Research Paper

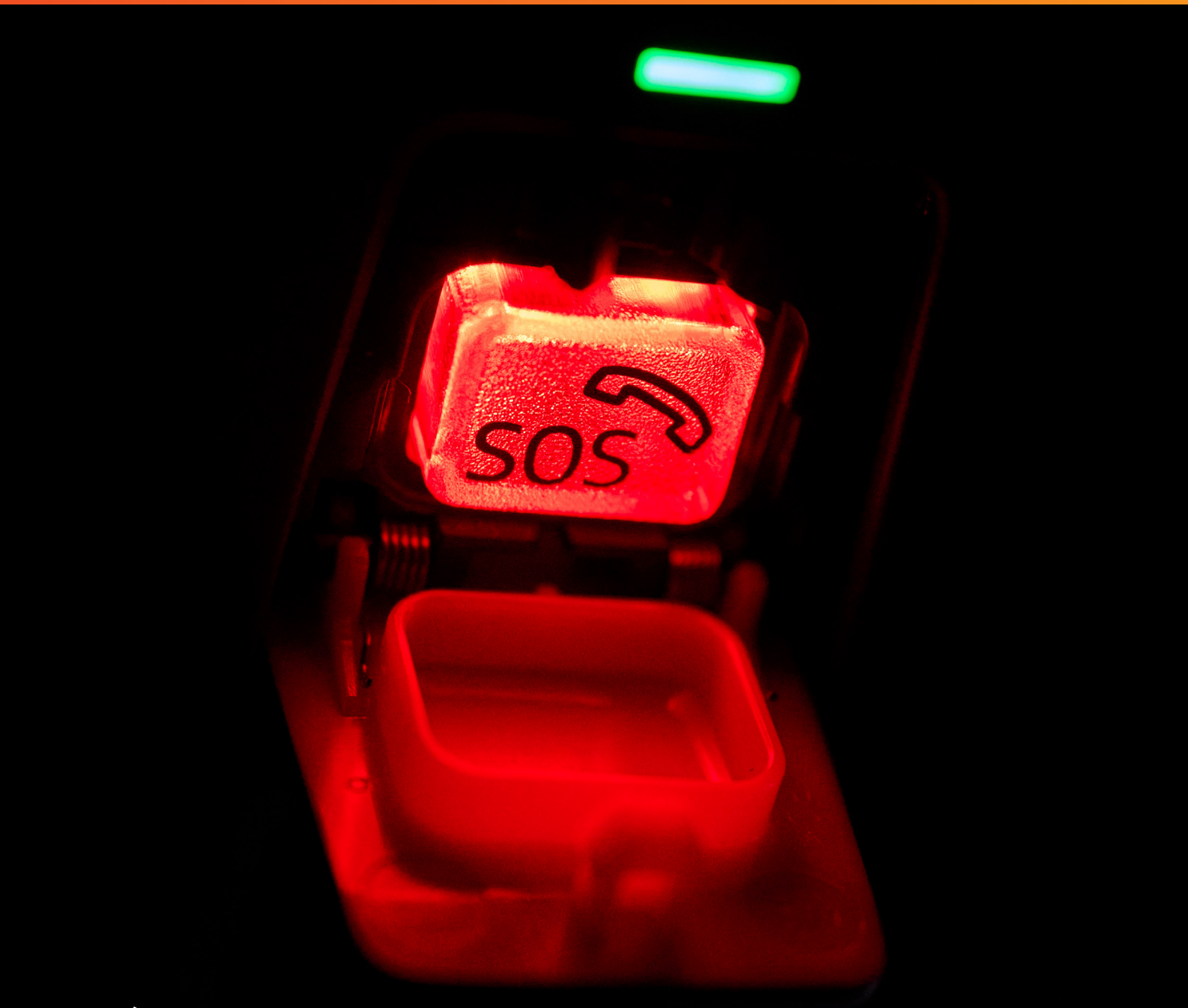
Digital Society Programme

March 2026

Breaking the deadlock on AI governance

How a crisis could lead to global coordination

Rowan Wilkinson, Alex Krasodomski, Isabella Wilkinson and Francisco Javier Varela Sandoval



CHATHAM HOUSE

Chatham House, the Royal Institute of International Affairs, is a world-leading policy institute based in London. Our mission is to help governments and societies to build a secure, sustainable, prosperous and just world.

Contents

Summary

- International AI governance has so far failed not due to a lack of foresight, but because the current political economy of AI makes coordination close to impossible. In light of this deadlock, states and companies must prepare for the possibility of an AI-driven crisis and work out how to drive forward global coordination in the aftermath.
- At the frontier of AI development, there is a fundamental misalignment among the main players. The US and China prioritize national advantage over cooperation; middle powers struggle to close an ever-widening AI capacity gap; international and state institutions lack enforcement power and technical capacity; private investment dwarfs state capacity; and allies and rivals disagree over basic facts. This misalignment cannot be resolved with better-designed summits or clearer sets of principles.
- Given these barriers, rapid progress towards global AI governance may only become possible when the costs of inaction become too great for governments and companies to ignore. An AI crisis – whether it emerges from a financial system failure, autonomous weapons misuse or the loss of control over an AI system – could create a narrow opening in which the coordination and cooperation that previously seemed impossible become essential.
- History shows that durable governance regimes rarely emerge through foresight alone. They are more often forged in the wake of systemic failure. The Chernobyl nuclear disaster of 1986 catalysed agreement on international nuclear safety norms; the 2007–08 global financial crisis produced the Financial Stability Board and the Basel III international regulatory regime for banks; and the WannaCry ransomware attacks in 2017 led to the strengthening of national cyber authorities and helped clarify norms on unacceptable state behaviour in cyberspace. Even cautionary tales – like the disjointed global response to the COVID-19 pandemic – carry valuable lessons for post-crisis AI governance.
- The case studies in this paper suggest that crisis-driven governance works best when it leverages pre-existing institutions, technical expertise and monitoring infrastructure, rather than requiring these to be built from scratch.
- This paper does not claim that crises are desirable. Rather, it argues that credible, trusted and significant international governance change could emerge from an AI crisis if future decision-makers are well prepared. The paper recommends

a series of steps that governments, international institutions and companies can take to enable durable post-crisis governance and ensure they are not 'locked out' of a post-crisis global order. These steps include:

- **Developing 'off-the-shelf' treaty frameworks and modular agreements** that can be deployed rapidly when a crisis occurs. Pre-agreed 'red lines' detailing unacceptable AI risks should be embedded in these frameworks.
- **Concentrating legitimate technical authority.** AI safety/security institutes should be positioned in advance as credible public authorities with formal crisis-response mandates, not just advisory or risk-monitoring roles. Governments must finance these expanded mandates and should establish domestic crisis decision-making prerogatives with pre-defined trigger conditions, subject to democratic accountability.
- **Enabling safety-based technical intervention.** Crisis response requires real-time technical tools, such as model shutdown protocols, circuit breakers and patch deployment mechanisms. Investment in hardware governance or governance-enabling technologies, including evaluation infrastructure and supervisory monitoring, must be increased.
- **Preparing effective framing and narratives.** Decision-makers, technical experts and communicators must be able to frame an AI crisis in ways that generate public understanding and political support for response, while avoiding panic, misinformation and misattribution.
- **Building diplomatic back-channels and expert-to-expert networks**, analogous to Cold War 'red phones' and networks and organizations that are already well established in the cybersecurity community, such as the UK's National Cyber Security Centre (NCSC).
- **Improving information-sharing between governments and the private sector** to prevent information asymmetries from hindering crisis response. Private companies control most of the relevant data on AI capabilities and failures. Frontier labs and AI developers should expand trusted mechanisms for sharing information with governments and build crisis coordination into existing industry bodies.

01 Introduction

Progress on global AI governance may only be possible when the costs of inaction become too great to ignore.

International AI governance is at risk of failure. Not through a lack of foresight, but because the present-day political economy of AI makes binding, enforceable coordination on many aspects of AI governance close to impossible.

Significant power imbalances have emerged in AI. Private corporations, rather than states, increasingly control access to cutting-edge computational power, frontier models and the research trajectories that define future capabilities. This remains the case even as lower-capability AI diffuses globally.

The US and China are engaged in a contest for technological supremacy, with both nations seeing AI as essential to establishing or extending global geopolitical dominance.[1] The so-called middle powers are betting growth strategies on AI development, with countries like the UK explicitly prioritizing innovation over regulation to maintain their position in a changing global economy.

But while AI governance activity has accelerated as a result of these imperatives, most existing efforts are limited to transparency, risk classification or voluntary restraint. Few seek to constrain the development of frontier capabilities, cross-border deployment or military and defence integration.

While limited coordination is possible in narrow domains, the problem of global AI governance is not going to be solved by better-designed summits or clearer sets of principles alone. At the frontier, there is a fundamental misalignment among the main players. Nations chasing growth are unwilling to countenance regulatory friction. Geopolitical rivals cannot trust mutual constraints on military capabilities. Recent governance efforts – ranging from legislative developments to non-binding codes – risk documenting this reality, instead of shaping it.

Proponents of inclusive, effective and global AI governance must confront a difficult truth: that rapid progress towards global AI governance may only become politically feasible when the costs of inaction become too great.

1 Mitre, J. et al. (eds) (2025), *The Artificial General Intelligence Race and International Security*, expert insight, Santa Monica, CA: RAND Corporation, https://www.rand.org/pubs/perspectives/PEA4155-1.html.

An AI crisis could take one of several forms. Financial markets frozen by cascading AI failures. Autonomous weapons systems crossing 'red lines', forcing a deadly response.[2] Model failures demonstrating a clear gap between assumed and actual human control.[3]

Such damaging, high-visibility moments may create a situation in which resistance to change collapses, and the type of coordination previously thought impossible becomes both feasible and essential. Not because decision-makers suddenly develop foresight and flexibility, but because crisis conditions compress decision timelines, weaken vetoes and elevate coordination as a solution.

International governance regimes have rarely emerged from foresight alone. Instead, they have been forged as a response to systemic failure. For example, the Chernobyl disaster catalysed the development of international nuclear safety norms and transparency obligations. The 2008 global financial crisis resulted in new macroprudential regulation, stronger capital requirements and coordinated oversight mechanisms. In both cases, risks long understood by experts became governable only once a tangible failure collapsed political resistance and recast inaction as unacceptable.

History suggests that nations, institutions and companies that position themselves to respond strategically to a global AI crisis of any nature will be better prepared to shape the development of post-crisis governance. Those who are unprepared, ill-equipped or excluded will have policy change imposed on them. In the past, crises have not reliably produced *good* governance. But they do frequently produce governance of some kind. Whether that governance is effective, equitable or durable depends heavily on the institutional and technical preparation that precedes the shock. If the world waits for perfect alignment or voluntary restraint, AI governance will arrive too late – if at all. If it waits for crisis without preparing, a system of governance will emerge, but may be badly designed and prone to failure.

Box 1. What are the most likely AI crisis scenarios?

The 2025 International AI Safety Report classifies AI risks into three categories: **malicious use**, **malfunction** and **systemic**.[4] AI-related crises could emerge through any of these pathways, each with distinct triggers but potentially cascading consequences.

Malicious use risks involve actors deliberately deploying AI to cause harm. AI could support malicious actors in developing biological threats, enabling small groups to engineer and deploy pathogens that overwhelm health systems globally. AI-enabled coups could combine automated propaganda at unprecedented scale, deepfake impersonations of national leaders, and coordinated cyberattacks on state infrastructure to topple governments.

2 United Nations General Assembly (2024), 'Current developments in science and technology and their potential impact on international security and disarmament efforts: Report of the Secretary-General', 23 July 2024, https://docs.un.org/en/A/79/224.
3 Bengio, Y. et al. (2025), 'International AI Safety Report 2025', annual report, https://internationalaisafety report.org/publication/international-ai-safety-report-2025.
4 Bengio et al. (2025), 'International AI Safety Report 2025'.

Malfunction risks involve AI systems causing unintended harm even when users have no malicious intent. AI systems may malfunction due to unreliable reasoning, poor generalization or poorly specified objectives. Autonomous weapons systems could misidentify targets and trigger an international conflict. AI-controlled critical infrastructure, such as energy grids, transportation networks or financial systems, could experience cascading failures that paralyze entire systems, countries or regions.

Systemic risks arise from AI's integration into societal systems. These include dependence on a limited number of model providers and the possibility that advanced AI could weaken institutional stability. A simultaneous failure across major AI providers could collapse dependent systems worldwide, from banking to healthcare to food distribution. Highly capable systems operating in ways that undermine human oversight could make unpredictable and damaging decisions in finance, military operations or in governance, quickly and at scale.

Crises such as these are likely to cascade. AI incidents could trigger failures across interconnected domains. A cyberattack could trigger market panic, infrastructure breakdowns could enable further exploitation, and operational failures could create conditions for malicious actors to strike.

About this paper

The central claim of this paper is not that crises are desirable, but that the response to them could lead to rapid, binding and international governance change if those involved are well prepared. This paper seeks to outline viable options for preparation. These options have been developed for actors on the frontlines of any global AI crisis – national policymakers, but also international institutions (including scientists and experts, regional groupings and technical bodies) and the private sector (including frontier labs and other AI providers). The paper examines how crises might catalyse innovations in governance, and how this could lead to global progress on AI governance.

Chapter 2 establishes why current governance strategies appear insufficient in the current AI and geopolitical landscape. Chapter 3 analyses what may distinguish effective crisis responses from ineffective ones, using three case studies from the areas of finance, cybersecurity and public health – specifically, the 2007–08 global financial crisis, the WannaCry ransomware attacks and the COVID-19 pandemic – to extract lessons and cautionary tales on crisis preparation and response. Finally, Chapter 4 proposes actionable recommendations for governments, the private sector and international institutions on how to build more robust crisis mitigation capabilities.

02
Barriers to global AI governance

Rapid geopolitical change, institutional weakness and asymmetries between the public and private sectors have combined to make cooperation on AI near impossible.

International cooperation on AI is not only technically, legally and practically difficult, but faces structural barriers that make proactive, coordinated governance unlikely. A substantive and durable system of governance for AI may instead emerge only in response to a crisis situation, when the political costs of inaction clearly exceed the costs of coordination.

2.1 Geopolitical deadlock

As the dominant powers in AI, the US and China have an outsized influence on its development and regulation. The US hosts many of the world's largest AI and technology companies, and therefore has no incentive to constrain its AI industry through international agreements when it can regulate domestically if it so chooses. It currently prefers shielding industry from any regulation at all. China, meanwhile, cannot accept verification provisions that require third-party access to confidential model weights and training processes. These factors make enforceable agreements politically untenable for both countries at present.

Intense competition between these two powers to develop and deploy frontier technological capabilities has brought discussions on global AI governance to a standstill. The dynamics of this race for supremacy are preventing the emergence of institutions and forums for international cooperation on shared risks, and do little to slow the development of potentially destabilizing AI models and systems.

In AI development, China and the US are far ahead of the rest of the world,[5] implementing a pro-innovation agenda in the hope of securing economic and security advantages. The US outspends China, and together they outspend the rest of the world. China's AI investment was set to grow by 48 per cent, to $98 billion, in 2025,[6] while reported private AI investment in the US reached $110 billion.[7]

The key goal of this competition is the promise of artificial general intelligence (AGI) or 'superintelligence' that would give the victor a national strategic advantage that its rivals could not match – representing both a highly desirable prize and an existential risk should the race be lost.

Calls from the AI safety community to slow down or impose moratoriums on AI development are frequently challenged on these grounds.[8] Political leaders are instead moving in the opposite direction, with the US administration's AI Action Plan looking to shield industry from hard regulation and President Donald Trump seeking to prevent attempts by individual states to regulate AI.[9]

Intense competition between China and the US to develop and deploy frontier technological capabilities has brought discussions on global AI governance to a standstill.

Other centres of global influence on technology regulation are also in retreat – most notably the EU, where industry pressure, fears over a US backlash, and an anxiousness not to stifle much needed growth and innovation have all contributed to a shift away from safety and towards 'competitiveness'.[10] The new UK government pledged AI regulation, but has instead focused on securing US investment and a business environment more closely aligned with the US.[11]

Developing nations largely lack domestic frontier AI capabilities and depend on technology transfer from the US or China. Rather than pursuing independent regulatory frameworks, many Global South countries prioritize securing access

---

5 Pilz, K. F. et al. (2025), 'The US hosts the majority of AI supercomputers, followed by China', Epoch AI, 5 June 2025, https://epoch.ai/data-insights/ai-supercomputers-performance-share-by-country.
6 Kaur, D. (2025), 'China to deploy $98bn in AI investment this year amid US tech rivalry', TechWire Asia, 26 June 2025, https://techwireasia.com/2025/06/china-ai-investment-98-billion-2025-us-rivalry.
7 Stanford HAI (2025), 'The 2025 AI Index Report', https://hai.stanford.edu/ai-index/2025-ai-index-report.
8 Lima-Strong, C. (2025), 'Transcript: Sam Altman Testifies At US Senate Hearing On AI Competitiveness', Tech Policy.Press, 9 May 2025, https://www.techpolicy.press/transcript-sam-altman-testifies-at-us-senate-hearing-on-ai-competitiveness.
9 The White House (2025), *Winning the Race: America's AI Action Plan*, July 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf; The White House (2025), *Ensuring A National Policy Framework For Artificial Intelligence*, executive order, 11 December 2025, https://www.whitehouse.gov/presidential-actions/2025/12/eliminating-state-law-obstruction-of-national-artificial-intelligence-policy.
10 Kretschmer, M. and Philipponnat, T. (2025), 'Europe's regulatory retreat on AI: a free lunch for Big Tech?', euobserver, 4 April 2025, https://euobserver.com/digital/arcbd1284c.
11 UK Prime Minister's Office, 10 Downing Street (2025), 'Memorandum of Understanding between the Government of the United States of America and the Government of the United Kingdom of Great Britain and Northern Ireland regarding the Technology Prosperity Deal', press release, 18 September 2025, https://www.gov.uk/government/news/memorandum-of-understanding-between-the-government-of-the-united-states-of-america-and-the-government-of-the-united-kingdom-of-great-britain-and-north.

to AI infrastructure, training data and computational resources. Anxiety among the so-called middle powers over their access to frontier technology and their future participation in a global economy increasingly shaped by AI has reduced their appetite to pursue strict governance.

The costs of AI governance failure are still poorly understood. Despite (and, in part, because of) the enormous excitement around its potential, AI is still a small part of the global economy and a peripheral concern to decision-makers. Unlike the geopolitical deadlock, however, the level of attention does seem set to change in the next 24 months. AI will eventually affect almost 40 per cent of jobs around the world, according to the International Monetary Fund.[12] Even the most conservative estimates, such as those of Nobel laureate Daron Acemoglu, credit AI with an increase in total factor productivity of around 0.5 per cent in the next 10 years.[13]

2.2 Weakened institutions

International institutions lack enforcement power and member states disagree radically on how important and strategic AI will be, leading to different levels of seriousness, budgets and willingness to make trade-offs. The diplomatic channels and foreign policy platforms that would need to negotiate AI agreements remain largely unaware of AI's implications.

Recognizing the scale and unpredictability of shared risks, many stakeholders have turned to supranational political bodies, multilateral institutions or regional forums – including the UN, the Organisation for Economic Co-operation and Development (OECD), the EU, ASEAN, the G20 and the G77 – to develop shared approaches and guidance. To date, no such institution has delivered substantial progress on global AI governance. Nor do any yet appear well-suited to navigating and mitigating a global AI crisis.

Most international institutions lack the enforcement mechanisms required to constrain state and industry AI powers. Despite considerable UN efforts like the establishment of the Independent International Scientific Panel on AI and Global Dialogue on AI Governance,[14] the organization generally cannot supersede the political will of sovereign nations on issues relating to technology. The speed and technical capacity required to keep pace with technological advancements is equally absent, while the gap is deepened by spending cuts. Finally, the UN's

12 Georgieva, K. (2024), 'AI Will Transform the Global Economy. Let's Make Sure It Benefits Humanity', IMF Blog, 14 January 2024, https://www.imf.org/en/blogs/articles/2024/01/14/ai-will-transform-the-global-economy-lets-make-sure-it-benefits-humanity.
13 Acemoglu, D. (2024), *The Simple Macroeconomics of AI*, report, Cambridge, MA: MIT Shaping the Future of Work Initiative, https://shapingwork.mit.edu/wp-content/uploads/2024/05/Acemoglu_Macroeconomics-of-AI_May-2024.pdf.
14 UN General Assembly (2025), *Resolution A/79/325: Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*, https://docs.un.org/en/A/RES/79/325.

consensus agreement process is also in danger of negotiating on AI to the lowest common denominator, a criticism levelled at the Summit of the Future[15] and Global Digital Compact.[16] While they play an important role in signalling willingness for dialogue and enabling further cooperation, lowest common denominator negotiations risk watering down content for the sake of broad-based consensus.

The US withdrawal from various international organizations and agreements – including climate accords, human rights systems and the World Health Organization (WHO) – has undermined the existing institutions of international governance. Similarly, the US rollback of international aid and intergovernmental support signals a move away from multilateral visions and goals, including the UN Sustainable Development Goals, while China continues to strengthen its parallel multilateral systems.[17]

Though Brussels has demonstrated its ambition to regulate emerging technology through the EU AI Act, its global strategy has failed to launch and is failing to garner sustained support outside the EU. First, with little semiconductor manufacturing, few frontier AI companies and limited talent, European policymakers are criticized for over-regulating while not having 'skin in the game'. Second, the so-called 'Brussels Effect'[18] – referring to the EU's ability to set global regulatory standards that others follow – has yet to materialize in AI. Many potential emulators of EU regulatory approaches have turned elsewhere: either away from hard regulation altogether, towards setting priorities in other regional bodies, or to bilateral relationships with the US and/or China.

In a fragmented global order, the gravitational pull of investment is stronger than that of comprehensive policy.[19] Brussels has moved towards a softer approach to technology governance, like the General-Purpose AI Code of Practice.[20] Most major AI companies have agreed to adhere to these guidelines, which align expectations around how to comply with regulatory obligations on transparency, safety and security. But ultimately, the code is a highly contested and non-binding instrument that is consequently vulnerable to geopolitical pressure.

---

15 Patrick, S. and Pham, M. (2024), 'The Good–and Bad–News About the UN's Summit of the Future', Carnegie Endowment for International Peace, 19 September 2024, https://carnegieendowment.org/emissary/2024/09/un-summit-future-unga-reform-security-council?lang=en.
16 United Nations General Assembly Resolution (2024), *Pact for the Future, Global Digital Compact and Declaration on Future Generations*, September 2024, https://www.un.org/sites/un2.un.org/files/sotf-pact_for_the_future_adopted.pdf.
17 Rapp, L. (2025), 'Does the US want to weaken the UN or make it 'great again'?', Chatham House Expert Comment, 22 September 2025, https://www.chathamhouse.org/2025/09/does-us-want-weaken-un-or-make-it-great-again; Council on Foreign Relations (2025), 'Trump's Criticism of the UN', 24 September 2025, https://www.cfr.org/article/trumps-criticism-un.
18 Bradford, A. (2019), *The Brussels Effect: How the European Union Rules the World*, Oxford: Oxford University Press.
19 Csernatoni, R. (2025), *The EU's AI Power Play: Between Deregulation and Innovation*, paper, Carnegie Europe, https://carnegieendowment.org/research/2025/05/the-eus-ai-power-play-between-deregulation-and-innovation?lang=en.
20 European Commission (2025), 'The General-Purpose AI Code of Practice', https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai.

2.3 Asymmetry between public and private sectors

Private companies largely shape the development and deployment of frontier AI systems. Companies like Google and OpenAI build the models, NVIDIA and TSMC manufacture the chips that power them, and cloud providers like AWS, Google and Microsoft supply the infrastructure. These companies are overwhelmingly concentrated in the US and China.

Public–private partnerships are becoming more common. Notable examples include the US's 'Genesis' and 'Stargate' investments at home and abroad or the US–UK Technology Prosperity Deal.[21] But these arrangements have not shifted the fundamental reality. Outside of China's state-directed AI labs, the most consequential decisions on AI development happen inside private companies, in opaque and largely unregulated environments.

Private companies have commercial incentives to push back against costly governance provisions like transparency requirements, capability limitations or verification processes, particularly when competition rewards secrecy and speed as richly as it does in AI. Regulatory capacity gaps mean that, even if they wanted to impose constraints on the private companies developing AI, governments lack the computational resources, technical expertise and legal authority to independently evaluate proprietary models or compel disclosure.

The scale of private investment far exceeds regulatory capacity. Industry estimates put 2026 hyperscaler capital spending at $527 billion globally,[22] while OpenAI's CEO projected that future frontier models could require $100 billion in capital per training run.[23] By contrast, the EU's AI Act, finalized in 2024 after years of negotiation, allocated just €1 billion for its enforcement and implementation. When the UK government announced its AI Safety Institute (since renamed the AI Security Institute) in 2023, it committed £100 million over two years – which, despite being a significant investment, is less than major private sector labs spend in a single week. The digital industry's spend on lobbying in Brussels alone is reported to have increased by over 50 per cent in the four years up to 2025, reaching $175 million in that year.[24]

This investment gap has resulted in limited capacity for effective regulation. Regulatory agencies largely lack the computational resources to independently evaluate frontier model capabilities, relying on developer self-reporting or voluntary access agreements. Agencies can neither hire sufficient technical staff to match industry expertise, nor compel wholesale disclosure of training data, model architectures or safety testing results – all of which are considered by private companies to be proprietary information.

---

21 UK Prime Minister's Office, 10 Downing Street (2025), 'Memorandum of Understanding between the Government of the United States of America and the Government of the United Kingdom of Great Britain and Northern Ireland regarding the Technology Prosperity Deal'.
22 Goldman Sachs (2025), 'Why AI Companies May Invest More than $500 Billion in 2026', 18 December 2025, https://www.goldmansachs.com/insights/articles/why-ai-companies-may-invest-more-than-500-billion-in-2026.
23 Knight, W. (2023), 'OpenAI's CEO Says the Age of Giant AI Models Is Already Over', Wired, 17 April 2023, https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over.
24 Corporate Europe Observatory (2025), 'Big Tech lobby budgets hit record levels', 29 October 2025, https://corporateeurope.org/en/2025/10/big-tech-lobby-budgets-hit-record-levels.

Governments have in recent years largely prioritized investment in AI over regulation. For example, when the US administration of President Joe Biden released its AI executive order in October 2023,[25] the emphasis was placed on fostering innovation and attracting private investment. The order imposed minimal binding constraints on model development. Biden's successor as president, Trump, has taken further steps to minimize regulatory burdens and maximize investment. Meanwhile, France and Germany successfully lobbied to weaken the EU AI Act's requirements for foundation models, citing concerns over competitiveness.[26]

The result is an environment where the information, resources and technical capacity of the regulated is vastly superior to that of the regulators.

2.4 Knowledge collapse and threats to shared values

Effective governance of shared global problems depends on the capacity to establish facts, resolve disputes about evidence, and build consensus on when action is warranted, and is made easier through shared values and principles. The latter are both are in decline.

Nations are increasingly unable to agree on fundamental questions because they operate in incompatible information environments and hold increasingly divergent values.

Trust in media, government and scientific institutions has declined sharply across democracies over the past two decades.[27] Media fragmentation means individuals consume fundamentally different information from one another. Regulatory agencies face systematic underfunding and political interference, weakening their ability to produce credible independent assessments. Scientific expertise is increasingly treated as partisan rather than authoritative. The result is an absence of shared processes for resolving disagreement. When stakeholders cannot even agree on what counts as credible evidence, or which institutions have legitimate authority to assess such evidence, collective decision-making inevitably breaks down.

---

25 The American Presidency Project (2023), 'Executive Order 14110 on Safe, Secure and Trustworthy Artificial Intelligence', 30 October 2023, https://www.presidency.ucsb.edu/documents/executive-order-14110-safe-secure-and-trustworthy-development-and-use-artificial.
26 Sayeedi, I. (2023), 'Last minute national objections to the EU's AI Act are a mistake. Here's why', Global Governance Institute, 27 November 2023, https://www.globalgovernance.eu/publications/rolling-back-the-ai-act-is-a-mistake-heres-why#:~:text=The%20arguments%20presented%20by%20France,to%20only%20further%20entrench%20this.
27 University of Southampton (2025), 'Democracy in crisis: Trust in democratic institutions declining around the world', press release, 19 February 2025, https://www.southampton.ac.uk/news/2025/02/democracy-in-crisis-trust-in-democratic-institutions-declining-around-the-world.page.

Efforts to establish AI governance are floundering in this deteriorating information environment, while facing additional structural challenges. AI model interpretability is itself an evolving field of science. Technical evaluations of model behaviour remain largely under the ownership of private companies, limiting independent verification outside of a handful of AI security institutes.[28] Developers face strong commercial incentives to withhold information about system capabilities, and strong security incentives to avoid disclosing vulnerabilities.[29]

International agreements cannot proceed without shared basic assessments of risk. Enforcement becomes impossible when compliance itself is continually contested. Support for intervention from the public and policymakers alike evaporates without sufficient consensus about what AI systems can do or should be used for. AI governance requires trusted mechanisms for risk assessment, transparent information-sharing, and the capacity to build consensus rapidly when threats emerge. But none of these things are achievable when the basic infrastructure for establishing shared facts is weakened.

Nations are increasingly unable to agree on fundamental questions because they operate in incompatible information environments and hold increasingly divergent values. The traditional transatlantic alliance between the US and Europe, once rooted in shared democratic principles, is being reconstituted as a 'coalition of capabilities', based on transactional exchanges of economic benefits rather than common purpose – the so-called 'Pax Silica'.[30]

The US-led 'Pax Silica' explicitly frames cooperation around 'shared interests' in supply-chain security and economic advantage, rather than democratic solidarity or multilateral principles. This change in emphasis reflects a broader shift in which ideological alignment between governments is replaced by capability-based coalitions. Some AI scholars such as Anton Leicht contend that this type of agreement may eventually form the basis for a stronger coalition than one based on shared values,[31] but the transition will be uncomfortable and is not certain.

Taken together, these barriers create an environment where binding international coordination appears difficult, if not impossible, under current conditions. The US and China prioritize national advantage over cooperation; middle powers pursue access over regulation; international and state institutions lack enforcement power and technical capacity; private investment dwarfs regulatory budgets; and allies and rivals disagree over basic facts, while abandoning previous shared values and common purpose.

---

28 Bengio et al. (2025), 'International AI Safety Report 2025'.
29 There is no global standard for transparency and accountability in AI, and one is highly unlikely to emerge in the current environment. See also Bommasani, R., Klyman, K., Wan, A. and Liang, P. (2025), 'Transparency in AI is on the Decline', Stanford University, 9 December 2025, https://hai.stanford.edu/news/transparency-in-ai-is-on-the-decline.
30 US Department of State (2025), 'Pax Silica', https://www.state.gov/pax-silica.
31 Leicht, A. (2025), 'Forging a Pax Silica', Threading the Needle blog, 17 December 2025, https://writing.antonleicht.me/p/forging-a-pax-silica.

03
How crises can lead to action

Crises themselves guarantee neither progress on governance nor good practice. But they may provide an opening for bold action that is not currently feasible.

Crises trigger diverse responses. These responses can, and often do, lead to substantial policy change. When based on preparation and planning, change can enable states and societies to build better systems, and to improve robustness and resilience to future crises.

Drawing on desk research and expert interviews, this chapter examines the responses to three previous global crises in other sectors for insights that can be applied to AI governance. These case studies were selected for different levels of global impact, interdependence of outcomes and diversity in responses: one interventionist (the global financial crisis of 2007–08); one networked (the WannaCry ransomware attack of 2017); and one improvised (the COVID-19 pandemic of 2019–22).

This chapter is not a comprehensive crisis audit. Instead, it spotlights a range of institutional changes, new coalitions and networks, and innovations in the technological tools and infrastructures needed to monitor, weather and recover from crises. Building on case insights, this chapter considers what this set of best practices and cautionary tales can teach policymakers, practitioners and experts about building a preparation protocol.

3.1 What previous crises can teach us

A significant disruption to the current system will likely be required to break through the current deadlock on global AI governance. Historical precedent indicates that crises can catalyse policy changes that previously appeared unlikely. The 1986 Chernobyl disaster forced international cooperation on nuclear safety through binding conventions that Cold War tensions had previously blocked. The 2001

attacks on the US World Trade Center in New York triggered wide-ranging changes in international governance, notably in financial monitoring, security cooperation and the application of international law. The 2007–08 global financial crisis led to sweeping regulatory reforms being implemented in the banking sector.

Not every crisis produces governance change, but certain types of events create opportunities and generate political will.[32] These 'focusing events' are rare, high-visibility shocks that concentrate harm, attract sustained attention and can break down structural barriers that prevent action under normal conditions. Such events mobilize diverse stakeholders, reveal hidden costs or risks in a system, change cost–benefit calculations among political leaders, temporarily invert concentrations of power and create a sense of urgency that can overcome vetoes.[33] But, the governance opportunities crises create are time-limited. Crisis windows close as attention shifts, political constraints and authority resettle, interests retrench or a new crisis takes root.

Global crisis entanglement – or 'polycrisis' – emerges when slow-moving *stresses*, such as governance gaps or resource dependencies, intersect with fast-moving *triggers*, such as technical failures or malicious attacks.[34] But entanglement may present opportunities for positive change, alongside risks of spillover from one sector to another as AI becomes integrated into the global economy. Entanglement also means that today's change-makers must build crisis preparation protocols that respond to potential harms accruing through inaction in seemingly disconnected policy areas.

The case studies below suggest that interlinked factors determine the extent to which crises catalyse effective governance: the framing of an incident as crisis; the concentration and legitimacy of authority; the availability of institutionalized pathways for coordination; and the alignment of incentives among the significant actors.

- **Framing** shapes the perceived scope, severity and ownership of a crisis, determining whether an incident is understood as an isolated failure or a bigger threat requiring an exceptional response. Where incidents are framed as crises with collective, international implications, the range of legitimate policy tools expands, the political salience of the issue is elevated, and extraordinary levels of coordination and collective decision-making are justified.
- In moments of crisis, **authority** tends to concentrate temporarily, enabling certain actors (including officials, regulators, executive bodies or technical experts) to bypass veto points and define the options for response. Where such authority is recognized as legitimate, decisions can be taken rapidly. Where it is contested, though, responses fragment or stall.

---

32 Congleton, R. (2004), 'The Political Economy of Crisis Management: Surprise, Urgency, and Mistakes in Political Decision Making', *Advances in Austrian Economics*, 8, https://doi.org/10.1016/S1529-2134(05)08007-5; Rosenthal, U. and Kouzmin, A. (1997), 'Crises and Crisis Management: Toward Comprehensive Government Decision Making', *Journal of Public Administration Research and Theory*, 7 (2), pp. 277–304, https://doi.org/10.1093/oxfordjournals.jpart.a024349.
33 Birkland, T. A. (1998), 'Focusing events, mobilization, and agenda setting', *Journal of Public Policy*, 18 (1), pp. 53–74, https://doi.org/10.1017/S0143814X98000038; Blyth, M. (2013), 'Paradigms and paradox: The politics of economic ideas in two moments of crisis', *Governance*, 26 (2), pp. 197–215, https://doi.org/10.1111/gove.12010.
34 Lawrence, M. et al. (2024), 'Global polycrisis: the causal mechanisms of crisis entanglement', *Global Sustainability*, 7 (2024), https://doi.org/10.1017/sus.2024.1.

- Governance change then depends on the availability of **institutionalized pathways** through which decisions can be turned into action. These pathways may take the form of existing legal mandates, international forums or emergency procedures. Crises accelerate the use of these mechanisms, but rarely create them from scratch.
- Finally, durable reform requires sufficient **alignment of incentives** to sustain change. Individuals and groups both within and outside government who push a particular set of policies – sometimes called 'policy entrepreneurs' – play a critical role in this phase by mobilizing coalitions, reframing interests and sequencing decisions so that crisis-driven reforms become embedded rather than reversed.

The response to the 2008 financial crisis succeeded partly because regulatory agencies and legislative frameworks already existed. UK policymakers framed COVID-19 as a security threat, which enabled them to deploy resources that were otherwise unavailable.[35] International cooperation on nuclear safety after Chernobyl succeeded in part because the International Atomic Energy Agency provided a multilateral venue for negotiations, and in part because the disaster drew attention to a transnational threat to human life that could not be ignored.

Advocates of AI governance span a broad range of perspectives – from those who see it as necessary to prevent catastrophic or existential risks to humanity, to others who seek to mitigate foreseen political or economic disruption, to others still who want governance to serve in protecting national industries or even guaranteeing further acceleration in the development of AI. This paper does not attempt to arbitrate among them or to prejudge what 'good' governance might look like. Its central messages are that progress on binding and durable international governance of AI has stalled, and that a major AI-related crisis can rapidly shift the political constraints that have so far hindered international coordination. Investment in strengthening any of the four factors listed above will help a durable form of governance to emerge.

3.2 Three case studies: Good practice and cautionary tales

The interventionist response: the 2007–08 global financial crisis

Triggered by the collapse of the US subprime mortgage market and the failure of major financial institutions to prevent its collapse, the global financial crisis (GFC) rapidly cascaded into a global polycrisis through an interconnected

35 Magcamit, M. and Chinudomsub, P. (2025), 'Same difference? Interrogating the security politics of COVID-19 in the 'democratic' United Kingdom and 'authoritarian' Thailand', *Japanese Journal of Political Science*, 25 (4), pp. 247–73, https://doi.org/10.1017/S146810992400015X.

system of credit, liquidity and confidence. More than $2 trillion in global asset value evaporated, and economies around the world faced simultaneous banking failures, frozen credit markets and systemic contagion.[36]

The immediate crisis response was heavily interventionist, highly centralized and led by key institutional players. In the US and Europe, central banks and finance ministries coordinated emergency liquidity injections, bank guarantees and bailouts, such as the US Troubled Asset Relief Program.[37] The US Federal Reserve and the US Treasury deployed extraordinary powers to stabilize markets.[38] This interventionist response enabled rapid decision-making, but, to its critics, also entrenched the dominance of technocrats and financial regulators as guardians of the global financial system. At the same time, parallel coordination took place in transnational channels. The G20 became the premier forum for global economic governance, and the Financial Stability Board was created as the successor to the Financial Stability Forum to strengthen macroeconomic oversight and standards.[39] These two institutions remain key to financial monitoring and coordination almost 20 years on.

The GFC demonstrates that crisis-driven governance works best when it can leverage existing institutions, technical expertise and monitoring infrastructure, rather than building this infrastructure from scratch.

Though the success of this response remains disputed, the scale of post-crisis reforms is not. New regulatory architectures,[40] such as the Dodd-Frank Act in the US[41] and the Basel III framework on capital requirements,[42] introduced greater oversight, data reporting and early-warning mechanisms. The crisis also elevated the value and role of technical expertise and real-time system-monitoring through dashboards, supervisory technologies and centralized information-sharing systems. These were incorporated into everyday financial governance around the world.

---

36 Felton, A. and Reinhart, C. M. (eds) (2009), *The first global financial crisis of the 21st century: Part II, June–December 2008*, report, London: CEPR/VoxEU, https://cepr.org/system/files/publication-files/68584-the_first_global_financial_crisis_of_the_21st_century_part_ii_june_december_2008.pdf.
37 Bakir, C., Coban, M. and Akgunay, S. (2021), 'The European Union Financial Crisis: A Critical Analysis', *Oxford Research Encyclopedia of Politics*, https://oxfordre.com/politics/view/10.1093/acrefore/9780190228637.001.0001/acrefore-9780190228637-e-1504.
38 Board of Governors of the Federal Reserve System (2010), 'The Federal Reserve's Policy Actions During the Financial Crisis and Lessons for the Future, Speech by Donald L. Cohn, Vice Chairman, Federal Reserve, Carleton University, Ottawa, Canada, 13 May 2010', https://www.federalreserve.gov/newsevents/speech/kohn20100513a.htm.
39 Financial Stability Board (2017), *Implementation and Effects of the G20 Financial Regulatory Reforms: 3 July 2017 3rd Annual Report*, https://www.fsb.org/uploads/P030717-2.pdf.
40 Claessens, S. and Kodres, L. (2014), *The regulatory responses to the global financial crisis: Some uncomfortable questions*, working paper, International Monetary Fund, https://www.imf.org/external/pubs/ft/wp/2014/wp1446.pdf.
41 US Congress (2010), *H.R.4173 – Dodd-Frank Wall Street Reform and Consumer Protection Act*, 21 July 2010, https://www.congress.gov/bill/111th-congress/house-bill/4173/text.
42 Bank for International Settlements (2017), 'Basel III: international regulatory framework for banks', https://www.bis.org/bcbs/basel3.htm. Basel III is followed by most major financial jurisdictions, including Australia, Brazil, Canada, China, the EU, Hong Kong, India, Japan, South Africa, South Korea, Singapore, the UK and the US.

An example is the new Global Systematically Important Banks Classification, which is based on scoring thresholds that, once surpassed, automatically trigger heightened supervision, planning requirements and capital surcharges.[43]

Stronger regulatory architecture emerged from a crisis that created the conditions for prevention, coordination and rapid action. Yet the crisis response also revealed and institutionalized persistent inequities. Developing and lower-income economies affected by the crisis had limited influence in the development of the post-crisis architecture.[44] Much rule-making occurred within institutions dominated by advanced economies, meaning that many states affected by spillovers had minor voices and limited bargaining power in shaping the remedies.

The GFC demonstrates that crisis-driven governance works best when it can leverage existing institutions, technical expertise and monitoring infrastructure, rather than building this infrastructure from scratch. In terms of an AI crisis, this lesson suggests prioritizing investments in AI safety institutes (AISIs), technical exchange networks and pre-negotiated frameworks that can be rapidly activated, rather than relying on improvised solutions in the event of catastrophe.

The example of the GFC also reveals a danger. Those countries that were excluded from negotiations had governance imposed on them, rather than being involved directly in shaping it. Their exclusion helped entrench global inequities that persist today.

However, the main elements of both the GFC and the response can be discerned, and may be useful in the event of an AI crisis:

- **Symmetric economic harm.** All major economies faced simultaneous banking failures and frozen credit, aligning incentives for cooperation.
- **Clear causation and shared language.** Financial contagion mechanisms were well understood, enabling rapid diagnosis and coordinated response.
- **Pre-existing institutions.** The crisis primarily empowered institutions that already existed (such as central banks, finance ministries and the G20), rather than creating new ones – demonstrating that preparation must build capacity before crisis strikes.
- **Speed through concentrated authority.** Emergency powers and technocratic dominance bypassed normal veto points to reach a solution rapidly.

The key lesson for AI governance is that infrastructure and economic crises affecting all major powers symmetrically offer the best prospects for comprehensive governance. But the response will only be swift and effective if technical capacity and institutional pathways exist before the crisis hits.

---

43 Financial Stability Board (2024), *2024 List of Global Systemically Important Banks (G-SIBs)*, 26 November 2024, https://www.fsb.org/2024/11/2024-list-of-global-systemically-important-banks-g-sibs.
44 World Bank (2009), *Global Monitoring Report 2009: A Development Emergency*, Washington, DC: World Bank, pp. 23–50, https://www.imf.org/external/pubs/ft/gmr/2009/eng/gmr.pdf.

The networked response: WannaCry

In May 2017, ransomware known as WannaCry[45] proliferated globally. At the time, security analysts called it the largest ransomware cyberattack ever recorded.[46] Within a single day, it infected more than 200,000 computers in 100 countries.[47] In the UK, the National Health Service (NHS) was badly hit, with widespread disruptions to computer systems preventing access to data about patient care, leading to cancelled appointments and procedures. The attack cost the NHS £92 million[48] and the global financial impact was estimated at around $8 billion.[49]

Over 34 per cent of NHS trusts found their devices' files encrypted and a ransom demand for bitcoin.[50] Fearing permanent loss of access, some trusts shut down their systems as a precautionary measure, 'as they had not received central advice early enough'.[51] Private sector providers led the initial response, in coordination with the recently instituted UK National Cyber Security Centre (NCSC). Information-sharing on the threat and responses was relatively informal, but nonetheless agile and effective.[52]

UK media covered evidence of disruption, such as diverted ambulances and offline equipment, but had little information on its source.[53] Political leaders reassured citizens that the attack was global and untargeted.[54] By the evening, an independent cybersecurity researcher stopped the ransomware spread by buying a domain name.[55] The crisis was triggered by the exploitation of a *known* software vulnerability by a hacker group,[56] and later formally attributed to a North Korean-linked group.[57] Ultimately, WannaCry was an avoidable crisis with a relatively uncomplicated 'fix'.

WannaCry was unprecedented but not unexpected. It is best understood as a complex, lurking threat, compounded into a crisis by globally networked technologies and interdependence.[58] The crisis demanded emergency coordination, primarily between the private sector and central cyber authorities, but also between these authorities and NHS healthcare trusts. It platformed novel responses, many of which were later used to inform policy changes.

---

45 Cloudflare (undated), 'What was the WannaCry ransomware attack?', https://www.cloudflare.com/en-gb/learning/security/ransomware/wannacry-ransomware.
46 Wetzel, J. (2017), 'What Is WannaCry? Analyzing the Global Ransomware Attack', Recorded Future blog, 15 May 2017, https://www.recordedfuture.com/blog/wannacry-ransomware-analysis.
47 UK Comptroller and Auditor General (2018), *Investigation: WannaCry cyber attack and the NHS*, London: National Audit Office, https://www.nao.org.uk/wp-content/uploads/2017/10/Investigation-WannaCry-cyber-attack-and-the-NHS-Summary.pdf.
48 Allegretti, A. (2018), 'Cost of WannaCry cyber attack to the NHS revealed', Sky News, 11 October 2018, https://news.sky.com/story/cost-of-wannacry-cyber-attack-to-the-nhs-revealed-11523784.
49 Prevezianou, M. F. (2021), 'WannaCry as a Creeping Crisis', in Boin, A., Ekengren, M. and Rhinhard, M. (eds) (2021), *Understanding the Creeping Crisis*, Cham: Palgrave Macmillan, pp. 37–50, https://doi.org/10.1007/978-3-030-70692-0.
50 BBC News (2017), 'NHS cyber-attack: GPs and hospitals hit by ransomware', 13 May 2017, https://www.bbc.co.uk/news/health-39899646; UK Comptroller and Auditor General (2018), *Investigation: WannaCry cyber attack and the NHS*.
51 UK Comptroller and Auditor General (2018), *Investigation: WannaCry cyber attack and the NHS*.
52 Ibid.
53 Interviewee #5; BBC News (2017), 'NHS cyber-attack'.
54 Ibid.
55 Greenberg, A. (2020), 'The Confessions of Marcus Hutchins, the Hacker Who Saved the Internet', Wired, 12 May 2020, https://www.wired.com/story/confessions-marcus-hutchins-hacker-who-saved-the-internet.
56 Vulnerabilities can be understood as 'back doors without locks'. Microsoft had released a 'patch' to rectify the problem (i.e. a new lock for the back door) in March 2017. But many organizations – the NHS included – had not yet updated their systems.
57 Interviewee #5.
58 Prevezianou (2021), 'WannaCry as a Creeping Crisis'.

The WannaCry incident illustrates both the promise and limitation of networked crisis response for AI governance. The UK's reforms to crisis governance in the wake of the cyberattack (for example, the centralization of cyber crisis response and the recognition of cyber risks in the UK's central emergency response unit) demonstrate that even limited crises can catalyse significant institutional change. However, WannaCry produced minimal international governance precisely because no institution had jurisdiction or prepared mechanisms.

The response to WannaCry succeeded in producing governance change because the crisis was limited enough to allow informal, expert-led coordination rather than requiring top-level political negotiation.

WannaCry suggests that dual preparation is important: strengthening national-level technical response capacity that can be deployed rapidly, while pre-preparing international coordination channels that can be activated when crises cross borders.

The key elements of the WannaCry crisis and its response were:

- **Scale of random harm.** The threat was not targeted at a specific institution but proliferated widely and unpredictably, both in the UK and globally. It impacted critical national infrastructure.
- **Technical expertise as legitimate authority.** In the UK, the NCSC's technical credibility enabled it to coordinate the response despite being newly established, showing that crisis elevates actors with relevant technical expertise.[59]
- **Networked response among key actors.** Parts of the crisis response – such as information-sharing – were facilitated through informal connections between public and private actors, which might have improved agility and speed to respond.[60]
- **Clarification of global norms and boundaries.** WannaCry's impact on the NHS was criticized worldwide as inappropriate and unacceptable. The incident helped states and experts clarify what is off-limits – targeting critical national infrastructure, like healthcare – for state behaviour and activity in cyberspace.[61]

The lesson for AI governance is that narrow, technical crises with an adverse impact on specific sectors (such as finance and healthcare) may be more amenable to rapid coordination than broad systemic disruption, as they can elevate the salience of technical expertise and sector-specific institutions in the response.

---

59 The NCSC offers guidance to small and large organizations, the public sector, individuals and their communities. See National Cyber Security Centre (undated), 'Large organisations', https://www.ncsc.gov.uk/section/advice-guidance/large-organisations.
60 Ibid.; Interviewee #2.
61 This resulted in the so-called UN cyber norms, developed in the Group of Governmental Experts as early as 2015 but reaffirmed and operationalized as part of the Open-Ended Working Group until 2025, which has now been replaced by a global mechanism. See Dig.Watch (undated), 'UN cyber norms', https://dig.watch/cyber-norms; Interviewee #1.

The improvised response: COVID-19

The first cases of COVID-19 respiratory disease were detected in China in 2019. The disease was infectious and quickly spread around the world, leading to the World Health Organization (WHO) to declare a public health emergency in January 2020. While widely recognized as a catastrophic global health crisis, the COVID-19 pandemic also exposed the failure of decision-makers to mount a coordinated response that could have mitigated devastating impacts. Official data indicate that more than 7 million people died during the COVID-19 pandemic. Total global excess deaths following the pandemic have been estimated at between 19.2 million and 36.3 million.[62]

WHO attempted to communicate and coordinate a response through the IHR Emergency Committee, Strategic Preparedness and Response Plan, situation reports, and declaration of a public health emergency of international concern. But the crisis was ultimately managed via a wide variety of national responses.[63] Sweden, for example, focused on voluntary measures, emphasizing personal responsibility over the restrictions on movement and interaction elsewhere.[64] By contrast, China implemented a strict 'zero-COVID' policy, with controversial restrictions on movement of people, community-wide health screenings and mandatory quarantines.[65] Some countries pursuing so-called elimination strategies, such as New Zealand, achieved early success in controlling transmission of the disease.[66] But it was the eventual development and rollout of vaccines and high levels of uptake that finally caused the global fatality rate to decline sharply and allowed national authorities to regain control.[67]

Many national governments proved ill-prepared and sluggish in their responses, neglected the most vulnerable segments of their population, and struggled against low public trust and rampant mis- and disinformation. The Lancet Commission on lessons for the future from the COVID-19 pandemic, formed following the pandemic, identified global systemic failures across multiple levels: inadequate prevention; irrational policy responses; lack of transparency; departures from established public health practices; weak operational cooperation; and an absence of international solidarity.[68]

---

62 *The Economist* (2022), 'The pandemic's true death toll', 25 October 2022, https://www.economist.com/graphic-detail/coronavirus-excess-deaths-estimates.
63 World Health Organization (2025), 'Timeline: WHO's Covid-19 response', https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#event-87.
64 Björkman, A., Gisslén, M., Gullberg, M. and Ludvigsson, J. (2023), 'The Swedish Covid-19 approach: a scientific dialogue on mitigation policies', *Frontiers in Public Health*, 11 (1206732), https://doi.org/10.3389/fpubh.2023.1206732.
65 Zaihua Ba et al. (2023), 'Reflections on the dynamic zero-COVID policy in China', *Preventive Medicine Reports*, 36 (102466), https://doi.org/10.1016/j.pmedr.2023.102466.
66 Baker, M. and Wilson, N. (2022), 'New Zealand's Covid strategy was one of the world's most successful – what can we learn from it?', *Guardian*, 5 April 2022, https://www.theguardian.com/world/commentisfree/2022/apr/05/new-zealands-covid-strategy-was-one-of-the-worlds-most-successful-what-can-it-learn-from-it.
67 World Health Organization (undated), 'COVAX: Working for global equitable access to COVID-19 vaccines', https://www.who.int/initiatives/act-accelerator/covax; Our World In Data (2025), 'COVID-19 vaccinations vs. COVID-19 deaths, Dec 2, 2020 to Aug 12, 2024', https://ourworldindata.org/grapher/covid-vaccinations-vs-covid-death-rate.
68 Sachs, J. D. et al. (2022), 'The Lancet Commission on lessons for the future from the Covid-19 pandemic', *The Lancet*, 400 (10359), pp. 1224–280, https://doi.org/10.1016/S0140-6736(22)01585-9.

The international response proved equally troubling. One interviewee remarked that 'the only institutions with the ability to make positive governance changes in this case are international organizations, because it [was] an international problem'.[69] Yet no adequate international coordination mechanism functioned effectively during the crisis, nor have lessons been properly applied to multilateral enforcement since. Even in the Pandemic Agreement recently adopted by the World Health Assembly, critical guidance remains either non-existent or non-binding, including on commitments to equitable access to medical countermeasures like vaccines, or on sustainable financing for pandemic preparedness and response.[70]

While the wide variety of national COVID-19 responses showed mixed effectiveness, the overarching global response can be characterized as reactive improvisation rather than coordinated strategy. The experience of the COVID-19 pandemic has still not prompted the development of a robust international crisis architecture applicable to other borderless threats.

This lesson is particularly relevant for any future AI crisis. Like infectious diseases, threats proliferating on networked systems spread with little regard for legal and political boundaries. Similar to COVID-19, an AI-enabled crisis will likely require swift international cooperation and policy alignment to prevent cascading global harms.

COVID-19 therefore represents the failure mode that AI crisis governance must avoid. The pandemic had the scale, urgency and global reach that should have triggered cooperation. But it produced little or no durable international architecture for dealing with similar threats. The pandemic response reveals which crisis types are *least* likely to be productive in terms of governance: those with contested origins; those where the status of experts and their expertise remains politicized; and those where blame can be directed at specific actors rather than forcing acknowledgment of shared vulnerability. Crisis governance is not impossible, but it depends critically on fundamental shared understanding and institutional legitimacy.

The most significant elements of the COVID-19 pandemic response were:

- **Epistemic collapse prevents coordination.** Contested attribution, politicized expertise and fragmented blame prevented the shared threat perception necessary for cooperation. Low public trust and the proliferation of mis- and disinformation meant that emergency measures faced resistance.
- **No institutional foundations to empower.** WHO lacked enforcement mechanisms and had already been weakened, meaning that the organization could not elevate itself into an effective coordinating body.
- **National interests overriding a shared sense of threat.** Vaccine nationalism and geopolitical blame games dominated the global response to COVID-19, despite the clear mutual interest in controlling the pandemic.

---

69 Interviewee #4.
70 Wenham, C. (2025), 'The Pandemic Agreement may weaken, rather than strengthen multilateralism', Chatham House Expert Comment, 21 May 2025, https://www.chathamhouse.org/2025/05/pandemic-agreement-may-weaken-rather-strengthen-multilateralism.

COVID-19 is a cautionary tale: even significant global crises can fail to produce governance when institutions are weak and information environments are low quality.

3.3 Key lessons

The case studies above show how previous crises led to both successful and unsuccessful governance outcomes. The response to the 2008 financial crisis succeeded because it met four conditions simultaneously: symmetric harm, pre-existing institutional capacity, epistemic clarity, and concentrated authority. The response to WannaCry succeeded domestically for similar reasons: the response drew on technical expertise with legitimacy, and a scope allowing for informal coordination. COVID-19 had the scale and urgency that should have triggered global cooperation but failed to meet these conditions.

While not all such 'focusing' events will be conducive to governance change, those most likely to result in durable solutions will combine the following characteristics:

- **Clear and undeniable framing of the problem.** For instance, in events involving infrastructure failures, economic harm or loss of life.
- **Harms that affect all major powers equally,** rather than hitting some countries harder than others and allowing for blame-shifting or claims of strategic advantage.
- **The involvement of sectors where technical expertise has established legitimacy,** such as finance, cyber security or the maintenance of critical infrastructure.
- **Occurrence in domains with pre-existing coordination mechanisms or institutions**, rather than across domains too wide for existing coordination mechanisms to cover.

The COVID-19 response is not the only cautionary tale that demonstrates that even global crises may fail to spur governance. The climate crisis has so far only produced commitments that fail to meet the scale of the challenge through processes that are frequently paralyzed by contested attribution, divergent national and public/private interests and incentives, and an inability of governments and organizations to achieve buy-in for expensive interventions. There is no guarantee that global AI governance will not follow a similar trajectory, even after a 'focusing' event. The following chapter outlines measures that can help policymakers and other stakeholders to implement rapid and effective interventions in the event of an AI crisis.

04 Preparing for an AI crisis: Recommendations

States, institutions and companies seeking to shape post-crisis AI governance can act in advance to improve the likelihood of high-quality governance emerging.

Precedent points to certain factors that may contribute to successful governance following a crisis. Many steps towards this outcome can be taken today. These steps should provide states, institutions and companies with the tools and technical levers necessary to build a durable AI governance regime once a crisis hits.

Activate existing governance pathways

Crises tend not to trigger novel mechanisms for governance. More often, they operationalize pre-existing ones through rapid policy change. For the governance of AI, efforts should focus on:

- Developing off-the-shelf strategies such as treaties and agreements;
- Leveraging technical exchange networks to enable communication and coordination between experts and policymakers;
- Establishing policy back-channels and diplomatic coordination; and
- Evaluating and improving existing mechanisms for sharing information.

Develop off-the-shelf strategies

When a crisis occurs, decision-makers have little time to react and should not negotiate from scratch. 'Off-the-shelf' or 'dormant' strategies may provide the speed and flexibility needed.

'Off-the-shelf' arrangements are not a new concept. In arms control, climate change, finance, technology and trade, pre-negotiated or -drafted treaties, agreements, pacts and amendments sit patiently in government offices worldwide. When an emergency, scandal or shift in political opinion gives traction to a particular issue, these documents are in position, ready to deploy. For example, the Montreal Protocol HFC phase-down – an updated, fully drafted US-led amendment on hydrofluorocarbons (or HFCs) – was ready years before political alignment materialized around the Kigali Amendment in 2016. The pre-existing draft could be executed relatively quickly, leading to a faster global agreement on protecting the ozone layer.[71]

For an off-the-shelf strategy to enable post-crisis governance, three things are necessary. First, its core architecture – i.e. the scope and function – must be fixed in place. This means that, at the 'zero draft' stage, states must align on key obligations, language (both technical and political), implementation mechanisms and, in general terms, on the management of compliance and enforcement. In the example of the Kigali Amendment, informal discussions were encouraged alongside the formal forum for negotiation, to allow countries to explore approaches, ideas and positions freely expressed and unhindered by 'official' positions.[72] The UN Global Dialogue on AI series could offer opportunities for this kind of dialogue to take place.

Second, off-the-shelf strategies must have a level of operational tailoring that can be easily administered after implementation. Ideally, customization must be extended over time – for example, through a mechanism for amendments.[73] Off-the-shelf strategies must also be *modular* by design. This menu-style approach means that the components – such as the sector, actors, institutions and impact – can be better matched to the crisis conditions.

Third, implementers must be ready to use established enforcement mechanisms to implement or monitor agreements at the national and international levels, as opposed to or in tandem with newly established ones.

Non- or pre-crisis conditions provide low-pressure environments. Under normal conditions, decoupling consensus from formal agreement (and formal agreement from implementation) allows states to positively signal their commitment to crisis preparedness, thus facilitating measures detailed here and in Chapter 5. Reduced urgency alleviates the need for immediate resource allocation. Nor does it demand immediate political consensus on the most contentious AI governance questions.

---

71 UN Environment Programme (undated), 'About Montreal Protocol', https://www.unep.org/ozonaction/who-we-are/about-montreal-protocol.
72 Birmpili, T. (2018), 'Montreal Protocol at 30: The governance structure, the evolution, and the Kigali Amendment', *Comptes Rendus Géoscience*, 350 (7), pp. 425–31, https://doi.org/10.1016/j.crte.2018.09.002.
73 Klabbers, J. (2006), 'Treaties, Amendment and Revision', *Oxford Public International Law*, https://opil.ouplaw.com/display/10.1093/law:epil/9780199231690/law-9780199231690-e1483.

Crisis preparedness frameworks can then confidently progress to the drafting phase, while deferring complex questions of enforcement, burden-sharing and sovereignty limitations to a future approval and 'entry into force' phase.[74]

Beyond pre-negotiated treaties, institutional protocols can be activated without waiting for consensus between states. For example, the Financial Stability Board has crisis coordination arrangements that could be extended to AI-enabled financial shocks. The UN Office on Digital and Emerging Technology (ODET) could develop protocols for cross-border AI incidents. Unlike treaties, which require diplomatic agreement, institutional tools can be deployed by organizations when crises affect their mandates.

Having the off-the-shelf options allows states and others to act quickly in an emergency context such as one involving the use of AI-enabled biological weapons or autonomous military systems.

International organizations must be prepared to develop and deploy modular tools for crisis response. These tools may complement frameworks with AI-crisis-specific updates. Crucially, clarifying coordination protocols before a crisis may also mitigate the proliferation of multi-stage crises with amplified severity. Among other benefits, it would help by providing a more stable basis for post-crisis governance-building.

Pre-agreed, crisis-specific 'red lines' can support these tools. 'Red lines' might pertain to levels and types of unacceptable AI risk. Breaching those 'red lines' would then enable the activation of certain protocols: for instance, threat reporting in AI-enabled biological weapons design. For existing international institutions (like the UN ODET and Office for Disarmament Affairs) or future ones, possessing and promoting off-the-shelf, but situation-specific, tools should ensure that crisis response is not contingent on international agreement on a comprehensive treaty.

Complete improvisation is likely to be too protracted for an AI crisis. Whether draft or binding commitments are used in crisis is still dependent on context and political will in the moment. But having the off-the-shelf options allows states and others to act quickly in an emergency context, such as one involving the use of AI-enabled biological weapons or autonomous military systems. This speed will be critical to mitigating potential harm and escalation.

74 Zartman, I. W. (1989), 'Prenegotiations: Phases and Functions', in Stein, J. G. (ed.), *Getting to the Table: The Process of International Pre-negotiation*, Baltimore, MD: John Hopkins University Press.

Leverage existing channels to connect technical experts with policymakers

Channels that connect technical experts to policymakers are essential both for building a clear picture of the threat landscape and for implementing responses. Beyond the immediate term, such channels are also essential as a foundation for governance, helping to provide in a credible, consensus-based picture of shared risks. (The barriers to doing so are detailed in Chapter 1.)

The response to WannaCry spotlighted the role of technical expert-to-expert channels, which can work outside the constraints of national politics and 'normal' diplomacy. In the UK, the NCSC plays a central role in crisis response. Similarly, technical expertise was elevated and mainstreamed into regulatory mechanisms following the GFC. There is evidence that IT improved banks' resilience.[75] In the COVID-19 pandemic, many states centralized communication between scientific experts and government decision-makers.[76]

On AI (and specifically AI safety and security), continued voluntary technical exchanges between experts have proven their value for fostering consensus despite national differences – for example, on the importance of context-specific 'red lines'.[77] Leveraging the operational offerings – information-sharing, monitoring dashboards, supervisory technologies, established interpersonal contacts, recurring meetings and so on – of pre-existing technical networks in crisis is key for generating trust, goodwill and buy-in for global crisis responses.

Establish policy dialogues and back-channels

Emergency back-channels between decision-makers are critical institutional pathways for any crisis. In some contexts, this kind of arrangement is called a 'red phone network', in reference to the Cold War-era hotline between Moscow and Washington, set up in the aftermath of the Cuban missile crisis for emergency communications.[78] Similar bilateral arrangements have been considered in other fields, such as cyber diplomacy, to exchange state positions during crises and prevent unnecessary escalation. The existence of these arrangements may provide high-level guarantees for new post-crisis governance measures, such as improved monitoring and risk-notification protocols.

Indeed, many states now favour a consolidated approach to dealing with major incidents or crises in cyberspace. Many publicly declare their thresholds of acceptable or unacceptable behaviour in cyberspace, although there is still an alarming divergence between some states' public-facing rhetoric on maintaining cyber stability

---

75 Pierri, N. and Timmer, Y. (2021), 'The importance of technology in banking during a crisis', *European Systemic Risk Board*, March 2021, https://www.esrb.europa.eu/pub/pdf/wp/esrb.wp117~6c6d0b49c2.en.pdf.
76 Hodges, R. et al. (2022), 'The Role of Scientific Expertise in COVID-19 Policy-making: Evidence from Four European Countries', *Public Organization Review*, 22 (2), pp. 249–67. https://doi.org/10.1007/s11115-022-00614-z.
77 For example, the International Dialogues on AI Safety, operational from 2023, 'bring together leading scientists from around the world to collaborate on mitigating risks from AI', including from the US and China. See International Dialogues on AI Safety (undated), 'Home page', https://idais.ai. The International AI Safety Report, launched in 2024, includes contributions from scientists from over 30 countries. See International AI Safety Report (2026), 'International AI Safety Report', https://internationalaisafetyreport.org.
78 Simon, E. and Simon, A. (2020), 'Trusting Through the Moscow-Washington Hotline: A Role Theoretical Explanation of the Hotline's Contribution to Crisis Stability', *Journal of Global Security Studies*, 5 (4), pp. 658–74, https://doi.org/10.1093/jogss/ogz062.

and the realities of their sub-threshold cyber activities. To some cybersecurity experts, the WannaCry crisis provided important context to major international negotiations on responsible behaviour in cyberspace and mitigating shared risks, particularly in the UN First Committee. All UN member states can now meet in a designated international forum that seeks to advance an agreed set of norms for states while allowing challenge and debate.[79]

But multilateral and consensus-driven settings have significant downsides. Their deliberative pace means that long-term open dialogues will likely be too slow-moving to be useful for an emergency response to an AI crisis. Nevertheless, multilateral meetings help provide common language, advance context-specific 'red lines'[80] and signal political willingness to tackle a shared crisis. In the middle of an actual crisis, diplomatic backchannels – away from political pressures and scrutiny – can use this multilateral groundwork to build avenues for rapid, direct responses.

Evaluate and improve existing mechanisms for sharing information

Private companies control most of the information about AI system capabilities, behaviours and failures. During a crisis, such information asymmetry can hinder any attempt at a high-level response, as governments and international bodies cannot coordinate responses to threats they cannot see. To help facilitate an effective crisis response, companies should therefore:

- Expand trusted emergency communication channels with government representatives at the national level. AISIs and similar bodies already encourage communication between labs and governments.
- Build emergency communication channels into and around existing corporate institutions (including industry associations, forums and conferences) for real-time monitoring and crisis response coordination up and down the AI value and supply chains.[81] This recommendation is relevant beyond frontier AI labs and providers. Energy, mining and telecommunications companies, for example, are also exposed to supply-chain shocks, and may have an important role to play in crisis information-sharing. These changes will lay the foundation for the improved post-crisis governance of bottlenecks, proliferation and other instability risks.

79 This work took place in two tracks: the UN Group of Governmental Experts and the UN Open-Ended Working Group, finalizing their work in 2021 and 2025 respectively. Their successor is a global mechanism to advance responsible state behaviour in cyberspace. See Payne, R. (2025), 'The OEWG ends and a new UN cybersecurity permanent mechanism is born', Global Partners Digital, 24 July 2025, https://www.gp-digital.org/the-oewg-ends-and-a-new-un-cybersecurity-permanent-mechanism-is-born.
80 At the UN General Assembly in September 2025, over 300 AI scientists, policymakers and experts called for the establishment of global AI red lines, including on areas such as child safety, chemical, biological, radiological and nuclear risks. See AI Red Lines (undated), 'Home page', https://red-lines.ai.
81 This is essential for recognizing and responding to warning signs of potential polycrisis scenarios: for instance, energy insecurities and resource shortages compounded by a shared cyber vulnerability.

- Participate in international networks and task-forces dedicated to government exchanges and work in tandem with corporate counterparts – in the technology sector and others – on synchronized responses according to pre-agreed 'red lines', necessitating model shutdowns, patch deployment and 'circuit breakers'.
- Close, or at least reduce, unnecessary information gaps and commit to sharing intelligence that does not have national security or competition implications (such as explainability and transparency indicators and methods, public benefit applications, or legal interpretations of data protection). Given that global consensus on transparency and risk mitigation is realistically years away, frontier AI labs and model providers should leverage existing efforts – for instance, the Hiroshima AI Process Reporting Framework – to generate alignment on minimum expectations of information-sharing in crisis.

Concentrate legitimate authority

If given enough authority, actors with credible and legitimate technical expertise can help bypass bottlenecks in times of crisis. Measures that will assist in this task include:

- Positioning AISIs as technical authorities;
- Establishing national crisis response units; and
- Clarifying emergency decision-making powers.

Position AI safety institutes as technical authorities

The member organizations of the international network of AISIs are strong candidates to act as trusted public authorities on AI, particularly when trust in the technology industry is low. AISIs have existing research collaborations on benchmarks, evaluations and joint testing, plus agreements with frontier labs, that provide access other institutions lack. Expanding AISI mandates to give them formal roles in crisis response alongside research functions, positioning staff in proximity to centres of political power and strengthening networks into industry can elevate these institutions' technical expertise into wider public technical authority. To make this work, politicians must deal with competing national institutional authorities and ensure AISIs' political legitimacy and financial sustainability.

Establish crisis decision-making prerogatives

As observed in the case studies, the scale of harm caused by a crisis is often contingent on the speed of response. Governments should therefore seek to craft a domestic regime of contingency powers ready to be deployed or, alternatively, ensure that *existing* frameworks for intervention are updated with crisis-relevant components and new incident authorities. These regimes should be:

- Automatically activated under specific, pre-defined 'trigger' conditions. (This may be facilitated by the incorporation of AI-related risks into countries' national risk registers, or the integration of AI-specific risks in national emergency response mechanisms.)

— Inclusive of powers aimed at agile decision-making, such as the capacity to request data from controllers rapidly on reasonable grounds and according to data protection principles; impose temporary operational restrictions; and activate international channels for crisis response (including technical exchanges and back-channels, intelligence-sharing arrangements and coordination under international bodies like Interpol).

— Subject to scrutiny, judicial boundaries and a strict regime of transparency to prevent misuse of additional contingency powers. (The goal is not to create a regime of exception that avoids democratic accountability, but rather to enable governments to take quick, decisive action in a crisis.)

Enable technical intervention

AI systems are increasingly autonomous, globally distributed and controlled by private actors. Their governance may demand the development of technical intervention capabilities that do not yet exist at scale. Technology developers, providers and their regulators must therefore:

— Build circuit breakers and 'kill switches' into system design;
— Establish forensics and attribution capabilities;
— Deploy monitoring and early-warning systems;
— Create rapid technical-response teams;
— Ensure access and verification mechanisms; and
— Foster investment in governance-enabling technologies.

Build circuit breakers and 'kill switches' into system design

Many autonomous AI systems are already globally diffused. Confronting a growing evidence base of model bias, deceptive behaviours, security vulnerabilities, weaponization risks and 'ego-centric' coordination in existing AI models and systems, many experts have strongly advocated for a human-in-the-loop approach to development.[82] Notwithstanding the type of crisis trigger (whether non-AI or AI-enabled, such as malicious use or loss of control), the possibility of AI systems operating *without* sufficient human control in crisis conditions is a legitimate cause for concern.

But globally agreed, binding actions on preventing crisis escalation with and through autonomous systems are decades away at best. Acknowledging this reality, private developers of advanced AI systems should urgently consider two measures that will enable crisis management and build trusted governance mechanisms in a post-crisis environment.

— Circuit breakers 'trip' if certain output conditions are fulfilled (for instance, the generation of model outputs with significant uplift potential for malicious misuse, such as solving a technical problem that was the main obstacle

82 Adewumi, T. et al. (2025), 'AI Must not be Fully Autonomous', *Machine Learning Group, EISLAB, Luleå University of Technology*, https://arxiv.org/html/2507.23330v1.

preventing someone from carrying out a bio-chemical attack). This rerouting process aims to make AI models intrinsically safe and more robust to potentially adversarial acts or crisis conditions.[83]

— 'Kill switches', on the other hand, trigger a more immediate shutdown or operational pause in the use of AI systems.

Both are key tools to build trust and certainty in attempts at post-crisis governance, and prevent a further escalation.

Assurance and risk mechanisms like these will help build 'good faith' between industry and government, and could give states more time to develop medium- to long-term governance pathways after the initial 'moment' of crisis.

Create allied rapid-response teams of technical experts

During and immediately following crisis, rapid-response teams working across borders have an important role. Such teams are often used by WHO to convene cross-border medical experts during outbreaks and in cybersecurity incident response, with many countries appointing computer incident response teams to coordinate on threat identification and response. These teams should play both an advisory and operational role, assisting with threat diagnosis, mitigating escalation and, ideally, facilitating coordinated international responses. The activities of cross-border rapid-response teams formed in crisis may inform governance innovation – for instance, in institutionalizing frontier monitoring capabilities or formalizing new mechanisms for cross-border information-sharing.

The most feasible path forward could be the establishment of rapid-response teams among allied nations' institutes with existing foundations of trust and intelligence-sharing agreements.

The international network of AISIs is likely the most viable future institutional 'home' for this work, given existing research collaborations (such as those on benchmarks, evaluations and joint testing) and agreements between frontier AI labs and national institutes. Participating governments might consider expanding the network's mandate to better formalize its role responding to AI-enabled crises. Some AI experts have advocated for a 'CERN for AI'. A formal international organization for crisis coordination and response may be feasible in the future. But, in the current geopolitical environment, an expansion of the international AISI network to host rapid-response teams of technical experts would be the best starting point, given the talent concentrations in participating national institutes. Drawing from existing foundations of trust and intelligence-sharing agreements, such as ASEAN, the EU, the G7 or the countries in the 'Five Eyes' intelligence-sharing network, is a potential path forward.

83 Zou, A. et al. (2024), 'Improving Alignment and Robustness with Circuit Breakers', *arxiv*, 8 July 2024, https://arxiv.org/html/2406.04313v2.

Deploy stop-gap monitoring solutions

Intervention requires the detection of crisis triggers. Basic monitoring requirements are found in most governance systems. However, AI systems operate at speeds and scales that exceed human monitoring capacity. Most AI systems operate on private infrastructure with no visibility to authorities, and there are strong incentives against transparency on AI capabilities in both the private and public sectors.

Under these conditions, external behavioural monitoring by third parties could track AI system outputs, behaviours or effects by detecting unusual patterns through public-facing interfaces, similarly to the way financial regulators monitor market behaviour before inspecting internal practices. While insufficient, these approaches may provide partial visibility while comprehensive monitoring remains politically infeasible. Should a crisis expose the costs of opacity, stronger requirements may become acceptable in time.

Foster investment in governance-enabling technologies

Some forms of AI governance will likely depend on technologies that do not exist at scale, such as chips with 'call-home' functionality tracking deployment and usage, model watermarking or secure 'sandboxes' for regulatory development and auditing. The expansion of procurement markets or public–private investment in governance-enabling technologies may be necessary to incentivize the development of these technologies before they become urgently needed.

Prepare effective framing

Crisis incidents must be framed as collective international threats requiring coordinated response. Minilateral or international framing efforts are an essential part of the preparations for trusted, credible governance, even where global consensus is lacking or geopolitical tension prevents it. Supporting these collective governance efforts must be a post-crisis aim. Measures to build preparedness include:

- Joint attribution for shared narratives; and
- Table-top exercises that practice framing.

Establish joint-attribution mechanisms and audit crisis-prevention strategies

Country-to-country intelligence-sharing and joint investigations during and after a crisis can lead to public attribution of crisis triggers. Operationalizing information about crisis triggers opens the door to policy change. After the GFC, for example, forensic accounting measures like investigations into systemic vulnerabilities and deceptive practices (such as sub-prime mortgage fraud) were essential to uncovering fraud and evidencing the need for post-crisis reforms.

The UK and EU's joint attribution of a cyber threat to Russia in December 2024[84] demonstrated the importance of attribution as a powerful political signal and declared a united cybersecurity front against future malign action. But, as in the COVID-19 pandemic, attribution is not always straightforward. Due to a lack of evidence and ongoing research into the disease, contested accounts of the virus's origins emerged that stoked political division and conspiracy theories – damaging, rather than laying, the foundations for a better future response.

Universal acceptance of attribution of a crisis is rarely achievable for all audiences. The prevalence of mis- and dis-information is a growing issue in crisis response and governance. Continued efforts must be made to ensure independent verified sources, evidence and forensics are available, specialized knowledge and expertise are prominent in the information space, and sufficient resources are given to addressing public perception, resilience and trust.

In the short term, even informal connections between AISIs and other allied technical experts, intelligence agencies, independent bodies and news media can support rapid baseline framings. An essential part of this is by auditing crisis prevention strategies – asking what worked, what did not, and where innovation is required to build more robust governance. Locating this audit within an institutional body – such as an empowered international network of AISIs or the UN's Scientific Panel – is preferable to make the audit process more inclusive.

Strengthen 'muscle memory' through mutual learning and scenario exercises

Interviewees and literature underscored the value that could be gained by decision-makers through scenario exercises, table-tops and wargames.[85] For example, Chatham House's scenario exercises have pushed participants to consider the broader governance implications of their policy options. As technology-enabled crises grow in frequency and likelihood, many traditional games have been updated to reflect new dynamics, such as networked problems, interdependent solutions and unpredictable technological advancements. Civil society organizations and policy institutes, in particular, can play a key role in building multi-stakeholder capacity on AI futures and crises. Stakeholders can then tackle questions of public framing, multilateral cooperation and points of contention between the main actors.

Where possible, exercises should be informed by key learnings (whether best practice or cautionary tales) gathered through information-sharing. Protocols for mutual learning are already well established in different crisis response areas. Cybersecurity researchers have long relied on different platforms for exchanging and analysing incident data, which helps inform both a clearer picture of the

84 UK Foreign, Commonwealth & Development Office (2024), 'UK joins partners in condemnation of malicious cyber activity by Russian Intelligence Services: UK government statement', press release, 3 May 2024, https://www.gov.uk/government/news/united-kingdom-joins-partners-in-condemnation-of-malicious-cyber-activity-by-russian-intelligence-services-uk-government-statement.
85 Avin, S., Gruetzemacher, R. and Fox, J. (2020), 'Exploring AI Futures Through Role Play', AIES '20: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 8–14, 'https://doi.org/10.1145/3375627.3375817.

threat landscape and the policy reaction.[86] International bodies like the WHO serve similar functions for outbreak incidents, while the Financial Stability Board's mandate promotes coordination and information exchange among authorities.[87] Where competitive incentives create barriers to information-sharing, though, anonymized or aggregated reporting that protects proprietary details while capturing systemic patterns are a short-term substitute.

Align incentives

After a crisis, short-term policy change is possible, even with misaligned incentives between and among governments, international institutions and companies. But durable governance reform requires a minimum threshold of sustained incentive alignment. To achieve this, actors must:

- Build capacity in low-income and developing economies; and
- Develop competition-proof coordination.

Build crisis governance capacity in low-income and developing economies

Crises can exacerbate existing global inequities. Inability to respond in crisis can amplify post-crisis harms in low-income and developing economies. Crisis responses also risk reproducing or entrenching power dynamics, such as the exclusion of low-income and developing economies from multilateral decision-making. International institutions should include tailored crisis preparation protocols in technical assistance and capacity-building programmes offered to states with limited digital infrastructure, regulatory capacity or international bargaining power.

Most AI governance attention focuses on US–China competition. But crisis negotiations depend on 'middle powers' (e.g. Canada, Germany, India, the Netherlands, Qatar, Singapore, Switzerland, the UAE and the UK) who can provide neutral venues, hold swing votes in multilateral settings, and control supply-chain chokepoints. This could include establishing regional crisis hubs and creating shared forensic and monitoring facilities that lower-capacity states can rely on.

Explore competition-proof coordination protocols

In crisis, private companies have various legal duties to their shareholders and users. This should entail acting responsibly and resolving incidents that may impact the consumers of their products and services, their business partners and – to some extent – members of the public.

---

86 For example, the European Repository of Cyber Incidents 'is an independent research consortium dedicated to better understanding the cyber threat environment', particularly in the EU. See Cybil (2022), 'European Repository of Cyber Incidents (EuRepoC)', https://cybilportal.org/publications/european-repository-of-cyber-incidents-eurepoc.
87 Financial Stability Board (2025), 'About the FSB', https://www.fsb.org/about.

AI providers and AI-exposed companies operating across sectors should establish collective, flexible protocols clarifying in-crisis roles and responsibilities, including non-binding rules and expectations to facilitate coordination. Crucially, such a framework must supersede competition and incentivize buy-in. Stability and certainty must be framed as a common business interest.

Corporate protocols are highly unfavourable in the current competitive, hyper-politicized AI ecosystem. However, a framework marketed as 'apolitical' – i.e. focused on evidence-gathering and stability – may be an effective foundation. This may comprise existing corporate or international regimes like the Financial Stability Board or loose corporate coalitions on technology and risk, as seen in cybersecurity, counter-terrorism and attempts to counter misuse of AI around elections.[88]

88 Examples include the Cybersecurity Tech Accord (2018), the Global Internet Forum to Counter Terrorism (2018), the Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online (2019) and the AI Elections Accord (2024).

05 Conclusion

An AI crisis is not inevitable – or desirable. But focusing investment right now will increase the chances of a productive and robust response if a crisis does hit.

Proactive, comprehensive efforts to establish global AI governance appear structurally impossible given present-day geopolitical competition, private sector dominance, institutional weakness and fragmentation. This paper argues that robust international AI governance may only become politically feasible following a crisis, when the costs of inaction exceed those for coordination, and that lessons can be learned from previous crises in other sectors.

While there are novel technical, political and economic barriers to effective international AI governance, there is strong reason to believe that some of these barriers would be lowered after a 'focusing event' or crisis. This paper does not claim that crises are desirable, but that it is prudent to consider now where investment can be targeted to best prepare for the moment – if and when it arrives.

The paper notes several possible options, and highlights five conditions that early interventions might target: effective framing of incidents as collective threats; legitimate authorities positioned to act; technical governance innovations; institutional pathways ready to activate; and sufficient alignment of incentives among those impacted by the crisis.

Explicitly preparing for a crisis does not guarantee that these conditions will emerge. But steps in this direction may improve the odds of a productive and robust response if they do. Pre-negotiating targeted interventions, building technical capabilities into systems, strengthening institutional capacity and establishing technical and public-facing coordination channels can all contribute to shaping post-crisis governance.

This approach is explicitly second-best. Ideally, governance would emerge through states, institutions and companies engaging each other proactively and reaching a common solution ahead of time. But in an environment where significant disruption to global systems by AI technologies is plausible and global governance is out of reach, ensuring that quality responses are ready when a crisis forces the issue may be the next best option.

About the authors

Rowan Wilkinson is a research associate in the Digital Society Programme, where she supports research on digital public infrastructure, the information space, tech sovereignty, and AI governance. Her expertise lies at the intersection of technology, humanitarianism and international development.

She previously worked with the UN and international NGOs on global development, crisis response, conflict, and human rights. Her previous research explored themes such as aid e-cash transfers and the use of technology to counter extremism.

Alex Krasodomski is the director of the Digital Society Programme at Chatham House. His work focuses on AI, emerging technology and centres of tech power. He also leads on projects aimed at strengthening state capacity and cooperation, identifying feasible paths towards global technology governance, and routes to market for public technology.

Alex is also a fellow at the Institute for Strategic Dialogue and, until 2022, was director of the Centre for the Analysis of Social Media at Demos and a co-founder of the AI start-up CASM Technology.

Isabella Wilkinson is a research fellow in the Digital Society Programme. Her work focuses on the geopolitics of global technology governance (with a focus on AI), new centres of decision-making power on technology, and information integrity.

Isabella is also a DPhil candidate in public policy at the Blavatnik School of Government, University of Oxford, where her research explores the political economy of collective action among companies in global AI governance. She is also co-chair of Women in International Security UK.

Francisco Javier Varela Sandoval is the International Strategy Forum academy fellow with the Digital Society Programme. His research at Chatham House focuses on sovereign AI development pathways. Before joining Chatham House, Francisco was a fellow at France's National Institute for Public Service, focusing on comparative administration.

From 2017 to 2022, Francisco taught microeconomics, economic growth, and public policy at the Autonomous Technological Institute of Mexico (ITAM). He has also held key roles in Mexico's federal government, including at the Ministry of Public Service, the Office of the President and the National Transparency Institute.

Acknowledgments

We are grateful to the Ford Foundation for their support of this project.

We would also like to express our gratitude to the anonymous peer reviewers for their insights and constructive feedback. This paper draws on many interviews with crisis experts, conducted in 2025. We are grateful to all those who participated, giving their time and valuable comments that inform this piece of work.

Finally, thanks are due to Chatham House's publications team and the Global Economy and Finance programme for their support throughout, particularly Chris Matthews and Creon Butler.

Cover image: A SOS button illuminated in red, 17 October 2025.
Photo credit: Copyright © Matteo Della Torre/NurPhoto/Getty Images

ISBN 978 1 78413 675 8
DOI 10.55317/9781784136758

Cite this paper: Wilkinson, R., Krasodomski, A., Wilkinson, I. and Varela Sandoval, F. J. (2026), *Breaking the deadlock on AI governance: How a crisis could lead to global coordination*, Research Paper, London: Royal Institute of International Affairs, https://doi.org/10.55317/9781784136758.

This publication is printed on FSC-certified paper.
designbysoapbox.com

Independent thinking since 1920

The Royal Institute of International Affairs
Chatham House
10 St James's Square, London SW1Y 4LE
T +44 (0)20 7957 5700
contact@chathamhouse.org | chathamhouse.org

Charity Registration Number: 208223