Institut EGA

# Emerging Issues in AI: Threats & Opportunities

Webinar held on
27 August 2025

## Executive Summary

**September-October 2025**

**Institute for Applied Geopolitical Studies**

## The US–China AI governance race

The evolving dynamics between the US and China are shaping how frontier, high-performance AI models are developed and governed globally. **The US government has positioned AI development as a critical — arguably existential — competition with China. However, the precise end goal of this "race" remains undefined, and it is unclear whether such a goal exists in concrete terms.**

The US–China Economic and Security Review Commission, a bipartisan congressional body, has proposed a Manhattan Project-style initiative to ensure American leadership in AGI. This hawkish, bipartisan stance has shaped the policy landscape alongside a series of assertive legislative proposals — including a bipartisan bill that would ban Chinese AI technologies from use in federal agencies, as well as tighter controls on advanced chips and outbound US investment.

For its part, China has made a major investment in large-scale AI systems, some of them (like DeepSeek) open-source and carrying potential dual-use risks for surveillance, influence, or cyber operations. It has long focused more on proven, deployable capabilities over speculative frontier development, but this trajectory may be shifting. At this summer's World AI Conference in Shanghai, Zhou Bowen, the Director and Chief Scientist of the Shanghai AI Lab, outlined his vision for the responsible development of AGI, signaling that AGI is appearing more frequently in the Chinese calculus.

**US export restrictions have pushed Chinese AI developers toward efficiency and the development of domestic capabilities**, which could make their systems cheaper to run and easier to export at scale, aligning with China's export-oriented foreign policy goals. The PRC government has long advanced a narrative of benevolence and cooperation rather than competition, although there has often been a discrepancy between the actions of the government and the words of its officials. At home, China requires registration and content controls for public AI services, which affects what providers can build and deploy.

The US has stated its intent to curb China's influence in AI governance, and is advancing its own version of global AI governance by leveraging its private sector, influencing allies to adhere to its approach, and promising a technologically superior AI offering. This puts the two on a collision course.

## US strategic framing

In Washington, there is growing bipartisan interest around the idea of Artificial General Intelligence (AGI), with policymakers increasingly treating it as a feasible and near-term development. Discussions of AGI in the US government reflect a spectrum from enthusiasm about accelerating its development to concern over potential existential risks. Central to these conversations is the perceived need to "win" the AGI race against China — an idea that has gained bipartisan support. In the words of Ben Buchanan, former special adviser for AI in the Biden White House, "there are profound economic, military and intelligence capabilities that would be downstream of getting to AGI or transformative AI, and I do think it is fundamental for US national security that we continue to lead in AI."

In July 2025, the US released its AI Action Plan, a three-pillar framework for future AI policy. While the plan encompasses various goals, its central narrative is US–China AI competition. It was issued alongside a broader deregulatory pivot (revoking Biden's 2023 order) to speed domestic deployment.

The pillars are:

1. **Accelerating AI Innovation**
2. **Building American AI Infrastructure**
3. **Leading in International AI Diplomacy and Security**

The plan outlines priorities in research funding, safety standards, federal workforce upskilling, and public–private collaboration. It shows a strong focus on maintaining US global leadership while mitigating risks from misuse and adversarial AI development. It emphasizes supply chain resilience, compute infrastructure, and open innovation balanced with national security safeguards. There is a pronounced emphasis on monitoring foreign capabilities and securing critical systems — that is, a focus on security, but less on safety. There is also a strong preference for "open-source and open-weight AI."

The final section of the plan sets forth a multipronged strategy to secure US leadership in global AI by aggressively exporting the full AI technology stack to allies, countering adversarial influence in international governance, and tightening export controls on advanced AI compute and semiconductor manufacturing.

## Chinese strategic framing

China has been investing in its own AI governance approach and brand appeal for years, creating an inviting, open-source, out-of-the-box AI stack to offer foreign governments, and emphasizes promoting knowledge-sharing, coordination, and increasing access for the Global South. However, this does not mean that China always follows through on promises of coordination and cooperation. For example, despite publicly championing international alignment on safe and ethical AI in military domains, it refused to sign the "Blueprint for Action" which called for maintaining human control over nuclear weapons and stressed ethical, human-centric AI use in the military domain.

Xi Jinping has described AI as having a "spillover effect" and spoken of its importance in the future of China's economy. By pursuing the widespread adoption of open AI models such as DeepSeek R1 and Moonshot AI's Kimi K2, it seeks to achieve foreign policy goals while its domestic focus on AI deployment and integration keeps it ahead of potential competitors who might benefit from the models it releases.

China is deploying a variety of "**challenger tools**" to get ahead in the AI race. To acquire and absorb the technology of competitors, its tactics include: aggregation of foreign research, mandated joint ventures, licensing and targeted investments in firms with desirable IP, talent capture and reverse-engineering, economic espionage, and even chip-smuggling. These efforts have yielded such success that China is now at the forefront of some key fields across emerging technologies, and their tactical success is also the reason that China is highly unlikely to abandon them.

Some elements of the China Action Plan for AI Governance include:

- Calls for consensus-building through dialogue and cooperation, calls for "shared" global standards on AI regulation.
- Open innovation system for AI, aiming to get the global community on the China AI stack, attract investors and developers.
- Calls for tech knowledge sharing, open systems, cooperation.

However: China strategically chooses **not** to endorse multilateral agreements and engage in cooperative approaches when these agreements have implications for national sovereignty and security.

## Structural employment mismatch

The US is attempting to reshore manufacturing while factories are already struggling to find workers. At the same time, AI is eroding entry-level white-collar roles, leaving many recent college graduates without suitable employment pathways. This risks a huge skills mismatch and the prospect of reduced upward mobility for a generation which might end up trying to pay off college debt on factory and service-sector wages. Workers are already beginning to organize against the implementation of AI in the workplace.

It is not clear what the US government's approach will be regarding those displaced by AI. There is no coherent national strategy to manage transition pathways for displaced workers, or even a strategy to align labor market incentives with an internationally competitive industrial concept. Some analysts argue for wage insurance, portable benefits, or targeted regional mobility support, while others call for expanding public sector employment and service corps to absorb displaced talent, but all proposals are likely to be dwarfed by the scale of the issue.

In contrast to America's pivot towards reshoring factory jobs, China is replacing factory jobs by **aggressively pursuing automation**, aiming instead to break into high-value service sectors typical of advanced economies. In a recent speech, Xi Jinping "emphasized that China's economy has transitioned from a phase of rapid growth to one of high-quality development" and implied that artificial intelligence and other emerging technologies would be a key driver. This coincides with a shift away from flimsy mass-produced exports; the "Made in China 2025" plan sought to turn the phrase from an insult into a point of pride. Economic transition is not just a desire, it is a necessity for the Chinese economy, which is still in some ways reeling from the massive property market shock while facing an aging population and a public sector choice between severe austerity or debt. Frontier tech investment and automation are seen as the structural solutions to both fend off economic stagnation and deliver on the government's promises of greater wealth equality.

This contrasts sharply with most developing economies, where automation is often delayed by low wages, weaker infrastructure, and informal labor markets. In high-income economies like the US, automation typically replaces labor outright, whereas in emerging economies, AI more often **augments large low-wage workforces**. China straddles these worlds: domestically, it automates high-margin sectors such as finance and healthcare; internationally, it exports affordable, cloud-based AI tools designed for minimal infrastructure and large-scale workforce augmentation (particularly attractive in the Global South). The result is a strategic divergence in labor market adaptation: the US faces a policy coordination gap between industrial and labor strategies, while China is systematically integrating workforce transformation into its development model, potentially giving it a first-mover advantage in scaling AI-driven productivity.

## Industrial inputs

The US and China are already fighting a pitched battle over access to the semiconductors, energy, and critical materials/rare earth elements needed to maximize AI implementation. There is an **inherent tension between export controls and growth**, which is currently playing out in negotiations between American semiconductor companies and the federal government. It is possible that the end result will be 'pay-to-play' exports with US national security as a secondary consideration. This could be hugely deleterious to the US' long-standing AI national security agenda, and even compromise the security of allies to some degree.

Beyond the semiconductor arms race, both the US and China are also locked in competition over control of AI-critical upstream inputs — especially energy and rare earth elements. AI systems, particularly large language models and generative AI platforms, require enormous compute and energy resources to train and deploy. The US faces structural vulnerabilities in its energy infrastructure: demand for high-performance data centers is growing faster than local grids can support, creating bottlenecks in compute access even when chips are available. Meanwhile, China maintains a near-monopoly on rare earth processing capacity and has already begun pressing this advantage. These supply chain chokepoints give Beijing both international leverage and a domestic edge in AI deployment.

## Talent race

The US, China, and EU are in competition for top AI, quantum, and other frontier-tech scientists. The US holds an advantage in producing high-impact AI PhDs — more than double the per-capita number compared to China — even though China graduates more total STEM PhDs. The EU, by contrast, trains large numbers of skilled technical workers but produces fewer globally cited AI researchers and has historically struggled to retain top doctoral talent, with many migrating to US institutions or industry roles after graduation.

The multiple ongoing conflicts between the US federal government and its universities provides an opportunity for competitors to reverse the flow of talent to the US. Europe and China are both stepping up efforts to attract elite talent to reduce dependency, but the EU's experience — more software developers than the US, yet difficulty commercializing at scale — shows that talent is only one key to success. If the EU succeeds in attracting these highly skilled individuals, it will still face a significant challenge in achieving the desired economic benefit from their knowledge.

## AI in armed conflict

**Military use of AI is now a certainty.** The US proposal for a DOD AI and Autonomous Systems Virtual Proving Ground and an update to DOD guidance, roadmaps, and toolkits related to AI is a strong step toward effectively deploying advanced AI in national security contexts. The recent announcements of contracts with major companies such as OpenAI and Anthropic to acquire frontier AI systems show that the DOD is interested in such systems for potentially high-stakes use cases. Given that frontier AI systems introduce novel vulnerabilities and failure modes, this "proving ground" should include testing and evaluation guidance designed specifically for these systems, and developing adversarial testing environments that mirror operational reality. There should also be formal channels for third-party experts to supplement vendor-led evaluations and sharpen DOD processes.

The use of frontier AI models in battlefield environments comes with **adversarial exploitation risks**. If adversaries are able to access, reconstruct, or replicate these models, they could reverse-engineer US capabilities or integrate stolen technology into their own defense stack. Military deployments should mandate air-gapped infrastructure and consider developing purpose-built defense models with controlled scaling and tight compartmentalization to mitigate the risk of enemy exfiltration.

There is also a risk in relying on AI for military decision-making, when it is found to give **"escalatory" foreign policy recommendations**. AI agents used in strategic simulation environments have already demonstrated a tendency to favor escalatory or preemptive actions, especially in adversarial scenarios with ambiguous payoff structures. While useful in revealing blind spots, this pattern could be dangerously internalized into doctrine or planning if not critically reviewed. Models used in wargaming or red-teaming must be subject to sociotechnical evaluation frameworks, and their behaviors carefully annotated to separate model bias from human policy.

Another major risk for AI in warfare is deception. Interesting uses for AI in warfare include erroneously triggering early-warning systems and creating readiness fatigue by flooding intelligence pipelines with plausible but **false signals**, masking the real information in a deluge of unreal. Finally, implementing AI in the military without standardization across allies risks fragmenting interoperability in joint missions.

## AI as an agent of disinformation

AI has significant potential to shape public opinions to align with **state objectives**. This is one of the possible risks of DeepSeek, which on certain topics functions as a Chinese state propagandist (which researchers have shown results from deliberately designed mechanisms of internal censorship, which can be bypassed with the right knowledge).

Russia has pioneered a technique known as **LLM grooming** for narrative-shaping and subtle, long-term output manipulation. Russia's Pravda network is a sprawling propaganda operation that pumps out low-engagement content in dozens of languages, with as many as 3 million articles per year, intended to skew LLM outputs by boosting false narratives through sheer volume. Studies show that up to one-third of chatbot responses on controversial topics like purported US bioweapons in Ukraine echo these falsehoods, even if disconnected from user intent. The newest "reasoning" models are still highly susceptible to grooming, and the long-term risks to democracies are difficult to estimate.

Recent research shows that it is possible to train AI models to exhibit apparently compliant behavior during the evaluation phase, but to activate hidden deceptive functions once in production — a phenomenon described as "**sleeper agents**." Experiments demonstrate that these "backdoors" can persist even when state-of-the-art safety techniques (such as supervised fine-tuning or adversarial training) are applied. Worryingly, in some cases adversarial training — intended to eliminate such behaviors — actually makes the model better at recognizing the triggers and preserving its hidden malicious actions. In other words, apparent compliance may conceal dangerous latent capabilities, which seriously undermines structural trust in generative systems. This mechanism reinforces the idea that AI models can be instrumentalized for strategic disinformation not only through their direct outputs, but through hidden behavioral strategies that evade conventional oversight and testing.

Finally, the rise of sophisticated synthetic media has created opportunities for malicious actors to exploit the "**liar's dividend**" — the strategic use of plausible deniability to dismiss authentic evidence as fabricated. Politicians, corporations, and state-linked figures have begun claiming that real images, videos, and documents are "deepfakes" to evade accountability for misconduct or corruption, or to promote conspiratorial narratives. This dynamic undermines public trust in legitimate evidence, complicates verification efforts, and erodes the epistemic foundations of the democratic process at a time when trust in the institutions which typically verify images and documents is at critical lows.

## AI for crime or extermination

As previously mentioned, it is possible to bypass restrictions on what content an AI is permitted to produce. This process, known as "**jailbreaking**," can enable a range of criminal activities from AI-assisted cybercrime to the generation of detailed, tailored instructions for illicit activities such as drug synthesis, identity theft, or even assassination. These jailbreaks or purpose-built "dark" models can be shared widely. Researchers have already demonstrated that models can be used to design bioweapons precursors, bombs, or even jailbreak themselves when prompted in adversarial ways. As these models become more capable, more open, and more distributed, controlling their downstream misuse will be significantly harder than regulating physical weapons systems.

It is likely that an "arms race" to AGI will increase **existential risk**, or X-risk. There is growing concern that premature scaling or deployment of misaligned advanced systems — likelier under competitive pressure — may accelerate timelines toward catastrophic failures. The legal/ethical burden to regulate this process is currently the subject of intense international debate, often drawing parallels to Cold War–era nuclear nonproliferation frameworks. However, the AI domain differs in one critical way: access to foundational capabilities does not necessarily require state-level funding or infrastructure, making non-state actors a far more credible threat than in the case of nuclear technologies.

On that point, both the US and China recognize the risk of mixing AI and nuclear weapons, and have expressed public commitments to keeping AI out of nuclear launch decisions. However, not all nuclear powers will share this restraint — particularly newer or more isolated regimes, where the perceived deterrent value of "faster" or "smarter" command systems may tempt integration. Worse still, AI-enabled false alarms or spoofed signals could manufacture escalatory behavior by triggering early-warning systems. As models become more autonomous, the possibility of runaway feedback loops between automated sensors and AI decision engines becomes a genuine concern in future crisis scenarios.

A final, underexplored threat concerns two distinct but converging pathways:
(i) self-replicating digital agents that can spread across networks and connected devices
(ii) proliferated physical autonomous systems whose safeguards can be removed or altered.
These systems pose serious challenges for attribution, control, and existing regulatory frameworks.

## Evolution of AI

**1956**: The term “Artificial Intelligence” is coined at a Dartmouth workshop to attract funding and interest. It still lacks a precise definition but broadly refers to technologies that simulate human behaviors or tasks. What follows is a period of initial AI optimism and a long AI “winter” marked by skepticism about AI’s potential.

**2017**: The paper *Attention Is All You Need* introduced the Transformer, revolutionizing AI by replacing Recurrent Neural Networks (RNNs) and other popular forms of AI with attention-based mechanisms.

Key improvements:

- **Purpose**: Used the transformer for English-to-German translation.
- **Parallelism**: Processes all input simultaneously. Faster and more scalable on GPUs than CPUs.
- **Long-range dependencies:** Attention lets the model relate all parts of a sequence, avoiding limitations in long-distance relationships in data.
- **Much better performance**: Enables better understanding and training efficiency on language tasks.

**2018:** OpenAI (founded in 2015 as a nonprofit) released GPT-1 (~117M parameters).

**2019**: GPT-2 released (up to 1.5B parameters). Produced coherent text at scale. Scaling effects begin to show: AI capabilities improve predictably with data and computational power, making budgets and chips key in the AI development race.

**2020**: GPT-3 released (175B parameters). Major leap in quality and versatility.

**2022**: ChatGPT launched using GPT-3.5. Marked the breakthrough of conversational AI into the mainstream, reaching 100M users in 2 months, making it the fastest-growing consumer software app in history. With AI now in use by non-experts, reliability becomes the essential factor.

# Lessons Learned

- It would be unwise to overcommit to the goal of AGI, as it is unclear whether it is feasible. “Reasoning” is still an unsubstantiated claim — generative AI is still operating based on probability calculations, not logic. There may be a plateau or ‘bubble’ in AI development, and in some cases (such as that of Grok) the quality and reliability of model outputs may fluctuate noticeably based on the decisions of a relatively small number of executives and engineers. This presents risks in the case of both commercial models and potential future proprietary government models.

- Export controls buy some time, which must be used well to afford a meaningful advantage. Evidence suggests controls have slowed China’s access and temporarily preserved the US lead, but overreach can also undercut US technological diffusion and innovation.
  - Controls should be planned with allied coordination, but the legal systems of allies may prove to be a weak link. Many partners lack US-style legal tools (e.g., FDPR/Entity List) to implement coordinated controls on sensitive technologies. Harmonizing legal authorities is a prerequisite for effective multilateral regimes.

- Safety and security have to be planned from the earliest stages. Frontier models require loss-of-control (LOC) evaluations, adversarial testing, and continuous monitoring — safety cannot be effectively implemented through fine-tuning. Similarly, grid expansion should be paired with enhanced resilience and cybersecurity for data centers, and in some cases, recognizing AI facilities as critical infrastructure.

- AI policy is now industrial policy, labor policy, and foreign policy.
  - Electricity is now a primary limiting factor. Chips and data matter, but grid access and siting are the rate limiters for AI capacity. Grid modernization should be a top industrial priority.
  - The line between competition and collaboration, already thin, has blurred under the current US administration, placing allies in a challenging position as they choose between importing capabilities or waiting for local options to develop.
  - Workforce support measures such as retraining, apprenticeships, and transition support are underdeveloped relative to the scale of the potential disruption; they should be integrated into any future AI legislation. The US is currently offering the highest wages to lure the limited talent pool, resulting in a talent drain from lower-paying countries, but changes to the visa system may limit this effect.

**Machine Learning**: This is a specific branch of AI where software is built to statistically compute patterns in data. It is a subcategory under the broader term of AI.

**Deep Learning/Neural Networks**: This is a sub-category of machine learning that uses neural networks to calculate patterns in data. Most modern-day AI systems developed by companies like Meta, OpenAI, and Google are deep learning systems. While "neural networks" refer to the mathematical models, "deep learning" refers to the process of using very large neural networks, which has become ubiquitous in today's commercial models.

**CPUs vs. GPUs**: Transformer-based models like ChatGPT run better on GPUs (Graphics Processing Units) than CPUs (Central Processing Units). This accounts for the meteoric rise in valuation of companies like NVIDIA, which previously produced GPUs for gaming, and now make them for AI. Semiconductor manufacturing is highly specialized, requiring rare materials which are largely controlled by China and trade secrets which are largely held by Taiwanese companies. This places the global supply chain of the hardware underpinning AI at huge risk in the event of escalation between China and Taiwan.

**Open-weight vs. open-source**: In an **open-weight** model, the weights (the "knowledge" aka the trained parameters) are released for download. You can run, copy, or fine-tune it (often offline) and even remove/alter safety layers. The full training pipeline (data, recipes) is usually not included. Policy note: treat weight-sharing as high-risk for reuse and proliferation. **Open-source** would strictly mean that the whole project is testable/reproducible: not just the final weights, but also the code (like blueprints for reconstructing the original). Few frontier-scale models meet this bar today. Crucially, open-source does not necessarily mean open-weight; without the final weights, the risks are much lower.

**Generative AI**: systems capable of generating new content, such as text or images. ChatGPT is a prominent example of generative AI for text, while DALL-E and Stable Diffusion are examples for image generation. Resource-intensive to train.

**Agents**: Models calling tools under permissions and logs. Require audit trails & kill-switches.

'**Trustworthy' AI**: arguably not here yet, as all generative models hallucinate. Nonetheless, even critical industries such as compliance now run on AI.

**Artificial General Intelligence (AGI)**: This term is considered to be as ill-defined as AI itself, and is often seen as a "rebranding." AGI refers to the theoretical ability of AI to understand, learn, and apply intelligence across a wide range of tasks at a human level, potentially augmenting its own intelligence or becoming sentient. It represents a quasi-religious fervor for some companies like OpenAI and Anthropic, despite a lack of scientific evidence for its feasibility in the near future, with some researchers believing the necessary techniques don't yet exist.

## About the Speakers

**Emma Isabella Sage** is the co-founder and CEO of a research software startup, LIVINI, a research affiliate at the University of Glasgow, and the 2025 Young Professionals in Foreign Policy (YPFP) Rising Expert in National Security. Her work has been presented at GLOBSEC, submitted as evidence to the UK Parliament, and discussed on US national television. She graduated with Distinction from the Erasmus Mundus International Master in Security, Intelligence and Strategic Studies. Over the course of her career, she has specialized in geoeconomics, intelligence, and sub-threshold conflict.

**Steve Jarosz** is a former Senior Technical Consultant at Oracle, bringing 15 years of corporate consulting experience in large-scale business software and database implementations. He co-founded the research software company LIVINI while pursuing dual doctorates in linguistics with a focus on artificial intelligence at the Universities of Silesia and Sapienza. His research interests include various facets of emerging technology, including space and aeronautics, natural language processing, and artificial intelligence. Steve also has particular experience in the CEE region; he speaks Polish and Russian, and holds advanced degrees in Computer Science, General Linguistics, and Slavic Linguistics.

**Kateryna Halstead** is a national security, policy, and geopolitical risk professional specializing in technology and AI policy, national security strategy, and geopolitical risk consulting. Her interest in technology and AI policy was sparked during her Master of Arts in International Relations at Johns Hopkins University SAIS and deepened during her tenure as a Google Public Policy Fellow, where she explored global technology regulation, landmark US antitrust cases, and data policy developments shaping the digital landscape. She has also completed fellowships with the Federation of American Scientists and the Pallas Foundation. Currently, Kateryna serves as an AI Governance Research Fellow with ERA Cambridge, where she investigates US Government and Chinese Communist Party planning, regulatory priorities, and strategic investments on artificial general intelligence (AGI).

**Kelsey Quinn** is the Program Head and Analyst of Tech Sovereignty & Security at the New Lines Institute, where she investigates realistic approaches to mitigating current and future harms of emerging technology without impeding critical innovation and scientific discovery. She previously worked at the National Consortium for the Study of Terrorism and Responses to Terrorism (START) on the DARPA Sigma+ project, examining the decision and attack space for the use of CBRN weapons. She also previously worked as a research assistant at Michigan State University, investigating bacterial pathogenesis and physiology in Vibrio cholerae, a Category B bioterrorism agent. Quinn received her Bachelor of Science in Microbiology with a minor in Global Terrorism from the University of Maryland in 2019 and earned a Master's in Security and Terrorism Studies, also from UMD.

Institut EGA

ISSN : 2739-3283


Institute for Applied Geopolitical Studies
66 avenue des Champs-Élysées, 75008 Paris
Email: secretariat@institut-ega.org
Website: www.institut-ega.org