APRIL 2020

# Why AI Chips Matter

CSET Policy Brief


**AUTHOR**
Saif M. Khan

As the demand for artificial intelligence applications increases and their complexity grows, the computational power necessary to develop and deploy these applications will become increasingly expensive.[1] Specially designed "AI chips"[2] fabricated using "state-of-the-art"[3] technology are necessary to achieve the combination of performance and cost-effectiveness needed for these AI applications. The United States and its allies have a strategic advantage in state-of-the-art AI chip production that should be maintained, if not increased.

**Cutting-edge AI applications are computationally expensive, bottlenecking progress in AI development and deployment.** Rising demand for developing AI applications—from autonomous drones to cybersecurity—will require massively expensive amounts of computation. Since 2012, the amount of computation needed to train cutting-edge AI algorithms has doubled every few months.[4] AI lab computing costs have skyrocketed as a result. Training the leading algorithms of Google's DeepMind, including AlphaGo, costs between $5 and $100 million.[5] At OpenAI, computation accounted for nearly 30 percent of costs in 2017.[6] Once trained, the ongoing operation of advanced AI algorithms also requires large amounts of computation.[7]

**The cost of computation is therefore a major differentiator of AI capability, determined by state-of-the-art manufacturing technology and chip design**. AI chips made with state-of-the-art technology have two important characteristics: they feature the smallest transistors and are optimized for AI-specific calculations. Such chips deliver cost-effectiveness as a result of lower production and operational costs per calculation. Chips manufactured using state-of-the-art technologies are at least 10 times more cost-effective than similar old technology chips over the three-year life-cycle of a typical chip. AI-specialized chip designs are an additional 10 to 1,000 more cost-effective for training AI algorithms than ordinary chips,[8] for a combined 100 to 10,000 times improvement in cost-effectiveness for state-of-the-art AI chips. As a result, competitiveness in AI development and deployment requires state-of-the-art AI chips.[9]

**The United States and a small number of allied democracies currently dominate state-of-the-art AI chip production—a competitive advantage that must be seized upon.** State-of-the-art AI chips comprise a small

percentage of total global chip production, and only a few factories—representing 8.5 percent of current global chip factory capacity—can be used to manufacture them.[10] If the United States and its allies maintain their competitive advantage in AI chip production, they can continue to lead in developing and fielding advanced AI applications. For these reasons, state-of-the-art AI chips—and the complex supply chains that produce them—must be treated as strategic assets to be maintained and grown.

**Without concerted action, this U.S. and allied strategic advantage may be time-bound.** For now, the United States and its allies dominate state-of-the-art AI chip production because the technologies needed to build such chips belong to U.S. and allied firms. U.S. firms, for example, lead in AI chip design; U.S., Taiwanese, and South Korean firms control essentially all of the factories capable of producing state-of-the-art AI chips; and U.S., Dutch, and Japanese firms collectively dominate the production of semiconductor manufacturing equipment used to fabricate chips. China is investing heavily in indigenizing its chip industry, but faces challenges, as chip technologies represent decades of accumulated expertise and hundreds of billions of dollars of investments by leading companies. A concerted effort by the United States and its allies to limit China's access to these technologies could prevent China from producing state-of-the-art AI chips for at least a decade, if not longer.[11]

## Acknowledgments

For helpful discussions, comments, and input, great thanks go to Carrick Flynn, Sue Gordon, Igor Mikolic-Torreira, Alexandra Vreeman, and Lynne Weil. The author is solely responsible for all mistakes.



Document Identifier: doi: 10.51593/20190015

## Endnotes

[1] In this paper, AI refers to cutting-edge computationally-intensive AI systems, especially deep neural networks. Building a DNN requires "training" it to learn from a large dataset, such as a set of images. Once trained, in a process called "inference," the DNN can classify new pieces of data similar to those it learned about during training.

[2] In this paper, "AI chips" includes graphics processing units (GPUs), field-programmable gate arrays (FPGAs), and certain types of application-specific integrated circuits (ASICs) specialized for AI calculations. "AI chips" also includes a GPU, FPGA, or AI-specific ASIC implemented as a core on system-on-a-chip (SoC). AI algorithms can run on other types of chips, including general-purpose chips like central processing units (CPUs), but we focus on GPUs, FPGAs, and AI-specific ASICs because of their necessity for training and running cutting-edge AI algorithms cost-effectively and quickly.

[3] In this paper, "state-of-the-art" means chips with ≤16 nm transistors, which includes chips within the four most advanced generations using up to five-year-old technology. The most advanced chips today have 5 nm transistors.

[4] Dario Amodei and Danny Hernandez, "AI and Compute," *OpenAI*, May 16, 2018, https://openai.com/blog/ai-and-compute/.

[5] Jeffrey Shek, "Takeaways from OpenAI Five (2019)," *Towards Data Science*, April 23, 2019, https://towardsdatascience.com/takeaways-from-openai-five-2019-f90a612fe5d. DeepMind has incurred financial losses of hundreds of millions of dollars as a result of increasing computing costs. Gary Marcus, "DeepMind's Losses and the Future of Artificial Intelligence," *Wired*, August 14, 2019, https://www.wired.com/story/deepminds-losses-future-artificial-intelligence/.

[6] OpenAI, Form 990 for fiscal year ending Dec. 2017, 11, https://projects.propublica.org/nonprofits/organizations/810861541/201920719349300822/IRS990.

[7] Amodei et al., "AI and Compute."

[8] By "ordinary chips," we refer to general-purpose chips such as central processing units.

[9] Further data and details appear in CSET's corresponding technical report: Saif M. Khan and Alexander W. Mann, "AI Chips: What They Are and Why They Matter" (Washington, DC: Center for Security and Emerging Technology, April 2020).

[10] SEMI, *World Fab Forecast*, May 2019 Edition.

[11] Ideas on how the United States can leverage semiconductor supply chains appear in other CSET reports: Saif M. Khan, "Maintaining the AI Chip Competitive Advantage of the United States and its Allies" (Washington, DC: Center for Security and Emerging Technology, December 2019), https://cset.georgetown.edu/wp-content/uploads/CSET-Maintaining-the-AI-Chip-Competitive-Advantage-of-the-United-States-and-its-Allies-20191206.pdf; Carrick Flynn, "Recommendations on Export Controls for Artificial Intelligence" (Washington, DC: Center for Security and Emerging Technology, February 2020),

https://cset.georgetown.edu/wp-content/uploads/Recommendations-on-Export-Controls-for-Artificial-Intelligence.pdf; Saif M. Khan and Carrick Flynn, "Maintaining China's Dependence on Democracies for Advanced Computer Chips" (Washington, DC: Center for Security and Emerging Technology and Brookings Institution, April 2020).