Testimony of Mr. Greg Brockman

Chairwoman Barbara Comstock, Chairman Randy Weber, Ranking Member Daniel Lipinski, Ranking Member Marc Veasey, members of both subcommittees, thank you for having me today to deliver testimony on this important topic.

I'm Greg Brockman, co-founder of OpenAI, a non-profit artificial intelligence development organization. Our mission is to ensure that artificial general intelligence (AGI) — by which we mean highly autonomous systems that outperform humans at most economically valuable work — benefits all of humanity.

OpenAI

OpenAI has three main arms: capabilities, safety, and policy. All three of these areas, working in concert, are crucial to achieve our mission. Our capabilities arm is one of the most advanced AI research and development teams in the world. Our safety arm is responsible for developing techniques to ensure that AGI-level systems will operate as their human operators intend. Our policy arm is responsible for researching AGI's social challenges and providing information to policymakers.

On the capabilities side, one milestone in the field is solving complex strategy games[1], which capture many of the aspects of the real world not seen in previous milestones like Chess or Go. We recently announced OpenAI Five[2], a system we've created which has reached the semi-professional level at one of the most complex games played by humans, a ten-player team strategy video game called Dota. This system devises long-term plans and navigates scenarios far too complex to be programmed in by any human. We are aiming to play against top professionals during the Dota world championships in August. OpenAI Five taught itself the rules of the game by playing 180 years worth of games against itself each day. (For comparison, top human professionals have at least 12,000 hours of gameplay, so our system sees as many games each day as 100 human professionals have seen in their lifetimes.)

On the safety side, we recently developed a proof-of-concept technique[3] for allowing humans to monitor the behavior of advanced AI systems. We have also collaborated with Alphabet's subsidiary DeepMind to design AI systems which can learn from the implicit preferences of human trainers.

---

[1] Gershgorn, Dave. The massive global race to teach an AI to beat Starcraft II is under way. https://qz.com/1051052/deepmind-goog-and-facebook-fb-have-started-the-global-sprint-for-ai-to-beat-starcraft-ii/

[2] OpenAI Five. https://blog.openai.com/openai-five/

[3] Irving, Geoffrey, et al. AI Safety via Debate. https://blog.openai.com/debate/

On the policy side, we recently co-authored a report[4] forecasting how malicious actors could misuse AI, including recommendations of how to mitigate these threats. We're helping to develop the AI Index, an AI measurement and analysis initiative, as part of the Stanford One Hundred Year Study on AI. Our goal is to use this experience to make recommendations about how policymakers can measure and analyze the impact of AI on society. And we are attempting to nurture the field of AI policy to ensure there is a deep bench of thinkers about AI across all important actors — companies, research labs, and governments.

Narrow vs general AI

People often talk about narrow vs general AI in terms of whether they apply to one task (narrow) or many tasks (general). But there's also a dimension of competence: can they solve only easy tasks, or can they solve hard tasks? In practice, to build AI systems that solve harder problems, we've ended up creating increasingly general learning systems — since we let the machine learn more on its own rather than having a human provide knowledge or guidance.

Specifically, in the past, AI-like technology was written by humans to solve one specific problem. It wasn't capable of adapting to solve new problems.

In contrast, today's AI is all based on one core technique: the artificial neural network, in a form devised in the 1980s. This is a single, simple idea that is, as it scales, is proving itself to be able to match a surprising amount of human capability. Our neural networks still have a lot of room to grow — to give a sense of scale, though the numbers are not directly comparable, today they usually have around the same number of artificial neurons as an insect has biological ones.

In the 1980s, computers could only run tiny neural networks, so the resulting systems couldn't solve interesting problems. In 2012, computers were fast enough for a team of researchers (including my co-founder Ilya Sutskever) to train a large enough neural network to perform well on the task of categorizing images — performing far better than any other method. The neural network learned to categorize images by being shown many examples of already-categorized images, and this is now the dominant approach in the field rather than the previously hand-crafted rules (which were limited in performance). Since then, neural networks have become the standard tool for solving problems in a variety of fields such as speech recognition and machine translation.

To give you a sense of progress, here are some AI advances from recent years:

- **Image recognition:** AI's ability to correctly categorize images has gone from 75% (pre-neural network, 2011) accuracy to around 98% accuracy (neural network, 2017) on a difficult standard benchmark on which human accuracy is around 95%.

[4] Clark, Jack, et al. Preparing for Malicious Uses of AI. https://blog.openai.com/preparing-for-malicious-uses-of-ai/

- **Fake images & videos:** AI techniques are increasingly able to generate convincing fake images and videos — including fakes of politicians, such as Vladimir Putin and President Trump. In 2014, the best generated images were low-resolution images of fake people; by 2017, they were photorealistic faces that humans have trouble distinguishing from real ones[5]. Also in 2017, free software became available allowing people to create their own “deepfake” images.
- **Translation:** In 2014, researchers developed “neural machine translation” — where computers learn to translate between languages using only large datasets, lacking any of the specific rules which human translators use to do their work. In 2016, Google Translate performance significantly improved by switching to it; in 2017 Facebook improved its site translation by doing the same.
- **Speech recognition:** Due to switching to neural networks, over the past few years speech recognition went from barely working (such as we’ve all experienced when calling an automated phone tree) to running on smartphones with much higher accuracy.
- **Sophistication:** The complexity of games where neural network-based AIs can rival the top human players has increased in complexity from 1970s Atari games like Space Invaders or Breakout (2013) to rich strategic games like the board game Go (2015) to modern real-time strategy games like Dota in both 1-versus-1 (2017) and 5-versus-5 (2018) formats.

These advances, with neural networks at the core of all of them, are more general than past systems, and when trained properly, can achieve unprecedented performance at one or more interesting tasks. A system that can learn image recognition at record-setting levels could also learn to do the same in speech recognition. The tools used to generate fake images of politicians could also be used to synthesize new artistic paintings, or imaginary architectural plans. Neural machine translation systems can learn to translate between any pair of languages — provided we have the training data.

The next step along the spectrum are future AI systems that can accomplish very hard, valuable real-world tasks such as:

- automatically devising and performing scientific experiments in chemistry or neuroscience
- helping us design better or cheaper drugs, cars, computer software and hardware, and public infrastructure like transport or logistics systems
- performing surgery with more precision, safety, and efficiency than is possible for human surgeons, to
- orchestrating the movements of thousands of self-driving cars and drones across a city or rural area, whether to deliver ordinary goods or provide emergency supplies during an extreme weather event.

[5] Brundage, Miles, et al. The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, page 15. https://arxiv.org/pdf/1802.07228.pdf

What is AGI?

AGI is even further along the spectrum of generality. Our working definition of AGI is systems that are sufficiently advanced that they can outperform humans at most economically valuable work — which includes starting companies, making business deals, and writing books. Such technology will need to be smart in a way unlike our traditionally literal, uncreative computers. The system's generality means it wouldn't be limited to commercial applications — it could also help with reasoning through complex international disputes, city planning, and even lawmaking or running countries. Rather than being developed for any one use-case, AGI would be developed for an entire spectrum of important tasks.

Why AGI could be developed sooner than commonly expected

AI systems are built on three foundations: algorithms, data, and compute (or amount of computational resource). Next generation AI systems being developed today are relying less on conventional datasets, since they can either consume freely-available unlabeled data (like a recent state-of-the-art language model we released which learns from an open dataset of books) or can expend compute to generate data. For example, by simulating a robot, we can create training data in quantities limited only by the number of computers available to run the simulation.

We recently released a study[6] showing that the amount of compute powering the largest AI training runs has been doubling every 3.5 months for the past six years (a total increase of 300,000x). This growth is significantly faster than the historic driver of hardware progress, Moore's Law, which had an 18 month doubling period (a 12x increase over the same period). AI compute progress is driven partly by faster computers, and partly by figuring out how to effectively train AI systems on many computers simultaneously. This means that data and compute are rapidly becoming less significant bottlenecks on AI progress.

We expect this trend to continue. We track over 45 hardware startups (most in the US) that are building next-generation AI computers. Most are building on proven technologies that do not require further breakthroughs like quantum or optical. As these computers hit the market, and as we figure out how to use many such computers at once, we expect the rate of breakthrough results to continue apace or even accelerate.

Our current computational paradigm allows for substantial increases in compute each year for at least the next five years. Will the incoming tsunami of compute (combined with near-term improvements in algorithmic understanding) be enough to develop AGI, or will we need to wait for some future algorithmic or hardware breakthrough? We don't know the answer to this question yet, but given today's rapid progress, it seems unwise to be too confident in either direction — at least before uncovering further evidence.

---

[6] Amodei, Dario, et al. AI and Compute. https://blog.openai.com/ai-and-compute

The post-AGI future

Investment in AI research is increasing rapidly due to how quickly AI advances can be deployed into products. Transformative applications on the horizon like self-driving cars promise to save lives, increase efficiency, and generate huge value, with the potential to add trillions of dollars to US GDP[7].

After AGI is created, we expect economic and technological growth to accelerate markedly — with the new growth driven primarily by teams of creative computers partnering with creative humans. We'll have the technological means to not just generate but also distribute essential resources (and hopefully much more) to ensure no one falls through the cracks, and will be able to concentrate on efforts like education, re-training, and re-skilling, to help people navigate the new economy. The benefits will be vast, and OpenAI believes those benefits should be equitably distributed, rather than locked up with any one entity.

Technological progress has been accelerating rapidly for the past few hundred years, and we expect the post-AGI world to add another jolt to the rate of progress. We should expect advances in curing disease, life extension, transportation and space travel, and communication.

Challenges of AGI development

Each stage of AI development will bring its own challenges.

Narrow AI challenges are easiest to understand and act on because they apply to existing systems. These are also the ones that today's corporations are most incentivized to fix. These challenges include issues such as fairness, transparency, privacy, and bias — all of which require serious attention if we want even more advanced AI technologies to have a positive impact. We expect narrow AI progress to increase the rate of technological progress across the board, further challenging today's policy machinery to keep pace; as we approach AGI, we expect the rate of change to increase further.

AGI challenges are harder to understand and foresee, partly because they apply to systems that have not yet been developed. OpenAI focuses on AGI challenges because we believe that they are simultaneously neglected and may happen sooner than is commonly believed.

The core danger with AGI is that it has the potential to cause rapid change. This means we could end up in an undesirable environment before we have a chance to realize where we're even heading. The exact way the post-AGI world will look is hard to predict — that world will likely be more different from today's world than today's is from the 1500s. Some open questions:

---

[7] Lanctot, Roger. Accelerating the Future: The Economic Impact of the Emerging Passenger Economy, page 5. https://newsroom.intel.com/newsroom/wp-content/uploads/sites/11/2017/05/passenger-economy.pdf?cid=em-elq-26916&utm_source=elq&utm_medium=email&utm_campaign=26916&elq_cid=1494219

- What is the nature of international society in the post-AGI world? We've already seen technology amplify the impact that states and small groups have in the world. We expect AGI to further this trend.
- What will people do with their time as economic work becomes an increasingly smaller part of one's life? How can we help people live meaningful, enjoyable lives in such a world?
- How can AGI be deployed into our economic and social systems in a way that amplifies human preferences? Will all humans remain meaningful actors in society? Only humans in countries with powerful AGI deployments? Only the humans who own a share in the technology itself?
- How do we ensure that AGIs operate in line with the values of their operators? How do we avoid creating systems that cause social harm in blind pursuit of a poorly-specified goal — the technological version of what happened to Bear Stearns in 2008.

AGI will affect every sector of global society, and given the difficulty of these questions, we don't have the luxury of waiting to see how AGI starts affecting society before addressing its challenges. One example issue worth considering today is the possibility of a military arms race toward AGI. A military arms race would put pressure on deploying an AGI without adequately verifying that it is safe. AGI deployment will be challenging enough without pressure to gamble with safety. (Similar considerations also apply to pre-AGI AI technologies.)

Safe and responsible AGI development

Our views on safe and responsible AGI development are captured in three of the four sections of our Charter[8]: "Broadly Distributed Benefits", "Long-Term Safety", and "Cooperative Orientation".

- **Safety**. We do not yet know how hard it will be to make sure AGIs act according to the values of their operators. Some people believe it will be easy; some people believe it'll be unimaginably difficult; but no one knows for sure — which is why OpenAI believes that safety research is critically important. At the very least, any AGI project should leave enough time to get safety right. This includes taking steps — well in advance of the development of AGI – to avoid an uncoordinated race. In this vein, our Charter commits us to assisting rather than competing with a value-aligned, safety-conscious project that comes close to building AGI before we do.
- **Broadly Distributed Benefits**. AGI will create unprecedented economic benefits. If AGI can truly produce not just a Microsoft-sized amount of value, but 100 Microsofts or more, then returns beyond some point should not exclusively belong to a small group of people. The rest of humanity will have assumed the risks of developing and deploying AGI, and everyone deserves a fair share in the post-AGI future.

[8] OpenAI Charter. https://blog.openai.com/openai-charter/

- **Cooperative Orientation**. AGI has the potential to be the most socially beneficial technology humans ever create. The world is bigger than any one project, and any society which successfully builds safe AGI will win collectively. Thus, it's important that value-aligned AGI projects view themselves in "friendly competition". Today, we are all competing for talent and prestige. But we need the ability to come together under one roof in some form before building such a powerful system, bringing together companies (and hopefully governments) to ensure the resulting technology benefits everyone.

Policy recommendations

1. **Measurement**. Many other established voices in the field have tried to combat panic about AGI by instead saying it not something to worry about or is unfathomably far off. We recommend neither panic nor a lack of caution. Instead, we recommend investing more resources into understanding where the field is, how quickly progress is accelerating, and what roadblocks might lie ahead. We're exploring this problem via our own research and support of initiatives like the AI Index. But there's much work to be done, and we are available to work with governments around the world to support their own measurement and assessment initiatives — for instance, we participated in a GAO-led study on AI last year.
2. **Foundation for international coordination.** AGI's impact, like that of the Internet before it, won't track national boundaries. Successfully using AGI to make the world better for people, while simultaneously preventing rogue actors from abusing it, will require international coordination of some form. Policymakers today should invest in creating the foundations for successful international coordination in AI, and recognize that the more adversarial the climate in which AGI is created, the less likely we are to achieve a good outcome. We think the most practical place to start is actually with the measurement initiatives: each government working on measurement will create teams of people who have a strong motivation to talk to their international counterparts to harmonize measurement schemes and develop global standards.

It's easy to imagine the post-AGI world as a destination — but it is more of an arbitrary marker denoting a world with transformative AI technologies. There are many open questions around AGI, and the more we can understand where the field is, how fast we are moving, and what is likely to happen in upcoming years, the better prepared we will be to answer them. And perhaps the most important question AGI raises is that once the world has been fundamentally transformed by systems that perform tasks we'd historically thought of as "human" — what then?