Global Challenges Foundation

# Recent progress in AI and efforts to ensure its safety


## Risk Experts

**Anthony Aguirre,** Co-founder, Future of Life Institute
**Max Tegmark,** President and Co-founder, Future of Life Institute
**Ariel Conn,** Director of Media and Outreach, Future of Life Institute
**Richard Mallah,** Director of AI Projects, Future of Life Institute
**Victoria Krakovna,** Co-founder, Future of Life Institute


After many decades of slow but continuous progress, the last few years have seen an explosion of artificial intelligence (AI) capabilities, leading to better data analysis, increased automation, more efficient machine learning systems, and more general research interest from academics, governments, and corporations.

Last year, Google DeepMind shocked the AI world when it revealed AlphaGo, a program that had learned to master the famous game of Go. This classic challenge to AI had been expected to require at least another decade. Concurrently, a slew of AI advances repeatedly surprised and impressed AI researchers. Google Translate became strikingly better. Machines learned to accurately describe what is taking place in a picture, and to create images based on minimal descriptions of a scene. Self-driving cars are closer to becoming a daily reality. Programs are being developed that can mimic someone's voice, which can then be added to an AI-generated video. More generally, AI is learning to do more and more with less and less data, highlighting AI's huge potential for solving humanity's greatest problems.

But there were also debacles, such as Microsoft's Twitter chatbot Tay, which learned to be racist and sexist in under 24 hours, and Google's image classifier, which identified dark-skinned people as gorillas. Indeed, artificial intelligence, like all powerful technologies, naturally has risks associated with it.

Most immediately, the World Economic Forum predicts that five million jobs will be automated by 2020, and many experts fear this number will grow too rapidly for society to adjust. Looking forward, as AI advances, there is potential for major disruption, both positive and negative. Humans are the most powerful species on the planet because of our intelligence, so machines smarter than us could pose opportunities and risks unlike anything previously seen with other technologies — which could unfold with stunning speed if AIs learn to create better AIs.

*"The World Economic Forum predicts five million jobs will be automated by 2020, and many experts fear this number will grow too rapidly for society to adjust."*

In 2014, Nick Bostrom's book Superintelligence raised public awareness of AI related risk, and prominent thinkers such as Elon Musk, Stephen Hawking, and Bill Gates expressed concern about AI. A groundbreaking 2015 meeting in Puerto Rico helped mainstream such concerns, after which thousands of AI researchers signed open letters supporting research on how to keep AI beneficial and opposing an arms-race in AI-powered weapons. This mainstreaming helped trigger seed funding for dozens of teams around the world to research how to keep AI safe and beneficial.

By 2016, a significant response to AI risk was underway. Multiple efforts were made to map out the landscape of research required to ensure AI safety, and to tackle some of the basic questions. For example, researchers at Google and the Future of Humanity Institute presented steps toward ensuring that, if an AI does something we don't like, we can safely turn it off without it acting to prevent us from doing so. But these efforts are just the beginning of what AI safety researchers predict will be major technical and intellectual challenges en route to beneficial AI.

Moreover, society will need to adapt to the rapidly changing AI landscape in order to manage it. Many governments, businesses, and non-profits started to take action in 2016. Perhaps the biggest news was the formation of the Partnership on AI, which currently includes the Association for the Advancement of Artificial Intelligence, the American Civil Liberties Union, Amazon, Apple, DeepMind, Google, Facebook, IBM, Microsoft, and OpenAI. The White House, Stanford, and the Institute of Electrical and Electronics Engineers all produced reports outlining how to tackle challenges that AI may pose. In 2017, these and other guidelines were distilled into the Asilomar AI Principles, signed by over 1,000 AI researchers from around the world, aimed at ensuring that AI development will benefit humanity as a whole. The rapid development of AI portends significant changes and possible dangers unfolding over the coming decades, but with careful management, research, and cooperation, AI has the potential to become the most beneficial technology ever developed.