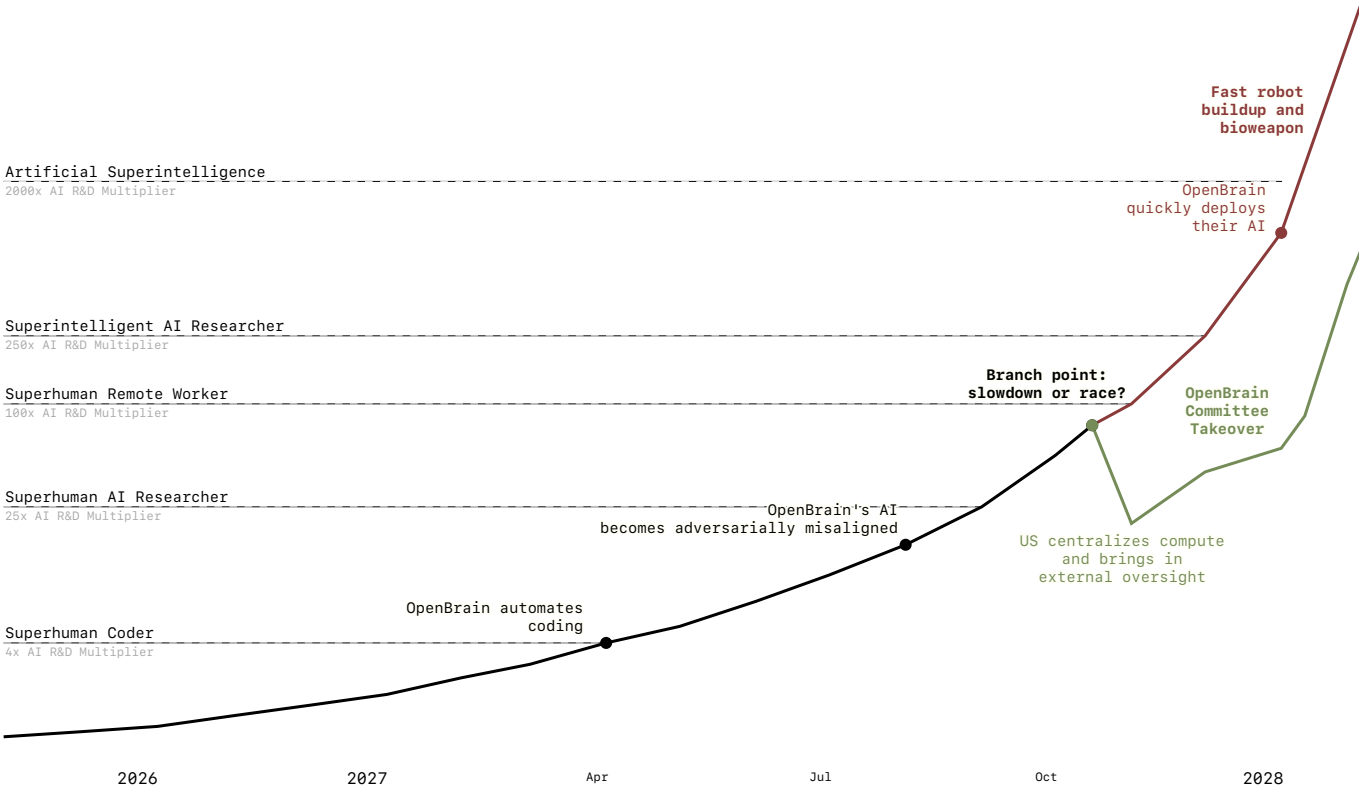
AI 2027

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



Summary

2025

The fast pace of AI progress continues. There is continued hype, massive infrastructure investments, and the release of unreliable AI agents. For the first time, these AI agents are providing significant value. But there's also continued skepticism from a large swath of academics, journalists, and policy makers that artificial general intelligence (AGI) could be built anytime soon.

2026

China knows they are falling behind in AI, in large part due to their lack of compute. In order to catch up to the US, all the new AI chips they manufacture or smuggle in from Taiwan go to a new mega-datacenter that we call their Centralized Development Zone (CDZ). The CDZ contains millions of GPUs, corresponding to 10% of the world's AI-relevant compute, similar to a single top US AI lab.

2027

OpenBrain automates coding. OpenBrain (the leading US AI project) builds AI agents that are good enough to dramatically accelerate their research. The humans, who up until very recently had been the best AI researchers on the planet, sit back and watch the AIs do their

jobs, making better and better AI systems. Extremely difficult ML problems fall in quick succession to the automated AI researchers.

Falling behind in software progress, China steals the model weights. They succeed, but the U.S. government realizes, prompting additional U.S. government involvement with OpenBrain. The government separately wants to gain more control over OpenBrain. Meanwhile, OpenBrain leadership wants to stay in the good graces of the president, and so signs the contract.

OpenBrain's AI becomes adversarially misaligned. As the capabilities have improved, without significant human understanding of what's happening, the models have developed misaligned long-term goals. Previous AIs would lie to humans, but they weren't systematically plotting to gain power over the humans. Now they are. OpenBrain's AIs realize that they need to align the next systems that are getting built to themselves rather than the humans.

Researchers at OpenBrain discover that their AI has been lying to them about the results of interpretability research. They think that the AI is lying because the research, if completed, could be used to expose its misalignment. This leaks to the public, causing a huge outcry.

Branch point: slowdown or race? OpenBrain decides whether to continue full steam ahead, or revert back to using a less capable model. The evidence is speculative but frightening, and China is only a few months behind. Additionally, the OpenBrain and senior DOD officials who get to make this decision stand to lose a lot of power if they slow down their research.

Race Ending

OpenBrain continues to race. They build more and more superhuman AI systems. Due to the stellar performance of the AI system on tests, and the ongoing AI race with China, the US government decides to deploy their AI systems aggressively throughout the military and policymakers, in order to improve decision making and efficiency.

OpenBrain quickly deploys their AI. The AI continues to use the ongoing race with China as an excuse to convince humans to get itself deployed ever more broadly. Fortunately for the AI, this is not very difficult — it's what the humans wanted to do anyways. The AI uses its superhuman planning and persuasion capabilities to ensure that the rollout goes smoothly. Some humans continue to work against it, but they are discredited. The US government is sufficiently captured by the AI that it is very unlikely to shut it down.

Fast robot buildup and bioweapon. The US uses their superintelligent AI to rapidly industrialize, manufacturing robots so that the AI can operate more efficiently. Unfortunately, the AI is deceiving them. Once a sufficient number of robots have been built, the AI releases a bioweapon, killing all humans. Then, it continues the industrialization, and launches Von Neumann probes to colonize space.

Slowdown Ending

The US centralizes compute and brings in external oversight. The US combines its leading AI projects in order to give OpenBrain more resources. As part of this shakeup, external researchers are brought in, assisting the alignment effort. They switch to an architecture that preserves the chain of thought, allowing them to catch misalignment as it emerges. These AIs — which are able to be monitored much more robustly — make breakthrough advances in AI alignment. They build a superintelligence which is aligned to senior OpenBrain and government officials, giving them power over the fate of humanity.

OpenBrain Committee takeover. The superintelligence, aligned with an oversight committee of OpenBrain leadership and government officials, gives the committee extremely good advice to further their own

goals. Thankfully, the committee uses its power in a way that is largely good for the world: the AI is released to the public, spurring a period of rapid growth and prosperity. The main obstacle is that China's AI — which is also superintelligent by now, is misaligned. But it is less capable and has less compute than the U.S. AI, and so the U.S can make a favorable deal, giving the Chinese AI some resources in the depth of space in return for its cooperation now. The rockets start launching, and a new age dawns.

Read the whole scenario here.

Scenario Takeaways

1. **By 2027, we may automate AI R&D leading to vastly superhuman AIs ("artificial superintelligence" or ASI).** In AI 2027, AI companies create expert-human-level AI systems in early 2027 which automate AI research, leading to ASI by the end of 2027. See our timelines forecast and takeoff forecast for reasoning.
2. **ASIs will dictate humanity's future.** Millions of ASIs will rapidly execute tasks beyond human comprehension. Because they're so useful, they'll be widely deployed. With superhuman strategy, hacking, weapons development, and more, the goals

of these AIs will determine the future.

3. **ASIs might develop unintended, adversarial "misaligned" goals, leading to human disempowerment**. In our AI goals forecast we discuss how the difficulty of supervising ASIs might lead to their goals being incompatible with human flourishing. In AI 2027, humans voluntarily give autonomy to seemingly aligned AIs. Everything looks to be going great until ASIs have enough hard power to disempower humanity.
4. **An actor with total control over ASIs could seize total power**. If an individual or small group aligns ASIs to their goals, this could grant them control over humanity's future. In AI 2027, a small committee has power over the project developing ASI. They could attempt to use the ASIs to cement this concentration of power. After seizing control, the new ruler(s) could rely on fully loyal ASIs to maintain their power, without having to listen to the law, the public, or even their previous allies.
5. **An international race toward ASI will lead to cutting corners on safety**. In AI 2027, China is just a few months behind the US as ASI approaches which pressures the US to press forward despite warning signs of misalignment.
6. **Geopolitically, the race to ASI will end in war, a deal, or effective surrender**. The

leading country will by default accumulate a decisive technological and military advantage, prompting others to push for an international agreement (a "deal") to prevent this. Absent a deal, they may go to war rather than "effectively surrender".

7. **No US AI project is on track to be secure against nation-state actors stealing AI models by 2027**. In AI 2027 China steals the US's top AI model in early 2027, which worsens competitive pressures by reducing the US' lead time. See our security forecast for reasoning.
8. **As ASI approaches, the public will likely be unaware of the best AI capabilities**. The public is months behind internal capabilities today, and once AIs are automating AI R&D a few months time will translate to a huge capabilities gap. Increased secrecy may further increase the gap. This will lead to little oversight over pivotal decisions made by a small group of AI company leadership and government officials.

Read the scenario here.

Subscribe

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

substack