George Randolph Dirth

July 29, 2026

Globalization Redux – via AI

From a world economic perspective, globalization has been a success. From an American perspective—and that of other high-income, market economy countries—the impact has been mixed, with some sectors and cohorts benefitting while others, even entire regions, enduring the opposite outcome and paying a heavy price.

The exporting country cohort enjoyed lower costs of production—due to lower overall cost of living and therefore labor, lower cost of assembling the means of production and due to simply acquiring as opposed to developing advanced technologies, at times illegally. Conversely, in the recipient country cohort, in addition to their overall higher cost of living, there was the not insignificant matter of societal infrastructure expense, in the form of health, pension and regulatory requirements imposed upon manufacturers—directly increasing cost of production. Collectively, this societal infrastructure manifested many of the very policies that were foundational to how the market economy democracies became high-income (coupled with rule of law and effective capital markets). A considerable portion of the gap in cost of production can be attributed to the disparity of societal infrastructure in the export vs market economy countries.

The high-income countries possessed exactly what the low-income, export countries lacked—robust consumer markets. Without access to these markets the export countries would have remained low-income, or at minimum, developed at a fraction of the pace. As any businessperson knows, it is not enough to have a good product—a willing and able market of end consumers is also requisite. This market access transformed the low-income countries, but it also transformed the high-income ones.

The era of globalization raised world economic development, lifted millions out of poverty and even benefitted sectors of the U.S. economy, such as technology and finance. America—along with Western European countries—also experienced

resultant economic instability—due to market share erosion in affected industries, leading to social and political instability, for which we are still paying the price.

Could the adverse consequences of globalization have been avoided? They certainly could have been mitigated via an import normalization tax determined country by country to capture the gap of societal infrastructure expense paid by American manufacturers vs those not paid by the exporting countries, ensuring a more level playing field. This inflow of funds could have been used to offset the shortfall in contributions from U.S. manufacturers to worker and government accounts—including tax revenue. (Note that employing variable tax or tariff rates by country would violate select WTO principles, but not doing so extends unfair trading advantages to the low-income, exporting cohort. WTO protocol would not be the first post WW2 treaty or legislation that is in need of revision given the ensuing change in the U.S. and world economies.)

Which brings us to today, and the rhetorical question of do we really want to go through the adverse consequences of globalization again, courtesy of AI?

Economically speaking, foreign trade falls into two sectors: goods and services. What we experienced before was disruption in the goods sector. AI falls into the service sector, which is the highest growing (measured by pure rate of growth—not total trade volume) sector in today's world economy.

Chinese AI is rightly "topic A" in government, financial and technology circles since the July 16, 2026 introduction of AI model Kimi K3. This follows what was known as the "DeepSeek moment" of early 2025 when the DeepSeek-R1 AI model launch disrupted financial markets by posing questions regarding its value for money and purported cost of development in contrast to U.S. models.

There are, amidst the tumult of the moment, two obvious questions to be answered a) how is it possible that a Chinese AI model could offer a lower cost with impressive performance, and b) what should we do about it. Let's briefly explore both.

Kimi K3 belongs to a genre of Chinese models based on MoE architecture, adopted aggressively by this cohort since 2024. What is MoE?

Even though we restricted Chinese access to our top-tier GPUs, we made available—via published research—a facsimile "cheat-code" architecture allowing creation of models that are less training intensive, less compute intensive but generally effective

for purposes of inference. This is commonly known as MoE, (mixture of experts) as introduced in the Google Brain research paper "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer", published in 2017. It posited that one need not query all billions of parameters per token, but perhaps only the few most relevant to the topic at hand. MoE models were deemed not as brilliant as the frontier models, but were quick, efficient and recognized as effective-enough for most use-cases with a considerably lower cost per token.

After its introduction in 2017, MoE architecture was mostly unused outside of Google for six years until December 2023 when French AI company Mistral released their Mixtral model, incorporating MoE. Coincidentally, in this same period of time, Chinese hedge fund High-Flyer released their first DeepSeek LLM model in November 2023—which did not use MoE. Then in January 2025—more than a year after Mistral had introduced their MoE model—High-Flyer released DeepSeek R1, to both critical and commercial acclaim—this time premised on MoE architecture. So, on one hand we limited Chinese access to our best GPUs but on the other hand we gave them the workaround via MoE. While MoE is efficient and cost-effective, it is also a different class of architecture than that used by closed U.S. AI models, but nonetheless a product of American research.

Chinese models are also known for being open-weight (OW), often confused with open-source.

OW refers to transparency in a model's trained weights—its learned parameters—giving the user the ability to adjust and customize a model's operational settings—as distinct from true open-source in which the training code is evident. This is useful if one downloads the model and runs it independently in one's own processing environment, increasingly known as "on-prem" for on premises. The industrial use case for on-prem is entities such as financial institutions, departments of defense or nuclear energy for whom absolute data security is non-negotiable, typically entailing wholly independent—even "air-gapped" servers, adhering to rigorous protocol. (Note that in Dec 2025 Congress passed and the president signed the FY 2026 National Defense Authorization Act which included strict bans on Chinese AI model use in defense agencies and government contractors.)

Contrary to public perception, four U.S. companies currently offer high-level OW models. These include MoE-based Llama 3 and 4 from Meta, MoE-based gpt-oss

from OpenAI launched August 2025, NemoTron 3 Super from Nvidia launched March 2026, and MoE-based Inkling from Thinking Machines launched in July 2026. Some of these, certainly Nvidia's NM3S which is a hybrid Mamba3/Transformer MoE model with a one million token context window, are serious product offerings—easily comparable to the best Chinese OW models. Conversely, most U.S. frontier models—such as Claude, GPT-5 and Gemini—are closed, meaning their internal code is proprietary and concealed (shielding their IP).

But why—all of a sudden—the fuss regarding Kimi K3 cost and apparent effectiveness, and how have Chinese OW models seemingly crept up on American closed AI frontier models? There are four main reasons.

First is simply the matter of cost of the actual model development. Adopting U.S. developed and publicly available MoE architecture is of course lower cost than developing a proprietary architecture. Add to this the fact that, just like in the goods sector, the cost of producing or developing anything in China remains substantially less than in the U.S.

Second—like it or not—there are persistent reports that restricted GPUs have found their way into China. There are simply too many anecdotal examples of this for it not to have been a factor.

Third is the process of training a model and purported use of "distillation". Both OpenAI and Anthropic have stated they have evidence of thousands of proxy queries originating in China used to extract their model training which in aggregate provides a deep set of inference responses. Their implicit point is that this is tantamount to technology theft, and part of the continuum of technology theft in both industrial and military sectors going back several decades. So, when one asks, "how could Kimi K3 spend so much less for training than Claude?", the answer presumably rests with the fact that it costs much more to create something than to copy it.

Fourth, and this is where one must simply give the Chinese their due: they now have several (globally ranked) top-tier technology universities, including Tsinghua and Peking Universities in Beijing and Zhejiang University in Hangzhou. This has led to the phenomenon of university / industry clusters in Beijing, Hangzhou and to lesser extents Shenzhen and Shanghai, following in the footsteps of Silicon Valley / Stanford, Berkeley and Boston / MIT, Harvard. Recall that in 1979 when Deng Xiaoping (the father of modern China) first visited the U.S. one of his main "asks"

was access to American universities for Chinese students. Decades later this focus has clearly paid off, with the work now being done by vast numbers of Chinese engineers serious and legitimately high-level by any standard.

Which brings us to how should we respond. The answer is via two tracks: track 1 is that which Secretary Bessent voiced July 21, 2026 on Fox News with respect to potential technology theft, which should be allowed to run its course; track 2 is taxation on the use of Chinese AI models within the US market, or used outside of the U.S. but with the U.S. as the end destination market of consumption.

In simple terms, this would involve taxation of tokens generated (supply-side) and consumed (demand-side) via Chinese AI models in or for use within the U.S. market. Whether these models are used from a Chinese API or a U.S. API, the AI service provider and the end consumer would pay a token tax assessed per million tokens generated and consumed. This taxation should be tiered to coincide with level of consumption, adhering to principles of progressivity in taxation. The enforcement mechanism would be equal for all: a certificate confirming payment of token taxation (by the Chinese or US entity) required by the IRS for business token expense to be tax deductible (with requisite use of OTel collectors within on-prem compute set-ups).

Globalization in the goods sector disrupted American manufacturing because exporting countries were allowed to avoid the societal infrastructure contributions required of American manufacturers. Allowing access to our digital markets without normalization for this disparity risks repeating the same mistake. With benefit of hindsight, the pathway to pre-empting globalization redux—via implementation of a Congressionally authorized normalization tax protocol, is clear. Free trade—with a level playing field—is the best guarantor of a sustainably inter-connected world.

The author in June 2026 published *The Fiscal Consequences of AI*, covering the AI revolution, the "high-productivity, low-tax-revenue economy", AI taxation, Chinese AI and the overall unfavorable prospective impact of AI on our fiscal stability.