**Mission**

[Company] is an AI research and deployment company. Our mission is to ensure that artificial general intelligence benefits all of humanity.

Our vision for the future of AGI

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

We are building safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome.

Our mission is to ensure that artificial general intelligence—AI systems that are generally smarter than humans—benefits all of humanity.

If AGI is successfully created, this technology could help us elevate humanity by increasing abundance, turbocharging the global economy, and aiding in the discovery of new scientific knowledge that changes the limits of possibility.

AGI has the potential to give everyone incredible new capabilities; we can imagine a world where all of us have access to help with almost any cognitive task, providing a great force multiplier for human ingenuity and creativity.

On the other hand, AGI would also come with serious risk of misuse, drastic accidents, and societal disruption. Because the upside of AGI is so great, we do not believe it is possible or desirable for society to stop its development forever; instead, society and the developers of AGI have to figure out how to get it right.

AGI could happen soon or far in the future; the takeoff speed from the initial AGI to more powerful successor systems could be slow or fast. Many of us think the safest quadrant in this two-by-two matrix is short timelines and slow takeoff speeds; shorter timelines seem more amenable to coordination and more likely to lead to a slower takeoff due to less of a compute overhang, and a slower takeoff gives us more time to figure out empirically how to solve the safety problem and how to adapt.

Although we cannot predict exactly what will happen, and of course our current progress could hit a wall, we can articulate the principles we care about most:

We want AGI to empower humanity to maximally flourish in the universe. We don’t expect the future to be an unqualified utopia, but we want to maximize the good and minimize the bad, and for AGI to be an amplifier of humanity.

We want the benefits of, access to, and governance of AGI to be widely and fairly shared.

We want to successfully navigate massive risks. In confronting these risks, we acknowledge that what seems right in theory often plays out more strangely than expected in practice. We believe we have to continuously learn and adapt by deploying less powerful versions of the technology in order to minimize “one shot to get it right” scenarios.

**The short term**

There are several things we think are important to do now to prepare for AGI.

First, as we create successively more powerful systems, we want to deploy them and gain experience with operating them in the real world. We believe this is the best way to carefully steward AGI into existence—a gradual transition to a world with AGI is better than a sudden one. We expect powerful AI to make the rate of progress in the world much faster, and we think it’s better to adjust to this incrementally.

A gradual transition gives people, policymakers, and institutions time to understand what’s happening, personally experience the benefits and downsides of these systems, adapt our economy, and to put regulation in place. It also allows for society and AI to co-evolve, and for people collectively to figure out what they want while the stakes are relatively low.

We currently believe the best way to successfully navigate AI deployment challenges is with a tight feedback loop of rapid learning and careful iteration. Society will face major questions about what AI systems are allowed to do, how to combat bias, how to deal with job displacement, and more. The optimal decisions will depend on the path the technology takes, and like any new field, most expert predictions have been wrong so far. This makes planning in a vacuum very difficult.

For example, when we first started [company], we didn’t expect scaling to be as important as it has turned out to be. When we realized it was going to be critical, we also realized our original structure wasn’t going to work—we simply wouldn’t be able to raise enough money to accomplish our mission as a nonprofit—and so we came up with a new structure.

As another example, we now believe we were wrong in our original thinking about openness, and have pivoted from thinking we should release everything (though we open source some things, and expect to open source more exciting things in the future!) to thinking that we should figure out how to safely share access to and benefits of the systems. We still believe the benefits of society understanding what is happening are huge and that enabling such understanding is the best way to make sure that what gets built is what society collectively wants (obviously there’s a lot of nuance and conflict here).

Generally speaking, we think more usage of AI in the world will lead to good, and want to promote it (by putting models in our API, open-sourcing them, etc.). We believe that democratized access will also lead to more and better research, decentralized power, more benefits, and a broader set of people contributing new ideas.

As our systems get closer to AGI, we are becoming increasingly cautious with the creation and deployment of our models. Our decisions will require much more caution than society usually applies to new technologies, and more caution than many users would like. Some people in the AI field think the risks of AGI (and successor systems) are fictitious; we would be delighted if they turn out to be right, but we are going to operate as if these risks are existential.

At some point, the balance between the upsides and downsides of deployments (such as empowering malicious actors, creating social and economic disruptions, and accelerating an unsafe race) could shift, in which case we would significantly change our plans around continuous deployment.

As our systems get closer to AGI, we are becoming increasingly cautious with the creation and deployment of our models.

Second, we are working towards creating increasingly aligned and steerable models. Our shift from models like the first version of [product] to [product] and [product] is an early example of this.

In particular, we think it’s important that society agree on extremely wide bounds of how AI can be used, but that within those bounds, individual users have a lot of discretion. Our eventual hope is that the institutions of the world agree on what these wide bounds should be; in the shorter term we plan to run experiments for external input. The institutions of the world will need to be strengthened with additional capabilities and experience to be prepared for complex decisions about AGI.

The “default setting” of our products will likely be quite constrained, but we plan to make It easy for users to change the behavior of the AI they’re using. We believe in empowering individuals to make their own decisions and the inherent power of diversity of ideas.

We will need to develop new alignment techniques as our models become more powerful (and tests to understand when our current techniques are failing). Our plan in the shorter term is to use AI to help humans evaluate the outputs of more complex models and monitor complex systems, and in the longer term to use AI to help us come up with new ideas for better alignment techniques.

Importantly, we think we often have to make progress on AI safety and capabilities together. It’s a false dichotomy to talk about them separately; they are correlated in many ways. Our best safety work has come from working with our most capable models. That said, it’s important that the ratio of safety progress to capability progress increases.

Third, we hope for a global conversation about three key questions: how to govern these systems, how to fairly distribute the benefits they generate, and how to fairly share access.

In addition to these three areas, we have attempted to set up our structure in a way that aligns our incentives with a good outcome. We have a clause in our Charter about assisting other organizations to advance safety instead of racing with them in late-stage AGI development. We have a cap on the returns our shareholders can earn so that we aren’t incentivized to attempt to capture value without bound and risk deploying something potentially catastrophically dangerous (and of course as a way to share the benefits with society). We have a nonprofit that governs us and lets us operate for the good of humanity (and can override any for-profit interests), including letting us do things like cancel our equity obligations to shareholders if needed for safety and sponsor the world’s most comprehensive UBI experiment.

We have attempted to set up our structure in a way that aligns our incentives with a good outcome.

We think it’s important that efforts like ours submit to independent audits before releasing new systems; we will talk about this in more detail later this year. At some point, it may be important to get independent review before starting to train future systems, and for the most advanced efforts to agree to limit the rate of growth of compute used for creating new models. We think public standards about when an AGI effort should stop a training run, decide a model is safe to release, or pull a model from production use are important. Finally, we think it’s important that major world governments have insight about training runs above a certain scale.

**The long term**

We believe that the future of humanity should be determined by humanity, and that it’s important to share information about progress with the public. There should be great scrutiny of all efforts attempting to build AGI and public consultation for major decisions.

The first AGI will be just a point along the continuum of intelligence. We think it’s likely that progress will continue from there, possibly sustaining the rate of progress we’ve seen over the past decade for a long period of time. If this is true, the world could become extremely different from how it is today, and the risks could be extraordinary. A misaligned superintelligent AGI could cause grievous harm to the world; an autocratic regime with a decisive superintelligence lead could do that too.

AI that can accelerate science is a special case worth thinking about, and perhaps more impactful than everything else. It’s possible that AGI capable enough to accelerate its own progress could cause major changes to happen surprisingly quickly (and even if the transition starts slowly, we expect it to happen pretty quickly in the final stages). We think a slower takeoff is easier to make safe, and coordination among AGI efforts to slow down at critical junctures will likely be important (even in a world where we don’t need to do this to solve technical alignment problems, slowing down may be important to give society enough time to adapt).

Successfully transitioning to a world with superintelligence is perhaps the most important—and hopeful, and scary—project in human history. Success is far from guaranteed, and the stakes (boundless downside and boundless upside) will hopefully unite all of us.

We can imagine a world in which humanity flourishes to a degree that is probably impossible for any of us to fully visualize yet. We hope to contribute to the world an AGI aligned with such flourishing.