NATIONAL INSTITUTE FOR PUBLIC POLICY

INFORMATION SERIES

Issue No. 608 December 12, 2024

AI is for Allies

Dr. Michael Hochberg
Dr. Michael Hochberg earned his PhD in Applied Physics from Caltech; he is currently a visiting scholar at the Centre for Geopolitics at Cambridge University. His writings on geopolitics can be found at longwalls.substack.com and his twitter is @TheHochberg.

Marcus Gomez
Marcus Gomez earned his BS in Computer Science from Stanford; he is currently the CEO of Luminous Computing, and advises CEOs and investors in the semiconductor and AI industries.

The U.S. Department of Commerce recently sought to enhance export controls on artificial intelligence (AI) hardware and added several People's Republic of China companies to the entity list.[1] This is a step in the right direction, aimed at containing China's ability to weaponize AI. But it's not enough.

Progress in AI models is moving at an astounding pace: The most advanced models are now producing glimmers of what may be considered strategic action and self-protection.[2] These models are like nuclear weapons were in the 1950s—a technology that will revolutionize warfighting in ways that we cannot yet anticipate. But unlike nuclear technology, which spread slowly, AI is being rapidly and widely weaponized, with attendant dramatic changes in warfare.

What is to be done? The liberal-democratic West (including Japan, Korea and Taiwan) have an effective monopoly on the best AI hardware, especially for training giant models. The West should now cut off China, Russia, Iran and North Korea's (i.e., the CRINKs) access to this hardware, and cripple their capabilities for building their own competing hardware based on

Western semiconductor technology. The next wave of AI improvement is going to be revolutionary, and there's no reason to enable our enemies to benefit.

Real-Time Threats

The threats posed by AI are not theoretical—they exist with today's technology. The combination of sophisticated AI chatbots with generative AI tools for creating audio, video, and images in real-time will enable a new personalized, scalable propaganda.[3] Such tools will target individuals; they will adapt to new circumstances; and their on-line manifestations will be indistinguishable from humans. These AI-driven audio and video chatbots will be extremely persuasive.

Large Language Models (i.e., LLMs) are already freakishly engaging: Users on Character.AI spend over 2 hours a day chatting with LLM's that they *know* are not human.[4] AI voice clones can be generated with <10 second recordings of speech; voice and video cloning scams are becoming commonplace.[5] For instance, employees have been convinced that they are on video calls with their colleagues, and parents convinced that they are on phone calls with their children, when in fact they are interacting with bots or AI-manipulated deepfakes.[6] The threat of these tools being used at-scale for political persuasion and the de-legitimation of regimes—a new kind of gray zone warfare—is dramatic.

Historically, it has been possible to disseminate disinformation at scale, through print, broadcast media and most recently through social media. But these interactive voice and video tools, combined with LLM's, will allow for personally customized, compelling interaction and persuasion to operate at scale on target populations for the first time.

Another example, in the realm of military affairs, deserves highlighting: Today, drones are largely dependent on GPS and off-board computation for navigation and targeting, making them vulnerable to jamming and communication disruption. Tomorrow, they will use video and inertial sensors to navigate autonomously as well as on-board computation to recognize landmarks, identify and prioritize targets, and act in concert as part of a swarm.[7] To stop such drones, new directed energy or kinetic systems will be needed; disrupting GPS and communications networks will not be effective.[8]

The technological backbone driving the proliferation of these tools is the development and deployment of giant AI models ("foundation models"). These foundation models—and the hardware used to develop and deploy them—are the essential technological and economic chokepoint that the United States can, for the moment, use to block our adversaries from access to future AI capabilities.

Future Utility of AI and Preventing Adversaries from Getting the Technology

The fundamental improvements in AI models over the past decade have come from dramatically scaling up the size of the models. This requires increasingly large amounts of high quality data, and more importantly, increasingly large compute clusters to develop said

models ("training"). A decade ago, training runs were done on small numbers of GPUs over the course of hours. Today's state-of-the-art training runs take multiple months, with the biggest runs occurring across 10,000 to 50,000 GPUs. These training runs occur in multi-billion-dollar facilities of 100,000+ GPU clusters that consume the energy output of entire power plants. Tomorrow's models will be trained on million GPU clusters and will require nuclear power plants to run.[9]

With the growth of these models, a near-human level of intelligence has been achieved: AI models are scoring in the top percentile on the bar exam, medical licensing exam, and dozens of other undergraduate- and graduate-level benchmarks.

After models are trained, they are then deployed—this is called inference.[10] Training is the creation of the model, whereas inference is the deployment of that model into the real world. Very roughly, one might think of training as analogous to software development, while inference is that software being deployed for an end-user. The difference is that both training and inference are automated, and both use large numbers of GPUs.

These inference deployments use different configurations of GPUs than training; each unit of hardware can only support a finite number of simultaneous users, so while demand for training hardware scales with the number of organizations developing foundation models, the demand for inference hardware scales with the number of users. It is widely suspected in the AI community that, in the long run, inference will be an even larger computational demand than training, due to the mass adoption of AI.

The ecosystem for the development of the hardware for AI is also evolving quickly. Today, the world's best and easiest-to-use GPUs are designed and sold by NVIDIA. AMD and Intel are racing to catch up, and Google, Amazon, Microsoft, Apple, and OpenAI have also developed or are developing internally-designed alternatives. Other companies are building more-specialized hardware for specific AI tasks, but these are a small part of the market. All these GPUs seem to be built, thus far, at TSMC. Only three companies in the world have factories that could fabricate modern GPU chips: Intel, Samsung, and TSMC. The network effects associated with the semiconductor IP ecosystem make TSMC a durable monopoly for smaller chip designers; however, it makes sense for NVIDIA and AMD to diversify their supply chains to include alternative foundries.[11]

China is desperate for access to this technology and American companies are eager to monetize this desperation. Thus far, China is having great success dodging western sanctions and export controls.[12] NVIDIA has built chips that skirt the very edges of US export controls so that they could be sold to China.[13] Chips are also being smuggled to China through organized networks and various states are setting up data centers for China, Russia, and Iran that can host remote computation for both training and for inference.[14] Recently, the export of AI chips below 7 nm to China has been severely curtailed, and the export of critical components of these AI chips, like high bandwidth memory (HBM) above a performance threshold, has also been capped.[15] But this is not enough; the current export regulations are so byzantine as to invite companies to make money in the regulatory grey areas. A different approach is needed.

There are several ways for the US to intervene aggressively over the coming months:

1. Models: Exporting trained models and making them available to adversary regimes needs to be banned, with strict liability applied to companies making these models available to such regimes. This ban should include access to model Application Programming Interfaces (i.e., APIs such as OpenAI's API, Anthropic's API, etc.), both to prevent access to the capability, and to prevent counterparties from fine-tuning their own models on the outputs of the APIs. This ban should also include the open-sourcing of such models (e.g. Meta's Llama series). Any attempt to measure capabilities of AI models can be gamed, so the ban should be done on a recency basis—models that have been trained within the prior 5 years should not be exported except to close allies and under a strict licensing regime.
2. People: Sharing AI-related knowledge developed in the West with adversaries needs to be treated as a crime. There is a need for preemptive export licenses before AI engineers can work for Chinese, Iranian, or Russian firms, regardless of their physical location.
3. Hardware: The performance of AI hardware more than doubles every 2 years.[16] The United States should ban the export of AI hardware to the CRINKs and any regimes that allow re-export to these regimes, if the hardware has been on the market for less than 10 years (older hardware can still be networked together fairly effectively, so the ban must have sufficiently long time scale). Modern hardware should be sold only to close and trusted allies. States caught reselling or renting this hardware to adversaries should be penalized with national loss of access to AI hardware and tools, and with severe economic penalties for the states involved. This needs to include chips, boards, cutting-edge optical interconnects, switches, and other advanced data center networking equipment. Export duties for the CRINK regimes should be imposed on any hardware that isn't banned, in order to make assembling clusters of old systems prohibitively expensive - this will mean 5x or more duties in some cases. Importantly, this will not meaningfully harm NVIDIA because they are production capacity limited (and have been for years).
4. Boundaries: The United States needs to ban Intel, TSMC, and Samsung from building chips or sharing Process Design Kits with entities based in, controlled by, or substantially tied to the CRINKs. The same needs to be done for advanced memory manufacturers like SK Hynix and Micron. There needs to be a positive duty for foundries and service providers to do know-your-customer work on all custom chip designs. CRINK access to processes within 10 years of general availability needs to be banned, again with strict liability for the companies involved. This should include processes for analog and III/V chips, which are not at the most advanced lithography nodes, but which progress on other metrics (i.e., bipolar and silicon photonic processes).[17]

5. Tools to make the tools: Software, materials, equipment, spares, specialty chemicals, IP, masks, and models need to be banned from export to the CRINKs for at least 10 (ideally 20) years from when they were released into the market. The United States may not be able to prevent China from developing their own semiconductor ecosystem, but it can certainly stop supporting their efforts by selling them the tools and equipment. If the United States cuts off access to spares and materials from the West, existing trailing-edge chip fabs in China will be crippled, along with their ability to advance to more deeply scaled nodes.
6. Imports: Western businesses should be banned from purchasing or integrating semiconductor chips or boards built in China or by CRINK-controlled companies. There is no telling what vulnerabilities are hiding within these chips and boards, and they constitute a national security threat if they are used in Western-produced electronics.
7. Security: The export bans and increase in regulation will increase the incentive for espionage. Western businesses that develop and deploy modern AI models should be required to maintain dramatically higher levels of security. Their technology stacks should be designed to withstand foreign cyber-attacks, their employees should be regularly trained to spot social engineering attacks, and independent organizations should be tasked with testing the security of these companies.

Once advanced chips (or fabrication/design tools) are exported to CRINK regimes, how they are used cannot be controlled. If China wants access to advanced semiconductors, they should be forced to spend immense amounts of money (trillions of dollars) to replicate the entire semiconductor ecosystem, independently and without Western help. Only very old chips and tools should be sold to China, if any, and at a high cost. Any state re-exporting prohibited chips or AI systems to China should be swiftly prohibited from purchasing advanced semiconductors, either in the form of chips or systems. The option of neutrality, and trading with both sides, needs to be taken off the table. An aggressive policing regime, including honey-pot false-flag sellers, will need to be implemented rapidly.

Dispelling the Myths

One counterproposal that cloud providers are sure to offer is that, if these regimes and the companies they control are allowed to use advanced AI only through cloud compute, and only with the data centers located in the West, a dependency on Western cloud infrastructure will result. That dependency could, in principle, be used as leverage against China in the event of escalation or war. Enforcing policies that require large-scale AI datacenters to be built only in the territory of strong U.S.-allied regimes (i.e., the Five Eyes, Japan, South Korea, etc.) could perhaps enable adversaries to legitimately use AI.

This is an appealing but spurious argument, similar to the one underlying the Atoms for Peace program, which offered nuclear capability to around 30 countries with the intention of

spreading peaceful applications of nuclear energy. In fact, this created enormous long-term proliferation risk, as many of these countries were primarily interested in the military applications of this new technology. Anything that the United States does to share AI and semiconductor technology with adversaries will be used, first and foremost, for the development of weapons, since that will be the first priority of the adversary regimes.

The AI chip industry will argue that these limitations will harm them and weaken their ability to compete in the world market. This is untrue: China accounts for only 12% of NVIDIA's revenues, and probably less of their margin dollars.[18] As for semiconductor equipment, the consequences will be more severe for Western companies, because China is frantically buying everything they can before sanctions are tightened. Radical policy changes are required to protect the semiconductor ecosystem from being moved to China, following the same recipe that the CCP has used in other industries over the past 20 years.

The AI developer community will argue that this will kill AI innovation, and that modern progress in AI is driven by open-source tools. Open sourcing the designs for nuclear weapons and reactors would certainly advance progress in the field of nuclear engineering. That doesn't make it a good idea to open these designs to the world.

Conclusion

Adversaries should not be allowed to take advantage of Western innovation in the realm of AI, thereby threatening to leapfrog ahead of Western developers; only allies should have access to this technology and its revolutionary potential.

The policies outlined above, if implemented quickly, will kneecap the Chinese AI industry, remove China's ability to supply Russia with modern chips for their ongoing war with Ukraine, and will fatally damage the emerging Chinese semiconductor ecosystem. The consequences will be disruptive to Western industry, and the United States and allied governments will need to step in, in some cases, to ensure that critical businesses survive the disruption. But out of this, a more vibrant Western semiconductor ecosystem will emerge, since there will be a rush to replace the lost capacity from China with semiconductor fabrication facilities in allied states, Europe, and the United States.

[1] Louise Matsakis, "The US Just Made It Way Harder for China to Build Its Own AI Chips," *Wired*, December 2, 2024, available at https://www.wired.com/story/2024-chips-export-controls-china/.

[2] Leopold Aschenbrenner, "SITUATIONAL AWARENESS: The Decade Ahead," June 2024, available at https://situational-awareness.ai/; and Shakeel [@ShakeelHashim], "OpenAI's new model tried to avoid being shut down. Safety evaluations on the model conducted by @apolloaisafety found that o1 "attempted to exfiltrate its weights" when it thought it might be shut down and replaced with a different model." Tweet. X, 2:09 PM, December 5, 2024, available at https://x.com/ShakeelHashim/status/1864748980908781642.

[3] "Introducing the Realtime API," *OpenAI*, October 1, 2024, available at https://openai.com/index/introducing-the-realtime-api/; and Timothy Lee, "I created my own deepfake—it took two weeks and cost $552," *ars technica*,

December 16, 2019, available at https://arstechnica.com/science/2019/12/how-i-created-a-deepfake-of-mark-zuckerberg-and-star-treks-data/.

[4] Michelle Cheng, "A startup founded by former Google employees claims that users spend two hours a day with its AI chatbots," *Quartz*, October 12, 2023, available at https://qz.com/a-startup-founded-by-former-google-employees-claims-tha-1850919360.

[5] Jake Peterson, "Microsoft Won't Let You Use Its New AI Voice Tool," *Life Hacker*, July 11, 2024, available at https://lifehacker.com/tech/microsoft-releases-vall-e-2-ai.

[6] Benj Edwards, "Deepfake scammer walks off with $25 million in first-of-its-kind AI heist," *ars technical*, February 5, 2024, available at https://arstechnica.com/information-technology/2024/02/deepfake-scammer-walks-off-with-25-million-in-first-of-its-kind-ai-heist/?utm_source=chatgpt.com; and Mahsa Saeidi, "Voice cloning scams are a growing threat. Here's how you can protect yourself," *CBS News*, May 17, 2024, available at https://www.cbsnews.com/newyork/news/ai-voice-clone-scam/.

[7] University of South Australia, "GPS alternative for drone navigation uses visual data from stars," *Techxplore*, December 3, 2024, available at https://techxplore.com/news/2024-12-gps-alternative-drone-visual-stars.html#google_vignette.

[8] Michael Hochberg, "US needs to double down on directed energy weapons," *Asia Times*, July 30, 2024, available at https://asiatimes.com/2024/07/us-needs-to-double-down-on-directed-energy-weapons/; and Trevor Phillips-Levine, "Return of the Gunfighters," *Behind the Front*, August 15, 2024, available at https://behindthefront.substack.com/p/return-of-the-gunfighters.

[9] Stephen Morris and Tabby Kinder, "Elon Musk plans to expand Colossus AI supercomputer tenfold," *The Financial Times*, December 4, 2024, available at https://www.ft.com/content/9c0516cf-dd12-4665-aa22-712de854fe2f; Josh Norem, "AMD Says an AI Cluster With 1.2 Million GPUs Could Be In the Cards," *Extreme Tech*, June 6, 2024, available at https://www.extremetech.com/computing/amd-says-an-ai-cluster-with-12-million-gpus-could-be-in-the-cards?utm_source=chatgpt.com; and Andrew Caballero-Reynolds, "Microsoft wants Three Mile Island to fuel its AI power needs," *The Verge*, September 20, 2024, available at https://www.theverge.com/2024/9/20/24249770/microsoft-three-mile-island-nuclear-power-plant-deal-ai-data-centers?utm_source=chatgpt.com.

[10] "AI inference vs. training: What is AI inference?," *Cloudfare*, available at https://www.cloudflare.com/learning/ai/inference-vs-training/.

[11] Michael Hochberg and Leonard Hochberg, "The chip industry and national security," *Asia Times*, December 7, 2022, available at https://asiatimes.com/2022/12/the-chip-industry-and-national-security/.

[12] Dylan Patel, Jeff Koch, and Sravan Kundojjala, " Fab Whack-A-Mole: Chinese Companies are Evading U.S. Sanctions," *Semianalysis*, October 28, 2024, available at https://semianalysis.substack.com/p/fab-whack-a-mole-chinese-companies.

[13] Fanny Potkin, "Exclusive: Nvidia preparing version of new flagship AI chip for Chinese market," *Reuters*, July 22, 2024, available at https://www.reuters.com/technology/nvidia-preparing-version-new-flaghip-ai-chip-chinese-market-sources-say-2024-07-22/.

[14] Stu Woo, "One of the Biggest AI Boomtowns Is Rising in a Tech-Industry Backwater," *The Wall Street Journal*, October 7, 2024, available at https://www.wsj.com/tech/ai/one-of-the-biggest-ai-boomtowns-is-rising-in-a-tech-industry-backwater-8dfcdfa1.

[15] Department of Commerce, *Foreign-Produced Direct Product Rule Additions, and Refinements to Controls for Advanced Computing and Semiconductor Manufacturing Items*, December 5, 2024, available at https://public-inspection.federalregister.gov/2024-28270.pdf; and Timothy Prickett Morgan, "US Curbs HBM Exports To China –

More For The Rest Of Us," *The Next Platform*, December 2, 2024, available at https://www.nextplatform.com/2024/12/02/us-curbs-hbm-exports-to-china-more-for-the-rest-of-us/.

[16] Jim Osman, "Can Nvidia's 'Hyper Moore's Law' Spark An AI Revolution?," *Forbes*, November 7, 2024, available at https://www.forbes.com/sites/jimosman/2024/11/07/can-nvidias-hyper-moores-law-spark-an-ai-revolution/.

[17] Alan Patterson, "Silicon Photonics Set for Takeoff," *EE Times*, October 30, 2024, available at https://www.eetimes.com/silicon-photonics-set-for-takeoff/.

[18] "China's Push Against Nvidia: How Discouraging Local Purchases Could Shape AI and Market Dynamics," *kavout*, September 29, 2024, available at https://www.kavout.com/market-lens/chinas-push-against-nvidia-how-discouraging-local-purchases-could-shape-ai-and-market-dynamics.

The National Institute for Public Policy's *Information Series* is a periodic publication focusing on contemporary strategic issues affecting U.S. foreign and defense policy. It is a forum for promoting critical thinking on the evolving international security environment and how the dynamic geostrategic landscape affects U.S. national security. Contributors are recognized experts in the field of national security. National Institute for Public Policy would like to thank the Sarah Scaife Foundation for the generous support that made this *Information Series* possible.