HKUST ICED
INSTITUTE FOR CHINA ECONOMIC DEVELOPMENT

FALL 2026 **no.004** iced.ust.hk

PDF & Additional Materials

THOUGHT LEADERSHIP BRIEF



The Half-Stack Rivalry

Alicia García Herrero

Photo byPachon in Motion on Pexels

In 2025, US private investment in AI reached $286 billion; reportedly, China's was $12 billion. Yet Chinese models such as DeepSeek and Kimi are now within touching distance of the American frontier, often at a fraction of the cost. China is close behind the US on models with roughly a twentieth of the private money. That paradox explains the US-China AI race, and why the rest of the world is likely to end up having to choose between them.

Two Incomplete Stacks

AI is a stack. At the bottom are chips, memory and fabs. Above them sit energy and compute, then models, then the applications that put AI into the economy. The US and China have each mastered a different half.

The US controls the bottom. Nvidia's chips, the CUDA software that locks developers in, and access to TSMC through allies make its hardware lead hard to challenge. Its capital matches that position: the five largest US hyperscalers are expected to spend around $750 billion on capex in 2026, roughly 0.6% of global GDP. The US constraints are now physical and financial: a power grid that cannot keep pace with data-centre build-out, and investors whose patience is not unlimited.

China has the upper half. Its strengths are cheap power, subsidised compute, a protected home market of a billion users, and above all open-weight models. Releasing weights lowers barriers at home, crowdsources improvements and reduces exposure to sanctionable foreign IP. This is how China stays close on models without spending American sums: it competes on diffusion and cost rather than on the frontier.

China's weakness is the bottom of the stack. Chips are its binding constraint, and export controls have made sure of that. Beijing is responding by building capacity, aiming to lift advanced-chip output from under 20,000 wafers a month to around 100,000 within two years. Even so, its chips lag a generation or more behind and high-bandwidth memory remains a bottleneck. China has built a parallel ecosystem, Huawei's Ascend chips and CANN software, but that is a substitute, not a peer.

The result is a mirror image. The US has the best hardware and the most capital but sells a premium, closed product. China has weaker hardware but sells a cheap, open one. Neither can dislodge the other, because each is strongest where the other is weakest.

The N-2 problem

The rest of the world is in a worse position than either. Measured by the ability to build and reproduce the full stack, only the US and, to a lesser extent, China are sovereign in AI. That is the N-2 problem: every other state is structurally dependent on one or both superpowers. Between them, the two hold roughly 90% of global AI computing capacity.

Europe shows the problem most clearly. It relies on US cloud providers and foundation models and on Taiwan for advanced chips. A second dependency is now forming: the most accessible near-frontier models are increasingly Chinese. Europe risks importing its compute from Washington and its affordable models from Beijing, while controlling neither.

Why the World Will Split, Even Though AI Can't Be Contained

Technology control has usually rested on keeping knowledge secret. AI knowledge cannot be kept secret: open weights, published papers and mobile researchers spread it faster than any export regime can stop it. So the split will not be in what countries *know* about AI. It will be in what they *use*.

Three forces push in that direction.

The stacks are becoming incompatible. Once Chinese models are optimised for Huawei hardware and American ones for Nvidia, choosing a model also means choosing chips, cloud, software tools and, eventually, standards. The split happens at the level of plumbing.

Both superpowers are turning their stacks into diplomatic offers. Washington's Pax Silica coalition, launched in December 2025 with seven members, secures the hardware supply chain for allies. Beijing's answer is the Shanghai-based World Artificial Intelligence Cooperation Organization (WAICO), pitched to the Global South with training, open models and a governance template. One side sells top-end hardware and conditions on alignment. The other sells cheap models with no conditions attached, at least for now.

Washington is forcing the choice. Export controls have become transactional, and access to US compute increasingly depends on keeping Chinese technology out. This leaves allies in a no-win position: the US is trying to force a split that the technology itself does not require.

The likely result is two bundles of AI use. Price-sensitive emerging markets will drift towards China's cheap, open bundle, and security-dependent allies towards America's. Middle powers such as Europe, India, the Gulf and ASEAN will be squeezed in between and asked to commit to one.

Programmable Sovereignty: A Third Option

Accepting the N-2 problem does not mean accepting a choice between two forms of dependence. The mistake lies in how sovereignty is measured. If sovereignty means building the whole stack, only two countries will ever have it. If it means being able to set the rules, and to walk away from a supplier, many more countries can have it.

This is programmable sovereignty: soft governance backed by just enough hard-stack control, above all pooled compute, to make rules credible and dependence reversible, without requiring full-stack autonomy. The key word is reversible. A country does not need to make its own chips. It needs enough control over its own compute that switching away from a supplier is a credible threat.

No middle power can do this alone, so the mechanism is coordinated specialisation: like-minded states pooling capabilities into a shared, credible stack, each contributing what it does best. There are precedents: EuroHPC, India Stack and the ASEAN Digital Economy Framework Agreement. In plainer terms, the AI cold war needs a non-aligned movement.

The Weak Points

There are three obvious objections. First, the split may be less clean than feared, because open weights run on any hardware and private firms will quietly use both stacks. Second, pooling compute across sovereign states is a classic collective-action problem. Third, any middle-power coalition still buys its most advanced chips from TSMC, which sits inside the US bloc. Programmable sovereignty reduces dependence, but it does not remove the reliance on TSMC at the bottom of the stack.

The Stakes

The choice for everyone outside the two superpowers is not whether to depend on someone. That is already settled. The choice is whether the dependence can be reversed. A world of two incompatible AI bundles would turn infrastructure choices into geopolitical alignments for decades. For Europe in particular, waiting means letting that choice be made for it, by Washington's licensing terms or by the price of Beijing's models. Pooling compute and governance capacity now is the cheapest way to keep the option of walking away from either.

Reference:

García-Herrero, A. and S. Dutta (2026). *Programmable sovereignty: coordinated specialisation as a response to the N-2 problem in artificial intelligence*. Bruegel. Working Paper 25/2026. https://www.bruegel.org/working-paper/programmable-sovereignty-coordinated-specialisation-response-n-2-problem-artificial

García-Herrero, A. and T. Storella (2026). *China's artificial intelligence goals and strategic choices for Europe*. Bruegel. https://www.bruegel.org/analysis/chinas-artificial-intelligence-goals-and-strategic-choices-europe

García-Herrero, A. (2026), *Tough choices ahead as the US pushes a global artificial-intelligence divide*. Bruegel. https://www.bruegel.org/first-glance/tough-choices-ahead-us-pushes-global-artificial-intelligence-divide



Alicia García Herrero is Adjunct Professor at the Hong Kong University of Science. Alicia also serves as a Senior Fellow at the Brussels-based European think-tank BRUEGEL and as Chief Economist for Asia Pacific at French investment bank NATIXIS. She is a non-resident Senior Follow at the East Asian Institute (EAI) of the National University Singapore (NUS) as well as at the Institute of Chinese Studies in Delhi. She is also a non-Executive Board member of listed-insurance group, AGEAS. Finally, Alicia is a Member of the Council of Advisors on Economic Affairs to the Spanish Government and an advisor to the Hong Kong Monetary Authority's research arm (HKIMR).

Alicia holds a PhD in Economics from George Washington University and has published extensively in refereed journals and books. She is also very active in international media.