The Gentle Singularity

We are past the event horizon; the takeoff has started. Humanity is close to building digital superintelligence, and at least so far it's much less weird than it seems like it should be.

Robots are not yet walking the streets, nor are most of us talking to AI all day. People still die of disease, we still can't easily go to space, and there is a lot about the universe we don't understand.

And yet, we have recently built systems that are smarter than people in many ways, and are able to significantly amplify the output of people using them. The least-likely part of the work is behind us; the scientific insights that got us to systems like GPT-4 and o3 were hard-won, but will take us very far.

AI will contribute to the world in many ways, but the gains to quality of life from AI driving faster scientific progress and increased productivity will be enormous; the future can be vastly better than the present. Scientific progress is the biggest driver of overall progress; it's hugely exciting to think about how much more we could have.

In some big sense, ChatGPT is already more powerful than any human who has ever lived. Hundreds of millions of people rely on it every day and for increasingly important tasks; a small new capability can create a hugely positive impact; a small misalignment multiplied by hundreds of millions of people can cause a great deal of negative impact.

2025 has seen the arrival of agents that can do real cognitive work; writing computer code will never be the same. 2026 will likely see the arrival of systems that can figure out novel insights. 2027 may see the

arrival of robots that can do tasks in the real world.

A lot more people will be able to create software, and art. But the world wants a lot more of both, and experts will probably still be much better than novices, as long as they embrace the new tools. Generally speaking, the ability for one person to get much more done in 2030 than they could in 2020 will be a striking change, and one many people will figure out how to benefit from.

In the most important ways, the 2030s may not be wildly different. People will still love their families, express their creativity, play games, and swim in lakes.

But in still-very-important-ways, the 2030s are likely going to be wildly different from any time that has come before. We do not know how far beyond human-level intelligence we can go, but we are about to find out.

In the 2030s, intelligence and energy—ideas, and the ability to make ideas happen—are going to become wildly abundant. These two have been the fundamental limiters on human progress for a long time; with abundant intelligence and energy (and good governance), we can theoretically have anything else.

Already we live with incredible digital intelligence, and after some initial shock, most of us are pretty used to it. Very quickly we go from being amazed that AI can generate a beautifully-written paragraph to wondering when it can generate a beautifully-written novel; or from being amazed that it can make live-saving medical diagnoses to wondering when it can develop the cures; or from being amazed it can create a small computer program to wondering when it can create an entire new company. This is how the singularity goes: wonders become

routine, and then table stakes.

We already hear from scientists that they are two or three times more productive than they were before AI. Advanced AI is interesting for many reasons, but perhaps nothing is quite as significant as the fact that we can use it to do faster AI research. We may be able to discover new computing substrates, better algorithms, and who knows what else. If we can do a decade's worth of research in a year, or a month, then the rate of progress will obviously be quite different.

From here on, the tools we have already built will help us find further scientific insights and aid us in creating better AI systems. Of course this isn't the same thing as an AI system completely autonomously updating its own code, but nevertheless this is a larval version of recursive self-improvement.

There are other self-reinforcing loops at play. The economic value creation has started a flywheel of compounding infrastructure buildout to run these increasingly-powerful AI systems. And robots that can build other robots (and in some sense, datacenters that can build other datacenters) aren't that far off.

If we have to make the first million humanoid robots the old-fashioned way, but then they can operate the entire supply chain—digging and refining minerals, driving trucks, running factories, etc.—to build more robots, which can build more chip fabrication facilities, data centers, etc, then the rate of progress will obviously be quite different.

As datacenter production gets automated, the cost of intelligence should eventually converge to near the cost of electricity. (People are often curious about how much energy a ChatGPT query uses; the average query uses about 0.34 watt-hours, about what an oven would

use in a little over one second, or a high-efficiency lightbulb would use in a couple of minutes. It also uses about 0.000085 gallons of water; roughly one fifteenth of a teaspoon.)

The rate of technological progress will keep accelerating, and it will continue to be the case that people are capable of adapting to almost anything. There will be very hard parts like whole classes of jobs going away, but on the other hand the world will be getting so much richer so quickly that we'll be able to seriously entertain new policy ideas we never could before. We probably won't adopt a new social contract all at once, but when we look back in a few decades, the gradual changes will have amounted to something big.

If history is any guide, we will figure out new things to do and new things to want, and assimilate new tools quickly (job change after the industrial revolution is a good recent example). Expectations will go up, but capabilities will go up equally quickly, and we'll all get better stuff. We will build ever-more-wonderful things for each other. People have a long-term important and curious advantage over AI: we are hard-wired to care about other people and what they think and do, and we don't care very much about machines.

A subsistence farmer from a thousand years ago would look at what many of us do and say we have fake jobs, and think that we are just playing games to entertain ourselves since we have plenty of food and unimaginable luxuries. I hope we will look at the jobs a thousand years in the future and think they are very fake jobs, and I have no doubt they will feel incredibly important and satisfying to the people doing them.

The rate of new wonders being achieved will be immense. It's hard to even imagine today what we will have discovered by 2035; maybe we

will go from solving high-energy physics one year to beginning space colonization the next year; or from a major materials science breakthrough one year to true high-bandwidth brain-computer interfaces the next year. Many people will choose to live their lives in much the same way, but at least some people will probably decide to "plug in".

Looking forward, this sounds hard to wrap our heads around. But probably living through it will feel impressive but manageable. From a relativistic perspective, the singularity happens bit by bit, and the merge happens slowly. We are climbing the long arc of exponential technological progress; it always looks vertical looking forward and flat going backwards, but it's one smooth curve. (Think back to 2020, and what it would have sounded like to have something close to AGI by 2025, versus what the last 5 years have actually been like.)

There are serious challenges to confront along with the huge upsides. We do need to solve the safety issues, technically and societally, but then it's critically important to widely distribute access to superintelligence given the economic implications. The best path forward might be something like:

1. Solve the alignment problem, meaning that we can robustly guarantee that we get AI systems to learn and act towards what we collectively really want over the long-term (social media feeds are an example of misaligned AI; the algorithms that power those are incredible at getting you to keep scrolling and clearly understand your short-term preferences, but they do so by exploiting something in your brain that overrides your long-term preference).

2. Then focus on making superintelligence cheap, widely available,

> and not too concentrated with any person, company, or country. Society is resilient, creative, and adapts quickly. If we can harness the collective will and wisdom of people, then although we'll make plenty of mistakes and some things will go really wrong, we will learn and adapt quickly and be able to use this technology to get maximum upside and minimal downside. Giving users a lot of freedom, within broad bounds society has to decide on, seems very important. The sooner the world can start a conversation about what these broad bounds are and how we define collective alignment, the better.

We (the whole industry, not just OpenAI) are building a brain for the world. It will be extremely personalized and easy for everyone to use; we will be limited by good ideas. For a long time, technical people in the startup industry have made fun of "the idea guys"; people who had an idea and were looking for a team to build it. It now looks to me like they are about to have their day in the sun.

OpenAI is a lot of things now, but before anything else, we are a superintelligence research company. We have a lot of work in front of us, but most of the path in front of us is now lit, and the dark areas are receding fast. We feel extraordinarily grateful to get to do what we do.

Intelligence too cheap to meter is well within grasp. This may sound crazy to say, but if we told you back in 2020 we were going to be where we are today, it probably sounded more crazy than our current predictions about 2030.

May we scale smoothly, exponentially and uneventfully through superintelligence.