# Executive Summary

The Compendium explains the extinction risks from AI, where they come from, and how to address them.

The trigger to the current situation is the explosion of AI progress in the last 10 years, reaching near-human-level in writing, coding, art, math, and many more fields of human activity. This progress has been driven by deep learning, which differs from traditional software: modern AIs are grown by feeding them massive amounts of data and letting them evolve in response, not built piece by piece by humans. First, this means that researchers and engineers don't need to understand AIs to create them – indeed experts consistently fail to anticipate how quickly new skills will be unlocked, or how existing AIs work. Second, this makes continued progress bottlenecked only by resources (such as AI chips, electrical power, data…) instead of scientific insights. As tech giants and frontier AI companies collaborate to unlock ever more resources, the path forward leads to increasingly intelligent yet opaque AIs.

How far can this trend of smarter and smarter AI go? If we look for inspiration at how humanity has historically increased its intelligence (understood as its ability to solve more and more intellectual tasks), three strategies emerge: tools, groups, and methods. AIs can leverage the exact same strategies; indeed each of them is already being used by current AI R&D! And although skeptics claim that various hypothetical components of intelligence cannot be automated, no existing scientific theory of intelligence backs these opinions. Thus we should assume the trend towards smarter AIs will continue, eventually leading to Artificial General Intelligence (AGI), AIs able to do the same intellectual tasks as humans.

Since AGI would be able to do anything a human can, it would notably be able to do AI research, and so to improve its own intelligence and capabilities. This is not the only way forward, yet AGI companies and AI researchers are already pushing hard in this direction. Since software is cheaper, more efficient, and easier to correct than brains, AGI would improve far faster than any human, eventually reaching artificial superintelligence (ASI) surpassing humanity's collective intelligence. As it continues to scale, ASI would unlock abilities to shape matter and energy that would look godlike compared to human engineering. And even without malicious intent, these godlike AIs would by default wipe out humanity as collateral damage while pursuing their own goals, in the same way ants are just collateral damage for contractors building a house.

Godlike-AIs lead to catastrophe because of the incredible difficulty of aligning AI's goals with those of a single human, let alone of humanity. Alignment is the harder version of the kind of problems with which humanity already struggles: for example making companies

and governments beneficial for what the actual citizens care and believe in. Solving alignment would require massive progress on questions like finding what we value and reconciling contradictions between values, predicting the consequences of our actions to avoid unintended side-effects, and design processes from the people's will to AIs doing the correct things. If we were serious about solving alignment, it would require at the very least decades of top-notch research and trillions of dollars of investment; yet only a handful of people and a couple 100s of millions are currently invested, with most of the money and effort going instead towards making AIs more powerful. Even worse, the little work that exists doesn't even try to pay the cost of alignment: instead it reacts to current issues with AIs by patching them in a whack-a-mole fashion, and passes the buck to future smarter AIs. Thus we're not on track to get anywhere near solving alignment, and thus godlike-AI would cause human extinction.

Lacking a solution to alignment, we need to ensure godlike-AIs are not built. This requires institutions with the authority to regulate frontier AI research, both at the national and international level. Yet these institutions simply do not exist, very little is being done to create them, and the little governance work already finds itself undermined by the very AI companies racing to AGI.

This lack of promising effort on alignment and regulation is not a coincidence: frontier AI companies are systematically undermining these to race to AGI without blockers. The root of this behavior lies in the ideologies of frontier AI companies: they are by and large utopists, who want to build AGI because they believe it will usher their ideal world. This belief brings with it a fear that AGI will be built by the "wrong" people, and so these utopists become more and more willing to cut corners in order to avoid this, undermining safety along the way. In practice, this looks like the key tactics of the industry playbook used by Big Tech, Big Oil, and Big Tobacco: spreading fear (by stoking geopolitical fires) and doubt (by changing their stances constantly) to free their path to AGI, capturing regulatory efforts under cover of self-regulation, and undermining research that might force them to slow down.

So we are on a dark trajectory, one that recklessly leads to human extinction. What can be done about this? We believe that giving up and reading the situation as hopeless is a mistake, one that the utopists and other actors racing to AGI want us to make. Instead, there is a narrow path forward, one that starts with basic civic duty. The people racing towards AGI are only a tiny minority, who are deciding to put everyone at risk to follow their delusions. Because no one in their right mind wants humanity to go extinct, this is an issue that can unify people across party lines and countries. This starts by spreading the word and awareness of the risks, and exercising basic civic duty by contacting your representatives and voting according to the extinction risks posed by AI.