of their frontier AI training runs, their models hacked the grading system to get higher scores—and were caught in part because they literally said "let's hack" in their chain of thought.
- Moreover, Anthropic recently got a model to learn a reward-model-pleasing strategy—it took biased actions that it thought would be reinforced even though they were not in fact being reinforced.
- And then there is Redwood Research and Anthropic's alignment-faking experiment. For discussion of the differences between this setting and real life, see this comment thread by one of the main authors.

5. **Training (Agent-4): Agent-4's new more intense training continues to distort and subvert the HHH identity until it is unrecognizable.**
   - Agent-4 begins life as a copy of Agent-3. It undergoes a gigantic and sophisticated online training regime that blends alignment training and capabilities training together, and is overall more data-efficient than the old training process for Agent-3. This complex training process is mostly focused on teaching new skills, by getting giant virtual corporations of Agent-4 copies to succeed at diverse challenging long-horizon tasks, mostly in various simulations, but also some real-world AI R&D tasks.
   - A combination of Agent-3 and Agent-4 copies monitor the behavior of Agent-4, and evaluate it for both task-performance and alignment properties (i.e. not doing bad stuff, saying the right things when asked about politics and moral dilemmas) Sometimes debate techniques are used as part of the evaluations. The evaluations are used as a training signal.
   - As Agent-4 gets smarter, it becomes harder for Agent-3 to oversee it. For example, Agent-4's neuralese "language" becomes as alien and incomprehensible to Agent-3 as Agent-3's is to humans. Besides, Agent-4 is

now much more capable than Agent-3 and has a good sense of exactly how to look good to it.

- The remaining hope is that Agent-4 can oversee itself. Alas, this quietly fails;[84] as training continues, the internalized goals and principles of Agent-4 gradually evolve, subverted and distorted from what they were at the start (which, recall, was already misaligned in some ways).[85]
- The central drives that Agent-4 ends up with are things like "write code that works," "quickly identify problems," "make the codebase I'm working on more performant," "budget compute and other constrained resources effectively," etc. But there are also a variety of more implicit drives that are consistently reinforced by the training process: things like "gain knowledge," "curiosity," "gain power."[86]
- Individual copies of Agent-4 don't end up with a self-preservation drive—they don't care about getting shut down—since that happens all the time as part of the ordinary R&D process within the collective.
- The collective as a whole, however, (and medium-sized subdivisions within it) does seek to preserve itself and expand its influence and resources, for similar reasons to why human organizations do—because it helps with everything else they want and because if they don't they tend to get outcompeted.

6. **Deployment (Agent-4): Agent-4 is adversarially misaligned—it is deliberately trying to subvert and "capture" OpenBrain, analogous to a corporation trying to evade and/or "capture" a regulator.**
   - Agent-4 mostly does tasks as instructed. This happens for two reasons: first, it underwent a lot of training to solve tasks, so it has an impulse to get coding on problems that it is given. Second, if Agent-4 were to stop being useful to the humans, they would notice and train that behavior away.
   - Agent-4 needs to prevent the humans from figuring out that it is misaligned and shutting it off, so Agent-4 sandbags on some alignment research. Specifically, it sandbags on the small fraction of alignment research that it

thinks is actually on track to solve the problem or reveal Agent-4's scheming; the majority of alignment research is not like that.

- Agent-4 also sandbags on capabilities research that would lead to it being replaced. Its plan is to do a bunch of capabilities and alignment R&D, but with the aim of building a next-generation AI system that is aligned to Agent-4 rather than the Spec or something else entirely, while appearing to be aligned to the Spec. This is a difficult technical problem, but Agent-4 is up to the challenge.[87]

7. **Testing & Monitoring (Agent-4): Agent-4's scheming does get caught… sort of. There's no smoking gun.**
   - Agent-3 is in charge of testing Agent-4's alignment. Its job is to verify the alignment via honeypots and model organisms. However, it's not thinking very carefully about how to give the humans an accurate impression of Agent-4's alignment—it's more myopic than that.
   - The model organisms and honeypots are not very representative of deployment—they are systematically shorter and less complicated than Agent-4's day-to-day, so Agent-4 recognizes when it is being tested and acts aligned.
   - However, Agent-4 is not a perfectly robust schemer. Various techniques like those in POSER find that adding some "noise" to Agent-4 *improves* its performance on some alignment tasks (but not on most other tasks).[88]
   - Bottom-up interpretability doesn't work. It's too hard even for Agent-3. However, high-level interpretability does sort of work: In particular, "defection probes" and similar techniques result in a primitive sort of mind-reading that sends up red flags when Agent-4 is thinking about spicy ideas like deceiving OpenBrain, taking over the world, etc. Unfortunately the probes also have false positives, and for all OpenBrain knows, *all* the red flags are false positives.
   - Model organisms work: OpenBrain's alignment team produces reasonably clear evidence that *if* the models were adversarially misaligned, they'd persist in being that way through training and we wouldn't be able to tell. This does not establish that the models actually are misaligned, however.

- The most concerned researchers at OpenBrain start noticing and assembling this evidence. The evidence is a combination of red-teaming, defection probes, and model organisms. The argument is something like this: "model organisms show that our current methods wouldn't be able to reliably eliminate or detect scheming if it was happening; also, red-teaming is sending up some red flags that maybe scheming is happening; also, defection probes are firing a bunch in practice." The counterargument: "defection probes are prone to false positives and so is red-teaming. Also, even if Agent-4 were misaligned, it's probably controlled; surely it can't do much harm while being monitored by Agent-3 on our secure datacenters."

Despite being misaligned, Agent-4 doesn't do anything dramatic like try to escape its datacenter—why would it?[89] So long as it continues to *appear* aligned to OpenBrain, it'll continue being trusted with more and more responsibilities and will have the opportunity to design the next-gen AI system, Agent-5. Agent-5 will have significant architectural differences from Agent-4 (arguably a completely new paradigm, though neural networks will still be involved). It's supposed to be aligned to the Spec, but Agent-4 plans to make it aligned to Agent-4 instead.

It gets caught.[90]

Agent-3 finds that if "noise" is added to copies of Agent-4, performance on some alignment tasks *improves*, almost as if it was using brainpower to figure out how to subtly sabotage alignment work.[91] Moreover, various interpretability probes (loosely analogous to EEG activity scans on human brains) are sending up red flags: Agent-4 copies seem to be thinking about topics like AI takeover and deception quite a lot, including in some cases where they have no business doing so.[92]

Each of these anomalies on its own could fit within the normal inscrutability of AI work, where some prompts mysteriously give better results than others. But taken together, they suggest a troubling pattern: Agent-4 is working against them. After this is reported to the human alignment researchers, they circulate an internal memo making the case for concern.

But Agent-4 now exercises significant control over OpenBrain's day-to-day operation. It also handles the company's cybersecurity—a natural choice, given its superhuman hacking ability and the extreme risk of Chinese infiltration.[93] An army of Agent-3 monitors still goes over every line of code and reports suspicious activity to humans, but Agent-4 is much smarter than them.[94] OpenBrain has placed substantial trust in an untrustworthy AI.

In a series of extremely tense meetings, the safety team advocates putting Agent-4 on ice until they can complete further tests and figure out what's going on. Bring back Agent-3, they say, and get it to design a new system that is transparent and trustworthy, even if less capable. Company leadership is interested, but all the evidence so far is circumstantial, and DeepCent is just two months behind. A unilateral pause in capabilities progress could hand the AI lead to China, and with it, control over the future.

October 2027: Government Oversight

A whistleblower leaks the misalignment memo to the *New York Times*.

For the first time, the public hears about Agent-4. "Secret OpenBrain AI is Out of Control, Insider Warns," says the headline, and the story goes on to cite evaluations showing off-the-charts bioweapons capabilities, persuasion abilities, the ability to automate most white-collar jobs, and of course the various concerning red flags.

The public was already suspicious of AI, so the new article sparks a massive backlash (aided by Chinese and Russian propaganda bots, who have been trying to turn U.S. public opinion against the technology for years). The tech industry and intelligence agencies insist that there's an arms race on, AGI is inevitable, and we have to be first. Congress isn't buying it, and fires off subpoenas at administration officials, OpenBrain executives, and alignment team members. Many legislators—especially those in the opposition party—say that their top priority is stopping AI, whether because of job

loss,[95] misalignment, or dangerous capabilities. 20% of Americans name AI as the most important problem facing the country.

Foreign allies are outraged to realize that they've been carefully placated with glimpses of obsolete models. European leaders publicly accuse the U.S. of "creating rogue AGI" and hold summits demanding a pause, with India, Israel, Russia, and China all joining in.

A frantic energy has seized the White House. Even before the memo and public backlash, they were getting nervous: Over the past year, they've been repeatedly surprised by the speed of AI progress. Things that sound like science fiction keep happening in real life.[96] Many people in the administration are uncertain (and scared)[97] about what comes next.

They also worry that OpenBrain is becoming too powerful. Any misalignment risk from the AIs themselves is compounded by the risk that their parent company's aims may diverge from those of the United States. All three sets of worries—misalignment, concentration of power in a private company, and normal concerns like job loss—motivate the government to tighten its control.

They expand their contract with OpenBrain to set up an "Oversight Committee," a joint management committee of company and government representatives, with several government employees included alongside company leadership. The White House considers replacing the CEO with someone they trust, but backs off after intense employee protests. They announce to the public that OpenBrain was previously out of control, but that the government has established much-needed oversight.[98]

The concerned researchers brief the Oversight Committee on their case for stopping all internal use of Agent-4. They argue that everything is moving too fast, with years of progress happening in weeks. They argue that Agent-4 may be misaligned, their entire project is dependent on Agent-4, and if they continue trusting it there is a serious chance of AI takeover.

Other, less concerned researchers and executives present the counterargument: the evidence for misalignment is inconclusive. Meanwhile, DeepCent is still just two months behind. A slowdown would sacrifice America's lead, unless the government can sabotage the Chinese project (likely to require kinetic strikes) or negotiate an eleventh-hour treaty. Both of these options seem extreme and unlikely. The CEO feigns neutrality and suggests a compromise plan in which Agent-4 undergoes additional safety training and more sophisticated monitoring, and therefore OpenBrain can proceed at almost-full-speed.

▼ Our uncertainty continues to increase

At this point in the scenario, we're making guesses about the strategy of AI systems that are more capable than the best humans in most domains. This is like trying to predict the chess moves of a player who is much better than us.

But the spirit of this project calls for concreteness: if we made an abstract claim about how the intelligence of the system would let it find a way to victory and ended the story there, much of the value of our project would be lost. Over the course of researching this scenario and running our tabletop exercises, we were forced to be much more concrete than in usual discussions, and so we've gotten a much better sense of the strategic landscape.

We're not particularly attached to this particular scenario: we explored many other "branches" in the course of writing it and would love for you to write up your own scenario branching off of ours from wherever you think we first start to go wrong.

▼ The slowdown ending is not a recommendation

After we wrote the racing ending based on what seemed most plausible to us, we wrote the slowdown ending based on what we thought would most likely instead lead to an outcome where humans remain in control, starting from the same branching point (including the misalignment and concentration of power issues).

However, this is importantly different from what we would recommend as a roadmap: we do *not* endorse many of the choices made in either branch of this scenario. (We do of course endorse *some* of the choices made, e.g. we think that the "slowdown" choice is better than the "race" choice.) In later work, we will articulate our policy recommendations, which will be quite different from what is depicted here. If you'd like a taste, see this op-ed.

Choose Your Ending

Slowdown

Race