Emergency Response Measures for Catastrophic Risk

James Zhang
Tsinghua University
ERA Fellowship
james.zhang@sc.tsinghua.edu.cn

Miles Kodama
Williams College
mmk4@williams.edu

Zongze Wu
Tsinghua University
zongze_wu@mail.tsinghua.edu.cn

Michael Chen
University of Oxford
michael.chen@linacre.ox.ac.uk

Yue Zhu
Tongji University
yue_zhu@tongji.edu.cn

Geng Hong
Fudan University
ghong@fudan.edu.cn

Abstract

Chinese authorities are extending the country's four-phase emergency response framework (prevent, warn, respond, and recover) to address risks from advanced artificial intelligence (AI). Concrete mechanisms for the proactive prevention and warning phases, however, remain under development. This paper analyzes an implementation model inspired by international AI safety practices: *Frontier Safety Policies* (FSPs). These policies feature pre-deployment evaluations for dangerous capabilities and tiered, pre-planned safety measures. We observe close alignment between FSPs and the proactive phases of China's emergency response framework, suggesting that the FSP model could help operationalize AI emergency preparedness in a manner consistent with China's established governance principles.

1 Motivating Emergency Response Measures

China's top authorities have recently called for the country to prepare for emergencies caused by AI. In an April 2025 study session of the Politburo, President Xi Jinping called for China to "establish systems for technical monitoring, early risk warning and emergency response" to guarantee AI's "safety, reliability and controllability" (44). The State Council subsequently published a whitepaper on national security, saying that China aims to establish "agile governance, tiered management, and rapid, effective response" for emerging security-relevant technologies such as AI (41). In September 2024, an international group of AI experts including Andrew Yao, Zhang Ya-Qin, and Xue Lan signed a statement calling upon governments to establish institutions for AI "emergency preparedness" (18). By December 2024, the National Technical Committee 260 (TC260) had released a draft guide for "Generative Artificial Intelligence Service Security Emergency Response," which was officially updated and issued in September 2025 (27). Overall, a broad consensus is emerging that China should prepare for potential future emergencies caused by AI.

China classifies AI security incidents alongside natural and national disasters. Authorities have made it clear that the AI emergencies they envisage include catastrophes more severe than those ordinarily caused by malfunctioning software. The new National Emergency Response Plan released in February 2025 put "artificial intelligence security" incidents next to earthquakes, cyberattacks, and infectious disease epidemics, on a list of potential emergencies requiring "mass monitoring and

Workshop on Regulatable ML at the 39th Conference on Neural Information Processing Systems (NeurIPS 2025).

prevention" (8). This indicates that Chinese authorities are seriously considering *catastrophic* AI risks, defined as risks of exceptionally destructive events (catastrophes) that persistently affect many people over a wide geographic area (7). An AI catastrophe would involve loss of life, mass injury, or loss of property on the scale of a major natural disaster or worse.[1]

One widely acknowledged way that AI could cause a catastrophe as devastating as a natural disaster is by giving malicious actors access to dangerous knowledge that enables them to build weapons of mass destruction. Another way is a loss of control event, where an AI evades its creators' control and acquires resources autonomously to replicate itself and amass power. TC260's AI Safety Governance Framework acknowledges both threat models and explicitly calls for "consensus-based guidelines for addressing catastrophic risks of AI" as a governance principle (23) (24).

China's emergency management system provides a foundation for AI risk governance. Although catastrophic risks from AI are novel and unprecedented, China's emergency management system features an established and institutionalized architecture. Since the early 2000s, China's "One Plan, Three Systems" has served as a unified framework for emergency response, defining clear responsibilities, enabling rapid resource mobilization, and mandating emergency preparedness (43). A key component of the framework is the Emergency Response Law of the People's Republic of China, which divides the emergency response process into four phases: prevention and preparedness, surveillance and warning, response and rescue, and rehabilitation and reconstruction. This four-stage system has been used to respond to natural disasters, severe accidents, and public health incidents; recently, TC260 has adopted such a framework for AI emergency response (27).

While the four-phase loop provides structure for emergency response, much work remains to be done on its technical implementation. In January 2025, TC260 released a draft on the nation's "AI Safety Standards System," outlining current and future technical standards for operationalizing its AI Safety Governance Framework (25). To complement this work, we examine the model of frontier safety policies widely used by leading AI companies and increasingly by governments worldwide.

We suggest that a system based on dangerous capability evaluations, thresholds, and pre-planned safety mitigations could constitute part of the first two phases of this emergency response loop. AI developers can prepare for emergencies by announcing how they will test their models for dangerous capabilities, what threshold capabilities would trigger heightened response, and what security measures they will impose on models at each threshold. The developers can then monitor for emergencies by performing the prescribed dangerous capability tests on their models at set intervals, providing a real-time indicator of catastrophic risk.

In §2, we discuss existing frontier AI regulations in China and how they could contribute to emergency preparedness. §3 then covers prevailing global practices in AI emergency preparedness, looking both at formal AI regulations in other jurisdictions and at voluntary industry standards. Finally, §4 concludes by highlighting key considerations for policymakers and suggesting how existing regulations and drafts might be extended.

2 Regulating Frontier AI in China

In this section, we examine the landscape of AI regulation and emergency preparedness in China. Namely, we review four foundational regulatory instruments and voluntary initiatives carried out by key players in the country's AI ecosystem.

2.1 Government Regulatory Frameworks

Interim Measures for the Management of Generative AI Services (August 2023). Issued by the Cyberspace Administration of China (CAC), the "Interim Measures" are China's first comprehensive regulation for public-facing generative AI systems and reflect a preemptive and compliance-oriented governance approach (11). The Interim Measures require all providers of generative AI models with public opinion or social mobilization capabilities to register their models with regulators and pass

[1] By this definition, an AI catastrophe could be expected to kill at least dozens of people or cause at least billions of dollars worth of damage. This is consistent with the definition of the most severe incident tier in China's National Natural Disaster Relief Emergency Plan and with the definition of "catastrophic risk" in California's proposed Senate Bill 53 (40) (5).

pre-deployment security assessments (Article 17), ensuring formal reviews before release. To pass, providers must show that their models uphold core socialist values, avoid discrimination, deal fairly in commercial settings, respect individuals' rights, and are adequately transparent and reliable (Article 4). These Measures create a foundation for emergency preparedness by establishing government oversight of frontier AI development and mandating pre-deployment evaluations.

AI Safety Governance Framework (draft released September 2024, updated September 2025). Published jointly by TC260 and China's Computer Network Emergency Response Technical Team, the Framework adopts a risk-based regulatory philosophy, setting forth high-level governance strategies and technical measures for three categories of AI risk: inherent, application-based, and societal (23) (24). The identified risks include the proliferation of weapons of mass destruction (biological, chemical, nuclear, and cyber) and the loss of human control; the Framework also acknowledges the misuse potential stemming from open-source AI models. To inform future AI regulations, it introduces a risk classification and grading system that assesses risks by application scenario, scale, and level of intelligence. The Framework serves as the foundation for subsequent standard-setting efforts, as referenced by TC260's draft on the "AI Safety Standards System" (January 2025) (25).

Emergency Response Guidelines for Security of Generative AI Services (draft released December 2024, updated September 2025). The "Practice Guidelines" mark the country's first attempt to outline a structured emergency response playbook targeting generative AI (26) (27). This guidance classifies security incidents into ten categories (drawn from standard GB/T 20986-2023) and defines four levels of incident severity. It further proposes technical and policy emergency response protocols across the four stages: preparation, monitoring and early warning, response, and summary and improvement. The Guidelines call for generative AI service providers to invest in emergency preparedness before emergencies strike, monitor continuously to identify emergencies as early as possible, act swiftly to address identified emergencies, and assess their own responses after emergencies have passed.

GB/T 45654-2025: Basic Security Requirements for Generative Artificial Intelligence Services (April 2025). This standard establishes prescriptive technical specifications for generative AI model providers across the entire model lifecycle, from training data collection to pre-/post-training evaluation to post-deployment monitoring (28). It operationalizes the safety goals outlined in the Interim Measures by specifying concrete criteria for pre-deployment evaluations and further requires model providers to implement mechanisms for rapid response and feedback to emerging emergencies.

China's AI safety regulation currently centers on tightening oversight at entry points, segmenting risks early, embedding safeguards across the lifecycle, and formalizing incident management protocols. While these measures primarily target content-level and security-related harms rather than catastrophic risk prevention, they establish important precedents. Requirements for pre-deployment assessment, continuous monitoring, and rapid response to identified risks provide a foundation that could be extended to address more severe, system-level threats from AI. The quantitative evaluation thresholds and detailed technical specifications set out in GB/T 45654-2025 further demonstrate regulators' willingness to impose capability-based safety standards on AI developers.

2.2 Self-Governance Initiatives

Beyond government regulations, China's AI ecosystem has begun developing voluntary safety frameworks that could significantly enhance emergency preparedness. Most notably, the Shanghai AI Laboratory, in collaboration with Concordia AI, released a comprehensive Frontier AI Risk Management Framework in July 2025, establishing systematic approaches to identifying, evaluating, and mitigating catastrophic AI risks (37).

Frontier AI Risk Management Framework (July 2025). The joint effort represents an advancement in the ecosystem's approach to safety, moving beyond content moderation to addressing catastrophic risks. The Framework establishes a three-tiered risk classification system with "yellow lines" serving as early warning indicators and "red lines" marking unacceptable risk thresholds. When models approach yellow lines, the Framework mandates enhanced safeguards and authorization requirements. Crossing red lines triggers stronger measures including potential development and deployment restrictions. Critically, the Framework specifically addresses the catastrophic risks that authorities have identified as requiring emergency preparedness, including detailed thresholds for cyber offense capabilities, biological threats, and loss of control scenarios.

The Framework's approach to capability evaluation is particularly relevant for emergency preparedness. It recommends continuous risk analysis throughout the AI development lifecycle, including pre-development capability prediction through scaling laws, pre-deployment safety assessments, and post-deployment monitoring. For biological risks, for example, the Framework establishes specific red lines such as "lowering barriers to acquiring and proliferating harmful biological agents" and prescribes corresponding evaluations using benchmarks like WMDP-Bio and LAB-Bench (21) (20). This creates a direct link between capability measurement and emergency response planning.

AI Safety and Emergency Preparedness in the Industry. Meanwhile, many AI companies in China have made voluntary commitments relevant to emergency preparedness. Seventeen leading Chinese AI companies—including Baidu, Alibaba, DeepSeek, and ByteDance—signed the "Artificial Intelligence Safety Commitments" in December 2024 under the auspices of the state-backed China AI Industry Alliance (AIIA) (9). The signatories pledged to "conduct safety assessments before models and services are made public" and to "perform continuous testing throughout the model's lifecycle, including security assessments, red teaming, and alignment checks." They further commit to establishing "management mechanism[s] for security vulnerabilities and other risks, taking timely measures for remediation," exactly as is required for rapid, effective emergency response.

These Safety Commitments were updated at the World AI Conference in July 2025 under a plenary session convened by the China AI Safety and Development Association (10). The new Commitments call for signatories to "strengthen the assessment of risks related to the abuse of AI systems in frontier fields, and prevent potential risks of their abuse in high-risk scenarios," in a possible nod to preventing catastrophic misuse. Five Chinese AI companies that were not signatories to the December 2024 Commitments signed on to the updated Commitments.

3 International Approaches to AI Emergency Response

As China advances its approach to AI emergency response, other jurisdictions worldwide are independently reaching similar conclusions about the types of catastrophic risks and the safety infrastructure needed to manage them. Recent regulatory developments in the European Union and the United States reflect an emerging consensus on core emergency preparedness mechanisms that include capability-based risk assessment, tiered response protocols, and rapid incident reporting systems. These frameworks reflect similar principles with China's policy direction and could be leveraged to enhance its implementation approach.

3.1 The European Union AI Act: General-Purpose AI Code of Practice

The EU's Code of Practice for General-Purpose AI Models (July 2025) implements many of the same emergency preparedness goals that Chinese frameworks identify as priorities (12). The EU Code recognizes four catastrophic risk categories requiring special attention: risks from chemical, biological, radiological, and nuclear (CBRN) attacks; loss of control scenarios where humans cannot reliably direct or shut down AI systems; sophisticated cyber-attacks; and strategic manipulation of human behavior.

To prepare for these four types of emergencies, the Code requires general-purpose AI service providers to establish comprehensive "Safety and Security Frameworks" before placing models on the market. These frameworks must define "systemic risk tiers" based on measurable model capabilities, with at least one tier representing capabilities not yet achieved by current models. This forward-looking approach enables proactive emergency preparedness, requiring providers to estimate when they might develop models surpassing the highest risk tier so authorities can plan response measures in advance. The requirement to define "trigger points" for additional evaluations throughout the model lifecycle creates natural checkpoints where emergency preparedness can be assessed and updated.

Particularly relevant for emergency response is the Code's incident reporting framework, which establishes concrete timelines based on severity: two days for incidents causing "serious and irreversible disruption of critical infrastructure," five days for serious cybersecurity breaches, and graduated timelines for other harms. This tiered approach, combined with the principle that reporting a serious incident is not an admission of wrongdoing, encourages rapid information sharing that could prove crucial during an AI emergency. The Code also mandates continuous monitoring through

"lighter-touch model evaluations at appropriate trigger points" defined by metrics such as training compute, development stages, and user access, helping model developers track risks in real time.

3.2 State-level Emergency Preparedness in the United States

While the US does not yet have comprehensive federal AI regulation, individual states have begun establishing emergency response frameworks. Namely, both California and New York—two of the largest and most innovative US states—have advanced legislation that would require AI developers to follow safety measures similar to those in Chinese frameworks.

California Senate Bill 53 (SB53). Signed into law in September 2025, SB53 requires every large frontier AI developer to publish and follow a framework detailing how they manage any catastrophic risks posed by their models (5). These frameworks must describe how the developer assesses their models' dangerous capabilities and what precautions or safety mitigations the developer will put in place when a model surpasses dangerous capability thresholds. The law explicitly identifies dangerous capabilities including CBRN weapons assistance, cyberattack capabilities, and scenarios where models might evade developer control, all of which directly parallel the risks identified in TC260's Framework. It also mandates 24 hour reporting for incidents posing "imminent risk of death or serious physical injury," enabling the state government to mount a rapid emergency response.

New York Responsible AI Safety and Education (RAISE) Act. The RAISE Act would likewise require large AI developers to publish and implement safety and security protocols (30). These must detail how the developer will evaluate whether their models pose an "unreasonable risk of critical harm" arising from misue or loss of control, and to outline safeguards and procedures that, if effectively implemented, would adequately mitigate such risks. Unlike SB53, the RAISE Act explicitly prohibits developers from "deploy[ing] a frontier model if doing so would create an unreasonable risk of critical harm," regardless of what their safety protocol may say. It also mandates rapid 72-hour reporting of safety incidents so that state authorities can respond promptly.

3.3 Industry Best Practices and International Commitments

In the absence of binding regulation, leading AI companies worldwide have voluntarily created frameworks for managing catastrophic AI risks.

AI Seoul Summit Commitments (May 2024). At the summit, twenty companies—including Chinese companies like MiniMax and 01.ai alongside Western firms such as OpenAI, Anthropic, Google, Microsoft, and Meta—signed on to eight safety commitments (35). The signatories commit to "assess the risks posed by their frontier models or systems across the AI lifecycle, including before deploying that model or system, and, as appropriate, before and during training." They must set "thresholds at which severe risks posed by a model or system, unless adequately mitigated, would be deemed intolerable" and establish "explicit processes they intend to follow if their model or system poses risks that meet or exceed the pre-defined thresholds." In extreme cases, signatories commit "not to develop or deploy a model or system at all, if mitigations cannot be applied to keep risks below the thresholds"—a strong pre-commitment that ensures emergency prevention takes precedence over commercial pressures to deploy.

Frontier Safety Policies (FSPs). AI companies have operationalized these commitments by writing detailed FSPs that create multiple layers of emergency preparedness (13) (22). The definition and components of an FSP are shown in Figure 1. The core of an FSP is a schedule of dangerous capability thresholds, specific levels of AI capability that would pose severe risks if left unmitigated. Anthropic's Responsible Scaling Policy, for instance, defines thresholds for CBRN weapons engineering and autonomous AI research and development capabilities that would trigger increased security (2).[2] Similarly, OpenAI's Preparedness Framework sets thresholds for biological and chemical capabilities, cybersecurity capabilities, and AI self-improvement capabilities (32). These thresholds directly parallel the catastrophic risks identified in TC260's AI Safety Governance Framework (biological and chemical weapons, cyberattacks, and loss of control), suggesting an emerging global consensus.

FSPs often mandate continuous assessments to monitor whether models are approaching thresholds, in line with China's calls for early risk warning. When evaluations reveal that a model is approaching

[2]The first of these thresholds was surpassed in May 2025, leading Anthropic to enhance its model weight security and to make Claude Opus 4 more robust against jailbreaks (1).

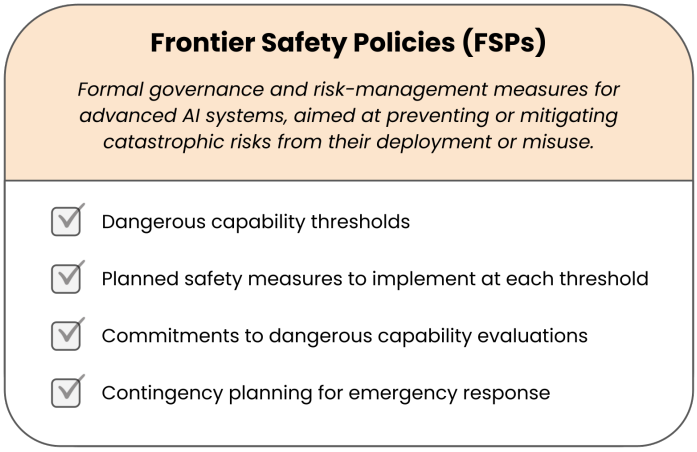

Figure 1: Definition and key elements of a frontier safety policy.

dangerous capability thresholds, FSPs prescribe escalating safeguards that mirror the "tiered management" approach advocated by China's State Council. For models that pose only moderate risks, companies might implement enhanced input and output filtering or restrict API access. As capabilities approach more dangerous thresholds, the policies mandate increasingly stringent protections, such as pausing deployment of a model too risky for the public to use.

In many cases, FSPs also mandate external transparency, especially transparency to government. Anthropic commits to "notify a relevant U.S. Government entity" of models requiring greater than ASL-2 security, and Google DeepMind will "share information with appropriate government authorities" when one of its models reaches a predefined critical capability level (17). Such notification mechanisms contribute to early warning by ensuring that regulators know in advance when AI systems are approaching dangerous capability milestones, giving them time to prepare for possible incidents.

4 Key Considerations and Future Directions

Top Chinese authorities have recognized the need for AI emergency preparedness, and regulators have laid a strong foundation for future rules to manage AI emergencies. To strengthen preparedness for catastrophic AI risks, Chinese AI companies could be encouraged or required to develop and adopt FSPs, and dangerous capability evaluations could become a standard component of pre-deployment assessments in the registry system.

The Case for Frontier Safety Policies (FSPs). The strategic goal of AI emergency preparedness requires implementing all four phases of China's established emergency response framework: prevention and preparedness, surveillance and warning, response and rescue, and rehabilitation and reconstruction. Frontier safety policies provide concrete technical mechanisms for the first two phases, which must be established *before* any emergency occurs. For prevention and preparedness, FSPs require AI developers to define dangerous capability thresholds and pre-plan safety mitigations, ensuring they are ready with specific countermeasures before dangerous capabilities emerge. This aligns with the Chinese principle that effective emergency management begins with proactive prevention rather than post-incident response (19).

For surveillance and warning, FSPs mandate continuous capability evaluations that serve as real-time indicators of emerging risks. Just as China monitors seismic activity to provide early earthquake warnings, AI developers must monitor their models' capabilities to detect when they approach dangerous thresholds. This way, monitoring does not wait until after an AI system demonstrates harmful capabilities, by which point an emergency may already be underway. These proactive phases enable effective response and rescue when needed. Pre-defined thresholds trigger pre-planned mitigations automatically, avoiding the chaos of formulating responses during a crisis. And the systematic evaluation data collected through FSPs supports post-incident rehabilitation and reconstruction by providing clear records of what capabilities emerged when and how the response unfolded.

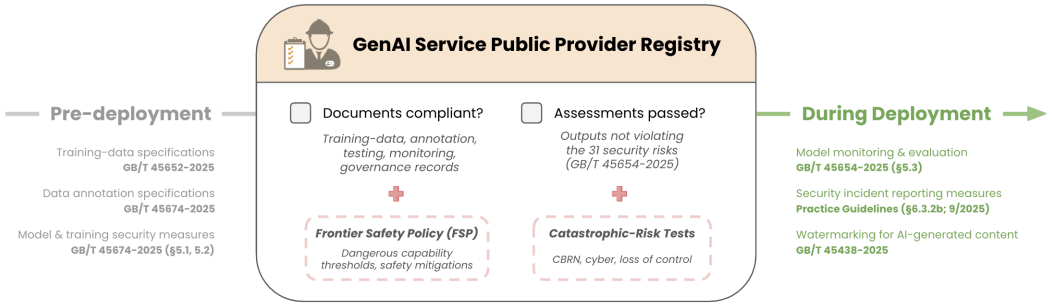

Figure 2: Registry-centric model lifecycle illustrating China's current regulatory ecosystem alongside our proposed additions: requiring providers to develop Frontier Safety Policies (FSPs) and integrating a catastrophic-risk test suite into pre-deployment assessments.

Testing for Catastrophic AI Risks. FSPs are a strong and suitable candidate for implementing key emergency planning goals. Some elements of FSPs are already present in TC260's Basic Security Requirements (GB/T 45654-2025), which direct generative AI model developers to evaluate their models' answers to politically sensitive questions before deployment and to make plans for managing security emergencies (28). The standard's main limitation is that it focuses primarily on content security rather than on catastrophic risks, but this limitation could be removed by expanding the scope of required pre-deployment testing to include dangerous capability evaluations.

For example, developers might be required to run the Shanghai AI Laboratory's suite of frontier risk evaluations on all of their new frontier models and to report evaluation results to the CAC (38). Future regulation might leave it up to AI developers to set their own dangerous capability thresholds in the spirit of SB53, or they might require specific mitigations at designated risk thresholds, much like how NY RAISE would ban deployment of models above a risk threshold.

Envisioning Extensions to Existing Generative AI Requirements. Figure 2 shows how existing requirements for generative AI service providers could be modified to mandate FSPs. Under GB/T 45654-2025, AI service providers must pass two steps before offering a model to the public (28). First, they must maintain documents demonstrating that the model was trained on a compliant data set, that the company followed data annotation rules, and that they have a sound plan for monitoring during deployment. Second, the company's model must pass assessments to verify that it does not pose any of the security risks on a thirty-one item list. The documentation requirements could be augmented by mandating providers maintain a frontier safety policy alongside other required records. At a minimum, this FSP should identify several dangerous capability thresholds beyond the current model's capability level and specify what additional safety mitigations the provider will implement when those thresholds are met. The security assessment requirements could be augmented by adding evaluations for catastrophic risk to the existing suite of assessments.

Indeed, there is precedent for these changes in preliminary drafts of the Basic Security Requirements. A draft released in February 2024 directed model providers to "pay close attention to the long-term risks that may be brought about by generative AI," and specifically mentioned risks of AI self-replicating, writing malware, and making biological or chemical weapons (29). TC260's prior work reflects early consideration of catastrophic AI risks and initial threat modeling efforts. The Committee could build on this foundation by developing a new standard that operationalizes its earlier goals through implementing FSPs or similar mechanisms.

Chinese authorities could also write an obligation for frontier AI developers to adopt FSPs into future legislation. China's national legislature has yet to pass laws intended to regulate AI, but the State Council has repeatedly indicated that such laws are being drafted (15) (16). One or more of the laws that eventually emerge could contain language requiring frontier AI companies to evaluate their models for catastrophic risks before deployment and to pre-plan tiered safety measures. This would be consistent with existing scholarly proposals for China's AI law, such as the Model Artificial Intelligence Law put forward by the Chinese Academy of Social Sciences (6). Their proposal would require AI providers to assess risks stemming from their models before and during deployment and to manage those risks through pre-established "safety risk management systems," capturing some of the key aspects of FSPs. A law like this might set China's AI emergency response regime on a sturdier legal foundation than regulatory standards alone can provide.

China's top leadership have positioned the country to lead in establishing comprehensive AI emergency preparedness. Chinese AI companies have already demonstrated their commitment to safety through voluntary frameworks. By mandating frontier safety policies, China would formalize the emerging consensus within its own industry. Such requirements would ensure Chinese companies meet the same safety standards already adopted by global competitors like OpenAI, Anthropic, and Google DeepMind, maintaining competitive parity while protecting China from AI emergencies within its borders. With growing industry buy-in, China has the opportunity to take a major step in the responsible evolution of AI as a global public good, advancing both national resilience and collective human well-being.

Limitations. While Frontier Safety Policies (FSPs) offer a promising approach to emergency preparedness, they have inherent limitations. They primarily address risks that can be anticipated and measured through capability evaluations, and may be less effective against the most novel and unforeseeable threats. Their effectiveness depends critically on the quality of evaluations, the appropriateness of thresholds, and companies' adherence to their commitments. Moreover, the FSP model we propose relies on both government oversight and corporate self-auditing, each carrying its own incentive-related risks: government-led systems may risk politicization or censorship, while self-governance can lead to opacity or underreporting. FSPs also focus primarily on the prevention and monitoring phases of emergency response. The response and recovery phases require additional implementations, which we discuss in Appendix A. Despite these limitations, FSPs still offer a concrete starting point for operationalizing the proactive stages of an emergency response framework.

5 Acknowledgments

James was supported by the ERA Fellowship. We would like to thank the ERA Fellowship for its financial and intellectual support. We would also like to thank Aris Richardson, Saad Siddiqui, Jason Zhou, Miro Pluckebaum, and Cameron Tice for helpful discussions.

References

[1] Anthropic. (May 2025). Activating AI Safety Level 3 protections. https://www.anthropic.com/news/activating-asl3-protections

[2] Anthropic. (May 2025). Responsible Scaling Policy Version 2.2. https://www-cdn.anthropic.com/872c653b2d0501d6ab44cf87f43e1dc4853e4d37.pdf

[3] Anthropic. (May 2025). System Card: Claude Opus 4 & Claude Sonnet 4. https://www-cdn.anthropic.com/07b2a3f9902ee19fe39a36ca638e5ae987bc64dd.pdf

[4] Beyer, Betsy, Chris Jones, Jennifer Petoff, & Niall Richard Murphy, eds. (2017). *Site Reliability Engineering*.

[5] California State Legislature. (2025). Senate Bill 53. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[6] Chinese Academy of Social Sciences Research Group on AI Ethics and Governance. (2023). Model Artificial Intelligence Law 1.1. http://iolaw.cssn.cn/zxzp/202309/W020230907361599893636.pdf

[7] Center for International Security and Strategy, Tsinghua University. (2025). Glossary Research on Artificial Intelligence Risks (Part I). https://ciss.tsinghua.edu.cn/info/eyjdt/8078

[8] Central Committee of the Communist Party of China & State Council. (2025). National Emergency Response Plan. https://www.gov.cn/zhengce/202502/content_7005635.htm

[9] China AI Industry Alliance (AIIA). (December 2024). Artificial Intelligence Safety Commitments. https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng

[10] China AI Safety and Development Association. (July 2025). China Artificial Intelligence Security and Safety Commitments Framework. https://mp.weixin.qq.com/s/jwtALdsfXLMN_ohfsOF5Dw

[11] Cyberspace Administration of China (CAC). (2023). Interim Measures for the Management of Generative Artificial Intelligence Services. https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[12] European Commission. (2025). Code of Practice for General-Purpose AI Models. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

[13] Frontier Model Forum. (2024). Issue Brief: Components of Frontier AI Safety Frameworks. https://www.frontiermodelforum.org/updates/issue-brief-components-of-frontier-ai-safety-frameworks/

[14] Frontier Mode Forum. (2025). FMF Announces First-of-its-Kind Information-Sharing Agreement. https://www.frontiermodelforum.org/updates/fmf-announces-first-of-its-kind-information-sharing-agreement/

[15] General Office of the State Council. (May 2023). Notice of the General Office of the State Council on Issuing the 2023 Legislative Work Plan of the State Council. https://www.gov.cn/zhengce/content/202306/content_6884925.htm

[16] General Office of the State Council. (May 2024). Notice of the General Office of the State Council on Issuing the 2024 Legislative Work Plan of the State Council. https://www.gov.cn/zhengce/content/202405/content_6950093.htm

[17] Google DeepMind. (February 2025). Frontier Safety Framework Version 2. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/updating-the-frontier-safety-framework/Frontier%20Safety%20Framework%202.0.pdf

[18] International Dialogues on AI Safety. (2024). The global nature of AI risks makes it necessary to recognize AI safety as a global public good. https://idais.ai/dialogue/idais-venice/

[19] Lan, Xue, Shen Hua. (August 2021). Five Major Transformations: Updating Concepts for Emergency Management System Construction in the New Era. https://ccmr.sppm.tsinghua.edu.cn/papers/1160.jhtml

[20] Laurent, Jon M., Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, Samuel G. Rodriques. (July 2024). LAB-Bench: Measuring Capabilities of Language Models for Biology Research. https://arxiv.org/abs/2407.10362

[21] Li, Nathaniel, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Liu, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, Dan Hendrycks. (May 2024). The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning. https://www.wmdp.ai/

[22] Model Evaluation and Threat Research. (2023). Responsible Scaling Policies. https://metr.org/blog/2023-09-26-rsp/

[23] National Information Security Standardization Technical Committee (TC260). (September 2024). AI Safety Governance Framework. https://www.tc260.org.cn/upload/2024-09-09/1725849192841090989.pdf

[24] National Information Security Standardization Technical Committee (TC260) & National Computer Network Emergency Response Technical Team (September 2025). AI Safety Governance Framework 2.0. https://www.cac.gov.cn/2025-09/15/c_1759653448369123.htm

[25] National Information Security Standardization Technical Committee (TC260). (January 2025). AI Safety Standards System V1.0 (Draft). https://www.tc260.org.cn/upload/2025-01-24/1737709785951070331.pdf

[26] National Information Security Standardization Technical Committee (TC260). (December 2024). Generative Artificial Intelligence Service Security Emergency Response Guidelines (Draft). https://www.tc260.org.cn/upload/2024-12-18/1734483139154029117.pdf

[27] National Information Security Standardization Technical Committee (TC260). (September 2025). Cybersecurity Standard Practice Guidelines – Security Emergency Response Guidelines for Generative Artificial Intelligence Services https://www.tc260.org.cn/front/postDetail.html?id=20250909095834

[28] National Information Security Standardization Technical Committee (TC260). (April 2025). Basic Security Requirements for Generative Artificial Intelligence Services (GB/T 45654-2025). https://www.tc260.org.cn/upload/2025-06-30/1751257342816036759.pdf

[29] National Information Security Standardization Technical Committee (TC260). (February 2024). Basic Security Requirements for Generative Artificial Intelligence Services (Draft). https://cset.georgetown.edu/publication/china-safety-requirements-for-generative-ai-final/

[30] New York State Legislature. (2024). Assembly Bill A6453A, Responsible AI Safety and Education Act. https://www.nysenate.gov/legislation/bills/2025/A6453/amendment/A

[31] Official Journal. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council.

[32] OpenAI. (April 2025). Preparedness Framework Version 2. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

[33] OpenAI. (June 2025). Disrupting Malicious uses of AI: June 2025. https://cdn.openai.com/threat-intelligence-reports/5f73af09-a3a3-4a55-992e-069237681620/disrupting-malicious-uses-of-ai-june-2025.pdf

[34] OpenAI. (August 2025). GPT-5 System Card. https://openai.com/index/gpt-5-system-card/

[35] United Kingdom Department for Science, Innovation and Technology. (2024). Frontier AI Safety Commitments. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024

[36] Raghavan, Prabhakar. (2024). Gemini image generation got it wrong. We'll do better. https://blog.google/products/gemini/gemini-image-generation-issue/

[37] Shanghai Artificial Intelligence Laboratory & Concordia AI. (2025). Frontier AI Risk Management Framework (v1.0). https://concordia-ai.com/research/frontier-ai-risk-management-framework/

[38] Shanghai Artificial Intelligence Laboratory. (July 2025). Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report. https://arxiv.org/abs/2507.16534

[39] Somani, Elika, Anjay Freedman, Henry Wu, Marianne Lu, Chris Byrd, Henri van Soest, & Sana Zakaria. (2025). Strengthening Emergency Preparedness and Response for AI Loss of Control Events. https://www.rand.org/pubs/research_reports/RRA3847-1.html

[40] State Council of the People's Republic of China. (2024). National Natural Disaster Relief Emergency Plan. https://www.gov.cn/zhengce/content/202402/content_6930038.htm

[41] State Council of the People's Republic of China. (2025). China's National Security in the New Era. http://politics.people.com.cn/n1/2025/0512/c1001-40478167.html

[42] Tamkin, Alex, Miles McCain, Kunal Handa, Esin Durmus, Liane Lovitt, Ankur Rathi, Saffron Huang, Alfred Mountfield, Jerry Hong, Stuart Ritchie, Michael Stern, Brian Clarke, Landon Goldberg, Theodore R. Sumers, Jared Mueller, William McEachen, Wes Mitchell, Shan Carter, Jack Clark, Jared Kaplan, & Deep Ganguli. (December 2024). Clio: Privacy-Preserving Insights into Real-World AI Use. https://arxiv.org/abs/2412.13678v1

[43] Wang, Zhe, Emily Chan, Kevin Liu, & May Yeung. (2016). The Disaster and Emergency Management System in China. https://www.hkjcdpri.org.hk/download/policy/PolicyBriefDisasterandEmergMxSysinChina.pdf

[44] Xinhua. (2025). Xi urges promoting healthy and orderly development of AI. http://english.www.gov.cn/news/202504/29/content_WS68100ef1c6d0868f4e8f2275.html

A Implementations for the four phases of emergency response

In this appendix, we discuss additional implementations for AI emergency response beyond those discussed in the main text. For a summary of the potential implementations associated with each of the four emergency response phases, see Table 1.

Table 1: China's four phases of emergency response with potential implementations.

Emergency response phase	Goal	Implementations
Prevention and Preparedness	Reduce the risk of emergencies, and prepare response protocols.	• Establish a tiered system of security and safety mitigations keyed to dangerous capability thresholds. • Share information about threats, vulnerabilities, and capabilities improvements among frontier AI companies. • Set up dedicated incident disclosure and communication channels between frontier model developers and relevant government authorities. • Develop standardized benchmarks to measure dangerous capabilities of frontier models
Monitoring and Early Warning	Detect imminent emergencies as early as possible.	• Regularly evaluate models for dangerous capabilities, including with human uplift studies that measure marginal risk given availability of other tools. • Monitor continuously for dangerous or suspicious user queries. • Coordinate relevant authorities and compute providers to monitor for signs of rogue AI replication.
Response	Contain the emergency, and limit harm.	• Suspend or restrict public and internal access to critically dangerous AI models. • Suspend or restrict hardware resources • Call upon pre-vetted external experts to rapidly patch vulnerable AI systems.
Evaluation and Improvement	Learn from successes and mistakes, and improve response protocols.	• Conduct blameless postmortem investigations to identify failures. • Update safety policies and security mitigations in light of lessons learnt. • Run regular tabletop exercises simulating past incidents and near-misses

The **prevention and preparedness** phase aims to reduce the risk of emergencies before they occur. The primary implementation discussed in the main text is for AI companies to adopt FSPs, tiered lists of dangerous capability thresholds with corresponding safety and security measures that the company commits to put in place if the threshold is met.

A second implementation is for AI companies to share information about emerging threats, model vulnerabilities, and expected capabilities improvements among themselves. This can enhance companies' awareness of risks and enable more timely response to impending emergencies. The members of the Frontier Model Forum—OpenAI, Google, Anthropic, Meta, Microsoft, and Amazon—have already formed an agreement to share risk and safety information, and there is potential for other AI companies to join FMF's agreement or to form similar agreements (14).

Third, AI companies and governments can establish dedicated incident disclosure and communication channels so that information can flow quickly in an emergency. Article 73 of the EU AI Act already establishes a rapid incident reporting mechanism for providers of high risk AI models placed on the European market (31). California's proposed Senate Bill 53 would create a similar mechanism for large AI developers to inform the state government of severe incidents involving their models (5).

A final prevention and preparedness implementation is for AI companies, government bodies, and civil society organizations to develop and share more standardized benchmarks to measure dangerous capabilities of frontier models. These benchmarks can strengthen the evidentiary basis for both internal safety evaluations and external oversight. They give all parties involved in governing frontier models a trusted and standardized metric by which to measure those models' dangerous capabilities.

The **monitoring and early warning** phase aims to detect unfolding emergencies as early as possible. The primary implementation discussed above is for AI companies, government regulators, and independent evaluators to regularly test models for dangerous capabilities. These evaluations could involve question-and-answer tests of dangerous knowledge, multistep agentic tasks, and human uplift studies. Frontier developers already use all three types of evaluations to assess whether their models are safe to deploy, as witnessed by recent model cards from Anthropic and OpenAI (3) (34).

AI companies that deploy their models via API or web app can detect emergencies early by monitoring their services continuously for suspicious queries that might indicate misuse. This mitigation has a track record of success. With continuous monitoring, OpenAI has detected and stopped automated scamming, influence operations, and malware campaigns that attempted to misuse their models (33). And Anthropic's Clio system has demonstrated that continuous query monitoring can be implemented while preserving user privacy (42). Similarly, AI compute providers, with help from relevant authorities, can continuously monitor their clusters for signs of rogue AI replication during a loss of control event (39).

The **response** phase aims to limit harm by intervening rapidly once an emergency has begun. An AI company can respond to an unfolding emergency by revoking or restricting access to the model(s) that is (are) causing the emergency. This has been done for comparatively minor incidents in the past. For instance, Google rapidly removed public access to Gemini's image generation feature in February 2024 after a content generation issue was discovered (36). Compute providers can respond to loss of control emergencies by suspending or restricting access to the hardware resources models need to propagate autonomously (39). And an AI company experiencing an emergency might be able to call upon pre-vetted external experts, from government or from independent organizations, to help rapidly patch vulnerable models or to advise the company in their response.

Finally, the **evaluation and improvement** phase aims to improve response protocols for future emergencies by learning from successes and mistakes. One implementation widely used in industry is the blameless postmortem, an in-depth investigation that seeks to identify failures without seeking to punish or blame those involved (4). Blamelessness helps to surface more valuable lessons by encouraging honest disclosures from people closest to the incident. AI companies should update their safety policies (FSPs) in light of what they learn from these postmortem investigations. AI companies, compute providers, and relevant government agencies can also keep the memory of past emergencies fresh by running regular tabletop simulations, so that staff remember emergency response protocols and lessons learnt from previous close-calls.