Microsoft wants to build artificial general intelligence: an AI better than humans at everything

A lot of startups in the San Francisco Bay Area claim that they're planning to transform the world. San-Francisco-based, Elon Musk-founded OpenAI has a stronger claim than most: It wants to build artificial general intelligence (AGI), an AI system that has, like humans, the capacity to reason across different domains and apply its skills to unfamiliar problems.

Today, it announced a billion dollar partnership with Microsoft to fund its work — the latest sign that AGI research is leaving the domain of science fiction and entering the realm of serious research.

"We believe that the creation of beneficial AGI will be the most important technological development in human history, with the potential to shape the trajectory of humanity," Greg Brockman, chief technology officer of OpenAI, said in a press release today.

Existing AI systems beat humans at lots of narrow tasks — chess, Go, Starcraft, image generation — and they're catching up to humans at others, like translation and news reporting. But an artificial general intelligence would be one system with the capacity to surpass us at all of those things. Enthusiasts argue that it would enable centuries of technological advances to arrive, effectively, all at once — transforming medicine, food production, green technologies, and everything else in sight.

Others warn that, if poorly designed, it could be a catastrophe for humans in a few different ways. A sufficiently advanced AI could pursue a goal that we hadn't intended — a recipe for catastrophe. It could turn out unexpectedly impossible to correct once running. Or it could be maliciously used by a small group of people to harm others. Or it could just make the rich richer and leave the rest of humanity even further in the dust.

Getting AGI right may be one of the most important challenges ahead for humanity. Microsoft's billion dollar investment has the potential to push the frontiers forward for AI development, but to get AGI right, investors have to be willing to prioritize safety concerns that might slow commercial development.

A transformative technology with enormous potential benefits — and real risks

Some analysts have compared the development of AGI to the development of electricity. It's not just one breakthrough; it enables countless other changes in the way we live our lives.

But the announcement also nods at the ways this could go wrong. OpenAI's team working on the safety and policy implications of AGI has been unafraid to articulate ways that AGI could be a disaster rather than a boon.

"To accomplish our mission of ensuring that AGI (whether built by us or not) benefits all of

humanity," Brockman says in the release, "we'll need to ensure that AGI is deployed safely and securely; that society is well-prepared for its implications; and that its economic upside is widely shared."

Those are hard problems. Current AI systems are vulnerable to adversarial examples — inputs designed to confuse them — and more advanced systems might be, too. Current systems faithfully do what we tell them to do, even if it's not exactly what we meant them to do.

And there are some reasons to think advanced systems will have problems that current systems don't. Some researchers have argued that an AGI system that appears to be performing well at a small scale might unexpectedly deteriorate in performance when it has more resources available to it, as the best route to achieving its goals changes. (You can imagine this by thinking about a company that follows the rules when it's small and scrutinized, but cheats on them or lobbies to get them changed once it has enough clout to do so.)

Even AGI's most enthusiastic proponents think there's a lot of potential for things to go wrong — they just think the benefits of developing AGI are worth it. A success with AGI could let us address climate change, extreme poverty, pandemic diseases, and whatever new challenges are around the corner, by identifying promising new drugs, optimizing our power grid, and speeding up the rate at which we develop new technologies.

So how far away is AGI? Here, experts disagree. Some estimate that we're only a decade away while others point out that there's been optimism that AGI is just around the corner for a long time and it has never arrived.

The disagreements don't fall along obvious lines. Some academics, such as MIT's Max Tegmark, are among those predicting AI soon, while some key figures in industry, such as Facebook's Yann LeCun, are among those who think it's likely fairly distant. But they do agree that it's possible and will happen someday, and that makes it one of the big open challenges of this century.

OpenAI shifted gears this year toward raising money from investors

Until this year, OpenAI was a nonprofit. (Musk, one of its founders, left the board in 2018, citing conflicts of interest with Tesla.) Earlier this year, that changed. Instead of a nonprofit, they announced they'll operate from now on as a new kind of company called OpenAI LP (the LP stands for "limited partnership").

Why the change, which critics interpreted as a betrayal of the nonprofit's egalitarian mission? OpenAI's leadership team had become convinced that they couldn't stay on the cutting edge of the field and help shape the direction of AGI without an infusion of billions of dollars, and that's hard for a nonprofit to get.

But taking investment money would be a slippery slope toward abandoning their mission: Once you have investors, you have obligations to maximize their profits, which is incompatible with ensuring that the benefits of AI are widely distributed.

OpenAI LP (the structure that was used to raise the Microsoft money) is meant to solve that

dilemma. It's a hybrid, OpenAI, says, of a for-profit and nonprofit, the company promises to pay shareholders a return on their investment, up to 100 times what they put in. Everything beyond that goes to the public. The OpenAI nonprofit board still oversees everything.

That sounds a bit ridiculous — after all, how much can possibly be left over after paying investors 100 times what they paid in? — but early investors in many tech companies have made far more than 100 times what they invested. Jeff Bezos **reportedly invested** $250,000 in Google back in 1998; if he held onto those shares, they'd be worth more than $3 billion today. If Google had adopted OpenAI LP's cap on returns, Bezos would've gotten $25 million dollars — a handsome return on his investment — and the rest would go to humankind.

If OpenAI makes it big, Microsoft will profit immensely — but, they say, so will the rest of us. "Both companies have very aligned missions," Brockman wrote me today in an email: "Microsoft to empower every person and every organization on the planet to achieve more; OpenAI to ensure that artificial general intelligence benefits all of humanity."

Whether such partnerships can drive advances that are good for humanity — or put the brakes on advances that are bad for humanity — increasingly looks like a question everyone should be very interested in answering.

Sign up for the Future Perfect newsletter. Twice a week, you'll get a roundup of ideas and solutions for tackling our biggest challenges: improving public health, decreasing human and animal suffering, easing catastrophic risks, and — to put it simply — getting better at doing good.