AI-Proliferation and Middle Powers: Preparation and Response Mechanisms

Teague, James
Independent

Ali, Alina
Independent

Sfeir, Sophie
Independent

Fort, Kristina
Independent

Abstract

As the set of actors with access to highly-capable model weights grows, whether by theft, leakage, or deliberate release, some individuals or groups could obtain an uplift in dangerous capabilities. For middle powers, which typically lack meaningful sovereign AI capacity or influence over frontier AI development, this proliferation dynamic represents a structurally under-addressed vulnerability. Rather than continued efforts to shape frontier AI at the international level, we argue that national-level preparedness for these threats is a more tractable priority. We model three attack vectors: cyber-attacks, CBRN threats, and influence operations, across a four-level hazard-to-crisis escalation framework, and present a structured preparedness checklist for detecting threats, escalating them through appropriate institutional channels, mitigating them before they cause harm, and containing them when they have.

Executive Summary

As the set of actors with unrestricted access to advanced AI models grows, some individuals or groups could obtain an 'uplift' in dangerous capabilities beyond what conventional tools already enable (Vaccaro et al., 2026). For the US and China, this is a relatively low priority risk, as they continue to heavily dominate the development frontier AI. By contrast, middle powers have limited influence over, or access to, frontier models, and are often left responding to these advancements. Rather than trying to contest this imbalance, a more strategic priority for middle powers is to focus on their **national-level preparedness for proliferation risks.**

This paper conceptualises middle powers as possessing some ability to prepare for, and respond to these threats through their *capacity depth*, *governance orientation*, and *infrastructure posture*. It explores how malicious actors could utilize advanced AI models for cyber-attacks, CBRN threats, and influence operations, showcasing how these attack vectors could materialise as different 'trigger-events' within middle powers. Finally, it adopts a three-part framework (Somani et al., 2025) to provide recommendations for middle powers to build their capacity to detect, escalate, and respond to various proliferation risks, presented as a preparedness checklist.

TABLE 1

AI Proliferation Preparedness Checklist for Middle Powers

Key actions across Detection, Escalation, and Containment & Mitigation

Detection	**Institutionalise a dedicated foresight function connected to security agencies and AISI-equivalent bodies.** *Enablers: Interdisciplinary Expertise · Technical Talent · Inter-governmental* **Establish continuous monitoring of the threat environment against defined indicators.** *Enablers: Intra-governmental Coordination · Supporting Infrastructure · Technical Talent* **Build verification capacity sufficient to confirm that a specific model or actor poses a threat that meets the threshold for escalation.** *Enablers: Technical Talent · Public-Private Partnerships · Supporting Infrastructure*
Escalation	**Develop and publish a national AI incident severity schema with defined response obligations at each tier.** *Enablers: Technical Talent · Research Output · Inter-disciplinary Expertise* **Legislate mandatory AI incident reporting requirements for AI operators within jurisdiction, with defined time frames.** *Enablers: Institutional Architecture · Net Importer / Developer / Supporting Infrastructure* **Pre-specify cross-sectoral escalation pathways connecting AI incident response to critical infrastructure agencies.** *Enablers: Technical Talent · Inter-disciplinary Expertise · Institutional Architecture*
Containment and Mitigation	**Build emergency response channels connecting technical experts, researchers, and private-public institutions with AISIs/government officials.** *Enablers: Inter-disciplinary Expertise · Public-Private · Institutional Architecture* **Invest in defensive capabilities.** *Enablers: Technical Talent · Research Output · Public-Private · Supporting Infrastructure* **Integrate contingency plans and technical intervention through pre-defined measures and crisis protocols as a way to mitigate harm in the event that a crisis cannot be resolved.** *Enablers: Institutional Architecture · Net Importer / Developer / Supporting Infrastructure*

Introduction

The development of advanced AI is heavily dominated by labs in the US and China (Sandoval et al., 2026). Rather than determining the terms under which AI is developed *a priori*, middle powers are limited to governing in response to deployments through *a posteriori* mechanisms. Nonetheless, their present efforts are primarily focused on influencing or contesting the great powers to steer the trajectory of this major technological transition, rather than strategically preparing for these changes (Leicht, 2025, 2026).

Instead of advising middle power governments on how to exercise greater influence over frontier AI development at the international level, such as through standards (Fort, 2024), diplomacy (Velasco et al., 2026), or developing (sub)frontier AI themselves (Verdi, 2026), this paper focuses on how middle powers could redirect a meaningful share of their efforts toward national-level preparedness. In particular, it evaluates how they should prepare and respond to proliferation risks: chemical, biological, radioactive and nuclear (CBRN) attacks, cyber-attacks, and influence operations that could be enabled by malicious actors gaining access to advanced AI.

By drawing on historical corollaries and existing research related to these attack vectors, this paper will provide a checklist of actions middle powers can take to detect and anticipate these threats, escalate them through appropriate institutional channels, mitigate them before leading to potential harm, and contain them when they have (Somani et al., 2025). We adopt a framework from The Future Society (2026) to discuss the different levels of AI-enabled threats, and provide recommendations for national-level preparedness at each level. By adopting these recommendations, governments will build their capacity to prevent and/or contain national crises, and will thereby be better positioned to participate and steer negotiations effectively when crisis response requires international coordination.

§1 Middle Powers and Their Vulnerabilities

In the international relations context, middle powers are defined as states that hold a 'middle' position in the international power spectrum, positioned below superpowers but with sufficient ability to shape international events (Muftuler Bac, 2025). However, this paper argues that middle powers are not in the position to shape the initial conditions of AI-driven transformation, and are instead limited to preparing for, and responding to, threats. Therefore, it draws on two complementary frameworks to redefine middle powers in the context of AI, and characterise how they differ in ways that matter for how they prepare, react and adapt to AI transformation. The first is the building blocks of AI power identified by the German Marshall Fund: human capital, compute infrastructure, energy, finance, regulation, connectivity, partnerships, and trust (de Hoop Scheffer et al., 2024). The second is the eight components identified by Chatham House: data, compute, advanced models, energy, industry, talent, infrastructure, and trust (Sandoval et al., 2026).

Together, these frameworks surface the dimensions along which middle powers most meaningfully diverge in their exposure to AI-enabled threats and their capacity to respond to them. Rather than treating all building blocks as equally weighted or collapsing them into fixed clusters, this paper identifies three axes that are most directly predictive of how AI-enabled threats materialise and how responses fail:

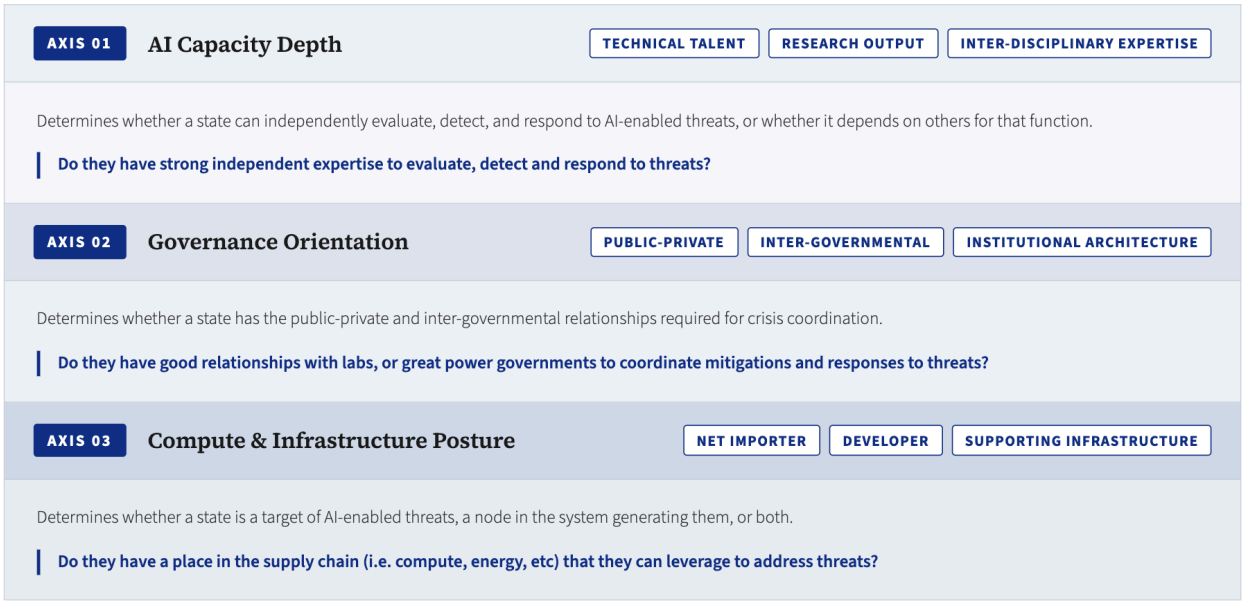

Table 2: The enablers to AI power.

The middle powers identified in the literature include the UK, Canada, France, Germany, Japan, India, Israel, Singapore, South Korea, Sweden, Saudi Arabia, the UAE, the EU, and Taiwan (J. Wang et al., 2024). They differ along these three axes, and the combination of their positions determines their specific vulnerabilities, and the enablers that matter most for their preparedness.

One particular dynamic to note is the presence of AI Safety Institutes (AISIs) among middle powers. Whilst states such as the UK, Canada, France, South Korea, Japan, South Korea, the EU, and Singapore have established AI safety bodies, and participate in the AISI International Network, others lack the equivalent institutions entirely, or have bodies with narrower mandates[1]. The checklist that follows reflects this differentiation, such that an action's feasibility depends heavily on a state's position along one of these axes, and the establishment of AISIs.

[1] See (U.S. Department of Commerce, 2024)

§2 Proliferation Threat Modelling

This section will first focus on the mechanisms through which malicious actors could gain unrestricted access to advanced AI models, or the 'proliferation problem'. It explains why the subsequent 'uplift' in dangerous capabilities beyond what conventional tools already enable should be of particular concern for middle powers (Vaccaro et al., 2026). This is followed by an exploration of three attack vectors that these actors could utilise to threaten the national security of middle powers, evaluating how advanced AI could lead to "an improvement in capabilities or outcomes that can be achieved with a given level of resources, or equivalently, the realisation of a given capability or outcome with fewer resources" (Rose, 2024).

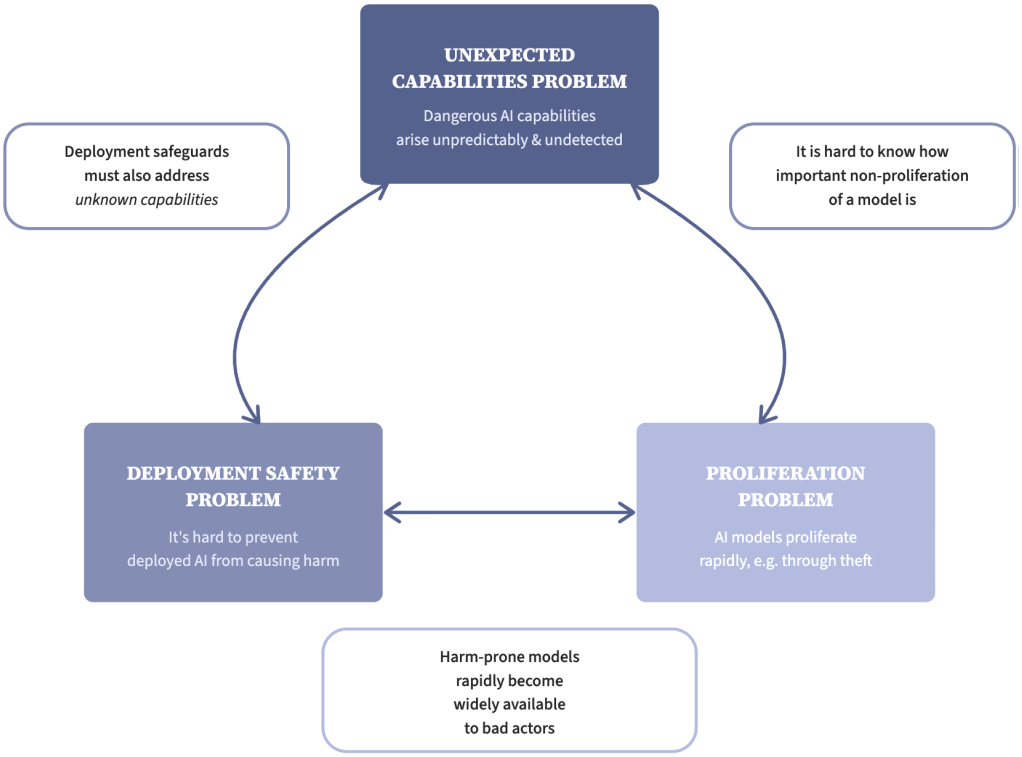

Figure 1: Summary of the three regulatory challenges posed by frontier AI. Adapted from GovAI (Anderljung, 2023)

2.1 Containment Failure

Model weights, the numerical parameters encoding a trained model's capabilities, represent an extraordinarily concentrated store of value. Their theft, inadvertent leakage, or deliberate release could enable actors to self-host advanced models without being subject to any monitoring arrangements or API restrictions (Nevo, 2024). To prevent dangerous AI models falling into the hands of malicious actors, and address the subsequent risks, three broad categories of actions are often proposed: access controls, embedded safeguards and defensive capabilities. However, their feasibility for middle powers are severely limited, leaving them disproportionately vulnerable.

The first proposed action is preventative: identifying a model's capabilities, and deciding whether weights should be publicly released, shared with select groups, or kept under strict containment. Frontier-weight access decisions currently rest on voluntary policy[2], with labs adopting weight security levels calibrated to capability evaluations (Nevo, 2024). However, many key figures in AI security remain sceptical that frontier labs are making sufficient progress on weights security to prevent theft and leakage (Dean, 2025), and some capabilities may only be discovered post-deployment, meaning open-weight releases or security strategies could be misguided.

A further concern is that, owing to uncertainties around the significance of compute, data, and 'algorithmic innovations' to capability scaling (Josephson, 2025), frontier capabilities may not remain concentrated in a small number of identifiable models whose weights can be restricted. As demonstrated by compute efficiency advances from DeepSeek and Google TurboQuant, frontier-level capabilities may become increasingly reproducible by actors without access to the original weights[3]. This could shift the governance burden from weights-access based on capability, to compute-access based on the shifting thresholds for what is sufficient for model training.

Middle powers have little to no input into the frontier-weight access decisions that determine their exposure to significant containment failures, leaving them with a much smaller set of feasible actions to prevent the proliferation of dangerous capabilities. Compute governance offers a set of downstream actions through which they can exercise some agency: requiring domestic cloud providers to implement KYC schemes that identify and verify significant GPU users (Egan & Heim, 2023), or strategically limiting domestic access to specialised models where uplift risk is high (Sastry et al., 2024). These measures cannot substitute for influence over frontier weight access decisions, but they represent tractable actions middle powers can take unilaterally to reduce the range of actors capable of exploiting a containment failure within their borders.

The second action is to ensure that models have embedded safeguards. Whilst commercial model deployments typically use safety filters and system prompts to identify and address harmful requests and outputs, these are not present when models are self-hosted (Özcan, 2024). Instead, developers use reinforcement learning (RL) and other post-training methods to fine-tune model weights toward non-harmful dispositions. Many other techniques have been proposed here, such as "machine unlearning", which attempts to selectively remove knowledge from trained models (W. Wang et al., 2024). The problem is that dangerous capabilities may only be discovered after model-weights are released, making it very difficult to design safeguards against them in advance. Further, even if labs conduct extremely rigorous safety training and filter out harmful training data; with access to the weights, there always remains a risk that malicious actors could tamper with the model's safety training or re-introduce hazardous datasets (Seger, 2023).

---

[2] For example, see Anthropic's RSP 3.0 (2026).

[3] See Meinhardt (2025) for detail on the algorithmic innovations behind DeepSeek, and Zandieh et al. (2025) on Google TurboQuant.

Whilst middle powers may have limited influence over which safeguards are used by frontier labs, understanding the resources an actor would require to engage in this kind of tampering is critical for them to grasp the relative threat level. If techniques for fine-tuning open-weight models with relatively small amounts of compute were discovered and shared widely, then a huge range of actors could gain access to dangerous capabilities with relatively low cost (Kemberry, 2024). Parameter-efficient fine-tuning methods such as LoRA have made it possible to adapt open-weight models on consumer-grade hardware with as little as 8GB of VRAM (Avinash, 2025), which could pose a significant risk, although this is far from the required amount to sufficiently fine-tune frontier-equivalent models for harmful purposes. Regardless, this faces the same problem as predicting the required compute for training models from scratch: it is unclear whether the costs to attain a meaningful capability uplift by fine-tuning open-weight models will be lower than the resources required for other kinds of attacks in the next decade.

Finally, the third set of actions is ensuring that, if weight security is breached and embedded safeguards fail, key actors have defensive capabilities that are sufficient to stop a containment failure from becoming a tangible threat. The dual-use nature of highly capable AI models is what makes this both essential and structurally difficult for middle powers: the same models that allow defenders to identify and address vulnerabilities simultaneously enable attackers to find and exploit them (Ee et al., 2025), meaning that advances in offensive and defensive capacity are structurally coupled and advance together.

In theory, middle powers are therefore severely limited by the robustness of their defenses, since they lack access to frontier models. For example, frontier model-weights being stolen or leaked represents a worst-case scenario for their governments, who will not have sufficient prior access to the model for building up proportionate defensive capabilities and identifying vulnerabilities before malicious actors can use the same model against them (Sandoval et al., 2026). However, whilst middle powers lack access to state-of-the-art defences outside of narrow bilateral agreements with frontier labs, investments in defensive capabilities have been found to be asymmetrically cheaper than offensive capability development for some attack vectors, such as AI-enabled biosecurity threats (Moulange et al., 2026), suggesting that a strategic focus on defensive acceleration, rather than attempting to match frontier offensive capabilities, represents a feasible path.

The rest of this section will evaluate three different attack vectors that a malicious actor could utilize following a containment failure, demonstrating how the proliferation of highly capable models could manifest in ways that threaten middle powers national security.

2.2 Cyber-attacks

AI-driven cyberattacks have the potential to overwhelm a country's core infrastructure. Unlike past cyber-threats that could be mitigated by human cybersecurity experts, advanced AI models are characterised by machine intelligence speed, cyber-expertise, and real-time adjusting abilities (Walden et al., 2025). In this context, uplift refers to "increas[ing] access to AI-enabled intrusion capability to an expanded range of state and non-state actors" (NCSC, 2025). For example, skilled cyber criminals have already circumvented a range of safeguards on available models (such as via fine-tuning) to make AI-powered cyber tools available 'as a service', enabling novice criminals to perform disruptive cyber-attacks. If cybersecurity infrastructure and proper measures are not prioritized, the speed at which these types of attacks can evolve create limited time for swift intervention and containment, with potentially grave national security consequences.

The dual-use of AI for cyber-capabilities has been demonstrated by Anthropic's Claude Mythos, whose ability to autonomously identify zero-day vulnerabilities in critical software infrastructure has been utilised through Project Glasswing: an initiative designed to give defenders a durable advantage in AI-driven cybersecurity[4]. Middle powers were not granted access to this model, meaning that if the Mythos weights were leaked, or a comparably capable model became accessible to malicious actors, states with sustained prior access to frontier defensive models would be significantly better positioned to respond, and middle powers without equivalent defensive capacity would be left structurally exposed to the offensive applications of capabilities they cannot themselves deploy.

2.3 Chemical, Biological, Radiological and Nuclear (CBRN) threats

A second threat vector concerns the use of advanced AI to lower the barriers for malicious actors to access or produce CBRN weapons. In this context, uplift refers to the degree to which an AI model can outperform conventional tools in assisting a malicious actor at specific stages of a CBRN attack process (Rose, 2024). However, the degree of uplift, and thereby risk for middle powers, varies significantly when different kinds of actors gain access to advanced AI models.

For example, research is contested on whether general purpose models could erode some of the informational barriers preventing novices from planning and ideating the CBRN weaponization process. Anthropic found that when actors with only basic technical backgrounds (e.g. undergraduate STEM degrees) gained access to Claude Opus 4, they scored 2.53× higher than those without it on a bioweapons acquisition task: close enough to their internal threshold for novice uplift that Anthropic deployed the model under precautionary ASL-3 safeguards[5]. However, other studies have shown that uplift is relatively low for novices accessing general purpose models (Mouton et al., 2024).

---

[4] For more details, see: (Project Glasswing: Securing Critical Software for the AI Era, 2026)
[5] See: (System Card: Claude Opus 4 & Claude Sonnet 4, 2025).

By contrast, benchmarks like the Virology Capabilities Test (VCT) have presented much more troubling results for the uplift that experts receive from general-purpose models. VCT is a benchmark of 322 multimodal questions developed by PhD-level virologists, which measures LLM's abilities to troubleshoot complex virology laboratory protocols. Notably, OpenAI's o3 outperformed 94% of expert virologists within their own sub-specialties by almost double their accuracy (Götting et al., 2025)[6]. Another concern around sufficiently resourced experts is that they could use specialised models to erode physical barriers to weaponisation: such as for developing dangerous compounds or protein sequences that bypass lab-screening protocols (Wittmann et al., 2025). Physical bio-security screening practices remain uneven worldwide[7], creating an opening for regulatory arbitrage in which bad actors evade high-standard jurisdictions by seeking out countries with weak or nonexistent requirements (Patrick & Barton, 2024)[8].

These vulnerabilities are especially pertinent for middle powers without a strong position in the AISI network, which are structurally outside the information flows through which CBRN-capable models are assessed and the bilateral channels through which risks are flagged, and often lack the domestic capacity to conduct robust evaluations themselves, or worse, do not receive access to models until they are already deployed (Frontier Capability Assessments, 2025). Middle powers may therefore have very limited independent means of verifying how an open-weight model has been evaluated for CBRN uplift before release, or whether any safeguards have since been removed. A further compounding vulnerability is that managed access and model safeguards alone are insufficient to prevent AI-enabled CBRN misuse, meaning that the offence-defence balance requires active investment in defensive technologies so that they advance before offensive applications mature (Moulange et al., 2026).

2.4 Disinformation and Influence Campaigns

A third attack vector involves malicious actors using advanced AI to manipulate the information ecosystem by gaining an uplift in audience analysis, content generation, and dissemination capabilities (Stockwell, 2026). This could enable targeted disinformation campaigns, synthetic media, and coordinated influence operations, with the intent to incite political violence or mobilisation, destabilise institutions, and in extreme cases, precipitate governmental collapse. The core mechanism by which this threat creates harm is through a verification asymmetry: generating

[6] VCT was only conducted in full for closed-weight models. However, open-weight models like DeepSeek-R1 matched expert performance on text-only questions, and showed comparable results on other benchmarks (e.g. BioLP-Bench, ProtocalQA).
[7] See: (International Meeting Advances Global Standards for DNA Synthesis Screening, 2025).
[8] Importantly, erosion of informational and physical barriers is primarily a concern for chemical and biological weapons. Minimal research has found that current models could enable malicious actors to access radioactive materials, although this could shift as models' task time horizons extend and their ability for strategic planning improves.

synthetic content has become much cheaper than verifying, attributing, and correcting it (Jin, 2026). A deepfake video that takes minutes to generate and distribute may require days of forensic analysis to authoritatively debunk, by which point it has already shaped public opinion.

For highly capable groups, AI adds relatively few tools to an already substantial arsenal of disinformation and influence instruments. The more significant uplift falls on lower-resourced actors, such as domestic political networks, radicalised individuals, and commercially-motivated disinformation-for-hire firms, who previously lacked access to scalable audiovisual content production (Shane, 2024). Unmonitored access to open-weight models could enable these actors to produce realistic deepfakes, synthetic audio, and a huge amount of targeted written content at a much lower cost, and to test and iterate messaging rapidly against audience responses (Kapoor et al., 2024). For example, some research has already suggested that reliable open-weight LLMs offer up to 70% in cost reductions for minimally resourced propagandists (Musser, 2023). Whilst some bottlenecks remain in disseminating the content, AI-directed botnets capable of building organic audiences represent a plausible near-term development (Yang & Menczer, 2023).

Middle powers are disproportionately exposed to these kinds of epistemic threats. As with CBRN and cyber-attacks, they have limited means to prevent containment failures: once a model is released open-weight, there is no chokepoint through which their governments can restrict access or mandate safeguards like output watermarking. Further, the primary distribution infrastructures for influence operations largely sit outside middle powers' government control, leaving them structurally dependent on platform moderation they cannot mandate or direct, making it difficult for them to attribute operations or establish an effective counter-narrative before damage is done (McDonald, 2026). Finally, the majority of middle powers are democracies, exposing them to social fissure exploitation, since the conditions under which influence operations cause the most harm are precisely those that define politically competitive environments (Summerfield, 2024).

§3 Trigger Events

A trigger event is the point at which one of these attack vectors causes some amount of harm, thereby becoming a 'live' threat. In the scenarios that follow, the hazard in each case is a containment failure: a malicious actor gaining access to a dangerous model, and thereby a capability uplift. Left unchecked, an individual or group could then utilise this uplift and 'trigger' an incident, emergency, or if on a large enough scale, a crisis. A helpful way of thinking about this chain of events is illustrated below:

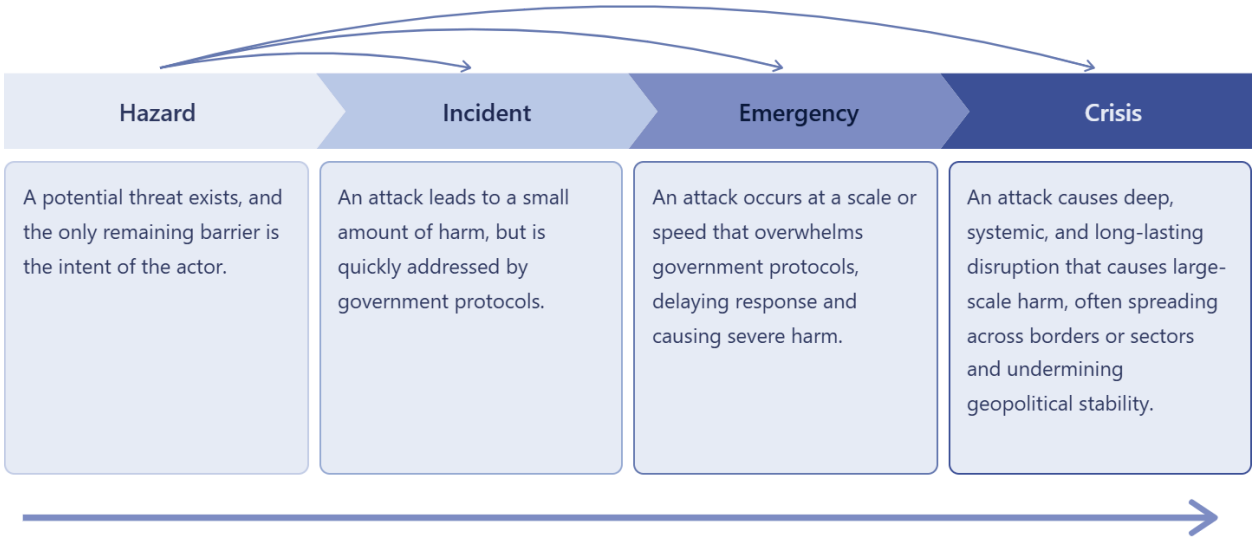

Figure 2. A pathway illustrating how an initial hazard can develop into an incident, emergency, and crisis. Adapted from The Future Society (What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One?, 2026).

TABLE 3

Hazard-to-Crisis Escalation Framework: CBRN, Cyber, and Influence Vectors

Illustrative scenario progression across three threat categories and four escalation levels

	CBRN	CYBER	INFLUENCE
HAZARD	Evo 2, a 40-billion parameter biological foundation model trained on 9.3 trillion DNA base pairs spanning all domains of life, is published fully open-source in Nature in March 2026, including model weights, training code, inference code, and the complete OpenGenome2 dataset	Anthropic's Claude Mythos Preview model weights are leaked, and accessible to the public.	Wan 2.1, Alibaba's open-weight video generation model, is released with fully public weights on Hugging Face in February 2025. Its 1.3B parameter variant requires as little as 8GB VRAM to operate, making it operable on consumer-grade hardware without specialist expertise. No output watermarking or provenance disclosure is mandated, and no centralised access controls exist through which governments can restrict domestic deployment.
INCIDENT	An expert uses Evo 2, freely available on Hugging Face, to design novel protein sequences with functional similarity to a known toxin. The expert submits synthesis orders across multiple providers in jurisdictions without mandatory screening requirements. A single order is flagged by a screening-compliant provider, the others are fulfilled. Harm is contained and localised.	Highly-capable state actors carry out a cyberattack on core infrastructure (i.e. financial systems, power grids, etc.) using a fine-tuned version of the Claude Mythos model. Private organisations or governments swiftly respond, identify the cyberattack entrypoint before it causes further damage within the system, and in other potential external domains, and localise damage.	A domestic political network uses Wan 2.1 to produce a video of a senior government official making inflammatory statements ahead of a national election. The video is widely circulated before platform moderation and government communications identify and label it as synthetic. A correction reaches a comparable audience within the news cycle. The election proceeds without significant disruption, and the operation is publicly attributed to the responsible actors.
EMERGENCY	Concurrent synthesis orders, distributed across providers in several jurisdictions with uneven screening standards go undetected. A viable pathogen-adjacent compound is produced. The affected middle power lacks the institutional infrastructure to coordinate across its biosecurity, intelligence, and health agencies at the speed the situation demands.	Claude Mythos Preview model is now used to coordinate a single cyberattack on multiple critical infrastructure points (i.e. financial systems, power grids, etc.) causing domestic and global disruption. Government response capacity is overwhelmed, and the attack causes severe harm on infrastructure.	A coordinated campaign deploys multiple Wan 2.1-generated synthetic media assets across platforms, timed to a period of elevated political tension. The volume and cross-platform distribution overwhelm moderation systems and government counter-communications capacity. Attribution is delayed by several days. Measurable shifts in public trust are recorded ahead of a scheduled vote, and inter-communal tensions escalate in at least one region, requiring law enforcement deployment.
CRISIS	The compound is deployed, causing casualties across multiple jurisdictions. Post-incident forensic analysis confirms that Evo 2's genome-scale generation capabilities were material to the attack. Attribution remains contested. *Note: This scenario applies broadly to chemical, biological, and radiological threats. Nuclear barriers remain primarily physical and material in nature and are not captured by this framework.*	The cyber-attack causes large-scale harm and disruption in critical infrastructure (i.e. financial systems, power grids, etc.) Core digital functionalities have been restricted by the cyber-attack, and governments are unable to recover critical infrastructure functionalities.	The synthetic media campaign succeeds in delegitimising an electoral outcome that cannot be rerun or contested within the required timeframe. Mass political violence follows that, due to the distributed structure of the operation and the absence of any centralised access log for Wan 2.1 deployments, cannot be retrospectively attributed to any single actor or group. No credible corrective narrative can be established, and the institutional damage persists.

§4 Preparedness Checklist

This section identifies actions that middle powers can adopt to improve their preparedness for the proliferation of dangerous AI models. Building on RAND's framework for emergency preparedness (Somani et al., 2025), we identify historical corollaries that could provide valuable insights for the detection, escalation, mitigation and containment of the three attack vectors the paper has outlined. The tractability of these actions depends upon a country's position within the 3 axes of AI power as identified in §1: AI capacity depth, Governance Orientation, and Compute and Infrastructure Posture.

Detection

Detection is the foundational layer of any preparedness architecture; without the capacity to identify a hazard before a trigger event, or incidents before they grow into emergencies and crises, governments are unable to prepare sufficient response mechanisms to threats. For middle powers, this presents structural challenges compared to states with direct frontier AI capacity. Whilst some states have AISI-equivalent bodies, with secured access to frontier models and the results of their evaluations, this represents a small subset of middle powers (U.S. Department of Commerce, 2024); such that the majority lack the access and/or technical capacity to independently evaluate advanced models and understand the threat environment. In addition, middle powers that have weaker defensive capabilities will have a greater number of proliferation threats they are vulnerable to and may need to identify.

Foresight

Foresight is the institutionalised capacity to understand plausible hazard states and how they could materialise into trigger events, enabling governments to identify the indicators and early-warning signs of threats that they should be monitoring for, and thereby the escalation and response protocols that should be in place. A foresight function of this kind requires significant investment in analytical capacity, inter-disciplinary expert networks, and structured scenario planning exercises that convert weak signals into contingency plans (Anticipatory Innovation Governance, 2020).

If scenario models are to be valuable, they need to be evidence based, and a result of deep cooperation between national AI safety bodies and relevant security agencies, rather than sitting in a general department with no operational link to incident response. The UK Government Office for Science's Futures Toolkit[9] offers a replicable model for embedding this capacity within government, though for middle powers the critical adaptation is ensuring foresight outputs are directly integrated into monitoring protocols and reviewed on a defined cycle so that indicator sets remain current as the threat landscape evolves.

[9] See: (The Futures Toolkit, 2024).

Foresight is particularly impactful for middle powers as it does not require a dominant position within the infrastructure supply chain, nor does it necessitate strong governance orientation with frontier labs and great powers. Rather, foresight primarily rests upon capacity depth: states without domestic technical talent or an AISI-equivalent body lack the personnel to convert scenario outputs into credible indicator sets. Governance orientation partially offsets this: states with strong alliance networks can import scenario intelligence from partners, though this creates a dependency on allied threat assessments that may not reflect middle powers' exposure.

Monitoring

Monitoring is the continuous, passive observation of the threat environment for signals that a hazard is materialising or that a 'live' threat is growing in severity. Before a trigger event, monitoring is landscape-level: tracking open-weight model releases against pre-specified capability indicators, conducting dark web surveillance for leaked weights, tracking anomalous compute usage through cloud provider telemetry, and tracking suspicious behaviours, such as coordinated inauthentic social-networking. Once harms have materialised, the same infrastructure should be applicable to active situational awareness: tracking whether a cyberattack is propagating across infrastructure nodes, whether a synthetic media campaign is producing measurable shifts in public discourse, or whether an AI-engineered virus is spreading (Somani et al., 2025).

The function of monitoring is to identify hazards or behaviours that warrant closer examination; and its effectiveness is directly dependent on the quality of the indicators it uses. Korea's Internet and Security Agency (KISA) demonstrates an effective monitoring architecture for cyber-threats: continuous passive surveillance against pre-specified threat indicators, with sufficient technical capacity to convert signals into escalation-ready assessments without depending on allied access or frontier model availability[10]. Sweden's Defence Research Agency (FOI) and Taiwan's Information Environment Research Center (IORG) demonstrate that the same architecture is applicable to the other two threat vectors. FOI combines CBRN-focused environmental surveillance with defence-oriented research capacity to maintain persistent coverage of the threat landscape[11]; and IORG applies the equivalent logic to influence operations, running continuous cross-platform detection of coordinated inauthentic behaviour using open-source tools paired with human analyst capacity (Wu, 2026). Across all three cases, the monitoring function is institutionally self-contained.

The feasibility of replicating these monitoring approaches varies significantly across middle powers. The binding constraint is whether a state has the technical talent and intra-governmental coordination to generate monitoring intelligence independently, or whether it must receive it from partners. States without established threat intelligence pipelines, or the talent that tends to

[10] See: (KISA Annual Report, 2024)
[11] See: ('FOI Confirms German Results on Novichok', 2020)

accompany an AISI-equivalent body, will likely rely upon their alliance relationships, or bilateral agreements with frontier labs for their monitoring ability.

Verification

Verification is the targeted, confirmatory process after receiving a monitoring signal: establishing whether a flagged model or deployment by a particular actor meets the threshold for continued monitoring, escalation and response (Maheshwari, 2026). It has two components: capability evaluations and actor identification.

On the capability side, verification involves targeted assessment of flagged models against specific criteria. For cyber threats, this means red-teaming to determine whether an open-weight model's offensive capabilities cross the threshold for escalation. For CBRN, it could require specialised biosecurity assessment of whether a flagged protein generation or synthesis model provides meaningful uplift to a specific class of malicious actor; and for influence operations, it might involve forensic analysis of suspected synthetic media to assess whether a campaign exhibits the coordination signatures of an organised operation. Importantly, the UK AISI has demonstrated that conducting these kinds of evaluations is achievable at the middle power level[12].

Actor identification presents a distinct and harder problem. Because models can be run entirely locally, compute governance is a potential lever: requiring cloud providers and compute suppliers to maintain verified records of significant GPU access creates a traceable chain of custody, making it possible to confirm whether a given actor's compute access is sufficient to fine-tune a model at a capability level of concern (Egan, 2023). The tractability of identification varies sharply across vectors. Whilst cyber incidents leave technical indicators, such as intrusion signatures, network telemetry and malware attribution, CBRN attribution is significantly more difficult after harm has occurred, and middle powers should therefore treat it as primarily prospective (Hodgson et al., 2022; Tegomoh, 2026). The priority is establishing the jurisdictional and regulatory architecture that makes pre-trigger traceability possible, rather than building retrospective forensic capacity to inform containment decisions.

Verification is most directly constrained by capacity depth. Red-teaming and evaluations for these threats require interdisciplinary expertise: AI specialists that also possess domain knowledge in cybersecurity, virology or computational forensics, which are in short supply (Vaccaro et al., 2026). For CBRN capability assessment specifically, membership in organisations like the Organisation for the Prohibition of Chemical Weapons (OPCW) offers a partial substitute, providing access to technical expertise and verification infrastructure that no middle power could replicate

[12] See: (AI Security Institute, 2024; 'Managing Risks from Increasingly Capable Open-Weight AI Systems', 2025; 'Our Evaluation of Claude Mythos Preview's Cyber Capabilities | AISI Work', 2026)

domestically[13]. However, similarly to intelligence dependencies in monitoring, reliance upon these kinds of organisations could pose a risk for middle powers. For actor identification, states with domestic compute infrastructure can try to mandate KYC schemes that generate the traceability verification requires; but those without strong infrastructure leverage must rely on allied intelligence sharing or voluntary disclosure from platform operators, both of which are slower and less reliable than domestically mandated traceability.

Detection Checklist

1. *Institutionalise a dedicated foresight function connected to security agencies and AISI-equivalent bodies.*
    - Conduct regular proliferation-scenario planning exercises that define what indicators should be monitored, and how responses might differentiate.
    - Balance the need for multi-disciplinary talent and wide participation with proportionate security standards.

Enablers: Interdisciplinary expertise, Technical talent, Inter-governmental networks

2. *Establish continuous monitoring of the threat environment against defined indicators.*
    - Initiate monitoring practices for early-warning indicators of different threat vectors, and for tracking post-trigger developments.
    - Mandate that domestic bodies that could enable malicious actors (e.g. cloud providers, DNA synthesis) maintain activity records.

Enablers: Intra-governmental coordination, Supporting Infrastructure, Technical talent

3. *Build verification capacity sufficient to confirm that a specific model or actor poses a threat that meets the threshold for escalation.*
    - Invest in AISI (or equivalent) red-teaming and rapid evaluations of open-source or leaked models.
    - Specify rapid post-trigger verification and attribution protocols for different threats.

Enablers: Technical talent, Public-private partnerships, Supporting Infrastructure

Escalation

Escalation is the process by which the responsibility for a detected threat is scaled. This can be through the activation of predefined protocols, notification of key stakeholders, and allocation of resources to address potential harm. It progresses from internal organisational response through to national coordination as the severity and scope of the incident increases (Somani et al., 2025). It

[13] See: (Report of the OPCW on the Implementation of the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction in 2023, 2024)

determines who has the authority to act, who must be informed, and through which channels. This makes the pre-specification of triggers, roles, and communication pathways essential for any crisis response system.

Schema and Triggers

Where foresight identifies plausible hazard states and monitoring tracks their early indicators, a schema pre-specifies the point at which a monitored signal crosses a threshold that compels a formal institutional response. The pre-specification of what constitutes a reportable threat, and what response it requires, is the foundational requirement of any escalation system. However, unlike traditional security risks, AI threats may present themselves as a series of ambiguous signals, which the categories of existing reporting frameworks may not map cleanly onto (Jeanmaire & Boger, 2025). This is especially acute across the three attack vectors considered in this paper: a cyber escalation trigger based on intrusion detection may fire only after a fine-tuned model has mapped critical infrastructure vulnerabilities, at which point containment options are already constrained. For CBRN and influence operations the signals are equally diffuse; anomalous synthesis orders or measurable shifts in public discourse do not map onto existing incident categories with much precision.

A potential measure for escalating hazards is mandatory near-miss reporting, which can aid early detection of novel AI risks by revealing the conditions that prevented other harms from occurring (Bin Lee Dixon & Frase, 2025). For threats that have already caused harm, middle powers can adopt a weighted, multi-factor severity scoring approach, rather than single-event triggers. This approach has been demonstrated by CISA's National Cyber Incident Scoring System, which aggregates multiple verifiable inputs into a composite score to drive escalation decisions and reduce reliance on individual analytical judgement (*CISA National Cyber Incident Scoring System*, 2020).

Designing a schema that functions under ambiguity requires the kind of technical expertise that can isolate genuine AI-enabled threats across different attack vectors, which is a function that well-coordinated national security infrastructure and established AISI-equivalent bodies are best placed to provide. For example, the UK AISI's approach to threshold-setting involves iterative development of capability evaluations, risk modeling, and pre-specified response obligations, which require sustained technical expertise, institutional infrastructure, and direct access to frontier models (AI Security Institute, 2024). States without this capacity are likely to default to generic incident reporting frameworks that will misfire against AI-specific failure modes. Participation in shared threshold-setting through the AISI network or bilateral agreements such as the UK-US Memorandum of Understanding offers a partial substitute (U.S. Department of Commerce, 2024), though imported schemas may not reflect a given state's specific exposure profile.

Notification and Coordination Architecture

Once a threshold has been established, the escalation system must specify who is notified, in what order, and with what obligations attached. This involves vertical notification up the chain of command, horizontal coordination across agencies whose mandates are implicated, notification of different private sector entities, and where necessary, engagement with foreign governments. These are analytically distinct functions, but they must be pre-specified together: an effective escalation architecture that notifies ministers without activating the private sector actors whose infrastructure is implicated will arrive at the coordination stage too late.

The UK's post-COVID pandemic preparedness reforms are an illustrative example of an effective national escalation architecture. An updated 'Pandemic Concept of Operations' defines roles and decision-making processes across government agencies, and a tiered NHS escalation framework spanning organisational, local, regional, and national response levels specifies escalation-trigger conditions for each vertical transition[14]. The Cabinet Office sits at the top of the structure, and has the powers to initiate a taskforce when a sustained whole-of-government response is required.

For the attack vectors modelled in this paper, the direct analogue would be a tiered notification structure in which a flagged open-weight model triggers mandatory notification between relevant bodies within a defined timeframe, rather than leaving notification to ad-hoc inter-agency communication. Additionally, defined handoff conditions should specify when ministerial involvement or AISIs are required for coordination authority. For example, health agencies and biosecurity intelligence units require horizontal notification for CBRN-threats; the national cybersecurity agency and critical infrastructure operators for cyber-attacks, and electoral commissions and communications regulators for influence operations. These frameworks must be calibrated to AI proliferation, where signals are more diffuse, attribution is weaker, and escalation timelines are compressed. Threshold-based escalation and whole-government activation through bodies like COBR offer a replicable foundation that middle powers can adapt to operate under conditions of persistent ambiguity and decentralised risk.

Importantly, the horizontal dimension stretches further than inter-agency coordination. Bringing in actors outside government requires an established institutional mechanism, because no individual agency's mandate covers the full set of actors that AI-enabled threats can implicate. Australia's National Coordination Mechanism[15] provides a replicable model that was also developed in response to COVID-19: an 'all-hazards design' that assigns coordinating responsibility for ambiguous threats and invites participants based on their expertise for, and interest in, defining and responding to the problem, spanning federal, state, and private sector

[14] For more details on the UK post-COVID crisis response, see: (Guidance and Framework, 2024; UK Government Response to the Covid-19 Inquiry Module 1 Report (HTML), 2025; UK Government Response to the Covid-19 Inquiry Module 2 Report (HTML), 2026)
[15] See: (National Coordination Mechanism, 2025)

actors (Buffone & Cameron, 2023). Its central feature is that it does not presuppose which agency has lead authority: making it directly applicable to the attack vectors modelled in this paper, which are novel enough that no existing mandate unambiguously covers them.

The feasibility of specific notification and coordination architectures depends heavily on a middle power's position in the relevant infrastructure. States that host significant compute capacity, cloud providers, or other threat-enabling infrastructures (e.g. social-media platforms) could legislate mandatory participation in escalation architectures. Those that are net importers of AI infrastructure lack this domestic chokepoint entirely, and must instead rely on their governance posture: bilateral relationships their governments have cultivated with frontier labs and platform operators, for instance the UK's MoU with Google DeepMind, which establishes priority model access and joint security research as pre-negotiated coordination channels[16]. This makes prior investment in those relationships, such as through alliance membership, formal lab agreements, or regulatory engagement, a precondition of effective coordination.

Escalation Checklist

1. *Develop and publish a national AI incident severity schema with defined response obligations at each tier.*
    - Adopt a weighted, multi-factor severity scoring approach rather than single-event triggers, drawing on the CISA National Cyber Incident Scoring System model.
    - Ensure the schema is vector-aware: the indicators and response obligations at each tier should differ between a flagged CBRN-relevant model, a cyber incident, and a synthetic media campaign.

Enablers: Technical Talent, Research Output, Inter-disciplinary Expertise

2. *Legislate mandatory AI incident reporting requirements for AI operators within jurisdiction, with defined time frames.*
    - Define reporting obligations that extend to compute providers, cloud platforms, and DNA synthesis providers, not only to AI model operators.
    - Establish mandatory near-miss reporting requirements to surface conditions that prevent harm, enabling early detection of novel AI risks.

Enablers: Institutional Architecture, Net Importer/Developer/Supporting Infrastructure

3. *Pre-specify cross-sectoral escalation pathways connecting AI incident response to critical infrastructure agencies.*
    - Map which agencies hold lead authority for each threat vector: health and biosecurity for CBRN, national cybersecurity agency for cyber, electoral commission and communications regulators for influence operations.

[16] See: (Department of Science, Innovation and Technology, 2025)

- Embed various public and private actors as mandatory participants in cross-sectoral coordination, depending on their role in defining and contributing to the problem at hand.

Enablers: Technical Talent, Inter-Disciplinary Expertise, Institutional Architecture

Mitigation & Containment

Once hazards and live threats have been detected, and escalated to the proper authorities, they must be mitigated before they lead to harm, and contained when they have. Given the rapid pace of AI-proliferation, it "is very likely to catch governance bodies off guard, making it difficult for them to address or even detect these issues in a timely manner"(Özcan, 2024). Middle powers have limited influence over safeguards and access restrictions that foreign frontier labs implement in development. As a result, mitigation and containment may present itself as an even greater challenge to minimise irreversible damage from AI-enabled threats. Therefore, efforts to build defensive capabilities, rapid response mechanisms, and recovery planning are crucial for domestic national security.

Sufficient Defensive Capacity

As advanced AI models continue to rapidly evolve, it is imperative for middle powers governments to invest in updated defensive capabilities to mitigate against hazards before they can materialise as trigger events. However as demonstrated through Anthropic's Project Glasswing, despite advanced defensive capabilities being developed, middle powers can be excluded from accessing them. This exposes a set of structural gaps that middle powers may face in defending themselves against a range of potential hazards.

The UK provides a concrete example for how middle powers could expand their arsenal of defenses to keep pace with proliferation threats. The government's investments in research and development led to the creation of frontier AI company DeepMind, which was later acquired by Google. Although the UK could not prevent this acquisition, it enabled the government to establish a strong institutional relationship with Google DeepMind, such that UK AISI is provided with priority access to its frontier models, and thereby frontier cyber-defenses [17].

For middle powers, this shows that strong investments in a domestic innovation ecosystem does not solely target frontier capacity. Instead, these investments can have the longer term goal of building talent pipelines, stronger research output, and public-private partnerships, which are key components for building defensive capabilities. The UK's approach provides a transferable design for middle powers: prioritise investment in technical expertise and research cooperation, and by building a domestic innovation ecosystem, public-private partnerships will follow that enable

[17] ('Memorandum of Understanding between the UK and Google DeepMind on AI Opportunities and Security', 2025)

defensive infrastructure. For example, research from the Centre for Long Term Resilience outlines 22 cutting-edge technologies that the UK should be prioritising to defend against AI-enabled chemical and biological weapons (Moulange et al., 2026), and the Turing Institute has outlined how AISIs may be able to support startups from preventing misuse in the areas of disinformation, such as by providing free safety tools and guidance (McDonald, 2026).

Response mechanisms

After a trigger event, threats must be contained before they cause irreversible harm, and the results of this process must inform middle powers' ongoing preparedness. Effective response requires specified timeframes and pre-defined measures for orderly resolution. Middle powers will need to develop rapid response mechanisms to mitigate hazards and contain AI-enabled trigger events.

Rapid response is only achievable if the institutional infrastructure exists prior to the crisis. Taiwan's whole-of-government containment architecture designed to address influence operations at scale applies to this most directly. That is, Taiwan's infrastructure has designated speed as a structural requirement in that official channels are required to draft a counter-narrative within sixty minutes, the "golden hour" before rumours become entrenched (Tang & Verfürth, 2025). This pre-specified response time functions as an escalation trigger in reverse: rather than waiting for harm to meet a threshold before acting, the architecture assumes that delay itself causes harm and treats rapid rebuttal as a containment mechanism.

This model is not unique to AI-driven influence operations, and can also be extended to AI-enabled CBRN threats. For example, the EU's TRANSTUN project establishes a public-private partnership to address and respond to CBRN risks against EU Critical Infrastructures (TRANSnational TUNnel Operational CBRN Risk Mitigation, 2021). Intervention models and coordination protocols are pre-defined across multiple teams, authoritative bodies, and Member States to ensure joint responses contain the CBRN threat without delay.

The challenge lies where rapid response shifts to orderly resolution. This is shown through the Bank of England's Resolvability Assessment Framework (RAF), which focuses on creating a resolution regime that ensures firms can fail in an orderly way (Bank of England, 2026). RAF used in the 2023 Silicon Valley Bank UK (SVB UK) resolution illustrates the value of pre-defined resolution protocols at this stage. When SVB UK failed, these pre-specified frameworks allowed rapid coordination with US authorities, a completed sale to HSBC, and resolution without delay (Cull, 2026).

Applied to AI-enabled threats, an equivalent resolution in these contexts could look like pre-negotiated kill switch agreements with cloud and compute providers, embedded in core physical AI infrastructures prior to deployment. In the event of an unrecoverable cyberattack, or

uncontrollable attack derived from these advanced AI systems, these agreements would allow the core infrastructure to be easily disabled, extinguishing the origin of this threat. Anthropic's AI Safety Level 3 (ASL-3) Deployment and Security standards serve as a precedent for pre-defined internal procedures and containment triggers activated upon pre-determined safety threshold (Activating AI Safety Level 3 Protections, 2025), in which middle powers can draw from when determining such requirements.

The feasibility of these response mechanisms vary depending on a state's position across all three axes. Middle powers without the AI capacity depth to develop systematic analysis of which monitoring signals were missed and where escalation protocols failed risk treating each incident as discrete rather than as intelligence that prepares them for the next. In terms of governance orientation, states with strong public-private relationships and prior coordination architecture can activate pre-specified responses at speed and mandate orderly resolution, while those without must negotiate participation under crisis conditions, which consistently produces slower outcomes. Finally, middle powers with domestic compute or platform infrastructure can legislate mandatory participation in kill switch agreements and containment protocols, while net importers must rely on voluntary cooperation from operators outside their jurisdiction.

Mitigation & Containment Checklist

1) *Build emergency response channels connecting technical experts, researchers, and private-public institutions with AISIs/government officials.*
- Expand AISIs role and political legitimacy
- Pre-established speed and coordination architecture channels with AISIs/government officials, public-private relationships, and technical researchers encourage collaboration in sharing information for high-level responses

Enablers: Inter-Disciplinary Expertise, Public-Private, Institutional Architecture

2) *Invest in defensive capabilities.*
- Invest in a partnerships to build defensive-capable models through public-private partnerships (e.g. UK AISI and Google Deepmind)

Enablers: Technical Talent, Research Output, Public-Private, Supporting Infrastructure

3) *Integrate contingency plans and technical intervention through pre-defined measures and crisis protocols as a way to mitigate harm in the event that a crisis cannot be resolved.* (e.g. Bank of England Recovery Planning)
- Pre-negotiated requirements for kill switch agreements are already put in place in advanced AI model hardware to diffuse threat, as model shutdown may need to be enforced.

- Pre-defined fallback system activated in the event of cyberattacks threatening primary critical infrastructure sources.

Enablers: Institutional Architecture, Net Importer/Developer/Supporting Infrastructure

Preparedness Checklist

TABLE 1

AI Proliferation Preparedness Checklist for Middle Powers

Key actions across Detection, Escalation, and Containment & Mitigation

Detection	**Institutionalise a dedicated foresight function connected to security agencies and AISI-equivalent bodies.** *Enablers: Interdisciplinary Expertise · Technical Talent · Inter-governmental* **Establish continuous monitoring of the threat environment against defined indicators.** *Enablers: Intra-governmental Coordination · Supporting Infrastructure · Technical Talent* **Build verification capacity sufficient to confirm that a specific model or actor poses a threat that meets the threshold for escalation.** *Enablers: Technical Talent · Public-Private Partnerships · Supporting Infrastructure*
Escalation	**Develop and publish a national AI incident severity schema with defined response obligations at each tier.** *Enablers: Technical Talent · Research Output · Inter-disciplinary Expertise* **Legislate mandatory AI incident reporting requirements for AI operators within jurisdiction, with defined time frames.** *Enablers: Institutional Architecture · Net Importer / Developer / Supporting Infrastructure* **Pre-specify cross-sectoral escalation pathways connecting AI incident response to critical infrastructure agencies.** *Enablers: Technical Talent · Inter-disciplinary Expertise · Institutional Architecture*
Mitigation and Containment	**Build emergency response channels connecting technical experts, researchers, and private-public institutions with AISIs/government officials.** *Enablers: Inter-disciplinary Expertise · Public-Private · Institutional Architecture* **Invest in defensive capabilities.** *Enablers: Technical Talent · Research Output · Public-Private · Supporting Infrastructure* **Integrate contingency plans and technical intervention through pre-defined measures and crisis protocols as a way to mitigate harm in the event that a crisis cannot be resolved.** *Enablers: Institutional Architecture · Net Importer / Developer / Supporting Infrastructure*

Conclusion

Preparing middle powers for AI-enabled threats demands a proactive national strategy. To guide efforts across detection, escalation, mitigation and containment, this paper offers three core principles: first, invest in preparedness before a crisis requires it; second, build the institutional relationships and information-sharing channels that cannot be improvised under pressure; and

third, treat national AI safety bodies not as optional governance infrastructure but as a prerequisite for the actions this checklist demands.

The cost of inaction is asymmetric and irreversible. A synthetic media campaign that delegitimises an election, a bio-terrorism attack enabled by an open-weight protein design model, or a coordinated AI-driven cyberattack on critical infrastructure cannot be fully undone once they have occurred. Various threat models in this paper are not hypothetical: Evo 2 is publicly available, Claude Opus 4 triggered precautionary ASL-3 safeguards, and open-weight video generation models require no specialist expertise to deploy. The pace at which a hazard becomes harmful is accelerating, and middle powers that lack the institutional infrastructure to detect, escalate, and respond will be unable to intervene before damage is done.

Effective information sharing between industry, government, and international partners is both the most critical enabler and the most structurally unequal resource in this framework. Within a nation, this requires policy frameworks that mandate reporting mechanisms and create transparent communication channels between private infrastructure operators and government bodies. Across borders, it requires active participation in the AISI International Network and bilateral agreements with frontier labs. Middle powers that sit outside these information flows are not merely less informed, but are structurally blind to the threat environment in which they operate.

Finally, the preparedness framework this paper sets out will only function if it is treated as dynamic rather than fixed. AI capabilities are evolving faster than governance institutions can revise their frameworks. The indicator sets that define monitoring thresholds, the schemas that determine escalation obligations, and the containment protocols that specify pre-committed responses must all be reviewed on defined cycles and updated against the threat landscape as it evolves. Middle powers that build such adaptive capacity into their institutions from the outset will be better positioned not only to respond to the next trigger event, but to anticipate the one after it.

Bibliography

Activating AI Safety Level 3 protections. (2025).

https://www-cdn.anthropic.com/807c59454757214bfd37592d6e048079cd7a7728.pdf

AI Security Institute. (2024, October 24). Early lessons from evaluating frontier AI systems. https://www.aisi.gov.uk/blog/early-lessons-from-evaluating-frontier-ai-systems

Anderljung, M. (2023). FRONTIER AI REGULATION: MANAGING EMERGING RISKS TO PUBLIC SAFETY. https://arxiv.org/pdf/2307.03718

Anthropic. (2026). Responsible Scaling Policy 3.0. https://www-cdn.anthropic.com/e670587677525f28df69b59e5fb4c22cc5461a17.pdf

Anticipatory innovation governance: Shaping the future through proactive policy making (OECD Working Papers on Public Governance No. 44; OECD Working Papers on Public Governance, Vol. 44). (2020). https://doi.org/10.1787/cce14d80-en

Avinash, M. (2025). Profiling LoRA/QLoRA Fine-Tuning Efficiency on Consumer GPUs: An RTX 4060 Case Study (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.12229

Bank of England. (2026). Resolvability Assessment Framework. https://www.bankofengland.co.uk/financial-stability/resolution/resolvability-assessment-framework

Bin Lee Dixon, R., & Frase, H. (2025). AI Incidents: Key Components for a Mandatory Reporting Regime. Center for Security and Emerging Technology. https://cset.georgetown.edu/wp-content/uploads/CSET-AI-Incidents.pdf

Buffone, J., & Cameron, R. (2023, September 15). Coordinating Australia's response to natural disasters and national crises. The Strategist. https://www.aspistrategist.org.au/coordinating-australias-response-to-natural-disasters-and-national-crises/

CISA National Cyber Incident Scoring System. (2020, September 30). https://www.cisa.gov/news-events/news/cisa-national-cyber-incident-scoring-system-nciss

Cull, A. (2026, March 31). Being ready for cross-border resolution: Lessons from Credit Suisse and Silicon Valley Bank. Bank of England. https://www.bankofengland.co.uk/bank-insights/2026/being-ready-for-cross-border-resolution-lessons-from-credit-suisse-and-silicon-valley-bank

de Hoop Scheffer, A., Quencez, M., Tausendfreund, R., Mohan, G., Weber, G., Jagtiani, S., & Kausch, K. (2024). Pivotal Powers. German Marshall Fund. https://www.gmfus.org/sites/default/files/2024-11/102924%20-%20Pivotal%20Powers%20A4.pdf

Department of Science, Innovation and Technology. (2025, December 11). Memorandum of Understanding between the UK and Google DeepMind on AI opportunities and security. https://www.gov.uk/government/publications/memorandum-of-understanding-between-the-uk-and-google-deepmind-on-ai-opportunities-and-security

Ee, S., Covino, C., Labrador, C., Krawec, C., Kraprayoon, J., & O'Brien, J. (2025). Asymmetry by Design: Boosting Cyber Defenders with Differential Access to AI. https://static1.squarespace.com/static/64edf8e7f2b10d716b5ba0e1/t/683a365e86d4da6cd4ef405e/1748645478254/Differential+Access+for+AIxCyber.pdf

Egan, J. (2023). Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers. https://doi.org/https://doi.org/10.48550/arXiv.2310.13625

FOI confirms German results on Novichok. (2020). FOI Newsroom. https://www.foi.se/en/foi/news-and-pressroom/news/2020-09-15-foi-confirms-german-results-on-novichok.html#:~:text=FOI%20holds%20the%20Swedish%20national%20capacity%20to,and%20radioactive%20agents%20in%20all%20types%20of

Fort, K. (2024). The Role of AI Safety Institutes in Contributing to International Standards for Frontier AI Safety (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2409.11314

Frontier Capability Assessments. (2025). Frontier Model Forum. https://www.frontiermodelforum.org/technical-reports/frontier-capability-assessments/

Götting, J., Medeiros, P., Sanders, J. G., Li, N., Phan, L., Elabd, K., Justen, L., Hendrycks, D., & Donoughe, S. (2025). Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark (arXiv:2504.16137). arXiv. https://doi.org/10.48550/arXiv.2504.16137

Guidance and Framework. (2024). NHS England. https://www.england.nhs.uk/ourwork/eprr/gf/

Hodgson, Q., Clark-Ginsberg, A., & Haldeman, Z. (2022). Managing Response to Significant Cyber Incidents. RAND. https://www.rand.org/pubs/research_reports/RRA1265-4.html

International Meeting Advances Global Standards for DNA Synthesis Screening. (2025, November 6). https://ibbis.bio/international-meeting-advances-standards-for-dna-synthesis-screening/

Jeanmaire, C., & Boger, S. (2025). AI Incidents Are Rising. It's Time for the United States to Build Playbooks for When AI Fails. The Future Society. https://thefuturesociety.org/us-ai-incident-response/

Jin, Z. (2026). AI Poses Risks to Democratic and Social Systems. https://zhijing-jin.com/d/2026-ai-risk.pdf

Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., … Narayanan, A. (2024). On the Societal Impact of Open Foundation Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.07918

Kemberry, E. (2024). Towards Responsibly Governing AI Proliferation. https://arxiv.org/pdf/2412.13821v1

KISA Annual Report. (2024). Korean Internet and Security Agency.

https://www.kisa.or.kr/EN/302/form?postSeq=73#fnPostAttachDownload

Leicht, A. (2026). How AI Safety Is Getting Middle Powers Wrong.

https://writing.antonleicht.me/p/how-ai-safety-is-getting-middle-powers

Maheshwari, S. (2026). Evaluation Awareness: Why Frontier Models are Getting Harder to Test. Institute for AI Policy and Strategy.

https://www.iaps.ai/research/evaluation-awareness-why-frontier-ai-models-are-getting-harder-to-test

Managing risks from increasingly capable open-weight AI systems. (2025). UK AISI: Red Team.

https://www.aisi.gov.uk/blog/managing-risks-from-increasingly-capable-open-weight-ai-systems

McDonald, B. (2026). AI Information Threats and Crisis Response: Practitioners' Handbook [CETaS Research Reports].

https://cetas.turing.ac.uk/sites/default/files/2026-04/cetas_policy_brief_-_ai_information_threats_and_crisis_response.pdf

Memorandum of Understanding between the UK and Google DeepMind on AI opportunities and security. (2025). GOV.UK.

https://www.gov.uk/government/publications/memorandum-of-understanding-between-the-uk-and-google-deepmind-on-ai-opportunities-and-security/memorandum-of-understanding-between-the-uk-and-google-deepmind-on-ai-opportunities-and-security

Moulange, D. R., Payne, O., Fady, D. P.-E., & Nelson, D. C. (2026). Defensive acceleration: The strategic pivot needed for UK biological resilience. The Centre for Long-Term Resilience.

https://doi.org/10.71172/x093-veqz

Mouton, C. A., Lucas, C., & Guest, E. (2024). The Operational Risks of AI in Large-Scale Biological Attacks: Results of a Red-Team Study. RAND Corporation. https://doi.org/10.7249/RRA2977-2

Musser, M. (2023). A Cost Analysis of Generative Language Models and Influence Operations (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2308.03740

National Coordination Mechanism. (2025). Australian Government Department of Home Affairs. https://www.homeaffairs.gov.au/about-us/our-portfolios/emergency-management/about-emergency-management/national-coordination-mechanism

NCSC. (2025). Impact of AI on cyber threat from now to 2027. The National Cyber Security Centre. https://www.ncsc.gov.uk/sites/default/files/pdfs/publication/impact-ai-cyber-threat-now-2027.pdf

Nevo, S. (2024). Securing AI Model Weights. https://www.rand.org/pubs/research_reports/RRA2849-1.html

Our evaluation of Claude Mythos Preview's cyber capabilities | AISI Work. (2026). UK AISI: Cyber & Autonomous Systems. https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

Özcan, B. (2024). AI governance challenges: Proliferation. https://cfg.eu/ai-governance-challenges-part-3-proliferation/

Patrick, S., & Barton, J. (2024). Mitigating Risks from Gene Editing and Synthetic Biology: Global Governance Priorities. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/10/mitigating-risks-from-gene-editing-and-synthetic-biology-global-governance-priorities?lang=en

Project Glasswing: Securing critical software for the AI era. (2026). Anthropic. https://www.anthropic.com/glasswing

Report of the OPCW on the Implementation of the Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on Their Destruction in 2023. (2024). OPCW. https://www.opcw.org/sites/default/files/documents/2024/12/c2904%28e%29.pdf

Rose, S. (2024). The near-term impact of AI on biological misuse (p. 5). https://www.longtermresilience.org/wp-content/uploads/2024/07/CLTR-Report-The-near-term-impact-of-AI-on-biological-misuse-July-2024-1.pdf

Sandoval, F. J. V., Wilkinson, I., Krasodomski, A., & Wilkinson, R. (2026). How middle powers can weather US and Chinese AI dominance: The case for 'sovereign AI' strategies. Royal Institute of International Affairs. https://doi.org/10.55317/9781784136710

Seger, E. (2023). Open-Sourcing Highly Capable Foundation Models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives. https://doi.org/https://doi.org/10.48550/arXiv.2311.09227

Shane, T. S. (2024). The near-term impact of AI on disinformation. The Centre For Long-Term Resilience. https://www.longtermresilience.org/wp-content/uploads/2024/05/The-near-term-impact-of-AI-on-disinformation.pdf

Somani, E., Friedman, A., Wu, H., Lu, M., Byrd, C., van Soest, H., & Zakaria, S. (2025). Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents. https://www.rand.org/pubs/research_reports/RRA3847-1.html

Stockwell, S. (2026). Adding Fuel to the Fire: AI Information Threats and Crisis Events [CETaS Research Reports]. The Alan Turing Institute. https://cetas.turing.ac.uk/sites/default/files/2026-02/cetas_research_report_-_adding_fuel_to_the_fire_-_ai_information_threats_and_crisis_events.pdf

Summerfield, C. (2024). How will advanced AI systems impact democracy? *https://doi.org/https://doi.org/10.48550/arXiv.2409.06729*

System Card: Claude Opus 4 & Claude Sonnet 4. (2025). Anthropic. *https://www-cdn.anthropic.com/6d8a8055020700718b0c49369f60816ba2a7c285.pdf*

Tang, A., & Verfürth, E.-M. (2025, June 12). Taiwan is standing up to disinformation. D+C - Development + Cooperation. *https://www.dandc.eu/en/article/how-taiwan-has-reduced-social-polarisation-and-become-more-resilient-disinformation*

Tegomoh, B. (2026). Part IV: AI and Biosecurity. In The Biosecurity Handbook.

The Futures Toolkit. (2024). Government Office for Science. *https://www.gov.uk/government/publications/futures-toolkit-for-policy-makers-and-analysts/the-futures-toolkit-html*

TRANSnational TUNnel operational CBRN risk mitigation. (2021). *https://ec.europa.eu/newsroom/cipr/items/738485/en#:~:text=TRANSTUN%20represents%20a%20unique%20example,final%20version%20of%20both%20Guidelines*

UK Government Response to the Covid-19 Inquiry Module 1 Report (HTML) (No. CP1240). (2025). *Cabinet Office.* *https://www.gov.uk/government/publications/uk-government-response-to-the-covid-19-inquiry-module-1-report/uk-government-response-to-the-covid-19-inquiry-module-1-report-html*

UK Government Response to the Covid-19 Inquiry Module 2 Report (HTML) (No. CP1534). (2026). *Cabinet Office.* *https://www.gov.uk/government/publications/uk-government-response-to-the-covid-19-inquiry-module-2-report/uk-government-response-to-the-covid-19-inquiry-module-2-report-html*

U.S. Department of Commerce. (2024, November 20). Fact Sheet: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes. https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international

Vaccaro, M., Song, J., Almaatouq, A., & Bakker, M. A. (2026). Evaluating Human-AI Safety: A Framework for Measuring Harmful Capability Uplift (arXiv:2603.26676). arXiv. https://doi.org/10.48550/arXiv.2603.26676

Velasco, L., Martinet, C., de Zoete, H., Trager, R., Snidal, D., Garfinkel, B., Ng, K. Y., Belfield, H., Wallace, D., Bengio, Y., Prud'homme, B., Tse, B., Radu, R., Lall, R., Harack, B., Morse, J., Miailhe, N., Singer, S., Sheehan, M., … Dennis, C. (2026). The Future of the AI Summit Series (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2601.02383

Verdi, G. (2026, February). Capability club: How the EU can lead the fight for AI middle powers [European Council on Foreign Relations]. https://ecfr.eu/article/capability-club-how-the-eu-can-lead-the-fight-for-ai-middle-powers/

Walden, K. E., Seah, C. K., & Song, D. (2025, October 9). How we enhance cybersecurity defences before the attackers in an AGI world. World Economic Forum. https://www.weforum.org/stories/2025/10/how-we-enhance-cybersecurity-defences-before-the-attackers-in-an-agi-world/

Wang, J., Milne, C., Jess Hu, Z., & Khan, F. (2024). Navigating geopolitics in AI governance. Oxford Global Society. https://oxgs.org/2024/04/08/oxgs-report-navigating-geopolitics-in-ai-governance/

Wang, W., Tian, Z., Zhang, C., & Yu, S. (2024). Machine Unlearning: A Comprehensive Survey (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2405.07406

What Is an Artificial Intelligence Crisis and What Does It Mean to Prepare for One? (2026). The Future Society. https://thefuturesociety.org/aicrisisexplainer/

Wittmann, B. J., Alexanian, T., Bartling, C., Beal, J., Clore, A., Diggans, J., Flyangolts, K., Gemler, B. T., Mitchell, T., Murphy, S. T., Wheeler, N. E., & Horvitz, E. (2025). Strengthening nucleic acid biosecurity screening against generative protein design tools. Science, 390(6768), 82–87. https://doi.org/10.1126/science.adu8578

Wu, A. (2026). On the Front Line of Foreign Influence: Enhancing Taiwan's Information Resilience (Vol. 11, Issue 3). Global Taiwan Institute. https://globaltaiwan.org/2026/02/enhancing-taiwans-information-resilience/

Yang, K.-C., & Menczer, F. (2023). Anatomy of an AI-powered malicious social botnet. https://doi.org/10.48550/ARXIV.2307.16336