Navigating the AI Policy Crisis

Eleven Case Studies

Disclaimer*

Contents

*Disclaimer: These Case Studies have been prepared through interaction with a Generative AI application. Users of this document should therefore check any parts of the content of it for accuracy before sharing it, and should ensure that if it is shared, it is done so with an equivalent disclaimer. The views expressed in the document are not necessarily those of the Social Cyber and Tech Academy or the Social Cyber Institute.

Three Case Briefs: Cross-Border Stakeholder Mobilisation on Child Protection, Election Interference and AI Capacity Building

Workshop material | Social Cyber Institute

These briefs describe three recent cases in which corporations, governments and civil society organisations mobilised across borders around a contested technology governance question: protecting children online, defending elections against foreign interference, and transferring AI knowledge and infrastructure to developing countries. The cases differ in important respects, and each raises its own questions about power, legitimacy and accountability. Read together, they pose a common puzzle: when stakeholders mobilise across borders in the name of protection or assistance, who actually ends up protected, and who ends up assisted?

CASE 1. Thorn, the European Commission and "Chat Control": Child Protection as Product

A US child-safety organisation helps shape the EU law it stands to profit from, while civil society splits down the middle over whether protecting children requires scanning everyone's private messages.

Overview

In May 2022, the European Commission proposed a Child Sexual Abuse Regulation that would have required online platforms, including encrypted messaging services, to scan private communications for abuse material on receipt of a detection order. Critics quickly labelled the proposal "chat control". Investigations by European news outlets later revealed that the proposal had been developed in close coordination between Home Affairs Commissioner Ylva Johansson's directorate and Thorn, a US organisation founded by the actors Ashton Kutcher and Demi Moore that sells CSAM-detection software while being registered in the EU lobby database as a charity (Fayed 2023; Fortune 2023).

Key facts

- Thorn markets Safer, a commercial detection product, and stood to benefit directly from a law mandating detection at scale. It nonetheless met EU lawmakers and commissioners under a charity classification, holding meetings with some 20 EU lawmakers in 2021 alone, and reported spending over $630,000 on lobbying in a single year (Fortune 2023).
- Emails obtained through freedom-of-information requests show Commissioner Johansson thanking Thorn for its help in crafting the proposal, describing the regulation as a response that "would not exist" without the organisation (Fayed 2023).
- Digital rights groups called the affair "ChatControlGate" and described it as a textbook case of regulatory capture: an entity that stood to profit from a law was deeply involved in writing it (European Digital Rights 2023).
- Hundreds of leading cryptographers and security researchers signed open letters warning that client-side scanning of encrypted communications cannot be done safely, and Signal stated publicly that it would withdraw from the EU market rather than undermine its encryption.
- The Council deadlocked on the proposal for more than three years. In late 2025 the Danish presidency dropped mandatory scanning, converting the scheme into a permanent voluntary framework with age-verification and risk-mitigation obligations attached (The Register 2026).
- In July 2026 the European Parliament extended the interim derogation that permits voluntary scanning for known abuse material until April 2028, while trilogue negotiations on the permanent regulation continued without agreement (The Register 2026).

The cross-border alliance dimension

The alliance here ran from Los Angeles through Brussels. A US organisation with a commercial product shaped the flagship child-protection law of a foreign jurisdiction, working alongside the WeProtect Global Alliance and law enforcement voices, and lent the Commission its moral authority in return for a seat at the drafting table. The proposed EU rules would in turn have set a

global template: legislators in other jurisdictions watched the EU debate closely, and detection infrastructure built for Europe would have been sold worldwide (Fayed 2023).

The protection tension

Nobody in this case argued against protecting children. The fight was over whether a mandate to scan billions of private messages is protection at all, or a surveillance architecture wearing protection's clothes. The moral force of the cause made ordinary scrutiny difficult: opposing the regulation invited the accusation of indifference to child abuse, which is precisely what made the file so attractive to those seeking expanded scanning powers (European Digital Rights 2023).

The civil society dimension

Civil society did not line up on one side of this case; it split. Digital rights organisations led by EDRi ran the "Stop Scanning Me" campaign, joined by more than a hundred organisations, and were reinforced by the scientific community's open letters. On the other side, child-protection organisations including ECPAT and Eurochild campaigned for the regulation, arguing that encrypted platforms had become safe harbours for abuse material and that privacy advocates were defending an intolerable status quo.

Both camps were sincere, and both mobilised across borders. The effect, though, was that civil society spent much of its energy contesting itself while the Commission–Thorn alliance defined the terms of the debate. The European Ombudsman and members of the European Parliament raised questions about the Commission's closeness to its outside advisers, and transparency complaints dogged the file for years (Fayed 2023). The eventual retreat from mandatory scanning owed much to the digital rights campaign, but the voluntary framework that survives still carries most of the machinery the campaign objected to.

Children themselves, in whose name the entire contest was conducted, appear nowhere in the record as participants. Survivor groups were cited by both sides and consulted by neither in any structured way.

Discussion questions

Q1: Does it matter that Thorn is a child-safety organisation rather than a conventional technology vendor? Would the same conduct by Microsoft have been treated differently?

Q2: When civil society splits into two sincere camps, who arbitrates? Did the split strengthen or weaken democratic scrutiny of the proposal?

Q3: Is there any institutional design that would let a body with commercial interests contribute expertise to lawmaking without capturing it?

Q4: The voluntary scanning framework survived even though the mandate did not. Is that a victory for the campaign against chat control, or a quieter version of the same outcome?

References

European Digital Rights (2023) 'How a Hollywood star lobbies the EU for more surveillance'. Available at: https://edri.org/our-work/how-a-hollywood-star-lobbies-the-eu-for-more-surveillance/ (Accessed: 10 July 2026).

Fayed, S. (2023) 'How Ashton Kutcher's "non-profit start-up" makes millions from the EU's fight against child abuse on the net', Follow the Money, 7 October. Available at: https://www.ftm.eu/articles/ashton-kutchers-non-profit-start-up-makes-millions-from-fighting-child-abuse-online (Accessed: 10 July 2026).

Fortune (2023) 'Reports: Kutcher-founded Thorn lobbied for EU CSAM law', 26 September. Available at: https://fortune.com/europe/2023/09/26/thorn-ashton-kutcher-ylva-johansson-csam-csa-regulation-european-commission-encryption-privacy-surveillance/ (Accessed: 10 July 2026).

The Register (2026) 'EU "Chat Control" snoopfest returns after vote to kill it falls short', 9 July. Available at: https://www.theregister.com/security/2026/07/09/meps-fail-to-prevent-chat-control-snoopfest-revival/ (Accessed: 10 July 2026).

CASE 2. TikTok and the Annulled Romanian Election: Foreign Interference and the Remedy Problem

A platform's recommender system becomes the suspected instrument of foreign political interference, and a constitutional court cancels a national election, leaving every stakeholder to argue about whether the cure damaged democracy more than the disease.

Overview

In November 2024, Călin Georgescu, a fringe nationalist candidate polling in single digits weeks earlier, won the first round of Romania's presidential election after a sudden surge of support on TikTok. On 6 December 2024, Romania's Constitutional Court annulled the result, relying in part on declassified intelligence assessments that a foreign influence operation, consistent with Russian methods, had manipulated social media in his favour. Eleven days later the European Commission opened formal proceedings against TikTok under the Digital Services Act, the first DSA case built squarely on election integrity (European Commission 2024).

Key facts

- Romanian intelligence assessments described roughly 25,000 pro-Georgescu TikTok accounts that became highly active in the two weeks before the vote. Nearly 800 of them had been created in 2016 and sat largely dormant until the campaign (DSA Observatory 2024).
- Global Witness ran experimental accounts in Romania and found TikTok's algorithm recommending pro-Georgescu content at between 4.6 and 14 times the rate of content supporting his run-off opponent (Global Witness 2024).
- Influencer payments for pro-Georgescu content were channelled in ways that evaded TikTok's prohibition on paid political content, while Georgescu himself declared campaign spending of zero.
- The Commission's proceedings focus on TikTok's recommender systems, its policies on political advertising and paid-for political content, and its mitigation of election-specific risks in national and linguistic context. A breach finding can carry a fine of up to 6 per cent of global turnover (European Commission 2024).
- The election was re-run in May 2025. Georgescu was barred from standing, the nationalist George Simion led the first round, and the independent Nicușor Dan won the run-off on 18 May 2025.
- The DSA proceedings remained open as of mid-2026, with no legal deadline for their conclusion. TikTok, meanwhile, strengthened its Romanian election measures: better labelling of political content, some 120 additional specialists on manipulation, and participation in the EU's rapid-alert arrangements (European Parliament 2025).

The cross-border alliance dimension

The alliance structure in this case is unusual because it formed after the fact and in response to a suspected hostile one. A foreign influence operation, a platform owned by a Chinese company, a national constitutional court, Romanian intelligence agencies and the European Commission all became entangled within a fortnight. Brussels used the Romanian crisis to assert that election integrity across all 27 member states is a matter for EU-level platform regulation, and other governments watched the precedent closely. So did other platforms: the case established that a recommender system can itself be the object of an interference investigation.

The legitimacy tension

Annulment was the strongest medicine available, and it cut both ways. If the intelligence assessments were right, the vote had been contaminated and cancelling it defended democracy. But the annulment also cancelled millions of authentic votes on the basis of largely classified evidence, and it handed every populist in Europe a ready-made narrative about establishments overturning elections they lose. The remedy for interference became, for a large part of the Romanian electorate, the interference (DSA Observatory 2024).

The civil society dimension

Romanian election-monitoring organisations, among them Expert Forum and Funky Citizens, did much of the early documentation of the TikTok campaign and pressed both the platform and the authorities to act. International organisations added investigative weight: Global Witness's algorithm experiments gave the Commission some of its most concrete evidence, and academic networks around the DSA supplied the analytical frame (Global Witness 2024; DSA Observatory 2024).

Yet civil society was divided over the annulment itself. Some of the same organisations that had documented the manipulation expressed unease that the Constitutional Court acted on evidence the public could not examine. International commentary split along similar lines, with some seeing a democracy defending itself and others a dangerous precedent. Civil society proved effective at detection and documentation, in other words, but had no agreed answer to the harder question of remedy.

The voters most affected, including the large Romanian diaspora that had voted heavily for Georgescu, were not represented in either the platform's risk assessments or the institutional response. Their grievance became a political resource for the next nationalist campaign.

Discussion questions

Q1: The evidence for interference was largely classified, yet the remedy was maximal. What standard of public proof should be required before an election result is annulled?

Q2: Is a recommender system that amplifies one candidate 4.6 to 14 times more than another a form of political interference even without foreign involvement?

Q3: The DSA gives the Commission investigative power but no deadline. Does slow enforcement protect elections, or does the deterrent evaporate if consequences arrive years later?

Q4: If you advised a smaller democracy outside the EU facing the same situation, without DSA-style leverage over platforms, what options would it realistically have?

References

DSA Observatory (2024) 'TikTok and the Romanian elections: a stress test for DSA enforcement', 20 December. Available at: https://dsa-observatory.eu/2024/12/20/tiktok-and-the-romanian-elections/ (Accessed: 10 July 2026).

European Commission (2024) 'Commission opens formal proceedings against TikTok on election risks under the Digital Services Act', press release IP/24/6487, 17 December. Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_6487 (Accessed: 10 July 2026).

European Parliament (2025) 'Publication of the report on the investigation into TikTok under the DSA ahead of the May 2025 presidential election campaign in Romania', parliamentary question P-000253/2025. Available at: https://www.europarl.europa.eu/doceo/document/P-10-2025-000253_EN.html (Accessed: 10 July 2026).

Global Witness (2024) 'What happened on TikTok around the annulled Romanian presidential election? An investigation and poll'. Available at: https://globalwitness.org/en/campaigns/digital-threats/what-happened-on-tiktok-around-the-annulled-romanian-presidential-election-an-investigation-and-poll/ (Accessed: 10 July 2026).

The Conversation (2024) 'Why Romania's election was annulled – and what happens next?'. Available at: https://theconversation.com/why-romanias-election-was-annulled-and-what-happens-next-245779 (Accessed: 10 July 2026).

CASE 3. Microsoft, G42 and Kenya: Capacity Building as Market Building

Global AI providers offer a developing country infrastructure, skills training and help with governance frameworks. The assistance is real, and so is the dependency that arrives bundled with it.

Overview

In May 2024, Microsoft and the Abu Dhabi AI firm G42 announced a $1 billion "comprehensive digital ecosystem initiative" for Kenya, agreed with Kenya's Ministry of Information, Communications and the Digital Economy and signed during President William Ruto's state visit to Washington. The package promised a geothermal-powered data centre at Olkaria running Microsoft Azure, a Swahili-and-English large language model, AI skills training for every government employee, connectivity investments, and collaboration with the Kenyan government on a "trusted data zone" and cloud governance. The letter of intent was crafted with the assistance of the US and UAE governments (Microsoft 2024). Two years later the flagship data centre had stalled, and Kenya's own draft AI policy had named dependence on US technology companies as a threat to national security (Bhattacharya 2026).

Key facts

- The initiative bundled infrastructure, skills and governance into one package: alongside the data centre came commitments to train government employees in AI, run cybersecurity skilling for more than 2,000 people a year, and work with the government on adopting a "cloud-first" policy encouraging state agencies to move their data and computing to the cloud the partners would operate (Microsoft 2024).
- Microsoft's vice chair Brad Smith described the deal as "a real opportunity to bring digital technology to the Global South in a safe and secure manner" (Microsoft 2024).
- By 2026 the project had stalled after the Kenyan government declined to guarantee the annual computing purchases the companies sought, with electricity capacity a further obstacle (Bhattacharya 2026; Tom's Hardware 2026).
- Nigeria, Egypt and Kenya have all released draft AI policies since January 2025 that identify dependence on US technology companies as a security risk. South Africa published a draft and withdrew it in April 2026 after the AI tools used to help write it generated fake citations (Bhattacharya 2026).
- Africa holds about 18 per cent of the world's population and less than 1 per cent of global data centre capacity. Several open-source African AI initiatives receive grants from Meta and run on Google Cloud, and comparable provider-to-government arrangements are spreading, including Anthropic's memorandum of understanding with Rwanda (Bhattacharya 2026).
- Collective responses are emerging: the African Union released a Continental AI Strategy in July 2024, a $60 billion Africa AI Fund was announced at the Kigali summit in April 2025, and Smart Africa established an Africa AI Council in November 2025 (African Union 2024; Bhattacharya 2026).

The cross-border alliance dimension

This alliance spanned four governments and two companies: a US corporation, an Emirati AI firm, and the governments of Kenya, the United States and the United Arab Emirates, each with distinct strategic interests in where African data lives and whose technology stack the region adopts. The offer to help Kenya govern AI arrived inside the same package as the infrastructure to be governed. When the provider of the technology also funds the skills training, drafts the governance framework and hosts the government's own data, the line between capacity building and market capture becomes difficult to draw (Microsoft 2024; Bhattacharya 2026).

The knowledge transfer tension

The assistance is not fake. Kenya gains real infrastructure, real training and real access to frontier technology it could not build alone, and regulation drafted with expert help may well be better than regulation drafted without it. The tension is that the same transfer creates what one researcher called the absorption of adoption costs: the provider makes dependence cheap and easy, and the

country's regulatory imagination forms around the technology it has already adopted. As a policy adviser observed of outsourced data centres elsewhere on the continent, the question is not just about having control but about having meaningful control (Bhattacharya 2026).

The civil society dimension

Civil society in this case operates mostly at expert level rather than in the streets. Research organisations such as the Global Center on AI Governance, advisory nonprofits like the Digital Impact Alliance, and university researchers have done the main work of naming the dependency problem and proposing procurement rules, data sovereignty measures and regional pooling as answers (Bhattacharya 2026). Kenyan digital rights groups and the data-labelling workforce have raised a parallel set of concerns about labour conditions in the AI supply chain, where workers in low-income countries carry out content moderation and annotation for global providers on poor wages.

What is missing is mobilisation with leverage. There is no Kenyan equivalent of the parliamentary committee pressure seen in the UK Palantir case, and no mass campaign comparable to the European fight over chat control. The stakeholders best resourced to shape the outcome remain the providers and the governments; the experts critique, the workers organise at the margins, and the citizens whose data and services are at stake are largely spectators. Notably, the most effective pushback so far has come from the Kenyan state itself, which stalled the deal by refusing to socialise the project's commercial risk (Bhattacharya 2026).

Discussion questions

Q1: Can a developing country accept AI infrastructure, skills and regulatory assistance from a global provider without its regulatory framework being shaped by that provider's interests? What safeguards would it take?

Q2: Kenya stalled the deal by refusing guaranteed capacity payments. Was that a failure of the partnership or the system working, a state pricing dependency accurately for once?

Q3: Is "AI sovereignty" a realistic goal for countries with less than 1 per cent of global data centre capacity, or should the goal be redefined as meaningful control within dependence?

Q4: Who should legitimately represent a country's citizens in negotiations like these: the government that signs, the experts who critique, the workers in the AI supply chain, or somebody else?

References

African Union (2024) Continental Artificial Intelligence Strategy, July. Available at: https://au.int/sites/default/files/documents/44004-doc-EN-_Continental_AI_Strategy_July_2024.pdf (Accessed: 10 July 2026).

Bhattacharya, A. (2026) 'Pushing back from Big Tech: Africa's hard road to AI sovereignty', Rest of World, 21 May. Available at: https://restofworld.org/2026/africa-ai-sovereignty-big-tech/ (Accessed: 10 July 2026).

Microsoft (2024) 'Microsoft and G42 announce $1 billion comprehensive digital ecosystem initiative for Kenya', press release, 22 May. Available at: https://news.microsoft.com/source/2024/05/22/microsoft-and-g42-announce-1-billion-comprehensive-digital-ecosystem-initiative-for-kenya/ (Accessed: 10 July 2026).

MIT Technology Review (2024) 'Africa's push to regulate AI starts now', 15 March. Available at: https://www.technologyreview.com/2024/03/15/1089844/africa-ai-artificial-intelligence-regulation-au-policy/ (Accessed: 10 July 2026).

Tom's Hardware (2026) 'Microsoft's $1 billion Kenya data center stalls over disagreements on power capacity'. Available at: https://www.tomshardware.com/tech-industry/microsofts-1-billion-kenya-data-center-stalls-over-disagreements-on-power-capacity (Accessed: 10 July 2026).

Comparative Note

The three cases represent distinct types of cross-border stakeholder mobilisation, and it helps to name the differences plainly.

- Case 1 (Thorn/Chat Control): mobilisation under a moral banner. A commercially interested actor borrows the authority of child protection to help write law, and civil society splits into two sincere camps that spend their energy contesting each other.
- Case 2 (TikTok/Romania): reactive mobilisation against a hostile operation. Detection and documentation succeed, but no stakeholder has a good answer to the remedy question, and the strongest response damages the legitimacy it was meant to defend.
- Case 3 (Microsoft/Kenya): mobilisation as benevolence. Assistance and dependency arrive in the same package, civil society critiques from the sidelines, and the most effective resistance comes from the recipient state's procurement decisions.

Across all three cases, the stakeholders with the strongest claim to be the point of the exercise – children, voters, the citizens of developing countries – are the ones least represented in the rooms where the outcomes were decided. Protection, defence and assistance were all conducted on their behalf and largely without them.

A second thread is worth drawing out: in each case the mobilising banner (child safety, election integrity, development) is genuine, and in each case it also does work for interests that have little to do with the banner. The analytical skill the session should build is the ability to hold both facts at once without collapsing into either cynicism or credulity.

Three Case Briefs: Corporate–Government Alliances on Algorithm Transparency

Workshop material

These briefs describe three recent cases in which a large corporation engaged in unusually intense cross-border communication and alliance building with foreign governments over algorithm transparency and trustworthiness. The cases differ in important respects, and each raises its own questions about power, legitimacy and accountability. Read together, though, they force a harder question: what does “transparency” actually amount to in practice, and for whom?

CASE 4. OpenAI and the EU AI Act: Lobbying Against Transparency Behind a Transparency Agenda

A US corporation publicly champions global AI regulation while quietly pressing the European Commission to soften the very transparency obligations it claims to support.

Overview

Between 2022 and 2023, OpenAI, backed by a $13 billion investment from Microsoft, lobbied intensively behind the scenes to influence the EU officials drafting the AI Act. Over the same period its chief executive, Sam Altman, toured world capitals urging governments to regulate AI. Documents obtained by TIME magazine from the European Commission under freedom of information rules revealed a considerable distance between the company’s public position and its private advocacy (Perrigo 2023).

Key facts

- OpenAI argued to European officials that its general-purpose AI systems, including GPT-3 and DALL-E 2, should not be classified as “high-risk”. That designation would have brought mandatory transparency, traceability and human oversight requirements with it (Perrigo 2023).
- The argument put OpenAI in the same camp as Microsoft and Google. Both had lobbied to shift transparency obligations away from the large foundation model providers and onto the downstream companies that apply the models to particular uses (Perrigo 2023).
- The outcome favoured the companies. The AI Act as enacted requires foundation model providers to comply with a reduced set of obligations rather than the more stringent requirements in the original proposal.
- At the same time, OpenAI launched “OpenAI for Countries”, framing partnerships with foreign governments around AI “trustworthiness” and “democratic values” and using those relationships to build international political support for US-led AI governance norms (OpenAI 2025).
- OpenAI’s US lobbying spend rose from $260,000 in 2023 to $1.76 million in 2024 (O’Donnell 2025). The wider industry has kept climbing: reporting in 2026 put combined lobbying expenditure by Meta, Nvidia and Alphabet at $47 million for 2025 alone (The New York Times 2026).

The cross-border alliance dimension

OpenAI, Microsoft and Google worked in parallel to shape EU regulation in ways that would also constrain regulatory ambition in other markets. The “OpenAI for Countries” initiative then invited governments to partner on AI infrastructure, creating structural dependencies that embed US corporate preferences into national AI policy frameworks (OpenAI 2025).

The transparency tension

OpenAI lobbied to control the definition of the very problem it claimed to be solving. It argued that its systems were not “high-risk” and so should not face the transparency and oversight requirements that the designation triggers, while publicly championing global AI safety (Perrigo 2023).

The civil society dimension

Civil society was active in this case but largely invisible to the public. More than 150 organisations, led by European Digital Rights (EDRi), Amnesty International, Algorithm Watch, Human Rights Watch and the Irish Council for Civil Liberties, signed open letters to EU legislators demanding that the high-risk classification loophole introduced through corporate lobbying be closed and the Commission's original language restored (European Digital Rights 2023). Their arguments were substantive and technically precise. They named Article 6 specifically and called for objective, externally determined risk classification in place of self-assessment by developers.

This activism ran almost entirely through Brussels policy channels and generated little public attention, let alone street-level mobilisation. The record is more complicated than a simple story of corporate capture, however. A 2024 Erasmus University study found that civil society groups were in fact more successful than industry in influencing the final text of the AI Act. In 2025, more than 50 organisations, with signatories including the Nobel laureates Daron Acemoglu and Geoffrey Hinton, wrote to the European Commission to oppose attempts to delay or weaken implementation (EU Perspectives 2025). Meanwhile the EU Advisory Forum for civil society participation, mandated by Article 67 of the Act, had still not been established as of mid-2026, despite design recommendations submitted by more than 30 organisations.

Where does that leave civil society? It shaped the text, but it never gained entry to the corporate-government alliance that drove the lobbying. Expertise and formal participation, this case suggests, are not the same thing as power.

Discussion questions

Q1: Is there a meaningful difference between lobbying against transparency requirements and lobbying against trustworthiness?

Q2: Which stakeholders were excluded from this alliance-building process, and how would you design a more inclusive governance process?

Q3: If you were advising a mid-tier government facing this situation, what leverage did it actually have?

Q4: Does the "democratic AI" framing of OpenAI for Countries change your assessment of OpenAI's motives?

References

EU Perspectives (2025) 'Reject AI deregulation: 52 civil society groups urge the Union'. Available at: https://euperspectives.eu/2025/07/reject-ai-deregulation-52-civil-society-groups-urge-the-union/ (Accessed: 10 July 2026).

European Digital Rights (2023) 'EU: Close the loophole in Article 6 of the AI Act : civil society statement'. Available at: https://edri.org/our-work/civil-society-statement-eu-close-loophole-article-6-ai-act-tech-lobby/ (Accessed: 10 July 2026).

O'Donnell, J. (2025) 'OpenAI has upped its lobbying efforts nearly sevenfold', MIT Technology Review, 21 January. Available at: https://www.technologyreview.com/2025/01/21/1110260/openai-ups-its-lobbying-efforts-nearly-seven-fold/ (Accessed: 10 July 2026).

OpenAI (2025) OpenAI for Countries: our approach to security. Available at: https://cdn.openai.com/global-affairs/ff86ec36-1741-40a0-ab2a-d46f8eec62a8/openai-for-countries-our-approach-to-security.pdf (Accessed: 10 July 2026).

Perrigo, B. (2023) 'Exclusive: OpenAI lobbied the EU to water down AI regulation', TIME, 20 June. Available at: https://time.com/6288245/openai-eu-lobbying-ai-act/ (Accessed: 10 July 2026).

The New York Times (2026) Report on AI industry lobbying in Washington, 13 May. Available at: https://www.nytimes.com/2026/05/13/technology/ai-lobbying-washington-openai-anthropic.html (Accessed: 10 July 2026).

CASE 5. Palantir and the UK/US Government Alliance: Opacity as a Feature, Transparency as a Sales Strategy

A data analytics company builds deep contractual relationships with allied governments through intensive lobbying and revolving-door hiring, while resisting the algorithmic transparency it markets as its competitive differentiator.

Overview

Over more than two decades, Palantir Technologies has built a dominant position in government data analytics, first in the US intelligence community and then across allied governments. Since 2023 it has secured contracts worth more than £500 million with the UK's NHS and Ministry of Defence, while its US federal contracts grew to over $970 million in 2025 (Tech Transparency Project 2025). The company's strategy has rested on intensive lobbying conducted across multiple jurisdictions at once, with the same playbook replicated in each allied capital.

Key facts

- Palantir's UK lobbying included cocktail parties for NHS leadership, donations to MPs, and engagement with the lobbying firm Global Counsel, founded by Peter Mandelson, who simultaneously served as UK Ambassador to the US (Prospect 2025).
- A 2025 Financial Times investigation found Palantir operating a "revolving door with Washington and Westminster", using current and former civil servants and ministers as a recruitment and influence pipeline.
- For months the UK government blocked MPs from examining Palantir's NHS contracts. It took an intervention from the Information Commissioner to open them up (The Guardian 2026).
- A Swiss army investigation found that Palantir's identity as a US enterprise raised concerns that data shared with the company could be accessed by US government and intelligence agencies.
- In the US, Palantir's ImmigrationOS contract with ICE, along with alleged links to a DOGE-initiated "mega API" at the IRS, raised concerns about algorithmic decision-making that has no public accountability attached to it (Tech Transparency Project 2025).
- Palantir markets its products on privacy-by-design and granular access controls. Its algorithms remain proprietary, however, which means neither the public nor elected representatives can verify how its systems actually evaluate data (Medact 2026).

The cross-border alliance dimension

The same influence tactics (revolving-door hiring, hospitality, strategic research funding) have been deployed in Washington and Westminster in parallel. CEO Alex Karp has said openly that his goal is for "all large institutions in the United States and its allies abroad" to run on Palantir systems. The effect would be a network of allied governments connected through shared proprietary infrastructure (Prospect 2025).

The transparency tension

Palantir's systems help determine who appears on deportation lists, which surveillance targets are prioritised, and which NHS patients receive what interventions. None of this can be independently audited. Palantir's answer is that its clients control the data. Critics respond that this dissolves accountability rather than allocating it: the government cannot explain the system it procured, and the public cannot scrutinise either party (Medact 2026; The Guardian 2026).

The civil society dimension

This is by some distance the most activist-intensive of the three cases, and the only one to have generated sustained street-level mobilisation. The campaign has been explicitly cross-issue, linking algorithmic opacity and patient privacy with Palantir's contracts with US Immigration and Customs Enforcement and with the Israeli military. That breadth has given it an unusually diverse coalition.

The key moments are worth recording. In December 2023, Healthcare Workers for Palestine blockaded Palantir's central London headquarters, with hundreds of NHS workers in scrubs chanting "no place for war profiteers in our NHS" (Ali 2023). The campaign group Pull the Plug coordinated protests outside a UK healthcare conference in June 2026, backed by Amnesty International and the Unison union (WIRED 2026). Later that month, over 100 medical workers and Amnesty supporters staged an action on Brighton beach (Amnesty International UK 2026). In June 2026 the cross-party House of Commons Science, Innovation and Technology Committee called for the NHS contract to be terminated. The contract comes up for renewal in February 2027.

Civil society here is highly visible, escalating, and equipped with real parliamentary leverage. The multi-front framing (privacy, Gaza, democratic accountability) has made the campaign hard for the government to contain as a narrowly technical governance debate. Of the three cases, this is the one where activist pressure may yet change the outcome.

Discussion questions

Q1: When a government cannot fully explain how an AI system it has procured makes decisions, who is accountable: the government or the vendor?

Q2: Is Palantir's lobbying strategy fundamentally different from that of any other large defence contractor, or does the algorithmic dimension change the analysis?

Q3: What procurement conditions would you recommend governments impose to ensure algorithmic accountability without deterring innovation?

Q4: How should a government respond when a vendor's cross-border operations create intelligence access risks, as alleged in the Swiss army report?

References

Ali, T. (2023) 'Health workers in the UK shut down the headquarters of a Gaza war profiteer', Jacobin, 27 December. Available at: https://jacobin.com/2023/12/health-workers-nhs-palantir-gaza-protest (Accessed: 10 July 2026).

Amnesty International UK (2026) 'UK health workers and activists brand Palantir a company "with blood on its hands" in Brighton beach protest'. Available at: https://www.amnesty.org.uk/latest/uk-health-workers-and-activists-brand-palantir-a-company-with-blood-on-its-hands-in-brighton-beach-protest/ (Accessed: 10 July 2026).

Medact (2026) Briefing: Palantir and the Federated Data Platform. Available at: https://www.medact.org/2026/resources/briefings/briefing-palantir-fdp/ (Accessed: 10 July 2026).

Prospect (2025) 'How Palantir infiltrated the state', Prospect Magazine. Available at: https://www.prospectmagazine.co.uk/politics/democracy/government/71511/how-palantir-infiltrated-the-state (Accessed: 10 July 2026).

Tech Transparency Project (2025) 'Inside Palantir's expanding influence operation'. Available at: https://www.techtransparencyproject.org/articles/inside-palantirs-expanding-influence-operation (Accessed: 10 July 2026).

The Guardian (2026) 'Calls to halt UK Palantir contracts grow amid lack of transparency over deals', 5 February. Available at: https://www.theguardian.com/politics/2026/feb/05/calls-to-halt-uk-palantir-contracts-grow-amid-lack-of-transparency-over-deals (Accessed: 10 July 2026).

WIRED (2026) 'Palantir protests at NHS conference in the UK'. Available at: https://www.wired.com/story/palantir-protests-nhs-conference-uk/ (Accessed: 10 July 2026).

CASE 6. TikTok/ByteDance and the Global Algorithm Transparency Gambit: Projects Texas and Clover

A Chinese-owned corporation attempts to negotiate algorithmic trust with several governments at once, investing $1.5 billion in transparency infrastructure that ultimately fails to resolve a structural conflict between corporate architecture and national security logic.

Overview

TikTok, owned by Beijing-based ByteDance, became the subject of the most intensive corporate–government negotiation over algorithm transparency in recent history. Facing pressure in the US, Europe, Australia and elsewhere, the company pursued parallel trustworthiness strategies: Project Texas in the US and Project Clover in Europe. It offered governments access to its source code and recommendation algorithm on a scale no competitor had contemplated. The effort ultimately failed in the US and remains contested in Europe (Perault 2024; Turvy and Scharlach 2026).

Key facts

- Project Texas (2022) was a $1.5 billion restructuring of TikTok's US operations. All US user data was migrated to Oracle Cloud servers. A new subsidiary, USDS, was staffed by US nationals holding government-grade background checks. Oracle was given authority to inspect TikTok's source code, and seven oversight entities, including CFIUS, were assigned monitoring roles (Perault 2024).
- Project Clover (2023) was the parallel European initiative: €1.2 billion in new data centres in Ireland and Norway, security gateways controlling employee access to data, and an independent security company monitoring data flows (BBC News 2023).
- TikTok opened physical Transparency and Accountability Centres in the US and UK, where regulators, journalists and civil society groups could review source code under non-disclosure agreements.
- The UK's National Cyber Security Centre spent years reviewing TikTok's source code in Oxfordshire and found no evidence of Chinese government interference. US-led political pressure nevertheless drove the European Commission to ban TikTok from staff devices.
- US lawmakers deemed Project Texas insufficient. The Protecting Americans from Foreign Adversary Controlled Applications Act was upheld by the Supreme Court 9-0 in January 2025, and a forced partial divestiture followed: a joint venture with Oracle, Silver Lake and MGX. ByteDance retains ownership of the recommendation algorithm, whose intellectual property remains in Beijing under a licensing arrangement (Turvy and Scharlach 2026).
- The key technical finding deserves emphasis. Even with US data held on US servers, source code and algorithmic updates still flowed from China. True separation of algorithmic control would have required full divestiture, and China's own technology export regulations effectively blocked that (Turvy and Scharlach 2026).

The cross-border alliance dimension

TikTok engaged CFIUS, GCHQ's cyber security arm, the European Commission, the Australian Parliament and regulatory bodies in several other jurisdictions simultaneously. In each market it briefed think tanks, academics, civil society groups and media, offering more algorithmic transparency than any Western competitor had ever provided.

The transparency tension

TikTok offered more transparency than Meta, Google or Snap have ever provided. Governments still concluded it was untrustworthy. The reason lay not in what the transparency revealed but in what it could not structurally guarantee: freedom from Chinese state direction. Transparency proved insufficient once the real question became ownership, jurisdiction and legal obligation.

The civil society dimension

Civil society was vocal here but ultimately ineffective, and its position was structurally awkward from the start. The ACLU, EFF, PEN America, the Knight First Amendment Institute, the Centre for

Democracy and Technology and Fight for the Future all opposed the US ban on First Amendment grounds and filed amicus briefs with the Supreme Court (PEN America 2023; The Verge 2024). A separate coalition of racial justice and civil rights organisations, including the Asian American Federation and the Hispanic Heritage Foundation, filed against the ban on anti-discrimination grounds, arguing that it was driven in part by anti-Asian animus (ABC News 2024).

The awkwardness could not be avoided. These groups found themselves defending a Chinese-owned corporation against a national security law on free speech grounds, and that framing limited their traction even with audiences normally sympathetic to free speech arguments. Human Rights Watch and Amnesty International, organisations with long records of criticising the Chinese government, still lined up against the ban. Their involvement underlined that the civil liberties concerns were real, but it did not change the political outcome. The Supreme Court upheld the ban 9-0.

Civil society opposition left almost no mark on the legislative or judicial outcome. The case illustrates a specific failure mode: once national security framing takes hold, civil society's usual leverage points (rights arguments, parliamentary pressure, public mobilisation) lose traction quickly. The 170 million US TikTok users most affected by the ban were not meaningfully represented in any part of the corporate-government negotiation.

Discussion questions

Q1: TikTok offered more algorithmic transparency than Meta or Google ever has. Why was it still banned in the US, and what does this tell us about the limits of transparency as a governance tool?

Q2: The UK's NCSC found no evidence of wrongdoing after years of code review. Was the eventual ban evidence-based policy or geopolitical signalling?

Q3: Is it possible to design a governance framework capable of assessing algorithm trustworthiness across jurisdictional boundaries? What would it require?

Q4: ByteDance retains the algorithm's intellectual property even under the divestiture deal. Does this mean the transparency effort ultimately succeeded, failed, or neither?

References

ABC News (2024) 'Racial justice, civil rights groups join fight against potential TikTok ban'. Available at: https://abcnews.go.com/US/racial-justice-civil-rights-groups-join-fight-potential/story?id=111481670 (Accessed: 10 July 2026).

BBC News (2023) Coverage of TikTok's 'Project Clover' European data security initiative, March. Available at: https://www.bbc.com/news/technology-64887704 (Accessed: 10 July 2026).

PEN America (2023) 'PEN America, with other free speech organizations, sends letter to Congress opposing proposed national TikTok ban'. Available at: https://pen.org/press-release/pen-america-with-other-free-speech-organizations-sends-letter-to-congress-opposing-proposed-national-tiktok-ban/ (Accessed: 10 July 2026).

Perault, M. (2024) 'What happened to TikTok's Project Texas?', Lawfare, 20 March. Available at: https://www.lawfaremedia.org/article/what-happened-to-tiktok-s-project-texas (Accessed: 10 July 2026).

The Verge (2024) 'Civil society groups to SCOTUS: stop the TikTok ban to preserve free speech', 17 December. Available at: https://www.theverge.com/2024/12/17/24323683/civil-society-groups-to-scotus-stop-the-tiktok-ban-to-preserve-free-speech (Accessed: 10 July 2026).

Turvy, A. and Scharlach, R. (2026) 'US power play over TikTok did nothing to protect Americans', Tech Policy Press, 30 January. Available at: https://techpolicy.press/us-power-play-over-tiktok-did-nothing-to-protect-americans (Accessed: 10 July 2026).

Comparative Note

The three cases represent distinct types of corporate-government alliance around algorithm transparency, and it helps to name the differences plainly.

- Case 1 (OpenAI): lobbying to weaken transparency obligations while publicly championing them. Transparency serves as a rhetorical asset; the regulatory burden is treated privately as a threat.
- Case 2 (Palantir): transparency as a sales pitch to governments, opacity as the operational reality. Accountability is displaced onto government clients who cannot explain the systems they have procured.
- Case 3 (TikTok): transparency as a good-faith structural offer that proves technically insufficient. The case shows the limits of transparency when the real question is jurisdictional control.

In all three cases, “algorithm transparency” is being contested or negotiated for very different purposes and under very different power dynamics. None of the three ends with genuine public accountability for how the algorithms actually work.

The civil society record across the cases follows a consistent pattern. Activism is most visible and energetic where the failure is most concrete, as with Palantir: a named company, identifiable contracts, tangible harms. It is active but channelled into policy processes where the failure is procedural and hidden from view, as with OpenAI and the EU AI Act. And it is structurally marginalised where national security framing displaces rights-based argument, as with TikTok. In none of the three cases were the affected communities themselves (patients, users, surveilled populations) meaningfully represented in the corporate-government alliances that determined the outcomes.

Navigating the AI Policy Crisis

Five Case Briefs: Existential Threat, Accelerating Urgency and Profound Uncertainty

The cases in Parts 1 and 2 examined who captures AI governance and how. This part moves to the strategic level, where the crisis in AI policy meets its three defining criteria: existential threat, accelerating urgency, and profound uncertainty amid fierce military and technological competition between states. The five cases that follow show the same dynamics documented earlier – competition beating cooperation, voluntary arrangements substituting for binding ones, affected populations absent from the room – repeating where the stakes are highest. None of these cases has an ending yet. That is part of what makes them worth studying.

CASE 7. AI and Nuclear Weapons: The Biden–Xi Understanding and Its Orphaning

The leaders of the two AI superpowers agree that humans, not machines, must control nuclear weapons. The agreement has no verification, no follow-up mechanism, and no successor.

Overview

At their November 2024 meeting in Lima, Presidents Biden and Xi affirmed the need to maintain human control over the decision to use nuclear weapons, and stressed the need for care in developing military applications of AI (Reuters 2024). It was the first statement of its kind between the two strategic competitors, and arguably the first genuine act of AI arms control anywhere. It was also, in formal terms, almost nothing: a sentence in a meeting readout, unverifiable, non-binding, and attached to no mechanism that would survive a change of administration. The military-to-military dialogue on AI that might have given it substance stalled as strategic competition intensified through 2025 and 2026.

Key facts

- The Lima statement addressed the most consequential AI risk that exists: the entanglement of machine-speed decision systems with nuclear command, control and communications, where automation bias, spoofed sensor data or compressed decision timelines could contribute to catastrophic error.
- Neither side defined what “human control” means in practice. AI already pervades early-warning, intelligence fusion and decision-support functions that shape the information a human decision-maker sees before any launch decision is reached.
- The wider diplomatic architecture remains declaratory. The US Political Declaration on Responsible Military Use of AI and Autonomy (2023) has gathered more than 50 state endorsements but creates no obligations (US Department of State 2023). The REAIM summit process (The Hague 2023, Seoul 2024) produced a Blueprint for Action that China declined to support.
- The follow-on US–China talks on AI risk, begun in Geneva in 2024, did not mature into a standing channel. No second leaders’ statement has followed.
- Both militaries meanwhile accelerated AI adoption, including in intelligence processing, targeting support and autonomous platforms, and both treat the other’s progress as justification for their own pace.

The cross-border alliance dimension

This case inverts the pattern of Parts 1 and 2: here the cross-border relationship that matters is between adversaries, not allies, and the mobilisation problem is how rivals build governance while competing. The Lima understanding suggests both governments recognise a shared existential interest. Everything since suggests neither will accept constraints that might advantage the other, and each accelerates partly because it assumes the other will. Alliance structures amplify the dynamic: allies of each power adopt its military AI posture, its declaratory frameworks and increasingly its systems.

The control tension

Both powers agree the risk is existential, and both are automating the systems that surround the human decision anyway. The formal commitment – a human finger on the button – may become progressively less meaningful as the button's context (warning, attribution, option generation, timing) is shaped by machines. The question the case poses is whether "human control" is a safeguard or a slogan, and what verification of it could even look like between adversaries.

The civil society dimension

Civil society and the scientific community have a long tradition here to draw on, from Pugwash to the nuclear winter studies, and organisations including the Bulletin of the Atomic Scientists, SIPRI and university programs have produced substantial analysis of AI–nuclear entanglement. But access is the problem. Nuclear command arrangements are the most closed policy domain that exists; there is no equivalent of the DSA investigation, the freedom-of-information request or the parliamentary committee. Civil society can frame the risk publicly, and did help create the climate in which the Lima statement was possible, but it cannot inspect, verify or litigate. The populations whose survival is at stake – everyone's – are represented in this system by exactly two individuals.

Discussion questions

Q1: Is a non-binding leaders' statement without verification better than nothing, or does it create false assurance? What would a verifiable version require?

Q2: "Human control" of nuclear use may be preserved formally while everything informing the human is automated. Where should the line actually be drawn?

Q3: Can adversaries build AI arms control while each believes falling behind is the greater risk? What historical precedents apply, and where do they break down?

Q4: What role, if any, can middle powers and civil society play in a domain this closed?

References

Reuters (2024) 'Biden, Xi agree that humans, not AI, should control nuclear arms', 16 November. Available at: https://www.reuters.com/world/biden-xi-agree-humans-not-ai-should-control-nuclear-arms-2024-11-16/ (Accessed: 12 July 2026).

US Department of State (2023) Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy. Available at: https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/ (Accessed: 12 July 2026).

Arms Control Association (2024) 'Guterres calls for autonomous weapons controls', Arms Control Today, October. Available at: https://www.armscontrol.org/act/2024-10/news/guterres-calls-autonomous-weapons-controls (Accessed: 12 July 2026).

CASE 8. Autonomous Weapons: Battlefield Acceleration Versus Diplomatic Deadlock

AI-enabled weapons scale monthly on real battlefields while the negotiation meant to govern them has never formally begun. The UN Secretary-General's 2026 treaty deadline expires as this workshop meets.

Overview

In October 2023 the UN Secretary-General and the President of the International Committee of the Red Cross jointly called on states to conclude a legally binding instrument on autonomous weapons by 2026 (UN News 2025). That deadline has now arrived. No treaty exists, and the formal negotiations that would produce one have never opened. In the same three years, autonomous and AI-assisted weapons moved from experiment to routine: terminal-guidance FPV drones and AI-supported strike planning in Ukraine, and investigative reporting on AI-generated target lists in Gaza (Abraham 2024). The gap between operational acceleration and diplomatic paralysis is the sharpest available measure of the crisis in AI policy.

Key facts

- The UN Convention on Certain Conventional Weapons has discussed lethal autonomous weapons since 2014 under consensus rules, which allow any single state to block a negotiating mandate. Russia and the United States, among others, have consistently done so.
- The UN General Assembly passed its first resolution on autonomous weapons in December 2023 and Resolution 79/62 in December 2024 with 166 votes in favour — large majorities, but resolutions are not treaties (Article 36 2025).
- By early 2026, roughly 127 countries had expressed support for a binding instrument, led by African and Latin American states. A smaller coalition including the United States, Russia, the United Kingdom, India, Israel and Australia maintains that existing international humanitarian law suffices (TechTimes 2026).
- In September 2025, a large group of states declared the GGE's rolling text "a sufficient basis for negotiations", and more than 40 states including France and Germany agreed the draft rules were ready to negotiate. The consensus rule held anyway (Stop Killer Robots 2025; Article 36 2025).
- Investigative reporting in 2024 described an Israeli AI system, Lavender, that generated tens of thousands of human targets in Gaza with brief human review, illustrating that the human-judgement questions at the centre of the treaty debate are operational realities, not hypotheticals (Abraham 2024).
- The CCW's Seventh Review Conference in 2026 is the next and perhaps last realistic opportunity for a negotiating mandate inside the consensus system; failing that, treaty advocates will likely move to a standalone process outside it, as happened with landmines and cluster munitions.

The cross-border alliance dimension

Two coalitions face each other across this case. A majority-world alliance of states, UN organs, the ICRC and a large NGO network seeks a binding instrument. A smaller alliance of the major military AI powers – strategic rivals cooperating tacitly with one another – uses procedure to prevent negotiation while their own capabilities mature. It is the clearest example in this collection of competition organising governance: the states blocking the treaty are precisely those racing each other.

The urgency tension

Every year of deadlock changes what a treaty would have to govern. In 2014 autonomous weapons were prospective; by 2026 they are proven, produced at scale, exported and combat-tested. Diplomatic time and technological time are running at different speeds, and the disparity compounds: the more embedded the weapons become, the higher the cost of prohibition and the lower the probability that their possessors will accept one. This is the Collingridge dilemma with munitions attached.

The civil society dimension

Civil society here is organised, global and experienced: the Stop Killer Robots campaign spans more than 250 organisations, and the ICRC provides institutional and legal weight that most campaigns lack. The movement has won the argument in numbers – two-thirds of UN member states – and lost it in mechanism, because the forum where numbers count cannot negotiate and the forum that can negotiate is blocked. Its next strategic choice, whether to abandon the CCW for an Ottawa-style standalone treaty that the major powers would simply not join, is a live question this workshop can debate. Meanwhile the populations most affected – those under the drones in Ukraine, Gaza and beyond – appear in the process as case studies rather than participants.

Discussion questions

Q1: The deadline expired with no negotiation ever opened. Was setting a deadline without a pathway a mobilising device or a strategic error?

Q2: Is a treaty without the major military powers worth having? What did the landmine and cluster munitions precedents achieve, and what didn't they?

Q3: Reporting from Gaza suggests the critical variable is not weapon autonomy but the compression of human judgement. Should governance target the machine or the decision process around it?

Q4: If you advised Australia – a member of the blocking coalition – what position would you recommend for the 2026 Review Conference, and why?

References

Abraham, Y. (2024) '"Lavender": the AI machine directing Israel's bombing spree in Gaza', +972 Magazine, 3 April. Available at: https://www.972mag.com/lavender-ai-israeli-army-gaza/ (Accessed: 12 July 2026).

Article 36 (2025) 'UK stays mute as France, Germany and 40 more states agree that draft rules on "killer robots" are ready to negotiate'. Available at: https://article36.org/updates/uk-stays-mute-as-france-germany-and-40-more-states-agree-that-draft-rules-on-killer-robots-are-ready-to-negotiate/ (Accessed: 12 July 2026).

Stop Killer Robots (2025) 'September 2025 GGE joint statement'. Available at: https://www.stopkillerrobots.org/news/september-2025-gge-joint-statement/ (Accessed: 12 July 2026).

TechTimes (2026) 'Killer robots ban deadline expires as Pentagon punished the only company that said no', 6 July. Available at: https://www.techtimes.com/articles/319795/20260706/killer-robots-ban-deadline-expires-pentagon-punished-only-company-that-said-no.htm (Accessed: 12 July 2026).

UN News (2025) '"Politically unacceptable, morally repugnant": UN chief calls for global ban on killer robots', May. Available at: https://news.un.org/en/story/2025/05/1163256 (Accessed: 12 July 2026).

CASE 9. From Bletchley to Paris: The Rise and Stall of the Frontier Safety Summit Process

For one moment in 2023, the AI superpowers signed the same page about frontier risks. Within eighteen months, safety had been rebranded as opportunity and the only global governance channel for frontier AI was buckling.

Overview

The Bletchley Declaration of November 2023 was a genuine anomaly: 28 countries and the EU, including both the United States and China, jointly acknowledged the potential for "serious, even catastrophic, harm" from frontier AI (UK Government 2023). It launched a summit series, national AI safety institutes, voluntary frontier safety commitments from sixteen companies at Seoul in May 2024, and an IPCC-style International AI Safety Report chaired by Yoshua Bengio, synthesising the views of some hundred experts from thirty countries (UK Government 2025). Then came Paris. At the February 2025 "AI Action Summit", safety was demoted from the agenda's centre, the US vice president warned Europe against "excessive regulation", and the United States and United Kingdom declined to sign the summit declaration. The process survives in name; whether it survives in function is the question.

Key facts

- The Bletchley process produced real institutions: the UK's AI Safety Institute (later AI Security Institute), the US equivalent (later CAISI), counterparts in Japan, Singapore and elsewhere, and a nascent international network of evaluators.
- The International AI Safety Report, published ahead of Paris in January 2025, formalised profound uncertainty: expert opinion ranges from manageable-risk to loss-of-control scenarios, with no scientific consensus on where the boundary lies (UK Government 2025).
- At Paris, the framing shifted from safety to competitiveness and investment. The US and UK refused the declaration; the "safety summit" became the "action summit", and its successor in India in early 2026 was branded an "impact summit".
- Capability meanwhile crossed thresholds the summits were created to anticipate. In April 2026 Anthropic limited the release of a model able to identify and exploit software vulnerabilities autonomously – a private company making, alone, the kind of judgement the international process was meant to institutionalise (NPR 2026).
- Government testing continues, but voluntarily and nationally: the US CAISI had completed more than 40 model evaluations by May 2026 under revocable agreements with the major laboratories, jointly with the UK institute in some cases (Nextgov/FCW 2026).

The cross-border alliance dimension

Bletchley briefly assembled the broadest alliance in this entire collection – rivals, allies and companies around one text. Its decay traces exactly how competition dissolves cooperation: each government concluded that restraint was a gift to its competitors, and the companies, courted by all sides for infrastructure and investment, found the opportunity framing more congenial than the safety one. What remains is bilateral and voluntary: evaluator-to-evaluator cooperation among allies, company-by-company agreements, and no channel at all that includes China.

The cooperation tension

The summit process rested on a premise the competition eventually falsified: that shared fear of catastrophic risk could motivate collective restraint among rivals. The fear proved real but subordinate. Every participant preferred a world with frontier governance to a world without it, and preferred falling behind in neither – and when forced to choose, chose the race. The unresolved question is whether the institutional residue (the safety institutes, the Bengio report, the evaluation agreements) is a foundation that survives the freeze, or debris.

The civil society dimension

The scientific community, rather than street-level civil society, is the key non-state actor in this case, and its instrument is the assessment: the Bengio report deliberately imported the IPCC

model of institutionalised scientific consensus into AI. The model's weakness imported with it: an assessment can define the problem but cannot compel a response, and governments facing competitive pressure treated the report as they often treat climate science – acknowledged, cited and set aside. Publics appear in this case chiefly as poll numbers; no mechanism connects popular concern about frontier AI to the summit process at all.

Discussion questions

Q1: What actually killed the safety framing between November 2023 and February 2025 – events, elections, lobbying, or something structural?

Q2: Are the safety institutes and evaluation agreements a durable foundation or hostages to the next political reversal? What would make them durable?

Q3: An IPCC for AI: does institutionalised scientific consensus work for a technology moving this fast, when the IPCC's subject moves over decades?

Q4: If the summit process cannot include China, what is it for? Design the minimum viable US–China frontier-risk channel you think both would accept.

References

NPR (2026) 'Trump's new AI safety order seeks voluntary review of new models', 2 June. Available at: https://www.npr.org/2026/06/02/nx-s1-5844347/ai-safety-trump-executive-order (Accessed: 12 July 2026).

Nextgov/FCW (2026) 'Commerce AI center will evaluate Google DeepMind, Microsoft and xAI models', May. Available at: https://www.nextgov.com/artificial-intelligence/2026/05/commerce-ai-center-will-evaluate-google-deepmind-microsoft-and-xai-models/413349/ (Accessed: 12 July 2026).

UK Government (2023) The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November. Available at: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration (Accessed: 12 July 2026).

UK Government (2025) International AI Safety Report 2025. Available at: https://www.gov.uk/government/publications/international-ai-safety-report-2025 (Accessed: 12 July 2026).

CASE 10. The DeepSeek Shock: Do Compute Controls Work At All?

A Chinese laboratory releases a near-frontier model trained at a fraction of the assumed cost, erases nearly $600 billion of market value in a day, and throws the West's central AI containment strategy into doubt.

Overview

On 20 January 2025, the Chinese company DeepSeek released R1, a reasoning model broadly comparable to the leading US systems, trained under US export controls that were designed to make exactly this impossible (DeepSeek-AI 2025). Within a week its app topped the US App Store and Nvidia lost close to $600 billion in market capitalisation in a single day, the largest one-day loss in US market history (CNBC 2025). The episode was instantly labelled a "Sputnik moment", but its real significance was epistemic: the entire Western strategy for governing the AI race rested on the assumption that compute access determines capability, and a single model release made that assumption uncertain.

Key facts

- US export controls since October 2022 had sought to deny China the advanced chips needed for frontier training. DeepSeek trained on legally acquired but deliberately hobbled hardware, plus algorithmic efficiency gains, and published its methods openly.
- The claimed training cost – around $6 million for the predecessor V3 base model – was contested and partial, but even sceptical estimates implied an order-of-magnitude efficiency surprise.
- The market reaction encoded the uncertainty: if capability no longer scales tightly with compute, both the chip-demand thesis and the export-control strategy weaken together (CNBC 2025).
- Policy responses pointed in opposite directions at once. One camp argued the shock proved controls were leaking and must tighten – the view argued by Anthropic's CEO days after the release (Amodei 2025). Another concluded controls merely spur Chinese efficiency innovation and commercial self-harm; within months Washington rescinded the Biden-era diffusion rule and later permitted downgraded Nvidia chip sales to China on commercial terms.
- R1's open weights spread worldwide within days, including into jurisdictions no export control reaches, and its efficiency techniques were rapidly absorbed by every major laboratory, accelerating the field it was meant to be excluded from (MIT Technology Review 2026).

The cross-border alliance dimension

Export controls are an alliance technology: they work only if the US, Netherlands (ASML), Japan, Korea and Taiwan hold the line together, and if the assumption behind them holds. DeepSeek attacked the assumption rather than the line. The result destabilised the alliance's internal bargain – each member bears commercial costs proportional to its belief in the strategy – and handed Beijing a demonstration that denial strategies breed workarounds. The subsequent US pivot to selling downgraded chips split the difference incoherently: containment as policy, diffusion as practice.

The uncertainty tension

Every actor in the AI competition must now act under radical uncertainty about the technology's core production function. If compute is destiny, controls are rational and the race is capital-intensive and legible. If algorithmic efficiency can substitute for compute, controls are theatre, open publication moves the frontier, and no state can reliably know its position in the race it believes it is running. Policy since January 2025 has oscillated between these worldviews because the evidence genuinely supports neither conclusively. This is profound uncertainty in its purest policy form: not uncertainty about values or intentions, but about the basic physics of the competition.

The civil society dimension

Civil society barely features in export-control policy, which is made by security agencies and a handful of firms. The nearest analogue is the open-source AI community, which functions as a

transnational actor with its own interests: R1's open weights were downloaded and built upon by researchers everywhere, making the diffusion irreversible regardless of what any government preferred. If Parts 1 and 2 showed corporations as the non-state governors of AI, this case adds a stranger one: a distributed global community that no jurisdiction contains, treating models as public goods while states treat them as strategic assets.

Discussion questions

Q1: After DeepSeek, what is the strongest argument that compute controls still work – and the strongest that they never did? Which do you find more convincing, and on what evidence?

Q2: Was publishing R1's weights and methods a scientific contribution, a commercial strategy, a geopolitical act, or all three? Does the answer change how states should respond?

Q3: How should a strategy cope with not knowing the production function of the technology it is trying to control? Are there precedents (cryptography, biotech) worth borrowing from?

Q4: Australia is a consumer, not a producer, of frontier AI. Does the DeepSeek shock strengthen or weaken the case for alignment with US technology controls?

References

Amodei, D. (2025) 'On DeepSeek and export controls', January. Available at: https://darioamodei.com/on-deepseek-and-export-controls (Accessed: 12 July 2026).

CNBC (2025) 'Nvidia sheds almost $600 billion in market cap, biggest one-day loss in US history', 27 January. Available at: https://www.cnbc.com/2025/01/27/nvidia-sheds-almost-600-billion-in-market-cap-biggest-drop-ever.html (Accessed: 12 July 2026).

DeepSeek-AI (2025) 'DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning', arXiv:2501.12948. Available at: https://arxiv.org/abs/2501.12948 (Accessed: 12 July 2026).

MIT Technology Review (2026) 'Three reasons why DeepSeek's new model matters', 24 April. Available at: https://www.technologyreview.com/2026/04/24/1136422/why-deepseeks-v4-matters/ (Accessed: 12 July 2026).

CASE 11. The Superintelligence Statement: Existential Alarm Meets the Race It Cannot Stop

Nobel laureates, AI pioneers, faith leaders, royalty and political operatives from left and right sign a call to prohibit superintelligence development. The laboratories racing toward it do not pause for a day.

Overview

On 22 October 2025, the Future of Life Institute published the Statement on Superintelligence: a call for "a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in" (Future of Life Institute 2025). The signatory list was engineered for breadth: AI pioneers Geoffrey Hinton and Yoshua Bengio, five Nobel laureates, Steve Wozniak and Richard Branson, former US national security principals including Susan Rice and Admiral Mike Mullen, Mary Robinson, Stephen Fry – and, in the same column, Steve Bannon and Meghan Markle (CNBC 2025; Forbes 2025). Accompanying polling reported that 64 per cent of Americans opposed developing superintelligence until it is provably safe and controllable, with only 5 per cent favouring maximum speed (Future of Life Institute 2025). No laboratory altered course, and no government proposed implementing the prohibition.

Key facts

- The statement deliberately echoed earlier interventions – the March 2023 pause letter and the May 2023 one-sentence extinction-risk statement – but escalated from pause to conditional prohibition, and from expert signatories to a coalition spanning the political spectrum from Bannon to Rice.

- Its two conditions, scientific consensus on safety and public buy-in, are precisely the two things the competitive race structurally prevents: no laboratory can wait for consensus its rivals ignore, and no mechanism exists for publics to grant or withhold consent.
- The polling gap is the case's core datum: a large majority of the public in the leading AI state opposes what its leading companies explicitly aim to build, and this preference has no institutional pathway into policy (Future of Life Institute 2025).
- The frontier laboratories' stated missions – OpenAI's pursuit of AGI, Meta's "superintelligence" lab, comparable ambitions elsewhere – were unchanged; several executives dismissed the statement within days.
- Within six months, events gave the statement's premise partial vindication from an unexpected quarter: a frontier laboratory itself restricted a model over autonomous cyber-exploitation capability, and Washington moved to voluntary pre-release testing – acknowledgements that dangerous capability is near, without any acceptance of prohibition (NPR 2026).

The cross-border alliance dimension

The statement is cross-border civil society mobilisation in its widest possible form: a deliberately transnational, trans-ideological coalition addressed to no single government, because no single government can grant what it asks. That is its strength as expression and its weakness as politics. A prohibition on superintelligence is a coordination problem among rivals – the same problem Cases 7, 8 and 9 show going unsolved for weapons far better understood – and the statement names the destination while specifying no route through the competition that stands in the way.

The consent tension

This case poses the collection's deepest question: by what right is the most consequential technology in human history developed without the consent of those it will affect? The statement's answer – require public buy-in – collides with the reality that the development is private, the competition is international, and no institution aggregates global public consent for anything. The 64 per cent has no vote. The case invites comparison with earlier technologies (nuclear power, GMOs, human germline editing) where publics did, eventually, constrain what science and industry could do – and asks why AI so far has escaped every such mechanism.

The civil society dimension

Here civil society is not a dimension of the case; it is the case. The statement represents the maximal version of the strategy: assemble unignorable breadth, quantify public opinion, and state the demand plainly. The result clarifies the limits of moral authority unattached to leverage. Unlike the Palantir campaigners of Case 5, who had a contract renewal date to target, or the chat control coalition of Case 1, which had legislators to lobby, the superintelligence statement has no pressure point: no procurement decision, no election, no regulator with jurisdiction over the thing it seeks to prohibit. Its plausible long-term significance is as a marker – establishing, before the fact, that whatever happens next happened against documented public objection.

Discussion questions

Q1: Bannon and Rice, Hinton and Markle: does ideological breadth make a statement more powerful or easier to dismiss? What was the theory of change?

Q2: The statement demands scientific consensus and public buy-in before superintelligence development. Are these coherent, achievable conditions – and who would certify them?

Q3: If 64 per cent of the public opposes a private undertaking of civilisational consequence and this has no policy effect, what does that say about where AI governance authority actually resides?

Q4: Compare this statement with the Guterres–ICRC autonomous weapons appeal (Case 8). Both name deadlines or conditions without mechanisms. Is that failed politics, or how norms begin?

References

CNBC (2025) 'Hundreds of public figures, including Apple co-founder Steve Wozniak and Virgin's Richard Branson, urge AI "superintelligence" ban', 22 October. Available at:

https://www.cnbc.com/2025/10/22/800-petition-signatures-apple-steve-wozniak-and-virgin-richard-branson-superintelligence-race.html (Accessed: 12 July 2026).

Forbes (2025) ‘Harry and Meghan, Steve Bannon and more sign petition urging AI “superintelligence” ban’, 22 October. Available at: https://www.forbes.com/sites/siladityaray/2025/10/22/public-figures-sign-petition-urging-ban-on-ai-superintelligence-including-harry-meghan-steve-bannon-and-richard-branson/ (Accessed: 12 July 2026).

Future of Life Institute (2025) ‘Prominent scientists, faith leaders, policymakers and artists call for a prohibition on superintelligence, as poll shows Americans don’t want it’, press release, 22 October. Available at: https://futureoflife.org/press-release/prominent-scientists-faith-leaders-policymakers-and-artists-call-for-a-prohibition-on-superintelligence/ (Accessed: 12 July 2026).

NPR (2026) ‘Trump’s new AI safety order seeks voluntary review of new models’, 2 June. Available at: https://www.npr.org/2026/06/02/nx-s1-5844347/ai-safety-trump-executive-order (Accessed: 12 July 2026).