UVA | NATIONAL SECURITY DATA AND POLICY INSTITUTE

BEYOND EXPORT CONTROLS: PROTECTING AMERICAN INVESTMENT IN THE SOFTWARE AGE

July 16, 2025

Robert J. May

NATIONALSECURITY.VIRGINIA.EDU | NSDPI@VIRGINIA.EDU

Beyond Export Controls: Protecting American Investment in the Software Age

July 16, 2025

Executive Summary

The rise of distilled open-source AI models and frontier reasoning LLMs positions itself as a two-pronged attack on the American AI curtain—exemplified by DeepSeek's progression towards DeepSeek-v3, a mixture-of-experts (MoE) frontier LLM. While export controls on high-performance chips and closed-model weights remain vital towards AI security, adversaries and competitors are finding ways to work around rules set by the U.S.. DeepSeek focused on hardware compliance and software-centric strategies, using distillation from previous LLMs to fine-tune and repurpose open-source research. The circumventing of U.S. export controls, leveraging of software innovation, and changing policy stances on GPU manufacturers' sale-controls all work together to lag policy responses to emerging foreign AI-development strategies. Implications for policy include:

- **Export Control Inconsistencies**–The transition from the Biden to the Trump administration saw the return of country-to-country negotiations for advanced computing technology and AI, rescinding President Biden's AI Diffusion rules. Policymakers should look into methods that allow for greater baseline permitting rules in order to streamline trade negotiations with other countries.
- **Software Protection Policy**–Regulating AI software has been largely unattempted. Currently, catch-all provisions theoretically require licenses for software designed for model training when there is knowledge of potential military use. However, demonstrating knowledge about end-use through digital code or model weight changes is near impossible. Code repositories on GitHub, open-weight LLMs, and distillation pipelines fall outside any licensing regime. Policymakers should look into methods to regulate model-weight transfers.
- **Responding to Industry Changes**–Abrupt changes in market competition and training strategies, seen by the arrival of DeepSeek, can be provoked through strict export controls and policymakers are slow to respond. Policymakers should look into the costs and benefits of strict export controls under the context of slow response times and policy workarounds.

DeepSeek-R1 Case Study

In January 2025, the DeepSeek lab publicized R1, a reasoning LLM with 671 billion parameters that matched or exceeded the performance of top closed models in benchmarks such as MATH-500 and PhD-level question answering, all while training in under three million GPU-hours at an estimated cost of $5.6 million (DeepSeek-AI et al., 2025). These successes rested upon four foundations: hardware procurement through bordering U.S. export controls with previously-acquired or legal GPUs, high-performance computing (HPC) orchestration, precise software optimization, and the utilization of distillation pipelines:

Hardware Procurement

DeepSeek's parent company, High-Flyer LP—a well-capitalized algorithmic trading firm—had already amassed 10,000 NVIDIA A100 GPUs in 2020, replicating the hardware investments of leading U.S. quant funds (National Security Data and Policy Institute, 2025). When U.S. export controls barred A100 and H100 GPUs to China, DeepSeek pre-emptively acquired an unknown but substantial number of H800 GPUs—a limited variant designed to comply with existing rules—before the October 2023 policy tightening closed that loophole (National Security Data and Policy Institute, 2025). This timing exemplifies the laxity of hardware-centric controls in the face of rapid vendor innovation.

High-Performance Computing

Beyond raw GPU count, DeepSeek distinguished itself through meticulous HPC engineering. Its Fire-Flyer II cluster interlinked thousands of GPUs using high-bandwidth InfiniBand and NVSwitch, achieving competitive performance to top GPUs at a much lower financial and energy cost.. Key to this was DualPipe pipeline parallelism, which overlapped forward and backward passes across micro-batches to eliminate idle time, and a hybrid CPU–GPU all-reduce that concealed communication overhead behind ongoing computation.

Software Optimization

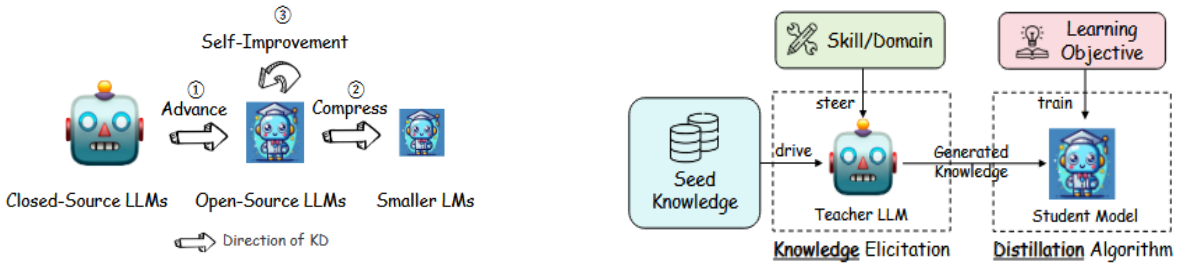
Deepseek advanced and utilized careful dequantization strategies and custom kernels to double throughput in tensor-core operations, slashed memory usage by half, and enabled R1 to converge in under 2.8 million GPU-hours (Liu et al., 2025). DeepSeek's software ingenuity provided them extreme control over memory allocation and processing speed, allowing them to use the allowed H800 GPUs for an unforeseen result.

Distillation Pipelines

R1's reasoning prowess emerged from a two-stage training pipeline. The first, dubbed R1-Zero, applied pure reinforcement learning via Group Relative Policy Optimization (GRPO)—eschewing a costly value network in favor of group-normalized reward advantages and per-token KL penalties (DeepSeek-AI et al., 2025; Shao et al., 2024). This stage uncovered emergent chain-of-thought and self-verification behaviors without a supervised warm-up. A

second cold-start fine-tuning phase used a modest curated dataset of expert-crafted long reasoning chains to stabilize output clarity and guard against reward hacking. Finally, R1's abilities were distilled into smaller parameter student models, trained off the previous larger LLM.

At its core, their distillation rests on three pillars. First, algorithms for elicitation extract another LLM's knowledge—ranging from simple instruction–response pairs to complex hidden state alignments—through methods such as supervised fine-tuning with chain-of-thought prompts, reinforcement learning from AI feedback, or iterative self-training loops. Second, skill transfer encompasses the discrete capabilities distilled: the ability to follow diverse instructions, to engage in coherent multi-turn dialogue, to solve mathematical and logical reasoning tasks, and to call external tools intelligently. Finally, verticalization tailors these distilled skills to high-value domains—legal reasoning, medical diagnosis, financial forecasting, and scientific problem solving—by fine-tuning on carefully curated, domain-specific corpora. Central to all these techniques is data augmentation, whereby a handful of parent-LLM examples blossom into millions of synthetic training instances, enabling distilled models to internalize nuanced reasoning patterns and ethical guardrails without requiring extensive human annotation.



Knowledge Distillation (KD) plays three key roles in LLMs: 1) Primarily enhancing capabilities, 2) offering traditional compression for efficiency, and 3) an emerging trend of self-improvement via self-generated knowledge. Source: (Xu et al., 2024).

Hardware Protections

The Export Administration Regulations (EAR) govern the export of advanced semiconductors under Export Control Classification Numbers (ECCNs) such as 3A090.a and 3A090.z, which cover GPUs and AI accelerators exceeding specified performance thresholds. Countries of concern—China and Macau (Country Group D:5)—face strict license requirements, while key allies enjoy broad exceptions (U.S. Department of Commerce, 2025; Bureau of Industry and Security, 2025).

In January 2025, an interim rule further refined these controls, streamlining licensing for low-risk orders, exempting purchases by 18 trusted allies entirely, and introducing Universal Verified End-User (UVEU) and National Verified End-User (VEU) statuses to allow qualified

entities to import thousands of GPUs under secure end-use certifications. Additionally, government-negotiated GPU cap increases were possible. However, these rules were rescinded in May 2025. NVIDIA has recently stated that they can once again develop and sell chips to China after a previous freeze from President Trump (Mickle, 2025).

Any sort of export measure faces acute challenges. Vendor-engineered variants like the H800 skirt category definitions before controls can catch up. Once hardware is in place, proving its use in prohibited AI applications is technically and legally fraught, as no on-chip provenance or chain-of-custody logs exist. Therefore, it is important to improve our informational network—power consumption metrics, procurement anomalies, and staff expertise profiles—to supplement export-control enforcement.

Software Protections

Regulating AI software transfers lags behind hardware. EAR Part 744 "catch-all" provisions theoretically require licenses for software "specially designed" for model training when there is "knowledge" of military or weapons of mass destruction end-uses (Bureau of Industry and Security, 2025). However, demonstrating such knowledge in purely digital exchanges—via code, model weights, or distilled small-model checkpoints—is near-impossible in practice. Code repositories on GitHub, open-weight LLMs on Hugging Face, and extensive open-source distillation pipelines fall outside any licensing regime. Even transfers of closed-weight models are often hidden in encrypted storage or distributed through private channels.

To detect software export violations, the Bureau of Industry and Security offers red-flag indicators (e.g., shell company addresses, sudden spikes in training data access) and due diligence steps—end-use certificates, on-site audits, and third-party infrastructure attestations (Bureau of Industry and Security, 2025).. Yet these are stopgaps at best. Without a transparent licensing or tracking mechanism for model artifacts, software dissemination continues unabated.

Policy Implications

The DeepSeek-R1 case study offers a compelling example of how adversaries are evolving beyond traditional hardware-based AI development constraints. Their success demonstrates that access to frontier AI capabilities no longer depends solely on massive compute or proprietary model weights. Instead, sophisticated distillation techniques, strategic software engineering, and stockpiling of permitted or pre-acquired hardware enable adversaries to replicate state-of-the-art reasoning capabilities at lower cost and risk. In light of these developments, U.S. policy must evolve beyond chip export controls to address software leakage and regulatory agility.

Currently, software and model-weight transfers are largely unregulated. While Part 744 catch-all provisions under EAR theoretically cover military end-use of AI software, enforcement remains

virtually nonexistent. Public code repositories, open-weight LLMs, and knowledge distillation pipelines do not require licenses and rarely leave a traceable footprint.

Software protection has not been understood as a critical element of export controls for AI and its input goods. To try and prevent further software and data spillover towards China, U.S. lawmakers are calling for a complete severance between the two countries on AI cooperation (Hong and Hu, 2025). While the efficacy of this foreign policy is outside the scope of this paper, there is room for concern when it comes towards innovation and acquisition. An attempt to try and remove China from any sort of American model-weight or teacher/input LLM for knowledge distillation could result in something just as unpredictable as DeepSeek surmounting hardware and software bottlenecks.

Policymakers should develop a model-weight and software artifact registry system that applies to any LLM checkpoint or distillation pipeline derived from American research. Such a system would enable post-distribution tracking and red-flag early warning systems without intruding upon open-source publication norms.

DeepSeek's rapid success with the Fire-Flyer II cluster and DualPipe parallelism underscores how innovation often outpaces regulation. The U.S. cannot rely on periodic rule changes to contain a market where AI development cycles are measured in weeks, not years. Policymakers should consider funding a dedicated AI intelligence division under the Office of the Director of National Intelligence to monitor weight dissemination, hardware procurement patterns, and research pipeline movements globally. Greater intelligence, along with a greater understanding of development pipelines, would give the U.S. more time and information to respond and alter current export controls and AI partners.

In sum, the evolving strategies employed by actors like DeepSeek reveal that America's existing export control regime—while necessary—is no longer sufficient. Hardware constraints are being outmaneuvered through legally ambiguous procurement strategies, while software protections remain virtually nonexistent despite their growing centrality in model replication and capability transfer. As AI development increasingly relies on software ingenuity and knowledge distillation, U.S. policy must broaden its scope beyond chips to address the full AI development stack. A forward-looking export control framework must integrate software tracking, international collaboration, and dynamic intelligence gathering, all while maintaining innovative leadership and ethical stewardship. Failing to act holistically risks allowing adversaries to match or exceed U.S. capabilities using the very tools American researchers pioneered.

References

Bureau of Industry and Security. (2025). *Department of Commerce Announces Rescission of Biden-Era Artificial Intelligence Diffusion Rule, Strengthens Chip-Related Export Controls* Retrieved July 8, 2025, from https://www.bis.gov/press-release/department-commerce-announces-recission-biden-era-artificial-intelligence-diffusion-rule-strengthens-chip\

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning (No. arXiv:2501.12948). arXiv. https://doi.org/10.48550/arXiv.2501.12948

Hong, T., & Hu, M. (2025). Opportunities, Challenges, and Regulatory Responses to China's AI Computing Power Development under DeepSeek's Changing Landscape. *International Journal of Digital Law and Governance*. https://doi.org/10.1515/ijdlg-2025-0002

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., … Pan, Z. (2025). *DeepSeek-V3 Technical Report* (No. arXiv:2412.19437). arXiv. https://doi.org/10.48550/arXiv.2412.19437

Mickle, T. (2025, July 15). Nvidia Says U.S. Has Lifted Restrictions on A.I. Chip Sales to China. *The New York Times*. https://www.nytimes.com/2025/07/14/technology/nvidia-ai-chip-sales-china.html

National Security Data and Policy Institute. (2025). *DeepSeek's Achievements and the Implications for Policy*. https://nationalsecurity.virginia.edu/sites/nationalsecurity/files/2025-06/00006_%2820250221%29_NSDPI_Deepseek%27s%20Achivements%20and%20the%20Implications%20for%20Policy%20%283%29.pdf

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y. K., Wu, Y., & Guo, D. (2024). DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models (No. arXiv:2402.03300). arXiv. https://doi.org/10.48550/arXiv.2402.03300

U.S. Department of Commerce. (2025). Export Administration Regulations 15.730-780. https://www.ecfr.gov/current/title-15/subtitle-B/chapter-VII

Xu, X., Li, M., Tao, C., Shen, T., Cheng, R., Li, J., Xu, C., Tao, D., & Zhou, T. (2024). *A Survey on Knowledge Distillation of Large Language Models* (No. arXiv:2402.13116). arXiv. https://doi.org/10.48550/arXiv.2402.13116