Scaling AI crisis diplomacy

How to use scenario exercises to build capacity and readiness

Policy report

September 2026

CHATHAM HOUSE

AISA
AI SAFETY ASIA

UNIVERSITÉ DE GENÈVE
GLOBAL STUDIES INSTITUTE

Executive summary

Scenario exercises are central to building preparedness for an AI-driven crisis: at the cross-border, regional and international levels.

- Governments, companies and technical experts are already using scenario exercises to improve crisis-ready AI diplomacy.
- International scientific reporting, technical standard bodies and incident-sharing networks offer solid foundations.
- In the face of unpredictable and cascading risks, capacity-building exercises rooted in fictional-but-feasible scenarios may have strategic value.

On the sidelines of the UN Global Dialogue on AI Governance, we ran a tabletop exercise on diplomatic responses to a regional AI incident.

- It generated a set of policy recommendations for how to build and scale preparedness functions at a regional and international level.
- Participants included policymakers, industry representatives and technical experts and researchers.

But more work is needed to connect practice to policy change.

- Appetite does not automatically translate to resources, institutions and political prioritisation.
- Participation in scenario exercises does not automatically create stronger diplomacy, networks and information-sharing.

Scenarios should be used as a strategic tool to stress-test preparedness. This policy report explains how.

- Increasingly advanced AI and crisis conditions challenge preparedness functions, like monitoring risks, reacting to escalation thresholds and activating cross-border and regional emergency backchannels.
- Scenarios should be used to find gaps in readiness, re-build and adapt capacity at a cross-border and regional level, and then scale it internationally at multilateral meetings.

Table of contents

Authors and contributors

Isabella Wilkinson[1] (Chatham House; Blavatnik School of Government, University of Oxford); **Alex Krasodomski** (Chatham House); **Alex Kelati**[2], **Ann Meceda**, **Edward Tsoi**, **Michael Bąk**, **Wing-Yee Lau** (AI Safety Asia); **Roland Bouffanais** (Université de Genève); **Roxana Radu** (Blavatnik School of Government, University of Oxford); **Shyam Krishna** (Member, UNESCO AI Ethics Experts Without Borders; Working Group Member, UN AI Governance for Humanity Lab; Research Leader, RAND Europe); **Anine Andresen, Aaron Halpern, Inbar Bobrovsky** and **Shahar Edgerton Avin** (Technology Strategy Roleplay); **Rowan Wilkinson**.

Cover image: Siavosh Hosseini via Unsplash.

[1] iwilkinson@chathamhouse.org
[2] alex@aisafety.asia

State of play

The preparedness paradox

Global efforts to govern advanced AI confront a paradox. On the one hand, the prospect of an AI-enabled crisis has never been so present. There is growing global consensus on the risks AI poses and how to mitigate them. Decision-makers and experts are taking crisis readiness seriously and taking steps to build resilience. But visibility, dialogue and evidence is not preparedness. Preparedness for a cross-border, regional or global AI crisis has never been more urgent nor so out of reach.

As the UN's preliminary Independent Scientific Report warned in June 2026, AI capabilities are developing faster than our capacity to govern them.[3] Increasingly powerful AI models and applications developed by a handful of US and Chinese providers have rapidly diffused globally and are increasingly an integral part of the digital infrastructure societies, economies and politics depend on. Regulation at any level - from companies,[4] governments and international institutions[5] - is not keeping pace. The release of models with agentic and cyber capabilities have triggered fresh anxiety,[6] as have reports of models-gone-rogue in testing.[7] Despite recent, promising steps on AI developer involvement in transparency frameworks[8] and voluntary sharing,[9] there is no clear global picture of frontier capabilities, particularly those developed in private or proprietary settings.

This uncertain, volatile and opaque technological reality undermines global preparedness, which is itself exacerbated by geopolitical fragmentation. A multilateral response to an AI-enabled crisis necessitates US and Chinese involvement. But despite some movement on intergovernmental talks[10] and Track 2 dialogues, the two great powers remain at loggerheads. They espouse

[3] United Nations Independent International Scientific Panel on AI, *Preliminary Report: Evidence-Based Assessment of Opportunities, Risks and Impacts of AI*, United Nations, 1 July 2026, https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report
[4] Jim VandeHei and Mike Allen, "Behind the Curtain: AI Godfathers Converge on Regulations," *Axios*, July 16, 2026, https://www.axios.com/2026/07/16/ai-regulations-openai-anthropic-google
[5] AI Governance for Humanity Lab, *Toward interoperability and inclusive participation in AI governance*, United Nations AI Governance for Humanity Lab, June 2026, https://www.un.org/digital-emerging-technologies/sites/www.un.org.digital-emerging-technologies/files/Toward_Interoperability_and_Inclusive_Participation_in_AI_Governance_White_Paper.pdf
[6] Alex Krasodomski, "What the UK should learn from Kimi K3, WAICO and the Hugging Face incident", *Chatham House,* 24 July 2026, https://www.chathamhouse.org/2026/07/what-uk-should-learn-kimi-k3-waico-and-hugging-face-incident
[7] Osmond Chia and iv McMahon, "Meta becomes latest firm to say its AI hacked another company", *BBC News*, 6 August 2026, https://www.bbc.co.uk/news/articles/c3ek3gvdnj3o
[8] See: https://oecd.ai/en/hiroshima
[9] See: https://www.anthropic.com/glasswing
[10] Laurie Chen, "US, China to hold AI talks in September, sources say", *Reuters*, 21 July 2026, https://www.reuters.com/world/china/us-china-hold-ai-talks-september-sources-say-2026-07-21/

minilateral and mutually exclusive visions for AI governance: the US' Pax Silica and China's World AI Cooperation Organisation.

But great power competition and bipolarity is a limiting narrative: it "risks leaving the toolbox of AI governance empty when it is needed most".[11] Far from the front of the AI race, middle powers - defined as states that lack sovereign frontier AI capabilities or cloud infrastructure, but have the institutional capacity to pool resources and set regional standards - express interest in coordination and co-developing capabilities to defend their foothold and interests.[12] Yet, as underlined by the Claude Mythos episode,[13] their lack of access to the frontier undercuts both their competitiveness and crisis preparedness. There is a widening gap in crisis readiness among countries, rooted in uneven access to capabilities and infrastructure, but also knowledge.

For example, the Global Majority faces the competing pressures of rapid adoption with limited safeguards (to enable digitisation and competitiveness) and a more cautious embrace of potentially destabilising emerging technologies. Building agency over AI infrastructure depends on "strategic collaboration": arrangements between countries with different capabilities but a mutual need for partnership.[14] But there are growing pains: partnership alone is not implementation, and most disproportionately focus on growth and prosperity, as opposed to safety and security. High-level policy interest in coordinating on AI development and governance can create useful entry points, but does not by itself establish standing arrangements for cross-border incident response.

Emerging solutions

Preparedness is not a single institution or technical safeguard: it is a coalition of functions at different points in the crisis lifecycle, including monitoring emerging risks, sharing information, establishing authority, coordinating decisions, preserving operational continuity, deploying technical interventions, and learning from incidents.

On a sectoral level and globally, elements of this architecture are beginning to emerge. For example, the Organisation for Economic Cooperation and

[11] Krasodomski, "What the UK should learn"
[12] Francisco Javier Varela Sandoval, Isabella Wilkison, Rowan Wilkinson and Alex Krasodomski, *How middle powers can weather US and Chinese AI dominance*, Chatham House, March 2026, www.chathamhouse.org/2026/02/how-middle-powers-can-weather-us-and-chinese-ai-dominance
[13] Isabella Wilkinson, "The US government's latest u-turn on Anthropic's Mythos sends mixed signals on AI governance", *Chatham House*, 2 July 2026, https://www.chathamhouse.org/2026/07/us-governments-latest-u-turn-anthropics-mythos-sends-mixed-signals-ai-governance
[14] Jane Munga, "A Kenya Technology Prosperity Deal Could Help Washington Secure Durable AI Partnerships with Africa", *Carnegie Endowment*, 23 June 2026, https://carnegieendowment.org/research/2026/06/united-states-kenya-technology-prosperity-deal-could-help-washington-secure-durable-ai-partnerships-with-africa

Development (OECD) Hiroshima AI Process reporting framework gives AI organisations a common structure for disclosing information like risk management practices;[15] the OECD also promotes a common framework for incident reporting[16] and hosts a public-facing Catalogue of Tools and Metrics for Trustworthy AI platforms.[17] The open-source AI Incident Database[18] and AI Incident Response Repository[19] are two of many sources for incident data. The international network of AI safety institutes builds shared evaluation science. In Southeast Asia, the ASEAN Responsible AI Roadmap, ASEAN AI Safety Network and Regional CERT provide foundations for regional capacity-building and incident cooperation.[20][21]

In other governance areas directly exposed to AI-related risks – like global health and finance – national and international institutions have crisis preparedness plans, testing and re-testing readiness at regular intervals, and updating resourcing and monitoring accordingly. While the architecture is comparably well-established, AI-enabled vulnerabilities present a new front and complications for 'traditional' readiness mechanisms. Despite efforts to improve interoperability, this patchwork of preparedness initiatives is fragmented. They are rarely tested together under crisis conditions, simulated or real-life. Crucially, establishing the level of risk is a different task to mitigating and then responding to it.

Middle powers face a particular challenge in operationalising preparedness: many have deep reliance on non-sovereign frontier models, cloud infrastructure and AI-enabled services, not to mention limited access to information about model capabilities from their developers, nor the ability to orchestrate unilateral technical remedies like kill switches or call home protocols.

Readiness is multi-faceted, and crisis forecasting, preparation and navigation is by no means a new practice. It is helpful to envisage a crisis lifecycle, with the connected aims and architectures included in *Table 1* located at different points:

- **Crisis risk identification** (including monitoring emerging risks and stress-testing readiness by identifying gaps practice, through scenario exercises)

[15] See: https://oecd.ai/en/transparency/overview
[16] OECD, *Towards a common reporting framework for AI incidents,* OECD, February 2025, https://www.oecd.org/en/publications/towards-a-common-reporting-framework-for-ai-incidents_f326d4ac-en.html
[17] See: https://oecd.ai/en/
[18] See: https://incidentdatabase.ai/
[19] See: https://www.aiaaic.org/aiaaic-repository
[20] The Association of Southeast Nations Secretariat, "Priority Areas of Cooperation", *ASEAN Secretariat,* n.d. https://asean.org/our-communities/economic-community/asean-digital-sector/priority-areas-of-cooperation/
[21] UK AI Security Institute, "International consensus and open questions on AI evaluations", *UK AI Security Institute* blog, 12 February 2026, https://www.aisi.gov.uk/blog/international-ai-network-consensus-and-open-questions

- **Crisis preparation** (such as sharing information about risk mitigation measures, appointing an authority for crisis response, and stress-testing readiness mechanisms through practice and learning from incidents)
- **Crisis response** (including deploying technical interventions, preserving operational continuity and coordinating responses across borders)

Table 1. Global patchwork of preparedness initiatives

Aim	Shared architecture (minilateral, regional or international level)	Other measures (sectoral or non-governmental)	Status of activation or readiness for an AI-related crisis
Monitor emerging risks	International Network for Advanced AI Measurement, Evaluation and Science (formerly the International Network of AI Safety Institutes) and the International AI Safety Report. UN Independent Scientific Panel on AI and its report.	Independent third-party evaluators such as METR and Apollo Research conduct pre-deployment capability and safety evaluations of frontier models. The UK AI Security Institute's open-source 'Inspect' evaluation framework.	**Operational.** INAAMES launched 2024, and now has 10 members (inc. EU) as of August 2026.[22] The UN Preliminary Scientific Report was published in June 2026.[23]
Share information about risk mitigation	OECD-G7 Hiroshima AI Process Reporting Framework (HAIPRF).	The AI Information Sharing and Analysis Center (AI-ISAC) is a public-private body coordinating threat intelligence on AI-specific risks across industry and critical infrastructure operators. The Coalition for Secure AI (CoSAI) publishes shared technical security standards across member companies.	**Operational.** HAIPRF builds on 2023 Hiroshima AI Process principles and commitments, and launched February 2025, with two reporting rounds and 25 reports as of August 2026.
Establish authority for crisis preparation and response	The EU AI Act requires providers of general purpose AI models to report "serious incidents" to the EU's AI Office from August 2026: the closest mechanism to a supranational crisis-authority mechanism currently in force, though regional rather than international.	Some national-level movement: China has a national plan for AI crisis response with designated authorities.[24]	**Not operational.**
Coordinate		While there is nothing	**Not operational**.

[22] However, proposals to formalise a crisis response mandate for the network remain just that: proposals.
[23] United Nations Independent International Scientific Panel on AI, *Preliminary Report*.
[24] Emily Grundy and Greg Sadler, "China has an AI crisis plan. Australia should, too", *Australian Strategic Policy Institute*, 16 February 2026, https://www.aspistrategist.org.au/china-has-an-ai-crisis-plan-australia-should-too/.

responses to crises across borders		specific to AI and crisis coordination at an international level yet, some regional or supranational AI initiatives focus on coordination and coherence between member states, such as the ASEAN Responsible AI Roadmap and, the EU AI Act.	
Preserve operational continuity		At a sectoral, national and regional level, some networks are in place for responses to cyber threats, e.g. regional Computer Emergency Response Teams (CERTs).	**Partially operational**.
Deploy technical intervention	The OECD Catalogue of Tools and Metrics platforms emerging best-practice at an international level. Project Glasswing reportedly enables its participating organisations - although they appear to be predominantly US-based - to use Anthropic's Claude Mythos Preview to secure their critical systems.[25] The US-based Frontier Model Forum (FMF) facilitates information- and best-practice-sharing about frontier AI risks, security.[26]	In financial services, the Reserve Bank of India's draft Model Risk Management guidance[27] would require banks and non-bank financial companies to build "kill switches" and mandatory human override into their AI systems. In Singapore, the Infocomm Media Development Authority (IMDA) Model AI Governance Framework for Agentic AI[28] recommends voluntarily technical controls including rollback mechanisms and least-privilege access, for autonomous AI agents.	**Operational.** The OECD catalogue is available and public-facing, but does not coordinate the deployment of technical interventions. Glasswing and the FMF are operational, but details about the technical interventions they coordinate is not public-facing.
Learn from incidents	OECD incident reporting framework. Open-source AI incident databases and repositories.		**Operational.** The frameworks and databases noted are publicly available and enable learning from incidents.
Stress-test responses through scenario-based		Backdoors and Breaches (created by Black Hills Information Security).[29] The US Cybersecurity and	**Not operational**. The exercises included demonstrate emerging approaches to AI crisis

[25] See: https://www.anthropic.com/glasswing
[26] See: https://www.frontiermodelforum.org/about-us/#scope
[27] Reserve Bank of India, "Draft Guidance on Regulatory Principles for Model Risk Management", *Reserve Bank of India, Mumbai*, 2026, https://www.rbi.org.in/Scripts/bs_viewcontent.aspx?Id=5089
[28] Infocomm Media Development Authority, "Model AI Governance Framework for Agentic AI. Version 1.0", *Infocomm Media Development Authority*, Singapore, 2026, https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/mgf-for-agentic-ai.pdf
[29] See: https://home.backdoorsandbreaches.com/

exercises		Infrastructure Agency / Joint Cyber Defence Collaborative AI Security Incident Tabletop Exercises.[30] Think tank exercises, e.g. RAND[31] and Alan Turing Institute.[32]	simulation, but there is no recurring international architecture that *systematically* tests governments, AI developers, cloud providers and other critical actors together.

There is fragmentation (at worst) or a lack of coordination (at best) in how countries, companies and other actors are preparing for a future AI crisis. Against this backdrop, middle powers are constrained, but not paralysed: they can still advance preparedness at an international, regional or cross-border level, although the barrier to doing so is high. In light of the capability asymmetry detailed above, middle powers' comparative advantage lies in pooling expertise, coordinating through regional institutions and establishing common expectations for information sharing and response.

As observed in other areas of governance - like global finance and cybersecurity - scenario exercises can connect otherwise separate preparedness functions. By encouraging participants to interpret incomplete evidence and make time-sensitive decisions, they can reveal the efficacy - or lack thereof - of the operational systems at work. They can expose missing relationships or unclear mandates. When they are designed intentionally and deployed repeatedly, exercises can help turn broad agreement - for example, on the need for safeguards - into practical capacity.

In AI governance, exercises are certainly not a substitute for technical evaluation, incident reporting, or other foundational tools, but instead a means of testing whether these tools can be mobilised collectively when systems and decision-makers are under stress.

[30] US Cybersecurity & Infrastructure Security Agency, "Joint Cyber Defence Collaborative (JCDC) Artificial Intelligence Cyber Tabletop Exercise Series", *CISA*, n.d., https://www.cisa.gov/topics/partnerships-and-collaboration/joint-cyber-defense-collaborative/Joint-Cyber-Defense-Collaborative-Artificial-Intelligence-Cyber-Tabletop-Exercise-Series

[31] See Section 2.

[32] Sam Stockwell, Ardi Janjeva, Broderick McDonald, *Adding Fuel to the Fire: AI Information Threats and Crisis Events*, The Alan Turing Institute's Centre for Emerging Technology and Security, February 2026, https://cetas.turing.ac.uk/publications/adding-fuel-to-fire.

About this report

- This policy report explains how scenario exercises[33] are central to building global preparedness for an AI-driven crisis.
- It highlights lessons about utilising scenario-based work for building awareness, networks and capacity from other areas of global governance.
- It also scans how governments, companies and technical experts are currently using scenario exercises to develop crisis-ready AI diplomacy.
- Reflecting on the design and outcomes of a tabletop exercise - designed and delivered by the report's contributing organisations in the sidelines of the UN Global Dialogue on AI Governance and Geneva Digital Week - it proposes a set of policy recommendations for empowering scenarios-as-a-tool for strengthening readiness mechanisms in regional and multilateral AI governance.
- Crisis preparedness is an inescapably cross-border and multi-stakeholder endeavour: while this policy report is intended for a policymaker and multilateral institutional audience, it also bears relevance for AI companies, businesses and technical experts from civil society, research institutes and other sectors.

[33] This report uses scenario exercises as an umbrella term. This includes simulation exercises, tabletop exercises and wargames.

Focus: key precedents

Scenario-based work - ranging from simple tabletops to multi-day wargames - has been a hallmark of strategic preparedness for centuries. Many governance domains closely related to AI have been defined by a cascading global crisis: whether in prospect or in reality. There are valuable lessons to be gleaned from how scenario- and future-focused work features in crisis preparation strategies across cybersecurity, global finance and maritime security.

These three, more mature and interdependent governance areas provide helpful analogues: each has decades of institutionalised scenario practice to draw design lessons from, though comparable exercise traditions exist in public health, nuclear security and aviation.

Across all three governance areas, scenarios make preparedness testable by revealing missing authority and information, and allowing different actors to rehearse coordination before a crisis. Importantly, their value is not in the exercise alone. The examples in *Table 2* depend on follow-through, assigned responsibilities for operationalising lessons, and repeat testing. Crucially, in each, preparedness strategies are informed by collective, transboundary experiences of crises.

Table 2. Scenarios in other governance areas

Domain	Example	Purpose
Cybersecurity	'Red teams' simulate a fictional-but-feasible cyberattack against an organisation's computer system; 'blue teams' respond to it.[34]	Improve attack readiness by identifying and patching vulnerabilities, and by testing emergency responses and coordination.
Financial stability	Key stakeholders across the financial sector participate in a simulation exercise to test market resilience to disruption.	Advance capacity for collective response, inform the design of crisis response tools and build familiarity between key stakeholders.
Maritime security	Port, maritime and security agencies use tabletop and operational exercises to work through fictional incidents affecting ships, ports and connected supply chains.[35]	Test security plans, clarify public-private roles, expose communication gaps and rehearse timely cross-border and inter-agency responses.

[34] Naming conventions vary: while US and NATO exercises traditionally assign 'Red' to the adversarial force and 'Blue' to the friendly one, Chinese military doctrine reverses this, assigning 'Red' to the friendly one and 'Blue' to the adversarial force.
[35] International Maritime Organization, "Global Maritime Security Integrated Technical Cooperation Programme", *IMO*, n.d., https://wwwcdn.imo.org/localresources/en/OurWork/Security/Documents/IMO%20Global%20Maritime%20Security%20Capacity%20Building%20-%202%20pager.pdf

Strategic value of AI scenarios

There is appetite for and momentum behind using scenario exercises to build preparedness for an AI-enabled crisis. In the face of unpredictable and cascading risks, scenarios may have strategic value and a measurable impact as part of AI crisis preparedness protocols. However, appetite and momentum do not automatically translate to resources, institutions and political prioritisation. State participation in scenario exercises does not automatically translate to stronger diplomacy, networks and information-sharing. Work is needed to connect practice to policy change.

Approaches to AI scenarios

Current AI scenario practice spans four broad approaches, although the field and its practice are ever-evolving.

- **Narrative foresight** develops multiple plausible futures in order for policymakers to test whether strategies remain robust across different technological, economic, social and geopolitical conditions. An example includes exercises run by Singapore's Centre for Strategic Futures, located in the Prime Minister's Office.
- **Strategic role play** - such as 'Intelligence Rising' by Technology Strategy Roleplay (TSR) and RAND's 'Day After [Artificial General Intelligence] AGI' exercises - places participants inside competitive or crisis dynamics to surface coordination challenges, escalation risks and difficult trade-offs between policy options.[36][37]
- **Operational tabletop exercises** go even further by asking participants to use *existing* authorities, evidence, information channels and technical capabilities in response to a concrete incident: as an example, RAND Europe, the UK AI Security Institute, and Mila partnered on three tabletop exercises on AI safety and cyber misuse for senior officials from France, Germany and the Netherlands.[38] The design was grounded in frontier risks identified by the UN's 2026 International AI Safety Report.

[36] RAND Global and Emerging Risks, "Day After AGI Games", *RAND*, n.d., https://www.rand.org/global-and-emerging-risks/centers/geopolitics-of-agi/focus-areas/day-after-games.html
[37] See: https://www.intelligencerising.org/
[38] Afek Shamier, Henri van Soest, Stephen Clare and Sana Zakaria, *Insights from table-top exercises in Europe on AI safety and cyber misuse*, RAND, 1 July 2026 https://www.rand.org/pubs/research_reports/RRA5082-1.html

- **Synthetic exercises** use computational agents - increasingly built on large language models - to rehearse scenario dynamics. For example, a 2024 Stanford study had language-model agents play out great power crises.[39] This approach can run far more iterations than a human-in-the-room exercise allows, but is a complement, rather than a substitute for human-led formats: 'believable' agent behaviour does not guarantee realistic macro-level outcomes, and simulated agents have been observed to escalate in sudden, hard-to-predict ways.

These approaches carry different types of strategic value for preparedness: with foresight expanding the range of futures considered, roleplay making strategic interactions tangible, tabletops testing whether institutions can act and synthetic exercises stress-testing scenario dynamics at a scale and speed no human exercise can match.

However, the field remains weighted towards capability jumps at the technological cutting-edge, great power competition and national security settings. Far fewer exercises examine crises across other sectors and domains: ones that may *become* catastrophic in nature in the absence of further capability jumps. Currently under-explored dynamics include widespread adoption, interacting systems or dependencies in countries that cannot influence or mediate frontier development.

Policy uptake

The 2026 UN Global Dialogue on AI Governance featured multiple scenario exercises attended by government representatives, indicating interest if not explicit support for scenarios-as-a-tool for readiness. The UN's recently established AI Governance for Humanity Lab's report on advancing interoperability in AI governance spotlights the value of scenario exercises "on the sidelines of global dialogues to [stress test] emergency backchannels and response mechanisms among regional policymakers".[40]

The scenario format has promise as a complement to traditional policy tools, but international uptake is uneven and unclear on two levels.

On one level, while there is growing trust in scenarios-as-a-tool, they are not a fixture of international gatherings on AI. International AI summits have produced

[39] Juan-Pablo Rivera, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith and Jacquelyn Schneider, "Escalation Risks from LLMs in Military and Diplomatic Contexts," Stanford HAI policy brief, 2 May 2024, https://hai.stanford.edu/policy/policy-brief-escalation-risks-llms-military-and-diplomatic-contexts

[40] AI Governance for Humanity Lab, *Toward interoperability*

safety commitments, evaluation networks and transparency frameworks; the UN Global Dialogue created a uniquely inclusive forum with a diversity of national and regional perspectives. Yet formal processes still devote limited attention to rehearsing cross-border incident response. Exercises risk remaining side events or one-off research activities, disconnected from international institutions and standing budgets.

On another level, the implementation of recommendations emerging from the patchwork of scenario exercises captured above is unclear. The central gap is the absence of repeatable mechanisms that convert scenario insights into assigned responsibilities, protocols and stronger regional-to-multilateral coordination. There is also a lack of trusted monitoring, evaluation and learning (MEL) indicators for gauging the impact of scenarios improved capacity on policy variables, like changes in behaviour (such as better and more sustained collaboration). Closing this gap is partly a matter of design: exercises should be built as data-generating research instruments, not only as events, generating structured and comparable measures round by round rather than relying on post-hoc impressions to guide future work.

Case study: tabletop exercise on regional AI crisis diplomacy

On 7 July 2026, alongside the inaugural UN Global Dialogue on AI Governance in Geneva, Chatham House and AI Safety Asia, in collaboration with Université de Genève, convened a tabletop exercise on diplomatic responses to a regional AI-enabled incident. The exercise was developed with consultation from Technology Strategy Roleplay.

This sub-section explores what the exercise's findings may suggest about crisis preparedness on regional and international levels. It also examines different design features intended to contribute to durable institutional and policy change. In order to protect the independence and integrity of future exercises using the same material, some exercise details have been omitted.

Design

Storyline

The exercise was designed around a fictional-but-feasible disruption to global shipping and trade. It reflects the technological state-of-play as of July 2026: increasingly autonomous AI with agentic capabilities, diffusion across sectors, and insufficient governance oversight, not to mention concerns about the malicious use of AI for disrupting critical infrastructure.

A port- and maritime logistics-based exercise offered an interdependent system with clear regional-to-global spillover consequences, while remaining accessible to participants without specialist AI or shipping knowledge. In order to preserve the integrity of future exercises using versions of the same storyline, this report intentionally does not comment on the specific trigger and signs of escalation.

The scenario did not depend on a speculative leap to AGI. Instead, it preserved uncertainty around the incident's origins, scale and likely trajectory. This enabled the exercise facilitators and participants to explore how decision-makers respond when familiar explanations remain plausible but incomplete, and when disruption - with unpredictable and cascading consequences - may arise across increasingly automated systems.

Participants

Convened by two expert facilitators and structured around two fictional taskforces, the exercise included over thirty participants (including convenors and observers): twenty-six total active participants, with ten in the North Atlantic taskforce and sixteen in the Asia-Pacific taskforce.

The exercise was multi-stakeholder: participants hailed from governments, international organisations, research institutions, technical organisations and industry.

There were no assigned roles in the taskforces: participants were encouraged to bring their own professional expertise to the table and represent themselves.

Structure

The taskforces experienced three crisis phases - trigger, escalation and global emergency. They also participated in regional briefings and a joint debrief.

Information was introduced incrementally, with each stage widening the geographic, economic and institutional implications of the incident. The exercise was designed to require participants to revise their assumptions as new evidence emerged, rather than identify a predetermined solution.

Facilitators controlled the pace through timed injects (pre-scripted pieces of information - such as fictional news bulletins or intelligence updates - released at set points to drive the scenario forward) and prompts, while allowing the discussions to remain participant-led.

Interactions

Facilitators ensured conversations were inclusive. Decisions - for example, on whether to request additional information - were made by consensus.

The exercise allowed for inter- and intra-group interactions, and unplanned adaptations. Delivery remained deliberately flexible. When the Asia-Pacific taskforce chose to appoint an envoy to the North Atlantic group and share information, facilitators allowed the move rather than treating the task forces as sealed. This unplanned adaptation suggested that in the absence of a formal cross-regional mechanism, participants improvised liaison and information-sharing. It also demonstrated why exercises need enough flexibility for participants to create, rather than merely discuss, coordination arrangements.

Information

Taskforces could request evidence from a range of stakeholders, including governmental, technical, commercial and international sources. No actor held a complete picture. This tested whom participants trusted, which channels they activated and whether they shared information across regions.

Facilitators enabled open-ended discussions about the information received whilst controlling the tempo through the injects and prompts.

Participants recorded their diagnostic clarity, level of concern and expected economic impact on visual scales at the start of each exercise phase. These measures were designed as indicators of how perceptions changed throughout the exercise. The joint debrief then surfaced practical recommendations and compared the two taskforces' approaches.

Outcomes and recommendations

The exercise's principal objective was to stress-test existing regional mechanisms for AI crisis response. It aimed to identify gaps in preparedness, engaging a broad range of decision-makers from across sectors to do so. It also served as a case study on the potential value of scenario exercises as complements to multilateral gatherings.

Following the exercise's delivery and evaluation, facilitators identified a set of key outcomes which can shape future capacity-building and awareness-raising efforts by participating organisations. They indicate promising pathways for connecting scenario-based work with the advancement of crisis readiness policies at the national, cross-border, regional and multilateral levels.

Improved awareness

Across both taskforces, participants deliberated on several plausible explanations for the disruption and revised their assessments as further information became available. The exercise broadened the discussion beyond a single technical or security diagnosis towards the systemic risks created when AI-enabled services become embedded across interconnected infrastructure.

This collective diagnosis and improved awareness of severe AI-enabled threats suggests further work is necessary, especially as frontier capability becomes more widely diffused. Decision-makers need to be able to distinguish between different sources and forms of AI-related disruption, and to act before attribution or causation is fully resolved.

Rising concerns

Throughout the exercise, participants' diagnostic clarity, perceived economic impact and level of concern increased, reflecting a growing appreciation of the systemic implications of AI-enabled disruption.

Crucially, increased understanding (diagnostic clarity) did not reduce individuals' reported levels of concern. This suggests that participants increasingly appreciated the complexity of responding to AI-enabled crises, even as they developed a better explanation of the underlying problem.

Governance analogues

During the debrief, participants drew from existing crisis response and risk mitigation measures from other governance areas, including maritime safety, cybersecurity and finance. Ideas included setting up risk monitoring frameworks, pre-agreed taskforces and channels, reversion to manually-operated systems and channels and business continuity, independent audits, transparent incident reporting and regulatory sandboxing.

Impact measures

There is an urgent need for trusted measures in order to evaluate and learn from the exercises at different levels: on an individual basis (like changes in decision-maker behaviour in crisis, not just their *awareness* of potential crisis risks), cross-sector basis (such as changes in network strength) and on a regional or international level (like multilateral institutional resourcing for new preparedness functions).

Complexity and breakdown

The exercise underlined the value of focusing on a networked system of systems - rather than an individual model or actor, evaluated in isolation.

This draws on complexity science. Cascading failures in interconnected networks typically emerge from interactions among components, none of which individually breaches a rule. This is why fixed thresholds, clean attribution and early warning all break down at the network level.

The exercise's maritime logistics scenario - a canonical complex adaptive system - produced a matching finding: diagnostic clarity and concern rose together as participants came to understand the system's interdependence, just as complexity theory would predict.

Continued interest

Post-exercise surveys indicated high agreement that regional coordination is important for AI incident response at regional and international levels, that better preparedness is needed, that standing regional coordination mechanisms would be valuable and that the discussions were directly relevant to participants' professional responsibilities.

Many respondents also volunteered to contribute to the written outputs following the exercise, whilst several participants expressed interest in continued collaboration.

The debrief surfaced several recommendations for improving regional diplomacy in the event of a severe transboundary AI-enabled crisis, analysed and streamlined by the report's authors and several expert contributors. The recommendations firmly point towards preparedness as a coordination challenge, rather than solely a technical capability issue.

Table 3. Participant recommendations on crisis diplomacy[41]

'Asia-Pacific Taskforce' team		
Recommendation	**Purpose**	**Key issues raised**
Risk and monitoring framework	Provide a common structure for detection, escalation and response that includes advanced AI capabilities.	Incident monitoring; escalation thresholds; coordinated roadmap for addressing risks; manual fallback; sector-specific risk mapping.
Coordinated crisis taskforce	Create a mechanism for preventative coordination and emergency response.	Use or adapt existing structures such as the International Maritime Organisation (IMO) and maritime coordination bodies; identify a lead convenor; identify regional-to-international links between bodies.
Cross-sector learning	Apply lessons from crisis to multiple sectors (beyond the exercise focus, maritime logistics).	Consult sectors undergoing large-scale AI adoption; test agent interaction, human oversight and systemic dependency in other domains.
'North Atlantic Taskforce' team		
Recommendation	**Purpose**	**Key issues raised**
Business continuity	Maintain essential operations (back up options) when automated systems fail.	Minimal viable tools and information; manual or degraded-mode operation; human in the loop protocols.
Regulatory sandboxing and pre-deployment testing	Test how interacting systems behave under stress before full deployment.	System interaction tests; realistic disruption scenarios; examination of actor type and sector context; red teaming.
Liability and accountable ownership	Ensure oversight responsibility is assigned for automated workflows.	Clear legal responsibility for firms and humans; empowered operational owner; authoritative source of accurate information in crisis.
Transparent incident and threat reporting	Reduce uncertainty and improve collective diagnosis.	Threat-level disclosure; full incident reporting; independent audit; clear emergency contacts.
Vertical and horizontal (cross-sector) coordination	Connect national, regional and sectoral response layers.	Government-to-government links; port authorities and industry coordination; cross-sector working groups, including in critical national infrastructure, finance, etc.

[41] Analysing the differences in the nature and scope of regional taskforce recommendations is beyond the scope of this report. Future work may consider how membership of a regional taskforce may impact policy optionality, measured based on factors like existing institutions, networks, etc.

Middle power agency

Many AI scenario exercises centre on near-future AI capabilities, with the US and China as the key players. This focus is essential: it reflects the global asymmetry in capabilities and difficult realities, like coordinating on crisis readiness across a geopolitical chasm and amid unpredictable frontier capabilities.

However, making the US and China the default protagonists narrows the preparedness agenda. It risks leaving other states and actors rehearsing roles in a crisis architecture they did not help design.

The Geneva exercise demonstrated the complementary value of middle power-focused work. Participants confronted a potentially debilitating risk with potentially catastrophic consequences: one enabled by the growing dependence of their economies on widely diffused and non-sovereign AI-enabled systems embedded across critical infrastructure. This shifted the question from *how* middle powers influence great power competition to *what* they can build and do themselves: shared monitoring, regional notification, continuity arrangements and collective leverage with providers.

The purpose of adopting a country-agnostic, regional approach was not to exclude particular states, but to create space for open-ended discussions about regional preparedness and coordination. Regional exercises can expand agency and develop practical building blocks that later connect to wider multilateral coordination.

Recommendations: scenarios for preparedness

AI scenario exercises should be treated as permanent but evolving capacity-building infrastructure, not as one-off convenings or static tools for collecting policy ideas. Their design process should begin by identifying the specific preparedness function that needs to be strengthened and work backwards to practice: the decisions, relationships, information, institutions and politics encountered by participants.

Developed for stakeholders designing and integrating scenario exercises into AI crisis readiness, *Table 4* identifies six preparedness functions that future exercises can build and specific ideas on how to do so in their design. Each includes measures and instruments, and each should have a named institutional owner, a post-exercise implementation plan and a timetable for repetition. A valuable exercise creates controlled friction around the precise mechanisms that determine how and whether participants can respond. The design features below can be adapted to different sectors and regions to maximise impact.

Regional exercises can generate interoperable practices and evidence that feeds upwards into multilateral processes, using summits to improve preparedness. This demands rethinking summits as policy platforms, not temporary windows, and investing in continuous intersessional work. To this end, *Table 4* includes a dedicated section on how and where to scale capacity.

Table 4. Stress-testing AI crisis preparedness functions

Component	Design guidance and capacity implications
1. Risk monitoring and escalation thresholds	
How to stress-test	Use conflicting signalling injects (for example, an early technical report setting out an incident's severity level, alongside a whistle-blower account contracting it), operational indicators and staged escalation thresholds before the cause is known. Require participants to decide when an anomaly becomes an incident, what evidence justifies broader notification and which risk indicators should be monitored continuously.
What AI changes	AI-enabled failures may emerge from unexpected or ungoverned interactions among models, agents, data and infrastructure without a single actor or system clearly breaching a rule. But failure can emerge through malicious (mis)use, too. Widespread AI integration can amplify existing vulnerabilities in systems, increasing the likelihood of failure and the gravity of a technical malfunction in just one part of the network. A clear picture of risk will be more difficult to gain: early signals may be weak, information about them may be proprietary or difficult to access, not to mention distributed across providers and sectors. Capability changes can make fixed risk thresholds obsolete, particularly during hybrid conflicts. Escalation and broader notification may therefore be necessary before attribution or causation is settled.
How and where to build and scale capacity	National AI authorities or equivalent bodies, sector regulators, cybersecurity agencies, Computer Emergency Response Teams (CERTs) and providers should define AI-specific risk indicators and provisional notification thresholds, secure access to relevant telemetry and operator data, and maintain shared regional risk watchlists. Repeated scenario exercises should review thresholds as systems and deployment patterns change. International organisations can support semantic, technical and institutional interoperability between approaches to risk monitoring, while regional bodies adapt the tools above to national and local contexts.
2. Information-sharing and a common operating picture	
How to stress-test	Distribute information asymmetrically across government, industry, technical and regional actors. Make some evidence available only upon request. Do not provide participants a list of possible information sources: encourage them to identify potentially valuable sources independently.

What AI changes	Rapidly evolving incidents leave little time to negotiate access or overcome interoperability hurdles to information- and evidence-sharing, and no single actor may hold a complete operating picture. Critical evidence may sit in proprietary model logs, evaluation results, cloud telemetry and downstream operational data. It may be commercially sensitive, security-classified or unintelligible to non-technical decision-makers.
How and where to build and scale capacity	Governments and regional bodies should establish flexible, pre-authorised emergency data-sharing frameworks, where possible, with providers and critical-infrastructure operators to enable fast, voluntary exchanges during an incident. This should define minimum incident data fields, designate technical and policy liaison points, and pre-authorise access to relevant logs and telemetry under clear confidentiality rules. Recurring exercises should test actual permissions, response times, communication between technical and policy decision-makers, and the handling of sensitive intelligence or commercial information.
3. Crisis authority and regional coordination	
How to stress-test	Assign participants to regional taskforces with explicit but imperfect mandates. Require them to select a convening location or institution and determine when to escalate to political leaders or international bodies. Run two concurrent regional taskforces (i.e. in the same room), but do not clarify whether or how they can contact each other.
What AI changes	Capacity to respond is distributed across model developers, cloud providers, deployers, sector operators and public authorities, often in different or multiple jurisdictions. A government may recognise the crisis but lack the legal reach to intervene, while a provider may possess the technical remedy but not the public mandate. An AI-enabled crisis may generate unprecedented epistemic uncertainty and mistrust in institutions traditionally tasked with crisis response. Existing emergency powers and sectoral structures may not map cleanly onto this landscape.
How and where to build and scale capacity	Governments should nominate a lead domestic authority and deputies, establish an inter-agency AI incident cell, agree escalation protocols and regularly review them with providers and sector operators. Regional organisations should maintain rapidly activatable taskforces, contact directories and cross-border notification procedures. On a regional or multilateral level, memoranda of cooperation and a small funded secretariat can preserve institutional memory, convene exercises and track follow-through.

4. Technical response and system-level testing	
How to stress-test	Force choices between pausing systems, changing incentives, isolating components, rolling back updates or maintaining degraded operation. Demonstrate that simply ‘pulling the plug’ will not mitigate the crisis. Encourage participants to deliberate on the technical and political feasibility of ‘kill switches’. Represent multiple AI providers and downstream deployers rather than a single controllable platform.
What AI changes	AI services are embedded in interconnected and globally diffused systems and may depend on shared models, data, cloud infrastructure and agent-to-agent interactions. Much technical infrastructure relies on automated decision-making processes and freshly integrated AI tools: whether basic or agentic. A local intervention can shift demand, move failure elsewhere or introduce vulnerabilities. The relevant unit of analysis is therefore the networked system of systems, not an individual model or ‘product’ evaluated in isolation.
How and where to build and scale capacity	AI safety institutes or equivalent technical bodies, providers, sector regulators and critical-infrastructure operators should develop shared test environments and emergency playbooks for pause, rollback, isolation and degraded operation. This requires provider participation, independent technical expertise, sufficient algorithmic transparency, access to logs and interfaces, and recurring cross-provider stress tests to gauge system-wide effects.
5. Business continuity and human override	
How to stress-test	Simulate the loss or suspension of different AI tools after organisations have become dependent on them: ranging from simple automation to complex agents. Ask participants to define minimum viable operations, identify the data and staff required to action it, and decide who has authority to override AI-driven workflows.
What AI changes	AI integration can erode manual skills, remove access to underlying data and concentrate dependency on a small number of providers. A nominal human-in-the-loop may lack the information, authority or practical ability to intervene: there is no automatic guarantee of accuracy and efficacy in crisis response.
How and where to build and scale capacity	Sector regulators and operators should set minimum service and recovery requirements, maintain manual or alternative workflows, identify staff with real override authority, and retain access to backup data and infrastructure. Procurement and continuity standards should cover portability, skills retention, regular degraded-mode drills and tested restoration criteria.

6. Accountability, communication and institutional learning	
How to stress-test	Require decisions on liability, public communication, incident reporting and evidence preservation while the evidence remains uncertain. End with a debrief to international authorities and include an after-action process assigning each gap to an institution, deadline and resource commitment; repeat the exercise after reforms.
What AI changes	Complex supply and value chains underlie an 'AI threat' as observed in a model, system or application. This makes attribution complicated: causation and responsibility may be distributed across developers, deployers, data suppliers and infrastructure operators. Relevant evidence can be proprietary, dynamic or difficult to reproduce. AI-driven mis- and disinformation can generate uncertainty and complicate public communication during crisis. Traditional institutional reviews may be too slow to keep pace with evolving systems and deployment patterns.
How and where to build and scale capacity	Governments should pre-assign legal, communications and evidence-preservation responsibilities, require timely incident reporting and independent after-action review, and track implementation against named owners and deadlines. Regional secretariats or independent evaluators can aggregate comparable lessons, fund repeat exercises and feed findings into standards bodies, ministerial processes and UN dialogues. On the multilateral stage, there are two major upcoming opportunities to facilitate AI crisis preparedness: at the Geneva AI Summit (June 2027) and alongside the UK's G20 presidency (2027). Organisers of the former should consider setting up a permanent track dedicated to stress-testing preparedness functions, with tie-ins to regional secretariats and technical institutes outlined above.

A bird’s eye view

There is no one-size-fits-all formula to scenarios and maximising their impact. But a function-based approach offers an opportunity for strategic sequencing, regional tailoring and leveraging windows of opportunity, like resourcing and political attention.

Scenario programmes should sequence the functions included in *Table 4* rather than attempting to test everything in one session. Early exercises can map actors, authorities and information routes; later iterations can test agreed protocols, simulate larger capability jumps and compare performance over time. Standardised measures for impact, evaluation and learning will facilitate this.

The six functions in *Table 4* are part of a crisis infrastructure. They describe what institutions must be able to do ahead, during and after a crisis: but they do *not* determine when an institution is actually fit for purpose. Post-exercise debriefs must also focus on institutional suitability: AI alignment research offers the RICE principles - robustness, interpretability, controllability, ethicality - which transpose naturally to crisis institutions, and give exercise designers a scoring lens for debriefs.

The aim is a cycle of exercise, reform and re-exercise that builds institutional memory and creates a credible pathway from regional preparedness to multilateral coordination.

It is hopeful that there is a growing wave of organisations - from government, the private sector, the technical community and civil society - experimenting with scenarios as a readiness tool. There is a need for improved coordination *among* scenario convenors to reduce the risks of duplication, share information on design, standardise impact measures, and co-strategise in advance of regional and multilateral gatherings.

AI scenario database

Looking ahead, the report's authors are scoping an open, standing repository for AI scenario exercises: a shared platform where governments, technical experts, industry and civil society could contribute case studies, methodologies and lessons from their own scenario work. The aim is to lower the barrier to entry for new exercise designers and reduce duplication across a fast-growing but fragmented field. It would provide AI scenario practice the kind of standing institutional home this report argues crisis preparedness infrastructure more broadly still lacks. Work to scope the platform's feasibility, governance and hosting arrangements is ongoing.