Leopold Aschenbrenner, June 2024

V. Parting Thoughts

What if we're right?

> I remember the spring of 1941 to this day. I realized then that a nuclear bomb was not only possible — it was inevitable. Sooner or later these ideas could not be peculiar to us. Everybody would think about them before long, and some country would put them into action. . . .
>
> And there was nobody to talk to about it, I had many sleepless nights. But I did realize how very very serious it could be. And I had then to start taking sleeping pills. It was the only remedy, I've never stopped since then. It's 28 years, and I don't think I've missed a single night in all those 28 years.
>
> JAMES CHADWICH
> (Physics Nobel Laureate and author of the 1941 British government report on the inevitability of an atomic bomb, which finally spurred the Manhattan Project into action)

BEFORE THE DECADE IS OUT, we will have built superintelligence. That is what most of this series has been about. For most people I talk to in SF, that's where the screen goes black. But the decade after—the 2030s—will be at least as eventful. By the end of it, the world will have been utterly, unrecognizably transformed. A new world order will have been forged. But alas—that's a story for another time.

We must come to a close, for now. Let me make a few final remarks.

AGI realism

This is all much to contemplate—and many cannot. "Deep learning is hitting a wall!" they proclaim, every year. It's just another tech boom, the pundits say confidently. But even among those at the SF-epicenter, the discourse has become polarized between two fundamentally unserious rallying cries.

On the one end there are the doomers. They have been obsessing over AGI for many years; I give them a lot of credit for their prescience. But their thinking has become ossified, untethered from the empirical realities of deep learning, their proposals naive and unworkable, and they fail to engage with the very real authoritarian threat. Rabid claims of 99% odds of doom, calls to indefinitely pause AI—they are clearly not the way.

On the other end are the e/accs. Narrowly, they have some good points: AI progress must continue. But beneath their shallow Twitter shitposting, they are a sham; dilettantes who just want to build their wrapper startups rather than stare AGI in the face. They claim to be ardent defenders of American freedom, but can't resist the siren song of unsavory dictators' cash. In truth, they are real stagnationists. In their attempt to deny the risks, they deny AGI; essentially, all we'll get is cool chatbots, which surely aren't dangerous. (That's some underwhelming accelerationism in my book.)

But as I see it, the smartest people in the space have converged on a different perspective, a third way, one I will dub **AGI Realism.** The core tenets are simple:

1. ***Superintelligence is a matter of national security.*** We are rapidly building machines smarter than the smartest humans. This is not another cool Silicon Valley boom; this isn't some random community of coders writing an innocent open source software package; this isn't fun and games. Superintelligence is going to be *wild*; it will be the most powerful weapon mankind has ever built. And for any of us involved, it'll be the most important thing we ever do.

2. ***America must lead.*** The torch of liberty will not survive Xi getting AGI first. (And, realistically, American leadership is the only path to safe AGI, too.) That means we can't simply "pause"; it means we need to rapidly scale up US power production to build the AGI clusters in the US. But it also means amateur startup security delivering the nuclear secrets to the CCP won't cut it anymore, and it means the core AGI infrastructure must be controlled by America, not some dictator in the Middle East. American AI labs must put the national interest first.

3. ***We need to not screw it up.*** Recognizing the power of superintelligence also means recognizing its peril. There are very real safety risks; very real risks this all goes awry—whether it be because mankind uses the destructive power brought forth for our mutual annihilation, or because, yes, the alien species we're summoning is one we cannot yet fully control. These are manageable—but improvising won't cut it. Navigating these perils will require good people bringing a level of seriousness to the table that has not yet been offered.

As the acceleration intensifies, I only expect the discourse to get more shrill. But my greatest hope is that there will be those who feel the weight of what is coming, and take it as a solemn call to duty.

What if we're right?

At this point, you may think that I and all the other SF-folk are totally crazy. But consider, just for a moment: *what if they're right?* These are the people who invented and built this technology; they think AGI will be developed this decade; and, though there's a fairly wide spectrum, many of them take very seriously the possibility that the road to superintelligence will play out as I've described in this series.

Almost certainly I've gotten important parts of the story wrong; if reality turns out to be anywhere near this crazy, the error bars will be very large. Moreover, as I said at the outset, I think there's a wide range of possibilities. But I think it is impor-

tant to be concrete. And in this series I've laid out what I currently believe is the single most likely scenario for the rest of the decade—the rest of *this* decade.

Because—it's starting to feel real, *very* real. A few years ago, at least for me, I took these ideas seriously—but they were abstract, quarantined in models and probability estimates. Now it feels extremely visceral. I can *see* it. I can *see* how AGI will be built. It's no longer about estimates of human brain size and hypotheticals and theoretical extrapolations and all that—I can basically tell you the cluster AGI will be trained on and when it will be built, the rough combination of algorithms we'll use, the unsolved problems and the path to solving them, the list of people that will matter. I can *see* it. It is extremely visceral. Sure, going all-in leveraged long Nvidia in early 2023 has been great and all, but the burdens of history are heavy. I would not choose this.

But the scariest realization is that *there is no crack team coming to handle this*. As a kid you have this glorified view of the world, that when things get real there are the heroic scientists, the uber-competent military men, the calm leaders who are on it, who will save the day. It is not so. The world is incredibly small; when the facade comes off, it's usually just a few folks behind the scenes who are the live players, who are desperately trying to keep things from falling apart.

Right now, there's perhaps a few hundred people in the world who realize what's about to hit us, who understand just how crazy things are about to get, who have situational awareness. I probably either personally know or am one degree of separation from everyone who could plausibly run The Project. The few folks behind the scenes who are desperately trying to keep things from falling apart are you and your buddies and their buddies. That's it. That's all there is.

Someday it will be out of our hands. But right now, at least for the next few years of midgame, the fate of the world rests on these people.

Will the free world prevail?

Will we tame superintelligence, or will it tame us?

Will humanity skirt self-destruction once more?

The stakes are no less.

These are great and honorable people. But they are just people. Soon, the AIs will be running the world, but we're in for one last rodeo. May their final stewardship bring honor to mankind.