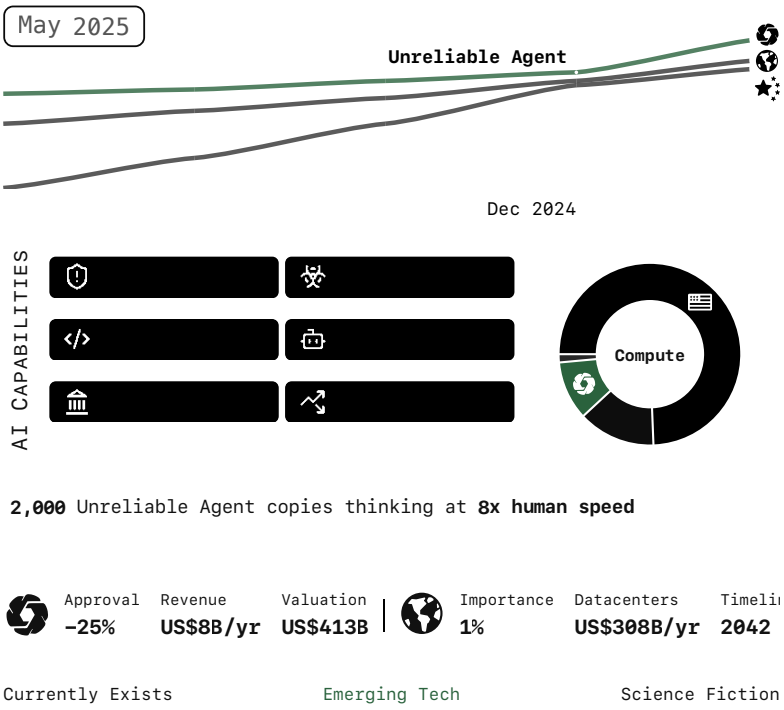
AI 2027

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?

Published April 3rd 2025

PDF Listen

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.

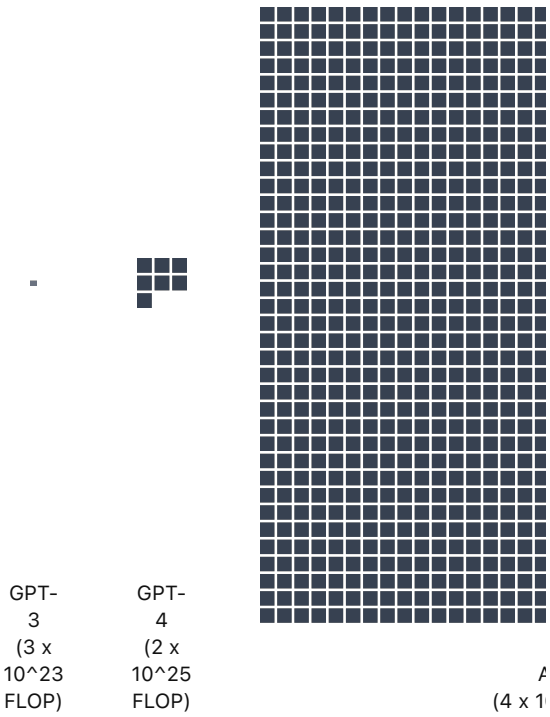
The AIs of 2024 could follow specific instructions: they could turn bullet points into emails, and simple requests into working code. In 2025, AIs function more like employees. Coding AIs increasingly look like autonomous agents rather than mere assistants: taking instructions via Slack or Teams and 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-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 US 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"

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?

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 US-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?

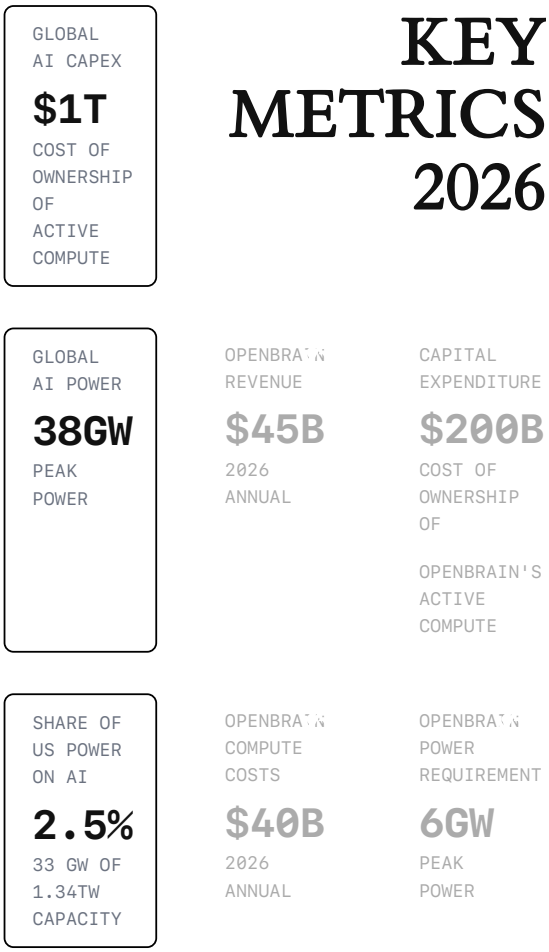
Late 2026: AI Takes Some Jobs

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]



▸ Why our uncertainty increases substantially beyond

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 US 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), Department of Defense (DOD), and US 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

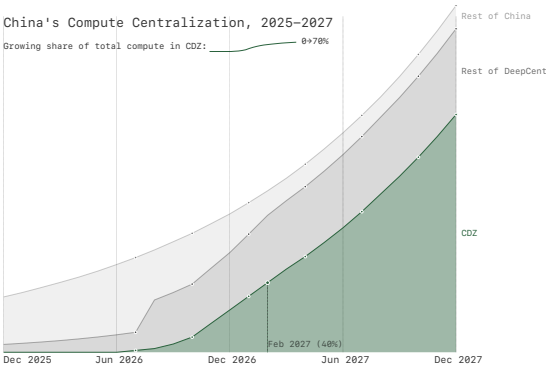
Defense considers this a critical advantage in cyberwarfare, and AI moves from #5 on the administration's priority list to #2.[49] Someone mentions the possibility of nationalizing OpenBrain, but other cabinet officials think that's premature. A staffer drafts a memo that presents the President with 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

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]



ai-2027.com

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

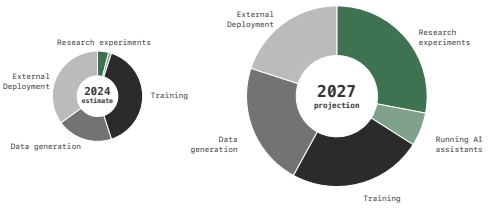
Three huge datacenters full of

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

▸ Iterated distillation and amplification (IDA)

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

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

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 US 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

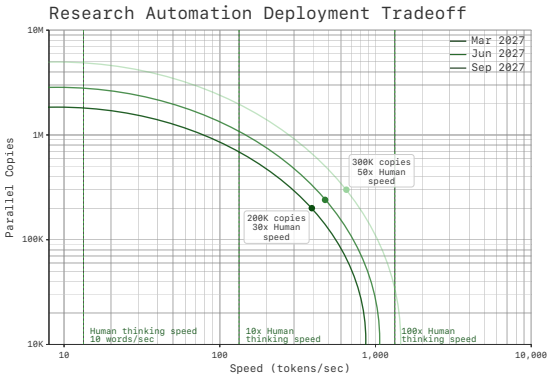
OpenBrain now has a "country of geniuses in a datacenter."

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

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

July 2027: The Cheap Remote Worker

Trailing US 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

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 US 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 US 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 US 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

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

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 US 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 US 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

▸ The slowdown ending is not a recommendation

Choose Your Ending

Slowdown Race

AI Futures Project

Ho Ab Su Co
Fo

Ti Th AI Se
Fo Fo Fo Fo

Design by
Lightcone Infrastructure

Subscribe

By subscribing you agree to Substack's Terms of Use, our Privacy Policy and our Information collection notice

substack