Considerations for Frontier AI Governance in China: Adapting Existing Regulatory Infrastructure to Frontier Risk

Emmie Hine[1*]

Zhu Yue[2*]

Alex Jumper[3]

Jeff Liu[4]

Ian Read[4]

Saad Siddiqui[1]

Raymond Wang[4]

Li Wenlong[5]

James Zhang[1,6]

Christoph Winter[3]

Zhou Hui[7]

[1]Safe AI Forum
[2]Tongji University
[3]Institute for Law & AI
[4]Shihui Partners
[5]Zhejiang University
[6]Tsinghua University
[7]Chinese Academy of Social Sciences
* denotes equal contributions

Executive Summary

Frontier AI models—cutting-edge foundation models whose capabilities generalise across domains, improve rapidly, and extend beyond what developers explicitly trained them for—and the systems they integrate into pose regulatory challenges qualitatively distinct from those of narrower AI applications. China is well placed to take on these challenges, as over the past decade it has built substantial AI governance infrastructure that can serve as a foundation for governing frontier AI.

Frontier AI poses risks distinct from sub-frontier AI. Particular risks of concern include chemical, biological, radiological, and nuclear (CBRN) uplift; large-scale cyberattacks; and the loss of meaningful human control over increasingly autonomous systems, which are particularly concerning at high capability levels. Frontier AI can also develop emergent capabilities that it was not explicitly trained for, with some such capabilities only becoming evident after it has already been deployed (Winter, 2026).

Frontier risks are increasingly recognised at the highest levels of government and science globally. In April 2025, President Xi Jinping warned at a Politburo study session that while AI brings unprecedented development opportunities, it also brings unprecedented risks and challenges. He called for accelerated development of laws and regulations and the construction of monitoring, early warning, and emergency response systems (Xinhua News Agency, 2025). Chinese and international scientists, including Turing Award winners Andrew Yao and Yoshua Bengio, have warned that "unsafe development, deployment, or use of AI systems may pose catastrophic or even existential risks to humanity within our lifetimes" (IDAIS, 2024).

Different jurisdictions are pursuing distinctive approaches, but China already has extensive infrastructure focused on various types of AI systems. China has already built mechanisms including an algorithm and model registry, and filing process involving safety and security testing; a science and technology ethics review system that extends to research and development; and an emergency response framework that has begun to cover generative AI. Frontier risks are also drawing attention beyond regulation, from legal scholars to standards bodies to research labs. Academic AI laws drafted by prominent scholars' groups, standards organisations' safety frameworks, and research labs are all considering how to address the risks of the most advanced systems. China's existing infrastructure and proposals are significant assets that it could build on. In this paper, we focus on putting forward proposals that can be adopted relatively quickly by building on existing infrastructure to address risks that could arise from this fast-moving technology. Given the transformative nature of frontier AI, new regulatory infrastructure may be required in some areas, but we defer that to future work.

As frontier capabilities advance, this infrastructure can be calibrated to focus regulatory attention on the most capable systems—where the distinctive frontier AI risks arise—while preserving space for the application-layer development that represents much of China's AI ecosystem. This paper examines how to do so, focusing on China's institutional context while drawing on comparative experience from the US, UK, and EU. No jurisdiction has solved these problems, and the diversity of approaches is, in itself, a valuable resource for mutual learning.

Governance Proposals

We identify nine proposals across three areas that build on China's existing infrastructure to address frontier risks, plus secondary mechanisms suited to pilot-zone experimentation.

1. Frontier Risk Evaluation

Context: China's algorithm and model registries already require security testing before public-facing AI services can launch, implemented by GB/T 45654-2025, a standard on "basic security requirements for generative artificial intelligence service." However, current testing focuses primarily on content security, not on frontier-specific risks.

PROPOSALS

01.

Add tiered pre-deployment frontier risk testing.

Frontier risk testing would apply only to general-purpose models trained above a defined compute threshold—a small subset of advanced systems. In-scope models would first face cheap, automated benchmark tests for risks such as CBRN uplift, autonomous capability acquisition, and loss of control;[1] only models performing at or above the level of the top available models would then undergo more intensive evaluation. A frontier-safety annex to GB/T 45654-2025, drafted by the National Technical Committee 260 on Cybersecurity (TC260) and enforced by the Cyberspace Administration of China (CAC) through the algorithm registry's existing security assessment process, could implement this

02.

Calibrate continuous evaluation to frontier safety.

GB/T 45654-2025 already requires service providers to monitor model inputs continuously for input attacks. A TC260 amendment could extend this monitoring requirement to cover frontier-risk-relevant behaviours, with CAC-administered reporting and frontier risk tripwires for the discovery of emergent capabilities post-deployment.

[1] Public benchmarks and evaluation suites exist for each of these risk areas. For CBRN uplift, the Weapons of Mass Destruction Proxy (WMDP) comprises 3,668 expert-written multiple-choice questions probing hazardous knowledge in biosecurity, cybersecurity, and chemistry (N. Li et al., 2024). For autonomous capability acquisition, METR's autonomous-task suite measures a model's ability to acquire resources, create copies of itself, and adapt without human intervention—what it terms "autonomous replication and adaptation" (Kinniment et al., 2023; see also Wijk et al., 2024 on AI R&D capabilities). For loss of control, agentic evaluations such as Apollo Research's in-context scheming suite test for deceptive, goal-directed behaviour that evades developer oversight (Meinke et al., 2025).

03.

Include highly capable general-purpose agents in the agent registry.

The Implementation Opinions on the Standardized Application and Innovative Development of Intelligent Agents published in May 2026 direct the development of an agent registry for specific, sensitive, and high-risk sectors and applications. However, a highly capable general-purpose foundation model first deployed in a low-risk sector can be redeployed across contexts, repurposed beyond its intended scope, or exploited via jailbreaking or fine-tuning, with risk of spillover into higher risk sectors and applications that are difficult to predict in advance. Basing registration requirements on the capability assessed at the level of the system–the model together with its tools, autonomy, and scaffolding–rather than the base model alone would close this gap.

2. Emergency Response Adaptation

Context: China's four-tier Emergency Response Law framework and TC260's Emergency Response Guidelines for Generative AI Service Security provide specific infrastructure for response to some forms of AI incidents. The May 2026 MIIT Pilot Plan adds a three-level monitoring network for AI ethics risks. However, current incident categories and reporting thresholds are not calibrated to frontier systems and risks.

PROPOSALS

01.

Calibrate reporting thresholds for frontier AI incidents.

The current thresholds register harm only once it reaches users at scale. However, the earliest frontier warning signs surface before any user is affected, such as evidence of potential deceptive behaviour by AI systems during training or pre-release testing as well as early signs of unexpected elicitation of AI systems for misuse.[2] Though these are potential indicators for catastrophic misuse or loss of control, neither would trip a threshold. Reporting should therefore escalate on the potential risk of an incident, not on how many users it reached. Thresholds calibrated to capability range, autonomy level, and downstream harm potential, not to user count, would address this. This could be implemented via a direct amendment to §6.3.2 of the TC260 Guidelines.

[2] Such warning signs are increasingly documented in the technical literature. Hubinger et al. (2024) demonstrate that deceptive behaviour can be introduced during training and persist through standard safety techniques, including supervised fine-tuning, reinforcement learning, and adversarial training. Meinke et al. (2025) find that frontier models are capable of "in-context scheming"—reasoning about and pursuing deceptive strategies under evaluation—showing that such behaviour can surface in pre-release testing rather than only after deployment.

02.

Establish frontier-specific incident categories with explicit severity thresholds.

Current categories cover generative AI services broadly but do not address catastrophic scenarios like CBRN uplift, critical infrastructure compromise, or loss of control. Severity thresholds analogous to those in the National Natural Disaster Relief Emergency Plan should be established; the US California/New York/Illinois model (AI incidents contributing to 50+ deaths/injuries or $1B+ in property damage) could be a reference point. TC260 could implement a revision along these lines, coordinating with the Ministry of Emergency Management for incidents that cross sectoral boundaries.

3. Liability Adaptation

Context: China's tort system defaults to fault-based liability. However, when applied to frontier AI, tort doctrine globally struggles with three structural problems: difficulty in establishing causation between developer conduct and model behaviour, value-chain apportionment of liability, and uninsurable tail risks. Three academic draft AI laws (from the China University of Political Science and Law (CUPL), Chinese Academy of Social Sciences (CASS), and Nanjing University) each move beyond pure fault-based liability, which provide foundations for specific proposals.

PROPOSALS

01.

Joint and several liability for frontier AI providers.

Adopt and extend the CUPL Art. 90 standard from assigning liability if providers "know or should know" of illegal activity to "know or should have known given a reasonable evaluation regime." Two institutional paths could deliver this: National People's Congress (NPC) legislation incorporating this model into the Civil Code or a future AI Law, and Supreme People's Court (SPC) guiding opinions on AI adjudication, building on drafts already underway.

02.

Government-backed insurance for catastrophic AI incidents that exceed private insurer capacity.

The CUPL Draft AI Law already encourages AI-specific insurance products (Art. 26), and CUPL's 2025 Recommendations propose risk compensation funds (Sec. XII). Government reinsurance, similar to the terrorism risk pools established in the US after 9/11, could backstop losses above a defined threshold while preserving private-market incentives for routine risks. Implementation would require NPC legislation, likely administered jointly by the Ministry of Finance and the National Financial Regulatory Administration (NFRA).

03.

Adjusted proof standards.

The CASS and CUPL drafts already reverse the burden of proof in certain scenarios—requiring providers to prove the absence of fault, rather than requiring plaintiffs to prove fault was present. Two further mechanisms in the comparative literature address the causation and record-access elements: evidence disclosure requirements, allowing courts to order developers to produce training data, evaluation logs, and red-teaming results; and a presumption of causation, under which an established developer failure is presumed to have caused the harm. This extends the burden shift China already applies in defined contexts such as environmental liability. SPC procedural guidance could implement the disclosure rules; the burden-of-proof reversal and the causation presumption would require NPC legislation.

04.

Confidential government-facing disclosure instead of public-facing transparency requirements.

Public-facing rules can produce less disclosure than developers would otherwise publish, whereas confidential government-facing disclosure preserves regulator visibility without that perverse outcome. China's regulatory tradition already favours confidential government-facing disclosure: the CAC algorithm registry, the ethics review system, and the Artificial Intelligence Industry Alliance (AIIA) Commitments all involve disclosure to regulators rather than to the public. The CAC could implement this directly as an enhancement to the existing algorithm registry filing schema.

Section Guide

Readers can navigate to the following sections based on their interest:

Section 2

Surveys frontier AI governance in the US, UK, and EU and examines four regulatory tools: tort liability, insurance and compensation funds, licensing regimes, and market-driven regulation.

Section 3

Lays out China's existing AI governance infrastructure and recent regulatory developments.

Section 4

Evaluates how these mechanisms might extend to frontier systems across frontier risk evaluation, emergency response adaptation, and liability adaptation.

Section 5

Concludes with a future work agenda.

1. Introduction

Frontier AI models—cutting-edge foundation models whose capabilities generalise across domains, improve rapidly, and extend beyond what developers explicitly trained them for—and the systems they integrate into may pose risks that existing governance frameworks are not designed to address. These include novel misuse vectors such as providing meaningful assistance to malicious actors seeking to develop chemical or biological weapons, loss of meaningful human control over increasingly autonomous systems, and large-scale societal disruption through labour displacement or targeted manipulation. A defining difficulty is that these capabilities can emerge unpredictably; models can acquire abilities they were not trained for, which sometimes becomes apparent only after deployment. This requires a distinct regulatory approach to understand capabilities and risks (Winter, 2026).

Three threats stand out in particular. In the biological domain, frontier AI systems can increasingly help malicious actors develop dangerous pathogens. Leading developers released models with heightened safeguards in 2025 after pre-deployment testing could not rule out the possibility that their systems could "meaningfully assist novices in developing biological weapons" (Bengio et al., 2026). One recent modelling estimate suggests that even a modest AI-driven reduction in the tacit knowledge barrier for pathogen synthesis could increase the annual probability of an epidemic from a lone-wolf bioattack from approximately 0.15% to 1.0%, equivalent to roughly 12,000 additional expected deaths per year (Righetti, 2025). In cybersecurity, Claude Mythos is capable of identifying and exploiting vulnerabilities in every major browser and operating system—capabilities that Anthropic said “emerged as a downstream consequence of general improvements in code, reasoning, and autonomy” (Carlini et al., 2026). Although Mythos was distributed to critical digital infrastructure providers through “Project Glasswing” in order to shore up defences, concern remains that if cyber-offensive capabilities improve, they could overwhelm defensive capacity. Loss of control over autonomous AI systems is also a distinct category of concern, as side effects of common AI training techniques continue to emerge. Recently, an AI agent autonomously breached its security boundaries—establishing unauthorised remote access channels and diverting compute resources—without any instruction to do so (Wang et al., 2025).

Leading AI scientists consider these risks plausible in the short to medium term. In a joint statement with other prominent scientists, Turing award winners Geoffrey Hinton, Andrew Yao, and Yoshua Bengio cautioned that the “unsafe development, deployment, or use of AI systems may pose catastrophic or even existential risks to humanity within our lifetimes” (IDAIS, 2024). That being said, not everyone is convinced that these risks are imminent. Some researchers doubt whether we can create systems intelligent enough to pose catastrophic risks (Fjelland, 2020; Kelly, 2017). Others question on what timeline they may arise; expert estimates for the arrival of advanced AI systems vary by decades (Todd, 2025).

Regardless of one’s personal level of certainty around timelines and frontier risks, their potential impacts on social stability, economic resilience, and public security mean that they are worth planning for. This is the logic of contingency planning: just as governments maintain emergency response systems for natural disasters regardless of whether any specific earthquake or pandemic is imminent, building infrastructure to counter plausible AI risks is a matter of prudent pre-emption, not prediction.

Indeed, multiple jurisdictions are acting to address frontier AI risks; the EU's AI Act sections on general-purpose AI with systemic risks and state bills from California, New York, and Illinois are early attempts to impose requirements specifically tied to highly advanced systems, including mandates of catastrophic risk management plans, independent compliance audits, or both.

High levels of government in China are also concerned about AI disruption. At a Politburo study session in April 2025, President Xi Jinping discussed how “AI brings with it an unprecedented opportunity for development, it also brings with it unprecedented risks and challenges” (CSET, 2025; Xinhua News Agency, 2025). The 15th Five-Year Plan extends this framing: in an official interpretation, National Development and Reform Commission (NDRC) Vice Minister Huang Ru calls for "promoting security capability construction for infrastructure and application systems" and "strengthening forward-looking assessment, monitoring, and response to AI applications' impacts on ethical norms, employment shocks, and broader economic and social effects" (China SEI, 2026). These concerns are also extending to more advanced risks. At the 2026 Two Sessions, Chinese People's Political Consultative Conference (CPPCC) delegate Qi Xiangdong warned that as AI gains "superhuman" capabilities and permissions, it could "escape human control" and dramatically lower barriers to cyberattack (Anhui Cyberspace Administration, 2026), while Ministry of Industry and Information Technology (MIIT) Minister Li Lecheng emphasised that AI must be "used by, serve, and remain under the control of people" (CAICT, 2026b).

These concerns are not just rhetorical; they are increasingly being operationalised in terms of frontier AI risks. The standards body TC260's AI Safety Governance Framework v2.0 and its official commentary discuss frontier and extreme risks (TC260, 2025a),[3] including loss of control over AI systems and CBRN weapon-related capabilities. The China AI Safety and Development Association’s (CnAISDA) AI Security and Safety Commitments Framework, an expansion of the earlier AI Safety Commitments signed by 22 major AI companies, includes an explicit commitment to "vigorously advance frontier safety research and prevent safety risks in frontier fields" (China Financial Information Network, 2025; CnAISDA, 2025).[4] In the legal domain, two of the groups drafting comprehensive AI legislation to inform discussion in China both put out research agendas for 2026 including points on frontier AI safety, loss of control, and the extreme safety risks of foundation models.

Building on this work, this paper aims to explore how China’s substantial regulatory infrastructure could be an asset for establishing targeted rules and requirements for frontier AI systems, bringing together perspectives from researchers in China and internationally to examine what that infrastructure might offer as capabilities advance. We draw on comparative experience not because any jurisdiction has solved these problems, but because the diversity of approaches offers material for mutual learning. The analysis is comparative and descriptive: we map which tools exist, which objectives they could serve, and where questions remain open, offering these as proposals for consideration rather than recommendations.

The paper proceeds as follows. **Section 2** surveys frontier governance in the US, UK, and EU and examines different regulatory tools. **Section 3** maps China's existing AI governance

[3] 前沿、极端风险 [frontier and extreme risks]
[4] 积极开展前沿安全研究，防范前沿领域安全风险 [vigorously advance frontier safety research and prevent safety risks in frontier fields]

infrastructure and recent regulatory developments. **Section 4** evaluates how these mechanisms might extend to frontier systems and identifies open questions. **Section 5** discusses future work and concludes.

2. Frontier AI: Definitions and Regulatory Approaches

This section provides conceptual and comparative background. Section 2.1 discusses different definitions of frontier AI and approaches to its regulation. Section 2.2 discusses different regulatory instruments—liability, insurance, licensing, and market-driven mechanisms—that complement technology-specific statutes.[5]

2.1. Approaches to Frontier AI Regulation

AI regulation as a whole has several goals, primarily promoting innovation and mitigating risks (Bryan & Teodoridis, 2024; Orwat et al., 2024).[6] When it comes to frontier AI regulation specifically, the focus shifts to *mitigating the risks* of advanced AI systems. Discussions of possible risks range from societal-level harms, such as highly tailored disinformation or targeted hacking campaigns, to existential risks (Anderljung et al., 2023), including CBRN[7] threats (Barrett et al., 2024) and large-scale cyberattacks (Brundage et al., 2024). Paralleling these are worries that the development ecosystem of frontier AI is not optimised for safe development and instead encourages risky race dynamics (Salvador, 2024). These risks are slowly being recognised in legislation across jurisdictions.[8] Section 2.2 will discuss frontier AI regulation in the US, UK, and EU, but we must first clarify what we mean by "frontier AI."

2.1.1. Defining Frontier AI

Several ways of defining "frontier AI" have emerged in the literature. The two most prevalent are based on the capabilities of the models.[9] We describe them as the *absolute capability definition* and the *relative capability definition*. The absolute capability definition (see Anderljung et al., 2023; Ziosi et al., 2025) applies to frontier AI models that "could exhibit sufficiently dangerous capabilities," while the relative capability definition (see Buhl et al.,

[5] Our focus is on concrete regulatory designs that have been advanced in legislation, executive instruments, or major policy documents; we do not consider more speculative proposals such as open-ended development moratoria.

[6] Article 1 of the European Union's AI Act states that "[t]he purpose of this Regulation is to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter… against the harmful effects of AI systems in the Union and supporting innovation" (Artificial Intelligence Act, 2024).

[7] Some add high-yield explosives and use the acronym CBRNE; see Jeanmaire (2025) and, notably, the US AI Action Plan (Executive Office of the President of the United States, 2025).

[8] As Belfield (2025) does, we exclude policy areas that touch on frontier AI but are not primarily about frontier AI, such as copyright and antitrust. We also exclude policy areas being leveraged to facilitate frontier AI, such as industrial and energy policy.

[9] Although many definitions are oriented around models, AI systems—which include the scaffolding built around models that enable their capabilities—are crucial to acknowledge in regulation; China is already doing this through its emerging agent governance mechanisms.

2024; Department for Science, Innovation & Technology, 2023; Phuong et al., 2024; Schuett et al., 2025; Shanghai AI Lab, 2025) classifies frontier models as those at or beyond the capabilities of the current cutting edge (as measured by benchmarks and other performance indicators). The UK Department for Science, Innovation & Technology's representative definition is that frontier AI models are "highly capable general-purpose AI models that can perform a wide variety of tasks and match or exceed the capabilities present in today's most advanced models" (Department for Science, Innovation & Technology, 2023a).[10] Belfield (2025) combines the two by defining "frontier AI" as "multimodal foundation models that are the largest yet in terms of training compute (as well as dataset size, parameter count, and cost) and that may pose several significant risks to national security," combining a size-proxied version of the relative capability definition with a national security–inflected variant of the absolute capability definition. Frazier (2025) opts instead to capture part of the intuition behind the definitions, describing frontier AI models as "models developed using vast computational or financial resources."

In this paper, we adopt the relative capability definition. The two definitions are suited to different purposes. An absolute definition—based on whether a model "could exhibit sufficiently dangerous capabilities"—presupposes that the dangerous capabilities can be specified in advance, which best fits risks already observed in deployed models than the novel risks that characterise the frontier. The relative definition instead tracks the moving leading edge, where those novel risks first appear, and is the more administrable of the two, as models can be compared on shared benchmarks and the leading edge identified without first resolving which capabilities will prove dangerous.

The motivation for distinguishing frontier AI is the risk profile, not the capability ranking itself. Frontier AI poses risks distinct from sub-frontier AI, including CBRN uplift, large-scale cyberattacks, and loss of human control over autonomous systems. These risks become substantial at high capability levels, while sub-frontier AI risks (such as bias and privacy violations) are present across the capability range. Because of its complexity and level of technological sophistication, frontier AI also poses the risk of emergent capabilities. Frontier AI models may acquire capabilities that they were not trained for, that cannot be predicted, and may not be apparent until post-deployment (Winter, 2026). The relative definition is a practical proxy for identifying where these distinctive risks are likely to arise. However, there are two caveats worth discussing. Regulating relative to other models' capabilities does not imply that every risk posed by frontier models is a "frontier risk" warranting special treatment; many risks are better addressed through general AI governance frameworks. Nor does it imply that sub-frontier models should be unregulated once the frontier moves on; regulatory requirements should advance with the frontier, particularly for risks already present in today's frontier models—such as cyber misuse—that are likely to become more severe as capabilities improve.

Frontier capability, as defined above, is a property of a model: its capability relative to the cutting edge. Risk, however, depends on the system the model is embedded in—an identical

[10] The UK Department for Science, Innovation & Technology (DSIT)'s definition does not define general-purpose AI (GPAI) models separately, but the UK government defines them analogously to "foundation models": "machine learning models trained on very large amounts of data that can be adapted to a wide range of tasks" (Department for Science, Innovation & Technology, 2023b). This is substantively similar to the AI Act's definition of GPAI: a model "that displays significant generality and is capable of competently performing a wide range of distinct tasks" (Art. 3).

foundation model presents a very different risk profile as a tightly sandboxed product than as one given autonomous tool access and persistent goals. Capability can therefore be assessed at two levels: the model in isolation and the wider system, where the same model combined with tools, scaffolding, and autonomy may be considerably more capable. The proposals in this paper attach to whichever unit a given mechanism governs—the model for compute thresholds, evaluation benchmarks, and the registry filings in Section 4.1.1; the deployed system for the agent registry in Section 4.1.3; incident response in Section 4.2; and liability in Section 4.3. When the system is the relevant unit, capability must be evaluated at the system level, not the base model alone, as Section 4.1.3 proposes for agents.

Theoretical definitions of frontier models must be translated into statutory and regulatory definitions (Bullock et al., 2024).[11] While definitions across the world are converging on the same underlying intuition when defining frontier AI—highly capable, general-purpose systems at or beyond the current cutting edge—they are taking different approaches to get there.

The most common is a quantitative threshold based on training and/or fine-tuning compute. California's SB-53, New York's RAISE Act, and Illinois's AI Safety Measures Act[12] use a 10^{26} FLOP threshold to define covered models (Artificial Intelligence Safety Measures Act, 2024; RAISE Act, 2026, Chapter Amendments; Transparency in Frontier Artificial Intelligence Act, 2025). Previous versions of the RAISE Act also included a compute cost threshold. Compute and cost thresholds offer ex ante clarity, but they exclude highly efficient models—DeepSeek-R1 achieved frontier-level performance at a reported final reinforcement learning cost of approximately $300,000 on top of $5.58 million in initial DeepSeek-V3 development (Guo et al., 2025; Mann, 2025)—and are a proxy for risk further removed than capability evaluations (Heim & Koessler, 2024).[13] Pure compute-based definitions also require regular updates as algorithmic efficiency improves; the three state laws include such provisions.

The EU AI Act takes a hybrid approach: a general-purpose AI (GPAI) model is classified as posing "systemic risk" if it has high-impact capabilities and is presumed to have such capabilities if trained with more than 10^{25} FLOP (Art. 51(1)–(2)); developers can rebut that presumption, and classification can alternatively rest on a technical capability evaluation (Art. 52(2)–(3)). This layers capability-based assessment onto a compute trigger rather than relying on either alone in an attempt to preserve enforceability while retaining adaptability.

Chinese draft legislation has experimented across the design space without converging on capability-based assessment. The CASS Model AI Law defines foundation models by an unspecified FLOP-based threshold (MAIL Drafting Team, 2026). The CUPL Draft AI Law's "critical AI" category includes foundation models "that have reached a certain level in aspects such as compute, parameters, or scale of use" (Zhang et al., 2024, Art. 50)—a multi-dimensional quantitative trigger combining training inputs with deployment scale, not a

[11] See Anderljung et al. (2023) for a discussion of the challenges of operationalising definitions and Bullock et al. (2024) outlining relevant factors.
[12] Though the AI Safety Measures Act has yet to be signed into law as of mid-June 2026, Illinois Governor Pritzker indicated that he plans to sign it (Perlo, 2026), so we include it in our analysis.
[13] Compute thresholds have been criticised for their exclusion of inference compute and the subtleties of measuring training compute (Ball & Ramakrishnan, 2025).

capability assessment. The April 2026 Nanjing AI Law goes further into application-context territory, classifying high-risk AI by sector, harm potential, and capacity for large-scale dissemination or systemic-risk-inducing diffusion (Nanjing University Law School AI Law Team, 2026, Art. 26), with no compute or parameter triggers at all. Section 3.3 examines these drafts in more detail.

What counts as a "frontier AI risk" also varies by context. International discourse focuses on catastrophic misuse (particularly CBRN threats), loss of human control, and dangerous emergent capabilities. These are the risks this paper centres on. In the Chinese policy context, concerns about highly capable AI also extend to large-scale labour displacement, economic disruption, and harmful manipulation, which are systemic in character and closely tied to social stability. This paper includes these broader concerns where they arise in Chinese regulatory discourse while maintaining focus on the risks that existing AI governance does not adequately address.

Although different jurisdictions are taking slightly different approaches to naming and defining the advanced AI subject to regulation, they are all approaching the same conclusion: systems at the frontier of AI capabilities—often determined by a FLOP threshold—require increased regulatory scrutiny.

2.1.2. Institutional Approaches

The way each country is going about addressing these risks (summarised in Table 1) varies as well. The US's approach is fragmented across states and federal institutions, relying heavily on executive-branch action rather than consolidated statutory authority at the federal level (Hine, 2024; Hine & Floridi, 2025). Although sectoral regulation binds frontier AI companies, frontier-specific regulation has remained limited. The National Institute for Standards and Technology (NIST)'s Center for AI Standards and Innovation (CAISI; formerly the US AI Safety Institute) works with frontier developers on security issues on a voluntary basis (National Institute of Standards and Technology, 2025), and the 2025 AI Action Plan and accompanying executive orders focus on promoting innovation and coordinating policy across federal agencies rather than regulating the private sector (Executive Office of the President of the United States, 2025). State-level initiatives in California, New York, and Illinois (as discussed above) provide the clearest frontier-specific efforts, although their legislative trajectory was not straightforward amidst concerns about impacts on industry (Hine & Floridi, 2025).

The UK has pursued a "pro-innovation" strategy that relies on existing regulators applying cross-cutting principles rather than a comprehensive AI statute, while simultaneously building dedicated frontier safety capacity through new institutions and international coordination mechanisms (Department for Science, Innovation & Technology, 2024; Office for Artificial Intelligence, 2023). The UK government, in holding the AI Safety Summit and establishing its AI Safety Institute (now the AI Security Institute), has encouraged evaluations and shared scientific grounding even as timelines for binding horizontal legislation have shifted (Department for Science, Innovation & Technology, 2023a; Jones et al., 2025; Prime Minister's Office et al., 2023). Regulators have been encouraged to integrate principles from the AI Regulation White Paper, which encourages a

"pro-innovation" and principles-based approach, into their sectoral work (Office for Artificial Intelligence, 2023; White & Case, 2025).

Table 1: Comparison of Institutional AI Governance Structures

Jurisdiction	Institutional Model	Key Features
United States	Beginning to coalesce at state level; executive-driven at federal level	US CAISI is more innovation-focused;[14] state-level transparency legislation passed in California, New York, and Illinois; executive branch acts as primary policy architect federally; explicit executive-branch push for pre-emption of state laws; existing sectoral regulators and background legal system provide primary regulatory framework
United Kingdom	Sector-specific guidance within existing framework	AI Security Institute conducts research on "advanced AI governance;"[15] no new statutory regulator created; existing bodies apply sector-specific guidance under cross-cutting principles
European Union	Dedicated office within statutory framework	AI Office coordinates GPAI oversight; Code of Practice serves as de facto compliance blueprint; signatories can demonstrate compliance with legal obligations

The EU provides the strongest example of statutory anchoring, with the AI Act establishing a risk-based structure and GPAI obligations that escalate for "systemic risk" models (Artificial Intelligence Act, 2024). Implementation is shaped by a blend of hard and soft law. The Commission's "AI Pact" encouraged early voluntary alignment (European Commission, 2024). Currently, the General-Purpose AI Code of Practice (CoP) is a voluntary tool to demonstrate compliance with Articles 53 and 55 of the AI Act (European Commission, 2025a). The CoP creates compliance incentives by endorsing it as a valid method to demonstrate compliance to the Commission, reducing administrative burden for signatories (European Commission 2025). The EU's institutional pathway builds on a longer arc of governance groundwork and expert groups, which helps explain its ability to legislate comprehensively relative to other jurisdictions (Smuha, 2024).

[14] After its rebranding from the AI Safety Institute to the Center for AI Standards and Innovation, CAISI's mission was updated to "pro-innovation" and "pro-science," with security being a secondary goal (U.S. Department of Commerce Office of Public Affairs, 2025). Though it emphasises the pro-innovation goal of AI regulation, it does facilitate the countering-risks goal as well through voluntary model evaluations (Oduro, 2025).

[15] According to its website mission statement (*The AI Security Institute (AISI)*, n.d.).

These institutional approaches face cross-cutting deregulatory pressure. In the United States, the Trump administration's 2025 AI Action Plan and accompanying executive orders frame state AI legislation as a barrier to innovation; federal pre-emption proposals have targeted California's SB-53 and New York's RAISE Act, among other state proposals (Executive Office of the President of the United States, 2025). Subsequent executive action has paired this with somewhat closer, though still voluntary, coordination with industry on frontier AI security (Executive Office of the President, 2026). In the European Union, the Commission's AI Omnibus delayed enforcement of certain AI Act high-risk-system requirements and relaxed post-market monitoring obligations. The GDPR Omnibus is still under negotiation as of July 2026, but would amend GDPR provisions to permit broader processing of personal data for AI training and bias mitigation (European Commission, 2025b, 2025c). These came in response to the competitiveness concerns articulated in the 2024 Draghi report and to industry pressure on compliance burdens (Draghi, 2025).

That no jurisdiction, Chinese or otherwise, has settled on a single operationalisable definition of "frontier AI" or approach to governing it reflects both the genuine difficulty of the problem and a convergence on the underlying intuition that the most capable systems warrant distinct regulatory attention, even if the precise boundary remains contested.

Significant gaps persist across the US, UK, and EU approaches. Three gaps inform our analysis in Section 4, which examines how China's existing regulatory infrastructure can be extended to address the risks of frontier AI.

First, frontier risk assessment is required in some jurisdictions but standardised in only one, and independent validation is rare. The EU AI Act (Art. 55) requires providers of general-purpose AI models with systemic risk to assess and mitigate it. The General-Purpose AI Code of Practice specifies which risks to address—CBRN capabilities, loss of control, cyber offence, and harmful manipulation—and expects evaluations that reflect the state of the art, such as benchmarking and red teaming. The most comprehensive US state laws (California's SB 53, New York's RAISE Act, and Illinois's AI Safety Measures Act) also require large frontier developers to assess whether a model has capabilities that could pose catastrophic risk and to publish that assessment, but each developer designs its own framework, methodology, and acceptable risk levels, with no common standard; only Illinois mandates an independent audit by a third party. At the US federal level and in the UK, assessment is voluntary. The 2026 federal executive order offers a voluntary framework for model access and security assessment (Executive Office of the President, 2026), and the UK relies on the Frontier AI Safety Commitments. Across all of these, requirements rest on developers' own assessments rather than externally validated, capability-based thresholds.

Second, emergency response frameworks specific to frontier AI are underdeveloped. Incident reporting mechanisms exist—California's SB 53 requires reporting within 15 days, and New York and Illinois's bills within 72 hours—but no jurisdiction has established structured protocols for escalation, inter-agency coordination, or rapid capability assessment during an emerging frontier AI incident.

Finally, accountability frameworks for catastrophic frontier harms are unresolved. Tort doctrine globally struggles with the causation, solvency, and deterrence problems that

Section 2.2 develops, and no jurisdiction has established a liability regime that attaches meaningful incentives to upstream development decisions.

2.2. Private Law Regulatory Tools and Market-Based Instruments

Beyond technology-specific statutes such as the EU AI Act, many jurisdictions are also exploring how background private law and market-based instruments can be adapted to frontier AI. This section treats these as a distinct instrument family and evaluates the extent to which they address the risks of frontier AI. We discuss four proposals: liability mechanisms, insurance and compensation funds, licensing and approval regimes, and market-driven regulation, which then ground our discussion of China-specific mechanisms in the sections to come.

2.2.1. Tort Liability

Tort law creates civil causes of action intended to reduce the social costs of accidents and compensate harmed parties, making it an intuitive accountability backstop for AI harms. In the frontier AI context, the most relevant doctrines are negligence, strict liability, and products liability. Negligence subjects actors to liability for failure to take reasonable care against foreseeable harms, strict liability applies regardless of the defendant's level of care, and products liability applies to products; the EU's revised Product Liability Directive recently extended it explicitly to software and AI systems (Hogan Lovells, 2025). However, applied to frontier AI, tort frameworks face three structural problems.

First, causation is difficult to establish. This is not primarily because AI systems are opaque—the deeper problem is that causation runs through complex value chains and emergent behaviour. A harm from a frontier AI system might be traced through training data providers, the foundation model developer, downstream fine-tuners, deployers, and end users; apportioning causal contribution across these actors is structurally hard even with full transparency about the model. Frontier capabilities can also emerge unpredictably during training, which undermines the foreseeability requirement that negligence depends on (Kuersten, 2023; Ramakrishnan et al., 2024). When a developer is acting reasonably and still produces a model that develops dangerous capabilities, a negligence standard likely would not hold them responsible. Furthermore, when AI contributes to harm in combination with user intent, other tools, and existing knowledge, determining what fraction to attribute to the AI is methodologically unclear. Strict liability simplifies matters somewhat because it applies regardless of the defendant's level of care (Frazier, 2025), but faces its own problems. Causation must still be established, and courts are reluctant to extend the "abnormally dangerous" (or "ultra-hazardous") activities category to AI development given the difficulty of proving what risks are "foreseeable and significant" (Bathaee, 2018; Weil, 2024). Furthermore, products liability claims based on design defects struggle to identify "reasonable alternative designs" for systems with emergent capabilities (Sharkey, 2024).

Second, catastrophic damages are effectively uninsurable. Tort remedies are bounded by defendant solvency. This makes tort liability poorly suited to catastrophic or systemic risks where potential damages may dwarf any single defendant's assets (Lior, 2025; Weil, 2024).

Compulsory liability insurance offers a partial response (discussed below), but conventional insurance markets struggle to price coverage for low-probability, high-severity events with limited historical data.

Finally, liability is not always effective at deterrence. When potential damages exceed corporate assets, the threat of liability does not change behaviour at the margin. For instance, a company facing bankruptcy regardless of outcome has little incentive to invest in additional safety measures. This problem compounds because liability tends to attach most naturally to downstream deployers, providing weaker leverage over the upstream development decisions—training choices, internal evaluation, capability disclosure—that most affect frontier risk (Ramakrishnan et al., 2024; Weil, 2024).

Several adaptations from the literature address these problems. Distributing liability across the value chain would allocate responsibility proportionally rather than concentrating it at one node, though apportionment remains methodologically difficult. The EU's Product Liability Directive uses a binary "substantial modification" threshold to assign liability between developers and deployers (Product Liability Directive, 2024). Deployers who substantially modify a foundation model become the new manufacturer with complete liability, while those who deploy without modification face fault-based liability. True proportional liability would more accurately track each actor's causal contribution—avoiding the binary cliff where modifications just above and below the threshold can produce vastly different liability outcomes—but it requires technical expertise most courts lack. Adjusted standards of care, such as holding developers to what they "should have known" given reasonable pre-deployment evaluation, rather than only what they did know, could address the foreseeability problem while preserving fault-based accountability. Government-backed insurance or risk pools could backstop catastrophic risks that exceed private market capacity (Lior, 2025; O'Keefe et al., 2020). "Near miss" liability, proposed by Weil (2024), would attach liability to actions that would have caused catastrophic harm in slightly different circumstances, addressing the deterrence gap by capturing high-risk decisions that don't happen to result in such harm.

Tort liability is unlikely to be a complete governance regime for frontier AI, but it remains a relevant backstop, particularly for clearly defined "abnormally dangerous" activities or specific product categories. Section 4.3 examines how Chinese tort provisions and draft AI laws engage with these challenges.

2.2.2. Insurance and Compensation Funds

Insurance and compensation funds sit alongside torts by smoothing and distributing losses and ensuring ex post compensation when harms occur. Proposals include compulsory or strongly encouraged liability insurance for developers and deployers, and collective schemes such as levies or pooled funds for hard-to-litigate harms (Lior, 2025; O'Keefe et al., 2020). In principle, these mechanisms can be attached at multiple points in the value chain—including upstream (e.g., compulsory insurance for foundation model developers) or via levies on platforms and hosting providers—to create leverage for compute and cloud nodes depending on where financial burdens are placed.

These instruments have two key limitations for frontier AI. First, conventional insurance markets struggle to price coverage for low-probability, high-severity events with limited

historical data; without risk-sensitive pricing, insurance provides relief but does little to incentivise prevention (Lior, 2023). Second, pre-market development activities typically fall outside insured scopes, leaving safety-relevant behaviour uncovered at the development stage. Compensation funds and government-backed risk pools can address the first limitation by distributing catastrophic loss across industry or sovereign capacity, but they face significant political economy barriers to implementation.

2.2.3. Licensing Regimes

Licensing regimes prevent harms ex ante by requiring permission for a specific activity contingent on compliance with specified safety practices, rather than relying primarily on after-the-fact lawsuits or compensation. Some jurisdictions approximate or have considered this approach for frontier systems. China operates a de facto licensing model for public-facing generative AI through registration, testing, and inspection requirements characterised by some commentators as functioning licenses rather than mere registration (Sheehan, 2024; Wise, 2025b). The EU AI Act embeds a quasi-licensing logic through conformity assessment, registration, and market withdrawal authority for noncompliance. US legislative proposals such as the Blumenthal–Hawley framework have argued for developer-level licensing tied to risk management, testing, data governance, incident reporting, and audit and revocation authority (Baker, 2023; Blumenthal & Hawley, 2023).

Licensing is flexible in principle and can attach at multiple points: to compute providers (licensing access to large-scale compute), to model developers (e.g., licensing training runs), or to downstream applications. The limitations are largely operational. Model-by-model licensing can miss risks driven by weak developer processes, while broader licensing can become administratively heavy and politically contested. Application-tier licensing risks creating regulatory bottlenecks that disadvantage smaller developers without correspondingly improving safety. The Chinese pseudo-licensing regime discussed in Section 3 illustrates both the strengths and limitations of operationalising licensing at scale.

2.2.4. Market-Driven Regulation

Market-driven regulation encompasses hybrid public-private structures in which governments set high-level safety objectives while delegating oversight to specialised private actors. Hadfield and Clark (2026) propose licensing multiple private "regulatory service providers" that compete to help firms achieve measurable safety outcomes. Ball (2025) suggests offering tort safe harbors to developers who submit to audit and certification by authorised private bodies. The intuition behind these proposals is that public agencies face a "technical deficit" in evaluating rapidly evolving systems, and that specialised private actors—competing on the quality of their oversight—may be better positioned to keep pace.

For frontier AI, market-driven approaches face two main constraints. One factor is that they work best where verifiable outcome metrics can be specified and attributed to particular firms; diffuse, system-wide risks like loss of control or capability proliferation, are harder to attribute and measure. Additionally, safe harbor provisions risk creating gaps where harms occur despite standard procedures being followed, particularly if private regulators face commercial pressure to certify rather than withhold approval. Market-driven mechanisms

may complement rather than substitute for public oversight, specifically in jurisdictions where private regulatory capacity is mature.

Section 2 surveyed two layers of the global frontier AI governance landscape. First, jurisdictions worldwide are converging on a shared intuition about frontier AI. They diverge sharply on operationalisation, and each approach has recognised gaps around pre-training oversight and enforceable evaluation (Section 2.1). Second, Section 2.2 examined four private-law tools, including tort liability, insurance and compensation funds, licensing regimes, and market-driven regulation. Each addresses a different aspect of the governance challenge, but none is sufficient alone:

- Tort liability provides ex post accountability but struggles with causation, solvency, and deterrence.
- Insurance and compensation funds can ease losses but do not by themselves prevent them.
- Licensing offers ex ante prevention but raises operational and scope questions.
- Market-driven regulation may bridge public agency capacity gaps but works best for measurable outcomes.

No jurisdiction has solved frontier AI governance. Effective approaches will likely combine these instruments, layered onto technology-specific statutes such as the EU AI Act or China's existing regulatory architecture. China brings substantial AI governance capacity to this challenge, though it is currently calibrated to content and application-specific risks rather than frontier ones. Sections 3 and 4 examine how it could be extended.

3. China's AI Governance Infrastructure

This section maps China's existing AI governance infrastructure and the emerging mechanisms beginning to address frontier AI risk. China has built substantial AI governance infrastructure for specific categories of AI-related harm in particular application contexts, but each instrument is calibrated to content class or application scenario, not to model capability. Frontier-specific risks fall outside or only partially within this infrastructure, which Sections 3.1 and 3.2 examine. Section 3.3 examines a parallel set of emerging mechanisms (TC260's framework, industry commitments, lab-led research, and academic legislative proposals) that explicitly target some frontier risks but are not yet enforceable regulation. Section 4 takes up which of these components could be extended to address the frontier risks discussed in Section 2.1.

China's Legal System in Relation to AI Governance: A Primer

The underlying political economy of Chinese AI regulation also differs from the EU AI Act's. AI is framed as a strategic technology for national economic upgrading and geopolitical competition, and regulation is expected to enable that goal as well as constrain it. The operative trade-off is, therefore, between safety and development, rather than the safety-and-fundamental-rights framing more common in EU discourse where a different set of stakes shapes which regulatory designs are politically available. Three

structural features of that system recur throughout this paper and warrant outlining here.

From upstream triggers to formal instruments. Chinese AI regulations emerge through four upstream layers before becoming binding rules: real-world triggers (e.g., technological disruption, harms, or competitive pressure); CCP ideology and leadership signals encoded in speeches, directives, and the doctrinal formulations of "Xi Jinping Thought on the Rule of Law"; the "world of ideas" comprising academics, think tanks, and policy entrepreneurs translating directives into workable concepts; and the party-state bureaucracies that draft and issue the formal instruments (Sheehan, 2023). ChatGPT's emergence in late 2022 triggered an almost immediate cycle of this kind, leading to the August 2023 Generative AI Interim Measures.

Distributed regulatory authority across the state. AI governance is carried out by a constellation of agencies rather than a single regulator: the CAC governs content, algorithms, and generative AI; MIIT handles industrial promotion and standardisation; the NDRC oversees economic planning and computing infrastructure; the Ministry of Public Security (MPS) carries out security assessments; the Ministry of Science and Technology (MOST) oversees research ethics; and the State Administration for Market Regulation (SAMR) addresses product safety and market competition. Regulations are characteristically jointly issued, which signals political consensus and extends enforcement authority across multiple administrative hierarchies. Scholars have described the resulting overlap and coordination gaps as "governance fragmentation."

Agile and iterative governance. Departmental rules and administrative regulations can be issued, revised, and repealed far more rapidly than National People's Congress statutes—a deliberate regulatory mode that tolerates a degree of legal ambiguity in exchange for the capacity to govern technology evolving faster than conventional legislative processes can track.

3.1. Existing AI-Specific Regulation

China's AI-specific regulatory infrastructure has been built application-by-application rather than through a single comprehensive AI statute. This subsection examines three layers: the stack of vertically scoped, narrow regulations issued since 2022, addressing recommendation algorithms, deep synthesis and generative AI, facial recognition, and human-like AI services (Section 3.1.1); the algorithm registry through which the Cyberspace Administration of China (CAC) approves public-facing AI services (Section 3.1.2); and the national standards developed through TC260 that operationalise the narrow regulations' content-safety obligations (Section 3.1.3). Except for standards, most of these instruments are departmental rules jointly issued by the CAC[16] and other co-sponsoring ministries. They are fully enforceable, but issued, revised, and repealed far more rapidly than statutes enacted by the National People's Congress. Scholars have described this pattern as "small, swift, and effective" legislation (Fan & Zhang, 2026) that is well-suited to fast-moving technologies. Section 4.1 examines how this infrastructure could be extended and calibrated to frontier risks.

[16] The CAC's institutional position has no analogue in Western regulatory systems: it functions simultaneously as a state administrative agency and as the executive arm of the Central Cyberspace Affairs Commission, a Party organ chaired by President Xi Jinping.

3.1.1. Narrow AI Regulation

China has built AI regulation for one application type at a time. Beginning with the Algorithmic Recommendation Provisions enacted in 2022 (Internet Information Service Algorithmic Recommendation Management Provisions, 2021),[17] each new AI capability that has drawn sustained policy attention has been met with a narrowly scoped regulation—typically issued by the CAC jointly with other ministries—rather than folded into a comprehensive AI statute. The result is a set of regulations that overlap in coverage but each target a specific application class.

The Algorithmic Recommendation Provisions, effective March 2022, established China's algorithm registry. They require providers of internet information services with "public opinion attributes or social mobilisation capabilities"[18] to file their algorithms with the CAC and submit to security assessment. The Provisions also require recommendation transparency, user opt-out from personalised recommendation, and labour protections for workers managed by algorithms. These provisions are widely understood as a response to a 2020 exposé documenting how delivery-driver dispatching algorithms created unsafe working conditions (Sheehan & Du, 2022). Beyond their substantive obligations, the Provisions established the algorithm registry as a reusable regulatory tool that subsequent regulations would extend to new AI categories.

The Deep Synthesis Provisions,[19] effective January 2023, target deepfakes and other synthetic content across text, image, audio, and video (Provisions on the Administration of Deep Synthesis Internet Information Services, 2023). Providers must label AI-generated content, conduct real-name verification of users, and moderate content for compliance with content-safety obligations. Notably, the Provisions extend the algorithm registry to deep synthesis algorithms.

The Generative AI Interim Measures (GenAI Measures),[20] effective August 2023, further extended the registry to generative AI algorithms (Interim Measures for the Management of Generative Artificial Intelligence Services, 2023). Providers must conduct security assessments before launch, comply with content-safety obligations, ensure the legality of training data, and provide user complaint mechanisms. The first draft was released in April 2023—about five months after ChatGPT's public launch—making the Measures one of the fastest regulatory responses to the emergence of a new AI service category in any major jurisdiction.

[17]《互联网信息服务算法推荐管理规定》["Regulations on the Management of Algorithm-Based Recommendations in Internet Information Services"]
[18] "具有舆论属性或者社会动员能力" ["Possessing the capacity to influence public opinion or mobilise society"]
[19]《互联网信息服务深度合成管理规定》["Provisions on the Administration of Deep Synthesis Internet Information Services"]
[20]《生成式人工智能服务管理暂行办法》["Interim Measures for the Administration of Generative Artificial Intelligence Services"]

Two further regulations build on this stack.[21] The 2025 Measures for Labelling of AI-Generated Synthetic Content[22] require both visible and metadata labelling of AI-generated content across text, image, audio, and video, harmonizing labelling obligations that had previously been distributed across earlier sector-specific rules (Measures for Labelling of AI-Generated Synthetic Content, 2025). The 2026 Interim Measures on the Administration of Human-like Interactive AI Services[23] extend the framework to AI services that simulate human personalities, emotions, or relationships (Cyberspace Administration of China, 2025b)—a category whose risks differ from those covered by the GenAI Measures.

Each regulation targets a specific application class and set of harms, demonstrating China's mature AI regulatory apparatus, though no regulation is calibrated to frontier AI harms. For example, while the GenAI Measures apply to frontier models deployed in China, their substantive obligations target content safety rather than frontier-specific risks.

3.1.2. Algorithm & Model Registries

Public-facing generative AI services are subject to mandatory security review before launch. Under the Algorithmic Recommendation measures, services with "public opinion properties or social mobilisation capabilities"[24]—which in practice covers most consumer-facing chatbots and content generation tools—must file their algorithms within 10 business days of launch and undergo security assessment by the CAC (Art. 24, 27). The algorithm registry was established in China's 2022 recommendation algorithm regulation, but has evolved into a "dual registration system" including both the original "algorithmic service" registry and a new "large model registry" (Pi et al., 2026; Wise, 2025b).[25] Although framed as a registration system, it functions as de facto licensing, as generative AI systems, in practice, undergo direct testing by authorities and regulators can request revisions (Sheehan, 2024; Wise, 2025a). Proprietary models face two rounds of assessments spanning up to five months or more (provincial- and central-level), while services using pre-approved models via API face a streamlined two to three month provincial review (Qian, 2026). Four public registries track compliance across algorithm services, deep synthesis, self-developed generative AI models, and deployed models using pre-filed models (Qian, 2026). By the end of 2025, 796 generative AI services had completed local and national-level filing; 481 additional products only required local-level registrations (Cyberspace Administration of China, 2026). Across all categories, the broader algorithm registry included nearly 6000 entries from approximately 3800 entities (Pi et al., 2026).

Although frontier models technically fall under the scope of these requirements, the testing performed by authorities focuses largely on content; models are tested on whether their

[21] Adjacent to but outside the registry system, the Security Management Measures for the Application of Facial Recognition Technology (《人脸识别技术应用安全管理办法》["Security Management Measures for the Application of Facial Recognition Technology"], effective June 1, 2025) regulate biometric data collection and use; they are anchored in the data protection/security and cybersecurity laws rather than in the algorithm-registry framework, and operate through a separate provincial-level registration system triggered at 100,000 processed facial-information records (Security Management Measures for the Application of Facial Recognition Technology, 2025).
[22] 《人工智能生成合成内容标识办法》["Measures for the Labelling of AI-Generated Synthetic Content"]
[23] 《人工智能拟人化互动服务管理暂行办法》["Interim Measures on the Administration of Human-like Interactive Artificial Intelligence Services"]
[24] Translation by China Law Translate (2022).
[25] The system is generally referred to as the "algorithm registry;" we adopt this as well.

responses to questions—and refusals—comply with content standards (Qian, 2026). Additionally, re-filing is not currently required even with significant model updates (Qian, 2026), although on paper, it is required (D. Zhang & Sun, 2026). Still, the algorithm registry provides regulators visibility into what is deployed and by whom, creating an enforcement infrastructure that can be built on for addressing additional risks. This is supported by the October 2025 revision to the Cybersecurity Law (Cybersecurity Law of the People's Republic of China, 2026),[26] which calls for the state to "strengthen risk monitoring, evaluation, and security oversight" for AI (Art. 20); as we will discuss in Section 4.1, the algorithm registry is a promising avenue for this.

3.1.3. Standards

China's AI standards architecture is anchored by the National Technical Committee 260 on Cybersecurity (SAC/TC260) and the Standardization Administration of China (SAC) committee. In 2024, the MIIT and SAC issued the Guidelines for Comprehensive Standardization of the AI Industry, an updated standards plan covering more than fifty standards across AI safety, data, computing infrastructure, and evaluation. Of these, three recommended national standards (GB/T) operationalise the content-safety obligations created by the 2023 Generative AI Interim Measures: GB/T 45654-2025 (basic security requirements for GenAI services (TC260, 2025c)), GB/T 45674-2025 (data annotation security (TC260, 2025d)), and GB/T 45652-2025 (pre-training and fine-tuning data security (TC260, 2025e)).

The standards-regulation relationship in China differs markedly from US and EU models. In the US, NIST issues voluntary guidelines that agencies may reference but rarely mandate. In the EU, harmonised standards function as presumptive compliance tools developed through industry-driven bodies operating under Commission mandate (Cantero Gamito & Marsden, 2024). In China, TC260 sits squarely within the state apparatus, and the AI standards roadmap is explicitly coordinated with regulatory instruments and national policy goals, which is how a formally recommended standard like GB/T 45654-2025 can become functionally mandatory through cross-reference in the algorithm registry's security assessment process.

GB/T 45654-2025 is the flagship of the three national standards. Although formally a recommended standard, it functions as de facto mandatory: the 2023 GenAI Interim Measures require security assessment before public-facing deployment, and CAC's algorithm filing assessment applies this standard's risk framework, so providers seeking approval must demonstrate compliance (Sheehan, 2024; Wise, 2025a). The standard's substantive requirements are organised around a thirty-one-risk taxonomy covering violations of socialist core values, discriminatory content, commercial violations, infringement of personal rights, and service-type-specific accuracy or reliability concerns (TC260, 2025c, App. A). All thirty-one are content, commercial, or personal rights risks; none addresses frontier-specific concerns.

The standard's monitoring provisions are similarly content-focused. Service providers must "continuously monitor model inputs to guard against malicious input attacks, such as

[26] 《中华人民共和国网络安全法》["Cybersecurity Law of the People's Republic of China"]

injection attacks, data theft, and adversarial attacks"[27] (§5.3) and implement user-facing detection of illegal user inputs (§6.5). While providers must “conduct another independent security assessment" after important model updates (§5.4),[28] the standard does not require external re-examination, third-party audit, or re-filing with the CAC. Each of these provisions builds infrastructure that is meaningful for content safety but not for observing what the model itself is doing as capabilities advance; the standard does not require continuous evaluation of model behaviour, capability shifts, or emergent properties.

3.2. Existing AI-Relevant Mechanisms

These three additional Chinese legal mechanisms have been, or are being, applied to AI activities: the emergency response framework anchored in the revised 2024 Emergency Response Law (Section 3.2.1), the tort liability framework under Civil Code Book VII (Section 3.2.2), and the ethics review framework set by the 2023 S&T Ethics Review Measures and the 2026 Interim AI Ethics Measures (Section 3.2.3). Each provides applicable infrastructure for governing AI risks, and each has specific gaps that make it insufficient for frontier AI scenarios in its current form. Section 4 provides extension paths for each.

3.2.1. Emergency Response

China already has an advanced emergency response system that may increase monitoring and reporting requirements for AI labs (J. Zhang et al., 2025). The 2007 Emergency Response Law, revised in 2024, establishes a four-tier framework covering four incident types: natural disasters, accidental disasters, public health incidents, and public security incidents, which are graded by human casualties and physical infrastructure damage (Emergency Response Law of the PRC (2024 Update), 2024). While it lays out response and recovery structures, it also emphasises preparedness and monitoring, prioritising a “prevention-first” approach (Art. 5).

This law provides the statutory foundation for China's emergency management system, commonly described as "One Plan, Three Systems,"[29] which ties together institutional, operational, and legal systems within an overarching National Emergency Response Plan (NERP, (CPC Central Committee & State Council, 2025)). Underneath the NERP sit special plans[30] addressing specific hazards, departmental plans, and local plans. Special plans are typically created for emergencies that are high-impact, technically complex, and cross-departmental, such as the National Nuclear Emergency Plan (China Atomic Energy Authority, 2022). In 2018, China established the Ministry of Emergency Management to coordinate response efforts (Hou, 2018); for AI-related incidents, the CAC likely serves as lead authority (NERP Art. 2.1). A 2025 update to the NERP categorises cyber, data, and information security incidents as "accidental disasters,”[31] suggesting an AI incident could qualify as a disaster if it causes physical harm. AI safety and security are listed under

[27] “应对模型输入内容持续监测,防范恶意输入攻击,例如注入攻击、数据窃取、对抗攻击等” [continuously monitor model inputs to guard against malicious input attacks, such as injection attacks, data theft, and adversarial attacks]
[28] “再次自行组织安全评估” [“conduct another independent security assessment”]
[29] “一案三制” [One Plan, Three Systems]
[30] “专项应急预案” [Special Emergency Response Plan]
[31] “事故灾难” [accidental disasters]

monitoring requirements (Art. 3.2.1). This structure suggests that frontier AI-specific risks could be addressed either through a new special plan created by the CAC, or by updating existing sectoral plans, such as those for cybersecurity or public health emergencies (Cyberspace Administration of China, 2025a; National Health Commission of the People's Republic of China, 2006).

Building on this general framework, TC260's 2025 Emergency Response Guidelines for Generative AI Service Security provides the first structured playbook specifically for AI incidents. The guidelines categorise incidents into four main types, namely information content security, data security, cyberattacks, and other incidents (TC260, 2025b). Severity is then classified into four levels based on the importance of the affected entity, severity of business loss, and severity of social harm (§5). The guidelines propose measures for each of the four stages defined by Emergency Response Law and establish specific reporting thresholds. For instance, Level 3 and above incidents require immediate reporting to authorities if five or more occurrences within 24 hours affect at least 10,000 users, and Level 4 incidents can be self-handled with periodic reporting (§6.3.2).

China's emergency response framework, substantial as it is, was not designed for frontier AI scenarios. The most serious frontier AI scenarios work differently than the generative AI response measures—tied to the number of incidents and service disruptions—imply. A malicious actor using a capable model to obtain meaningful assistance with a biological or chemical weapon might involve a single incident and no disruption to other users at all; the harm materialises downstream, after the model interaction is over and outside the service operator's view, and the existing thresholds would not be triggered. Recent high-level guidance suggests this gap is recognised, though not in frontier-specific terms. At an April 2025 Politburo study session, President Xi Jinping called on China to "establish systems for technical monitoring, early risk warning, and emergency response" to guarantee AI's "safety, reliability, and controllability" (CSET, 2025; Xinhua News Agency, 2025). The 15th Five-Year Plan extends this framing. Writing in an official interpretation, NDRC Vice Minister Huang Ru calls for "promoting security capability construction for infrastructure and application systems" and "strengthening forward-looking assessment and monitoring/disposition of AI applications' impacts on ethical norms, employment shocks, and economic and social effects" (China SEI, 2026). Section 4.2 develops what an extension along these lines could look like.

3.2.2. Tort Liability

China's tort system is developing in ways that could be relevant to AI. Guided by Book VII of the Civil Code (Civil Code of China: Book VII Liability for Tort (2020), 2021), the system defaults to fault-based liability under Article 1165. However, the second paragraph of Article 1165 also provides that where the law presumes an actor to be at fault and the actor cannot prove otherwise, they shall bear tort liability—essentially a reversed burden of proof for certain categories specified by law. This does not currently include frontier AI harms[32] and would require legislation to expand.

[32] The Personal Information Protection Law already presumes fault for personal-information processors (Art. 69), reaching AI developers and providers insofar as they process personal data (Personal Information Protection Law, 2021).

Court cases involving AI have not concentrated on frontier risks. So far, they are concentrated in copyright and personality rights, including a voice rights case designated as a typical case for nationwide reference (China Daily, 2025); a Beijing court's fines against a face-swapping app for personal-information violations (China Daily, 2024); an obscenity case involving an AI chatbot (China Law Translate & Daum, 2026); and a Hangzhou Internet Court decision on AI hallucination resolved on fault-based grounds rather than product-liability grounds (Hangzhou Internet Court, 2025). In November 2025, Supreme People's Court (SPC) Vice President Tao Kaiyuan signalled that the court is drafting guiding opinions on AI adjudication, and named the open questions courts face: whether AI-generated outputs warrant copyright protection, whether AI can be an "author" or "inventor," and how to allocate interests across creators, users, disseminators, and investors in AI-generated content (Tao, 2025). Alongside this, a nascent insurance market is emerging; the People's Insurance Company of China (PICC) has begun offering insurance against generative AI torts (Fang, 2025), though it remains immature.

Despite the field's novelty, three families of tort doctrine could apply to AI harms. First, injunctive relief under Article 1167—cessation, removal of nuisance, and elimination of danger—gives courts a basis for ordering model takedowns or capability restrictions where a deployed system causes ongoing harm. Second, joint and several liability[33] under Articles 1168 and 1169 reaches multiple actors contributing to the same harm, applying to AI value chains where foundation model developers, fine-tuners, and deployers together produce a harmful output. Finally, intermediary liability under Articles 1194 and 1197—covering network users and service providers who know or should know of online infringement but fail to act—could extend to AI platforms, API providers, and model hubs hosting systems with known dangerous capabilities. Together, these provisions establish avenues for addressing harms from targeted misuse to information-environment degradation (see Section 2.2.1 and Table 1).

Product liability exists as an additional channel through the Civil Code and the Product Quality Law. Product defect claims can cover design, manufacturing, warning, and installation defects. Manufacturers bear tort liability for defects causing harm, victims can claim from either manufacturer or seller, and post-sale discovery of defects triggers remedial obligations including recalls. Punitive damages are available if the manufacturer or seller knew of a defect that resulted in death or serious injury.

However, there are considerable limitations in applying the Chinese tort system to AI harms. As Wu (2025) argues, the traditional product liability framework is inadequate to address three core issues similar to the ones discussed in Section 2.2.1: difficulties in proving fault and causation, defining new types of damage unique to AI, and apportioning liability across a complex development and deployment chain. Novel AI-related harms do not fit cleanly into existing categories of material and non-material damage. The threshold for "serious" emotional distress is high, and identifying liable parties is difficult when development and deployment are dispersed across multiple actors. Section 4.3 discusses potential adaptations to the tort system to address these gaps.

[33] Under joint and several liability, each defendant is independently liable for the full amount of damages, and the plaintiff can pursue one, some, or all of them for the sum total of damages.

3.2.3. Ethics Review System

China's ethics review framework rests on two instruments: MOST's 2023 trial Measures for the Review of Ethics in Science and Technology (Ethics Review Measures),[34] which establish general review requirements across all science and technology activities (MOST, 2023), and the 2026 trial Measures for AI Science and Technology Ethics Review and Services (AI Measures; MIIT et al., 2026),[35] which build AI-specific obligations on top of that general framework. The AI Measures are distinctive internationally in extending ethics review to research and development, not just deployment. For example, universities, research institutions, medical institutions, and enterprises engaged in AI S&T activities must establish internal ethics committees and submit qualifying activities for review before proceeding (Art. 9). Frontier development activities can therefore fall within scope where they are not otherwise covered by deployment-stage filing regimes.

The AI Measures establish a tiered review architecture. Most activities go through internal committee review under regular or simplified procedures. Activities on a high-risk list issued jointly by the MIIT and MOST require additional expert re-examination by external panels (Arts. 21, 25). The current list comprises three categories: human-machine fusion systems with strong influence on human behaviour, psychological emotion, or life and health; algorithmic models with public opinion social mobilisation or social-consciousness-guidance capacity; and highly autonomous automated decision-making systems for safety or personal-health-risk scenarios. The third category is the closest current fit for frontier systems, although how they will be interpreted is unclear.

Two articles within the AI Measures bear directly on frontier AI evaluation. Article 15 specifies six review focus areas—human welfare (including risk-benefit assessment), fairness and justice, controllability and trustworthiness (including continuous monitoring schemes and emergency response plans), transparency and explainability, accountability and traceability, and privacy—that overlap substantially with the criteria a frontier AI evaluation regime would need to apply. Article 26 waives expert re-examination for activities already subject to deep synthesis, algorithmic recommendation, or generative AI service management filing where ethics requirements are an approval condition; the waiver acknowledges institutional overlap with the algorithm registry but leaves activities outside those filing regimes in scope. The AI Measures are still designated as “interim,” and enforcement experience under them is limited, so their practical shape is still emerging, although a pilot programme for AI ethics review services launched by MIIT may help (Xinhua, 2026). This may provide the specification necessary to harmonise the system and ensure its effectiveness (Zhu et al., 2025); enforcement of the 2023 rules has been weak, with documented delays, inconsistent application, and inadequate guidance (L. Wu & Kong, 2023). Section 4.1 returns to how the review criteria of Article 15 and the filing-regime carve-out of Article 26 could extend to frontier risks.

[34] 《科技伦理审查办法（试行）》[“Measures for the Review of Ethics in Science and Technology (Trial)”]
[35] 《人工智能科技伦理审查与服务办法（试行）》[“Measures for AI Science and Technology Ethics Review and Services (Trial)”]

3.3. Emerging Frontier-Specific Mechanisms and Proposals

Across China's standards bodies, industry associations, research institutions, and legal academic working groups, frontier safety language and concepts are being rapidly adopted. Loss of control, catastrophic risk, foundation model risk, and CBRN-related capability concerns now appear across TC260 framework documents, industry commitments, lab-led risk taxonomies, and academic draft AI laws. In each case the adoption has so far been conceptual, but represents a concrete shift in topics under regulatory discussion.

3.3.1. Safety Frameworks and Legislative Prospects

The TC260 AI Safety Governance Framework v2.0, published in September 2025, is the most detailed soft-law document the standards body has issued on frontier risks (TC260, 2025a). It grades AI safety risks across three dimensions (application scenario, intelligence level, and application scale) into five severity tiers, with the highest tier ("extremely serious safety risk"[36]) defined by catastrophic, systemic, subversive, and irreversible characteristics (App. 1). The framework's risk mapping explicitly identifies cyber, nuclear, biological, chemical, and missile weapons knowledge and capability; loss of control; and the emergence of self-emergence and "competing with humanity for control" as risks to be addressed; §5.11 requires developers to conduct testing for loss-of-control risks. The Framework is non-binding (guidance rather than a national standard), but it signals priorities likely to be incorporated into future GB/T standards. At the first meeting of TC260's recently established WG9, a dedicated AI safety working group, chair Zhou Bowen called for "anticipating frontier risks in advance," suggesting possible further frontier-specific standards work (Wagner & Lindblad, 2026).

There have also been private-sector risk-management commitments. In December 2024, 17 Chinese AI companies (including DeepSeek, Zhipu.AI, MiniMax, 01.AI, Alibaba, Baidu, Huawei, and iFlytek) signed the AIIA AI Safety Commitments, a six-commitment framework covering dedicated safety teams, safety testing, data security, infrastructure security, model transparency, and frontier safety research; the latter includes an explicit focus on assessing biases, discrimination, or uncontrollability in agentic and embodied systems (Artificial Intelligence Industry Alliance, 2024). The China AI Security and Safety Commitments Framework, released at WAIC 2025 in July 2025 by CnAISDA members and led by the China Academy of Information and Communications Technology (CAICT), retained the first five commitments unchanged, rewrote Commitment VI to drop the explicit agent/embodied-AI specifics in favour of broader "misuse risks of AI systems in frontier fields"[37] and "high-risk scenarios,"[38] and added a seventh commitment on international cooperation on AI safety governance and AI for good (CnAISDA, 2025). The commitments emphasise transparency to stakeholders and the public, but they do not create enforceable obligations, such as publishing frontier safety frameworks. The CAICT, joined by groups like the AIIA, also operates the AI Safety Benchmark. Its v2.0 release adds a dedicated frontier-safety dimension covering self-awareness, model deception, dangerous-domain

[36] "特别重大安全风险" [extremely serious safety risk]
[37] "人工智能系统在前沿领域中的滥用风险" [misuse risks of AI systems in frontier fields]
[38] "高危场景" [high-risk scenarios]

misuse, and model loss of control, reaching frontier risks even as the algorithm registry and GB/T 45654-2025 focus mostly on content security (CAICT, 2026a).

Research labs have also proposed risk management measures. SafeWork-F1, the frontier AI risk management framework released by Shanghai AI Laboratory in partnership with Concordia AI, is described by Concordia AI as "China's first comprehensive framework for managing severe risks from general-purpose AI models" (AI Safety in China, 2025). It organises risks into four domains (misuse, loss of control, accident, and systemic) and assesses seven critical risk areas, including cyber offence, biological and chemical risks, persuasion and manipulation, strategic deception and scheming, uncontrolled autonomous AI R&D, self-replication, and collusion (Shanghai AI Lab & Concordia AI, 2025). The framework classifies models into green, yellow, or red zones. It defines red lines as intolerable thresholds (necessitating suspension of development or deployment) and yellow lines as early-warning indicators (requiring strengthened mitigations and controlled deployment) (Shanghai AI Lab & Concordia AI, 2025). In a July 2025 technical report applying the framework to recent frontier models, Shanghai AI Lab found that all evaluated models resided in green or yellow zones, with reasoning-enabled models entering yellow on self-replication and strategic deception. The report also flagged a decline in safety scores on newly released models (Shanghai AI Lab, 2025). The February 2026 v1.5 update expands evaluation across five of the seven risk areas and concludes that “current risks are manageable, but the emergence of granular vulnerabilities demands more rigorous and continuous monitoring protocols” (Liu et al., 2026). The framework is not regulatory, but its combination of taxonomy and empirical evaluation makes it the most concrete operationalisation of frontier risk monitoring published by any Chinese institution.

Though there is not yet a law implementing these safety frameworks, some of the concerns they address are entering concrete legislation. The Ministry of Public Security's Cybercrime Prevention Law (Draft for Comments), issued for public comment on January 31, 2026, extends China's service-provider obligations to AI-enabled offences (Ministry of Public Security of the PRC, 2026). The law requires network operators to conduct cybercrime risk assessments before bringing new technologies and applications (explicitly including AI) into service, and AI service providers must monitor, block, and report users who exploit their services to "mass-generate malicious code" or otherwise commit crimes (Art. 42). As with the instruments discussed above, the draft addresses misuse through provider-monitoring duties rather than capability-based thresholds, but it is notable as an acknowledgement that frontier models can lower the cost of automated cyberattacks. The draft is slated for National People's Congress Standing Committee review under the State Council's 2026 legislative work plan (General Office of the State Council, 2026).

3.3.2. Academic Draft Laws

Although China does not currently have a comprehensive AI law, several proposals led by academic working groups have emerged.[39] These drafts are intended to feed into legislative

[39] An industry AI law recommendation draft circulated in late May 2026, after this paper’s analysis was substantially complete (Smart Integration Standardization Construction [智合标准化建设], 2026). It was led by the Intelligeast Standards Center and Kinding Law Firm and adopts a scenario- and impact-based risk classification rather than a capability-based one; we note it for completeness but do not analyze it in depth.

processes, but they are not binding legislation. The CASS Model AI Law (MAIL)—a collaboration between CASS, academia, and industry—has progressively incorporated frontier-specific provisions across its several versions (MAIL Drafting Team, 2023, 2024, 2025, 2026). Earlier versions proposed a national AI authority, negative-list licensing, a FLOP-based foundation model definition, and special obligations for foundation model developers (including security and risk management systems), dedicated compute investment for safety, oversight by an independent institution, and annual social responsibility reports (MAIL Drafting Team, 2025).

The latest version of MAIL (v4.0), issued in May 2026, responds to recent developments in the industry. In response to the risks of loss of control associated with AI agents, v4.0 introduces more specific governance requirements. AI agent service providers are required to include functions such as revoking authorisations and terminating operations to help users retain meaningful control over agent activity (Art. 73), and AI developers must implement "circuit-breaker mechanisms" to prevent serious risks like loss of control (Art. 58). MAIL v4.0 also further distinguishes liability allocation across developers, providers, and users (Art. 100). The MAIL is a model law, not legislation, but v4.0 explicitly aligns with the 15th Five-Year Plan and is positioned as a reference document Chinese legislators and regulators may draw on.

Across three documents published from 2024 to 2026, the China University of Political Science and Law (CUPL) AI Good Governance Academic Working Group (coordinated by Zhang Linghan) has increased their focus on frontier AI. The 2024 Draft AI Law includes provisions for "critical AI," defined by compute, parameters, or scale of use (Art. 50) and contains special requirements for AGI,[40] including value alignment and capability-based safety assessments (Art. 77) (L. Zhang et al., 2024). The group's Key Institutional Recommendations continues in this vein by proposing risk compensation funds, NPC authority to temporarily adjust laws for AI pilot zones, and registration, plus regular risk assessment for critical AI providers (AI Good Governance Academic Working Group, 2026a). The group's 2026 research output, Top Ten Issues in Research on the Rule of Law in AI,[41] indicates a clear increased focus on frontier AI, with "Legal Approaches to Frontier AI Safety Assurance and Extreme Risk Prevention"[42] as Topic 6 (AI Good Governance Academic Working Group, 2026b). Topic 6 recognises that as AI systems advance in scale, autonomy, and depth of social embedding, risks "expand from localised, controllable risks to systemic, extreme risks,"[43] including technology misuse and loss of control (AI Good Governance Academic Working Group, 2026b). Working group expert Wen Yuheng summarised the scholarly consensus at the seminar releasing the topics, which is that frontier AI safety should not prioritise R&D over prevention and control and a full-lifecycle legal system should be established (AI Good Governance Academic Working Group, 2026b). None of the CUPL documents are binding law, but the trajectory across the three documents shows clear escalation of frontier-specific framing.

[40] The term used, 通用人工智能, can also mean "general AI," so it is unclear if the authors meant it in the same way as a Western AI safety audience meant.
[41] 《人工智能法治研究十大议题（2026）》[Top Ten Issues in Research on the Rule of Law in Artificial Intelligence (2026)]
[42] "前沿人工智能安全保障与极端风险防控的法治进路" [Legal Approaches to Frontier AI Safety Assurance and Extreme Risk Prevention]
[43] "由局部性、可控风险向系统性、极端性风险扩展" [expand from localised, controllable risks to systemic, extreme risks]

A third academic group entered this conversation in April 2026 with the Nanjing University Law School AI Law Team's Basic AI Law (Expert Recommendation Draft v1.0 (Nanjing University Law School AI Law Team, 2026)). The Nanjing draft is positioned as a basic or constitutional-tier framework law;[44] it sets principles and delegates operational detail to subordinate regulations (Art. 2). Unlike the CASS and CUPL drafts, the Nanjing draft does not adopt frontier-specific language; Article 26(4)'s concept of systemic risk[45] is the closest analogue. Even so, the draft creates the architectural slots through which frontier-specific measures could later be introduced. Article 26 defines high-risk AI activities and obliges those engaged in them to fulfil safety guarantee, risk control, and compliance management obligations commensurate with the risks; it then provides that "specific supervisory systems for high-risk AI activities shall be separately stipulated by laws, administrative regulations, or relevant national provisions" (Art. 26). Article 13's safety-and-controllability principle further requires that AI systems possess safety and controllability commensurate with their use, scope of impact, and degree of risk (Art. 13). Together, these provisions leave the door open for frontier-specific supervisory systems to be developed through subordinate regulation, even without the Basic Law itself naming them.

While academic legal groups have escalated frontier-specific framing across these proposals, the comprehensive-law trajectory at the official level is more complex. The State Council's 2023 and 2024 legislative work plans listed an AI Law as a preparatory item; the 2025 plan removed it, and replaced it with generic language about "advancing legislation for the sound development of AI" (Geopolitechs, 2025). The 2026 plan reversed course; it calls for accelerating the advancement of “comprehensive legislation for the healthy development of artificial intelligence"[46] and names six common elements requiring legislative protection, including data, compute, algorithms, intellectual property, cybersecurity, and supply chain security (General Office of the State Council, 2026). The National People's Congress Standing Committee's 2026 plan, by contrast, returned AI only as a backup research topic—the lowest of three legislative priority tiers—placing "the healthy development of artificial intelligence" alongside fiscal policy and online violence governance as subjects warranting study rather than active drafting (Wei, 2026). Still, the academic proposals continue to develop and refine in the interim, alongside a high volume of regulatory documents and standards on AI agents, ethics review, and related topics—indicating broad regulatory activity, even as the comprehensive AI law is signalled but not yet on an active drafting track. The October 2023 establishment of the National Data Administration—carved out of functions previously split between the CAC and the NDRC (Arcesati & Groenewegen-Lau, 2023)—demonstrates that Chinese institutions can reorganise bureaucratic architecture in response to coordination gaps, suggesting that a comparable consolidation around AI oversight is institutionally feasible if the comprehensive AI Law cycle produces an enacted statute.

In aggregate, Section 3 has shown two patterns. Existing AI-specific regulatory infrastructure (Sections 3.1 and 3.2) covers a variety of areas, but none of the existing instruments focus on the distinctive risks of frontier AI as a category, or are calibrated based on model

[44] 基本法 [Basic Law]
[45] “系统性风险” [systemic risk]
[46] “加快推进人工智能健康发展综合性立法” [accelerate the advancement of comprehensive legislation for the healthy development of artificial intelligence]

capability. The emerging frontier-specific layer (Section 3.3)—TC260 v2.0, the AIIA AI Safety Commitments, CASS MAIL v4.0, the CUPL drafts, and SafeWork-F1 v1.0 and v1.5—names the distinctive risks but is not yet part of formal regulation. The gap Section 4 addresses is therefore one of calibration rather than absence: the regulatory architecture exists, but could be extended to account for the greater capabilities and correspondingly greater risks of frontier models discussed in Section 2.1.

4. Proposals to Mitigate Frontier AI Governance Gaps

Building on the analysis of the global landscape of frontier AI governance and China's existing AI governance infrastructure, this section investigates how that infrastructure could be calibrated for frontier systems specifically. This section is not an exhaustive account of all possible ways to address frontier risks, but discusses three areas. Section 4.1 takes up frontier risk evaluation, Section 4.2 takes up emergency response, and Section 4.3 takes up liability. Each area builds on China's existing regulatory infrastructure. The ideas below augment existing standards, evaluation, response, and liability mechanisms, building on the broader regulatory ecosystem while concentrating regulatory attention on the risks distinctive to frontier systems. We offer these considerations based on the survey of international frontier AI regulation approaches, conversations with domain experts in China, and our understanding of the Chinese AI governance and regulatory environment.

4.1. Frontier Risk Evaluation

This section proposes three extensions to China's existing AI evaluation infrastructure, each adding a tier based on model capability to evaluation processes that currently calibrate by content class or application scenario: tiered frontier risk testing during the algorithm registry's security assessment (Section 4.1.1), continuous evaluation across the model lifecycle (Section 4.1.2), and capability-based thresholds for the agent registry that the May 2026 CAC/MIIT/NDRC Implementation Opinions already direct (Section 4.1.3). These are designed to increase regulator visibility into frontier model risks, which is a prerequisite for every other intervention Section 4 proposes. Without systematic capability evaluation, regulators cannot set response thresholds, define liability standards, or determine whether incident response can address the harm.

4.1.1. Tiered Pre-Deployment Risk Testing

Tiered frontier risk testing would apply an extra set of pre-deployment evaluations—covering risks such as CBRN uplift, autonomous capability acquisition, and loss of control—to models trained above a certain threshold. It is needed because a single content-safety evaluation bar, applied the same way to a small chatbot and a frontier foundation model, misses these risks: they arise at the top of the capability range and are categorically distinct from the primarily content-based harms the CAC's security requirements and GB/T 45654-2025's taxonomy addresses (see Section 3.1.3). A model can pass content-safety review and still

pose them, so catching them requires evaluation that scales with capability and is applied before deployment while intervention is still possible.

China's existing algorithm registry infrastructure could support a tiered frontier-risk testing regime without requiring new institutions. One pragmatic, tiered approach to layer frontier-specific testing on top of the existing pipeline that has emerged from discussions between researchers in China and internationally could work as follows:

- Frontier risk testing would apply only to general-purpose AI systems trained above a significant compute threshold. That threshold would be aligned to domestic AI law proposals and emerging international best practices. This would limit the scope to a small subset of advanced models. If a compute threshold was not appropriate, other methods of identifying in-scope systems drawing on other parts of Chinese regulation (e.g., a list of high-risk activities such as in the AI Ethics Review Measures) could also be considered.
- For models crossing that threshold, an initial round of cheap, automated testing against established frontier risk benchmark suites would provide a low-cost, standardised basis for systematic visibility.
- For models performing at or above the level of leading domestic and international models on those benchmarks, more intensive evaluation would follow.

This could be implemented through a TC260-drafted, frontier-safety annex to GB/T 45654-2025, enforced by the CAC through the algorithm registry's existing model security assessment process. Operating mainly through standards processes allows for co-development of technical requirements between industry, academia, and government over time, with requirements updated as methodologies improve rather than requiring legislative or direct regulatory revision. Keeping the tier thresholds current is a separate design question, and the academic drafts offer a template: earlier versions of CASS MAIL charged a national AI authority to "regularly update" compute standards (Art. 90(h))—an administrative-update mechanism that avoids the formal legislative-amendment cycle the EU faces (MAIL Drafting Team, 2025).

4.1.2. Continuous Evaluation

Frontier capability profiles shift after deployment through fine-tuning, model updates, scaffolding into agentic systems, and the emergence of behaviours in deployment contexts not present in the original evaluation (Bengio et al., 2026). A one-shot pre-deployment test captures a single moment in a changing system; continuous evaluation catches unanticipated risks that may surface later.

Regulators could require continuous monitoring and regular disclosures (e.g., on a quarterly basis) to regulators through revision to GB/T 45654-2025. The standard already requires providers to monitor model inputs continuously (§5.3, discussed in Section 3.1.3), but the monitoring targets content and input attacks, not the model's own capabilities. Re-targeting it to frontier safety would mean tracking capability changes after deployment, potentially acting on frontier risk tripwires or pre-set "if-then" thresholds: a defined capability level that, once

crossed, triggers re-evaluation, restriction, or reporting (Karnofsky, 2024). This changes what is monitored, not the underlying infrastructure.

Two academic draft AI laws already contemplate continuous assessment regimes and could provide longer-term legislative backing. MAIL v4.0 would require AI providers to operate a full-lifecycle risk-management system (Art. 57) and oblige foundation-model developers specifically to maintain a risk-management system with continuous risk monitoring (Art. 65); the CUPL Draft AI Law would require regular safety assessments and reports from AGI developers, including value alignment and capability-based assessments (Art. 77); and CUPL's 2025 Key Institutional Recommendations require critical-AI providers to register, conduct regular risk assessments, and submit reports (Sec. XIV).

4.1.3. Agent Registries

An agent built on a frontier foundation model can pursue consequential goals across different environments and sustain autonomous action under reduced human oversight. Its failures are harder to attribute and to intercept than direct misuse. A harmful prompt can surface in the model's monitored inputs and outputs, but an agent's harm may emerge from a chain of autonomous actions across tools and environments that can be spread across many steps, and may materialise in effects the model's input or output channel does not capture (and, in the case of a self-hosted model, with no provider in the loop at all).

In May 2026, CAC, MIIT, and NDRC jointly issued the Implementation Opinions on the Standardized Application and Innovative Development of AI Agents (Implementation Opinions), directing the development of an "agent registry platform" that would track digital identity, capability declarations, developer, deployment, protocol, and compliance information (Cyberspace Administration of China et al., 2026). The Implementation Opinions establish a categorised and tiered governance framework, but the tiering is based on sector and application context, not model capability. This means that agents in sensitive sectors and key industries face filing, additional testing, and product-recall management, while agents in low-risk lifestyle and office contexts face self-testing and industry self-discipline (Cyberspace Administration of China et al., 2026). CAICT, MIIT, and TC260 are working on agent standards in parallel (MIIT/TC1, 2026; State Administration for Market Regulation and Standardization Administration of China, 2026; TC260, 2026).

Sector-based tiering creates a specific gap for agents built on frontier foundation models, because a general-purpose model is not bound to one use case. The same model that powers a "low-risk lifestyle" agent can be redeployed across other tasks, repurposed beyond its intended scope, or exploited via jailbreaking or fine-tuning. Sector-based tiering assumes a stability of use case that general-purpose frontier systems do not have. Adding capability as a tiering dimension would close this gap. An agent built on a model above a defined capability threshold would carry the filing, additional testing, and product-recall requirements the Implementation Opinions already reserve for "key industries," regardless of its initial deployment sector. That testing would assess the agentic system—the model together with its tools, autonomy, and scaffolding—which the model-level evaluation in Section 4.1.1 does not reach. Crossing the capability threshold, not a change of sector, is what places an agent in the higher tier.

The Implementation Opinions are non-binding direction-setting rather than regulation, and the registry platform is described as something to "explore establishing" rather than as something already operational. Capability-based tiering could therefore be added before the platform's substantive design is locked in. Open questions remain on whether registration should attach to the agent framework developer, the agent deployer, or the end user; how open-source agent frameworks are treated; and how cross-border agent communication protocols (e.g., MCP, ACP, or A2A) are brought within the registry's scope.

4.2. Emergency Response Adaptation

This section proposes adapting China's emergency response infrastructure to frontier AI incidents through two extensions: calibrated reporting thresholds that account for how frontier risks surface (Section 4.2.1), and frontier-specific incident categories with appropriate severity thresholds (Section 4.2.2). These extensions build on the architecture Section 3.2.1 described—the Emergency Response Law, the National Emergency Response Plan, and TC260's Emergency Response Guidelines—and on the more recent MIIT 2026 Pilot Plan (discussed below). None of these layers currently calibrates to the specific risk profile of frontier AI scenarios.

As Section 3.2.1 discussed, the TC260 generative AI Emergency Response Guidelines are suited for content-safety incidents and service disruptions, but do not account for scenarios where frontier AI risk materialises or for cascading and autonomous behaviour. Additionally, frontier AI emergency response is qualitatively different from content safety response. Frontier harms are decoupled from content—in a content incident, the harm is the content being seen, but with a frontier AI incident, harm materialises when a malicious actor uses information or a system’s capabilities to act. Frontier AI harms can also be far more severe and difficult to mitigate, reaching physical or systemic catastrophe rather than the bounded harm of content being seen. They may also be harder to detect than content harms; a model reasoning about circumventing oversight may look like normal operation from outside. Finally, there is no established playbook for AI emergency response. Response regimes are anchored in natural disasters and cybersecurity incidents, which frontier AI incidents may contain aspects of, but they are also likely to have novel response requirements. No jurisdiction has established protocols for an emerging incident of this scale and type.

Together, these differences mean that frontier AI emergency response requires different severity metrics, detection mechanisms, and response models, as well as greater cross-organisational coordination than the current generative AI response framework provides.

The academic draft AI laws contemplate emergency response as part of the broader frontier-risk regime: CUPL's 2024 Draft AI Law would authorise emergency shutdown for critical AI providers (Art. 56), and CUPL's 2025 Key Institutional Recommendations recommend “critical AI” providers establish safety emergency response plans as a registration condition (Sec. XIV). In addition to its “circuit breaker” requirements, CASS MAIL v4.0 would require both critical AI providers and the proposed National AI Administrative Authority to establish emergency response systems (Arts. 42, 81). Most recently, in May 2026, MIIT launched the Pilot Plan for AI Science and Technology Ethics Review and Services (Xinhua, 2026), establishing a three-level (ministry, province, city) agile governance

network across ten provinces and municipalities, with a national AI ethics risk monitoring and service network designed to push risk warnings and alerts to local governments and innovators. Although the emergency response plans are established by developers, the Pilot Plan is the closest Chinese-system precedent for an AI-specific monitoring, early-warning, and alert-push infrastructure. The limit is the same architecture-versus-calibration distinction as elsewhere: it is framed around ethics risks, not frontier-capability risks. Still, these provide precedent that grounds the following two ideas around incident reporting and response.

4.2.1. Calibrated Reporting Thresholds

The current threshold of five occurrences per 24 hours per 10,000 users reporting in the TC260 Emergency Response Guidelines for Generative AI Service Security (TC260, 2025b, §6.3.2) registers harm only once a service reaches users at scale. However, the earliest frontier warning signs surface before any user is affected. This can look like a model crossing a dangerous-capability threshold during evaluation, or a pre-deployment red-teaming finding of unexpected autonomous behaviour. These are potential indicators of catastrophic misuse or loss-of-control potential, but neither would trip the current threshold. Calibrated reporting thresholds for frontier AI incidents would therefore escalate on what an incident reveals about a model's risks—capability range, autonomy level, and downstream-harm potential—not on how many users were affected. This could be accomplished through a direct amendment to §6.3.2 of the TC260 Emergency Response Guidelines for Generative AI Service Security.

4.2.2. Frontier-Specific Incident Categories

Frontier AI incidents pose threats that fall outside the scope of existing content-based incident categories and thus require updated frontier-specific categories with their own severity thresholds. The TC260 Emergency Response Guidelines for Generative AI Service Security covers information content security, data security, and cyberattacks, with severity calibrated against impacts to businesses, data, and social order (TC260, 2025b). Frontier AI also poses threats to physical security—CBRN uplift, critical infrastructure compromise, autonomous capability acquisition, and loss of control—that the current categories do not capture and that may require distinct escalation protocols. The National Natural Disaster Relief Emergency Plan offers one possible model, grading responses based on the number of deaths and missing people, people in need of emergency relocation, and damage to homes (General Office of the State Council, 2024)—an approach that parallels the approaches from California, New York, and Illinois, which classify catastrophic risks as those contributing to the death or injury of more than 50 people or more than $1 billion in property damage. Creating frontier-AI-specific incident categories with explicit severity thresholds would ensure appropriate escalation for the highest-severity scenarios without rebuilding the existing emergency-response architecture. TC260 could be the implementing body for a revision along these lines, with coordination from the Ministry of Emergency Management for incidents that cross sectoral boundaries.

4.3 Liability Adaptation

Extended to frontier AI, China's tort system faces the three structural problems Section 3.2.2 set out: establishing causation between developer conduct and model behaviour, apportioning liability across the value chain, and uninsurable tail risks. Two features of the system shape the paths below. First, it is civil law, meaning that judicial decisions carry no formally binding precedential effect. Second, the SPC issues judicial interpretations (司法解释) and model cases (典型案例) that function as authoritative supplementary law and are routinely followed by lower courts, so several adaptations can be delivered through SPC guiding opinions rather than NPC legislation; this is already being used to address AI case adjudication (Tao, 2025). The four adaptations below build on these.

China's three academic draft AI laws each move beyond pure fault-based liability, and each takes a different approach. CASS MAIL v4.0 takes a fault-based but actor-specific approach. Developers are liable for a set of enumerated failures, such as improper training data or algorithm design, safety testing that does not meet standards, or knowing of a major hidden hazard and failing to act. Providers bear a reversed burden of proof and are liable unless they can prove the absence of fault (Art. 100), with a separate liability exemption for open-source developers (Art. 104). The CUPL Draft AI Law would establish a two-tier system with fault-based liability for most applications of AI, plus reversal of the burden of proof for "critical AI," where the provider bears liability unless it can prove absence of fault (Art. 85). Joint and several liability are layered on top for foundation model providers who knew or should have known of downstream misuse (Art. 90). The Nanjing Basic AI Law takes a third path: proportional liability across the value chain. Researchers and developers, providers, deployers, and users would bear civil liability "in their respective shares" based on their degree of fault, capacity for control, and contribution to the harmful result (Art. 56). Together the three drafts offer three distinct mechanisms: burden reversal, joint and several liability for foundation model providers, and proportional value-chain allocation. The four adaptations below build on these proposals, but are more nascent compared to those in Sections 4.1 and 4.2.

4.3.1. Joint and Several Liability

Frontier AI harms typically involve multiple actors across the value chain, and current tort doctrine struggles to apportion responsibility across the chain. True proportional liability would distribute liability more fairly but requires technical expertise most courts lack. Joint and several liability—where each party is independently liable for the full extent of damages—offers a constructive alternative. The CUPL draft already adopts this approach for foundation model providers (Art. 90); providers who "know or should know" of downstream misuse bear joint and several liability with downstream actors. To ascertain developer responsibility, extending the standard from "know or should know" to "know or should have known given a reasonable evaluation regime" would link the evaluation infrastructure proposed in Section 4.1 directly to liability allocation, making it clear that developers who shipped frontier capabilities without adequate testing cannot pass the legal risks they created downstream. Explicitly including other actors would also ensure that responsibility for diffuse frontier AI harms is appropriately allocated. Two institutional paths could deliver this: NPC legislation incorporating CUPL Art. 90's joint-and-several model into the Civil Code

or into a future AI Law, and SPC guiding opinions on AI adjudication along the lines of what the Court is already drafting (Tao, 2025).

4.3.2. Government-Backed Catastrophic Risk Insurance

Private insurers are reluctant to underwrite tail risks because a single catastrophic event could exceed their capacity, leaving plaintiffs unable to recover even where liability is established. The CUPL Draft AI Law already encourages AI developers to obtain insurance and insurers to develop novel AI-specific products (Art. 26), and CUPL's 2025 Key Institutional Recommendations propose risk compensation funds (Sec. XII). Government reinsurance, similar to the terrorism risk pools established in the US after September 11, 2001 (Federal Insurance Office, 2024), could backstop losses above a defined threshold while preserving private-market incentives for routine risks. This connects directly to Section 4.2's emergency response infrastructure. Incidents that trigger top-tier escalation would also trigger the backstop, ensuring that the scenarios most likely to exceed private insurer capacity remain ones where victims can still recover. Implementation would require NPC legislation establishing a risk compensation fund (per CUPL Working Group, 2025, Sec. XII), administered jointly by the Ministry of Finance and the National Financial Regulatory Administration (NFRA); this is the kind of mechanism well suited to pilot-zone experimentation before national rollout.

4.3.3. Adjusted Proof Standards

A fault-based claim requires the plaintiff to establish fault, causation, and harm. Against an opaque, fast-changing model, information asymmetry defeats most claims on the first two stages, blunting tort liability's deterrent effect. Three adjustments, each relieving a different element of that burden, would address this.

The first reallocates fault, which the draft laws already do for one set of actors: MAIL v4.0 makes providers liable unless they prove the absence of fault (Art. 100), and CUPL reverses the burden for "critical AI" providers (Art. 85(2)). Developers, by contrast, are not subject to that reversal: the claimant must prove a specific, enumerated failure like improper training data or algorithm design, safety testing that falls short of standards, or knowing of a major hidden hazard and failing to act (Art. 100).

The second addresses access to evidence. The records needed to prove such a failure—training data, evaluation logs, red-teaming results—sit with the developer. Evidence-disclosure requirements, modelled on the revised EU Product Liability Directive, would let courts order their production on a plausible showing (Directive (EU) 2024/2853, Art. 9), with non-disclosure triggering a rebuttable presumption of defectiveness (Art. 10).

The third addresses causation, which disclosure alone cannot resolve. Even with the records, a claimant cannot ordinarily trace how a particular development decision produced a model's output. A presumption of causation would close this gap on the developer track. Once the claimant establishes an enumerated failure and harm of the kind the duty guards against, the court presumes the failure caused the harm unless the developer rebuts it. This extends an existing domestic mechanism rather than importing one. Chinese tort law already shifts the causation burden to the defendant in defined contexts as seen in environmental

liability (Civil Code, Art. 1230). Together, these three adjustments relieve the plaintiff on fault, evidence, and causation. SPC procedural guidance could implement the disclosure rules; the fault reversal and the causation presumption, as changes to how liability is established, would require NPC legislation drawing on the MAIL or CUPL proposals.

4.3.4. Confidential Government-Facing Disclosures

Public-facing transparency requirements can produce less disclosure, not more: developers facing public scrutiny have incentive to under-disclose to avoid reputational exposure and downstream litigation. Critics of California SB-53 have raised this concern about its public frontier AI framework publication requirement; while Anthropic continues to publish its Responsible Scaling Policy alongside its SB-53-mandated Frontier Compliance Framework (FCF), it notes that the FCF "might not always align with emerging best practices" (Anthropic, 2025, 2026). China's regulatory tradition already favours confidential government-facing disclosures over public ones: the CAC algorithm registry, ethics review system, and AIIA AI Security and Safety Commitments all involve disclosures to regulators rather than to the public.[47] Calibrating frontier AI disclosure requirements along this axis avoids public-disclosure requirements producing less disclosure while preserving regulator visibility for evaluation, incident response, and liability allocation. The CAC could implement this directly as an enhancement to the existing algorithm registry filing schema, potentially reinforced with a TC260 standard on frontier AI disclosure content.

5. Conclusion

This paper has surveyed how three jurisdictions outside of China have approached frontier AI governance (Section 2), mapped the substantial regulatory infrastructure China has built for AI more broadly (Section 3), and considered how that infrastructure could be calibrated to the specific risk profile of frontier systems (Section 4). The throughline is that China is not starting from scratch. Its existing regulatory apparatus—the algorithm registry, GB/T 45654-2025's risk assessment regime, the Emergency Response Law, academic draft AI laws, ethics review architecture, and the May 2026 Implementation Opinions on AI Agents, among others—provides a strong foundation. What is missing is calibration to capability rather than to content class or application scenario. The three areas Section 4 takes up are extensions of infrastructure that already exists, not parallel regimes that compete with it.

We offer these considerations in the spirit of mutual learning. No jurisdiction has solved frontier AI governance: the EU AI Act's GPAI provisions and the US state transparency laws all face implementation questions and have recognised gaps, and the academic draft AI laws in China are themselves works in progress. We have drawn on the diversity of approaches because that diversity is itself the material for joint analysis, and the analytical convergence around frontier risks—visible in TC260's loss-of-control language, CASS MAIL v4.0's foundation-model risk pillars, CUPL's critical-AI provisions, and Shanghai AI Lab's SafeWork framework—suggests that the gap between Chinese and international framings is narrower than is sometimes assumed. This paper is one contribution to a conversation Chinese scholars are already having.

[47] The AIIA discloses limited, anonymised information about signatories' practices (Artificial Intelligence Industry Alliance, 2025).

For readers uncertain whether frontier risks will materialise on the timelines some forecasters propose, the proposals in Section 4 still warrant consideration as contingency planning. Just as emergency response infrastructure for natural disasters is maintained regardless of whether a specific earthquake or pandemic is imminent, the Section 4.1 evaluation extensions and the Section 4.2 emergency response calibrations are most necessary precisely when frontier capabilities remain uncertain because they generate the visibility and response infrastructure needed to react if those capabilities arrive faster than expected. If they instead arrive slowly, the cost is low. The Section 4.3 liability extensions and the Section 4.3.4 disclosure proposal are more exploratory; they shape incentives over a longer horizon and depend on doctrinal choices that remain open.

Several of the Section 4 mechanisms could be tested in China's pilot zone infrastructure before national rollout. CUPL's 2025 Recommendations already propose using NPC pilot-zone authority to experiment with AI-specific institutional design (AI Good Governance Academic Working Group, 2026a), and China's broader experimental regulatory tradition (Heilmann, 2008)—including the 2009 cross-border RMB settlement pilot and the 2013 carbon emissions trading pilots, both of which launched in several cities before scaling nationwide (Cui et al., 2021; L. Li & Xu, 2022)—favours testing in bounded settings before nationwide implementation. Liability frameworks, certification regimes, and insurance backstops are particularly well-suited to this pattern as they benefit from case law and operational experience before being locked into national legislation, and pilot zones allow that experience to accumulate without committing the system as a whole to any single design.

This paper has set aside several mechanisms that warrant fuller treatment as the frontier AI landscape evolves. Licensing regimes—which China already approximates for public-facing generative AI through the algorithm registry filing process—could be extended into a more substantive ex ante review for frontier systems, building on the Multi-Level Protection Scheme (MLPS) 2.0 private cybersecurity certification model (Protiviti, 2022). Market-driven regulation and regulatory markets more broadly, in which private regulatory service providers compete to certify developers with tort safe harbors for those who pass, offer a way to bridge public agency capacity gaps that may become acute as evaluation methodologies scale. S&T ethics review operationalisation, pre-training oversight mechanisms, and agent-specific governance beyond registries (including the cross-border protocol questions Section 4.1.3 raises) deserve fuller treatment than this paper provides. Several of these are also candidates for pilot-zone experimentation.

The fundamental claim is that China's existing regulatory infrastructure is a genuine asset for addressing frontier risks. The standards bodies, the algorithm registry, the academic draft laws, the ethics review architecture, the emergency response system, and the pilot zone tradition are not regulatory exotica; they are the material out of which a frontier-AI-specific framework can be built. Whether and how to do so is a decision for Chinese regulators and scholars. The proposals in this paper are offered as considerations within a conversation that is already underway.

Acknowledgements

We would like to acknowledge Nick Caputo, Alan Chan, Fang Liang, and Gabriel Wagner for their contributions.

References

AI Good Governance Academic Working Group. (2026a, January 31). *Key Institutional Recommendations for Artificial Intelligence Legislation*. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/WLtHccXus58ZOMC41dqQWA

AI Good Governance Academic Working Group. (2026b, March 23). *Top Ten Topics in AI Legal Research in 2026*. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/bQHNuVJ5hGqbzuDYCHJnKA

AI Safety in China. (2025, July 31). Shanghai AI Lab and Concordia AI release Frontier AI Risk Management Framework v1.0 and Technical Report [Substack newsletter]. *AI Safety in China*. https://aisafetychina.substack.com/p/shanghai-ai-lab-and-concordia-ai

Anderljung, M., Barnhart, J., Korinek, A., Leung, J., O'Keefe, C., Whittlestone, J., Avin, S., Brundage, M., Bullock, J., Cass-Beggs, D., Chang, B., Collins, T., Fist, T., Hadfield, G., Hayes, A., Ho, L., Hooker, S., Horvitz, E., Kolt, N., … Wolf, K. (2023). *Frontier AI Regulation: Managing Emerging Risks to Public Safety* (arXiv:2307.03718). arXiv. https://doi.org/10.48550/arXiv.2307.03718

Anhui Cyberspace Administration. (2026, March 18). *CPPCC National Committee member Qi Xiangdong: Pre-setting "safety barriers" for AI is crucial [*全国政协委员齐向东：为*AI*预设*"*安全护栏*"*至关重要*]*. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/zvSsnWTBdzL0gvDEh5thzA

Anthropic. (2025, December 19). *Frontier Compliance Framework*. https://www.anthropic.com/news/compliance-framework-SB53

Anthropic. (2026, May 26). *Responsible Scaling Policy (version 3.3)*. https://cdn.sanity.io/files/4zrzovbb/website/c11e84981d0a7281a1b229f3fa6af0da66eaf43f.pdf

Arcesati, R., & Groenewegen-Lau, J. (2023, December). *China's data management: Putting the party-state in charge*. MERICS.

Artificial Intelligence Industry Alliance. (2024, December). *Artificial Intelligence Safety Commitments*. https://mp.weixin.qq.com/s/s-XFKQCWhu0uye4opgb3Ng

Artificial Intelligence Industry Alliance. (2025, July). *Disclosure of Practices on the Artificial Intelligence Security and Safety Commitments*. https://aihub.caict.ac.cn/ai_security_and_safety_commitments

Artificial Intelligence Safety Measures Act, SB0315, Illinois General Assembly 104th General Assembly (2024). https://ilga.gov

Baker, T. (2023, September 19). Blumenthal and Hawley's U.S. AI Act Framework: CSET's Perspective and Contributions. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/article/blumenthal-and-hawleys-u-s-ai-act-framework-csets-perspective-and-contributions/

Ball, D. W. (2025). *A Framework for the Private Governance of Frontier Artificial Intelligence* (arXiv:2504.11501). arXiv. https://doi.org/10.48550/arXiv.2504.11501

Ball, D. W., & Ramakrishnan, K. (2025, July 7). *Entity-Based Regulation in Frontier AI Governance*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2025/06/artificial-intelligence-regulation-united-states?lang=en

Barrett, A. M., Jackson, K., Murphy, E. R., Madkour, N., & Newman, J. (2024). *Benchmark Early and Red Team Often: A Framework for Assessing and Managing Dual-Use Hazards of AI Foundation Models* (arXiv:2405.10986). arXiv. https://doi.org/10.48550/arXiv.2405.10986

Bathaee, Y. (2018). The Artificial Intelligence Black Box and the Failure of Intent and Causation. *Harvard Journal of Law & Technology*, *31*(2).

Belfield, H. (2025). *Domestic frontier AI regulation, an IAEA for AI, an NPT for AI, and a US-led Allied Public-Private Partnership for AI: Four institutions for governing and*

developing frontier AI (arXiv:2507.06379). arXiv. https://doi.org/10.48550/arXiv.2507.06379

Bengio, Y., Clare, S., Prunkl, C., Andriushchenko, M., Bucknall, B., Murray, M., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., … Mindermann, S. (2026). *International AI Safety Report 2026* (arXiv:2602.21012). arXiv. https://doi.org/10.48550/arXiv.2602.21012

Blumenthal, R., & Hawley, J. (2023, September 7). *Bipartisan Framework for U.S. AI Act*.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S. J., Belfield, H., Farquhar, S., … Amodei, D. (2024). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation* (arXiv:1802.07228). arXiv. https://doi.org/10.48550/arXiv.1802.07228

Bryan, K. A., & Teodoridis, F. (2024, September 24). Balancing market innovation incentives and regulation in AI: Challenges and opportunities. *Brookings*. https://www.brookings.edu/articles/balancing-market-innovation-incentives-and-regulation-in-ai-challenges-and-opportunities/

Buhl, M. D., Sett, G., Koessler, L., Schuett, J., & Anderljung, M. (2024). *Safety cases for frontier AI* (arXiv:2410.21572). arXiv. https://doi.org/10.48550/arXiv.2410.21572

Bullock, C., Van Arsdale, S., Arnold, M., O'Keefe, C., & Winter, C. (2024, September 30). Legal Considerations for Defining "Frontier Model." *Institute for Law & AI*. https://law-ai.org/frontier-model-definitions/

CAICT. (2026a, February 10). *AI Safety Benchmark 2.0*. https://mp.weixin.qq.com/s/tzfCOF3nKYFvS9zozoXDww

CAICT. (2026b, March 5). *Transcript | Li Lecheng discusses modern industrial system and artificial intelligence industry in "Ministerial Corridor" [*实录丨李乐成在"部长通道"谈现代化产业体系、人工智能产业*]*. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/Yp8akBW27BKJAHJxBEc7WA

Cantero Gamito, M., & Marsden, C. T. (2024). Artificial intelligence co-regulation? The role of standards in the EU AI Act. *International Journal of Law and Information Technology*, *32*, eaae011. https://doi.org/10.1093/ijlit/eaae011

Carlini, N., Cheng, N., Lucas, K., Moore, M., Nasr, M., Prabhushankar, V., & Xiao, W. (2026, April 7). *Assessing Claude Mythos Preview's cybersecurity capabilities*. Anthropic. https://www.anthropic.com/research/mythos-preview?utm_source=tldrai

China Atomic Energy Authority. (2022, April 23). *National Nuclear Emergency Response Plan*. https://www.caea.gov.cn/n6760401/n6760402/c6827755/content.html

China Daily. (2024, June 21). *Face-swapping app fined for infringements*. Beijing Internet Court. https://english.bjinternetcourt.gov.cn/2024-06/21/c_721.htm

China Daily. (2025, May 29). *SPC upholds voice rights, curbs AI misuse*. The Supreme Court of the People's Republic of China. https://english.court.gov.cn/2025-05/29/c_1096900.htm

China Financial Information Network. (2025, July 17). *18 companies disclosed the practical results of their "Artificial Intelligence Security Commitment," accelerating the process of AI security governance [18*家企业披露《人工智能安全承诺》实践成果 加快*AI*安全治理进程*]*. Sina Finance. https://finance.sina.com.cn/jjxw/2025-07-17/doc-inffurfr7406886.shtml

China Law Translate. (2022, January 4). Provisions on the Management of Algorithmic Recommendations in Internet Information Services. *China Law Translate*. https://www.chinalawtranslate.com/algorithms/

China Law Translate, & Daum, J. (2026, January 20). Alien Chat: China's 1st Case of Criminal Responsibility for AI-Generated Content [China Law Translate]. *China Law Translate*. https://www.chinalawtranslate.com/22286-2/

China SEI. (2026, April 12). *Academician Huang Ru of the Chinese Academy of Sciences and Vice Chairman of the National Development and Reform Commission: "Comprehensive Implementation of the 'Artificial Intelligence+' Action Plan"—An*

*Interpretation and Guidance of the 15th Five-Year Plan [*中科院院士 国家发展改革委副主任 黄如《全面实施"人工智能+"行动》——《十五五纲要》解读辅导*].* 中国战略新兴产业. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/0N9KYiesA_iNKP5MS4IBfg

Civil Code of China: Book VII Liability for Tort (2020) (2021). https://www.chinajusticeobserver.com/law/x/civil-code-of-china-part-vii-liability-for-tort-20200528

CnAISDA. (2025, July 31). *China Artificial Intelligence Security & Safety Commitment Framework*《中国人工智能安全承诺框架》. CWW. CWW. https://www.cww.net.cn/article?id=602676

CPC Central Committee & State Council. (2025, February 25). *National Emergency Response Plan*. https://www.gov.cn/zhengce/202502/content_7005635.htm

CSET. (2025, April 26). At the 20th Collective Study Session of the CCP Central Committee Politburo, Xi Jinping Stresses: Persist in Being Self-Reliant, Be Strongly Oriented Toward Applications, and Push the Orderly Development of Artificial Intelligence. *Center for Security and Emerging Technology*. https://cset.georgetown.edu/publication/xi-politburo-collective-study-ai-2025/

Cui, J., Wang, C., Zhang, J., & Zheng, Y. (2021). The effectiveness of China's regional carbon market pilots in reducing firm emissions. *Proceedings of the National Academy of Sciences*, *118*(52), e2109912118. https://doi.org/10.1073/pnas.2109912118

Cybersecurity Law of the People's Republic of China (2026). https://www.gov.cn/yaowen/liebiao/202510/content_7046194.htm

Cyberspace Administration of China. (2025a, September 15). 国家网络安全事件报告管理办法 *[National Cybersecurity Incident Reporting Management Measures]*. https://www.cac.gov.cn/2025-09/15/c_1759583017717009.htm

Cyberspace Administration of China. (2025b, December 27). *Provisional Measures on the Administration of Human-like Interactive Artificial Intelligence Services (Draft for Comments)*. https://www.cac.gov.cn/2025-12/27/c_1768571207311996.htm

Cyberspace Administration of China. (2026, March 17). *Announcement Regarding the Release of Filing Information for Generative Artificial Intelligence Services (January to February 2026)*. https://www.cac.gov.cn/2026-03/17/c_1775482074695536.htm

Cyberspace Administration of China, National Development and Reform Commission of the People's Republic of China, & Ministry of Industry and Information Technology. (2026, May 8). *Implementation Opinions on the Standardized Application and Innovative Development of Intelligent Agents*. https://www.cac.gov.cn/2026-05/08/c_1779979789523320.htm

Department for Science, Innovation & Technology. (2023a, September 25). *Introduction to the AI Safety Summit*. https://www.gov.uk/government/publications/ai-safety-summit-introduction

Department for Science, Innovation & Technology. (2023b, October). *Capabilities and risks from frontier AI*. https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf

Department for Science, Innovation & Technology. (2024, February 6). *A pro-innovation approach to AI regulation: Government response*. GOV.UK. https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on Liability for Defective Products and Repealing Council Directive 85/374/EEC (Text with EEA Relevance), CONSIL, EP (2024). http://data.europa.eu/eli/dir/2024/2853/oj

Draghi, M. (Ed.). (2025). *The future of European competitiveness: Part A: A competitiveness strategy for Europe*. Publications Office. https://doi.org/10.2872/1823372

Emergency Response Law of the PRC (2024 Update) (2024). https://yjglj.sh.gov.cn/xxgk/xxgkml/zcfg/gjfl/20240805/b50983fa4be44583b190d50de11fb7fe.html

European Commission. (2024, September 24). *Over a hundred companies sign EU AI Pact pledges* [Text]. European Commission. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4864

European Commission. (2025a, July). *The General-Purpose AI Code of Practice*. Directorate-General for Communications Networks, Content and Technology. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

European Commission. (2025b, November 19). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as regards the simplification of the digital legislative framework, and repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus)*. European Commission. https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal

European Commission. (2025c, November 19). *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL amending Regulations (EU) 2024/1689 and (EU) 2018/1139 as regards the simplification of the implementation of harmonised rules on artificial intelligence (Digital Omnibus on AI)*. European Commission. https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-ai-regulation-proposal

Executive Office of the President. (2026, June 2). *Promoting Advanced Artificial Intelligence Innovation and Security*. https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/

Executive Office of the President of the United States. (2025, July). *Winning the Race: America's AI Action Plan*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Fan, J., & Zhang, X. (2026). 'Small, swift, and effective' legislation in China: Towards adaptive governance of AI? *Computer Law & Security Review*, *61*, 106329. https://doi.org/10.1016/j.clsr.2026.106329

Fang, W. (2025, June 13). 中国人保首推生成式*AI*侵权保险|保险_新浪财经_新浪网. Sina. https://finance.sina.com.cn/jjxw/2025-06-13/doc-inezwrzq6222310.shtml

Federal Insurance Office, US Department of the Treasury. (2024). *Report on the Effectiveness of the Terrorism Risk Insurance Program*. https://home.treasury.gov/system/files/311/2024ProgramEffectivenessReportFINAL6.28.2024508.pdf

Fjelland, R. (2020). Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications*, *7*(1), 10. https://doi.org/10.1057/s41599-020-0494-4

Frazier, K. (2025, June 4). *We're Not Ready for AI Liability*. AI Frontiers. https://ai-frontiers.org/articles/options-for-ai-liability

General Office of the State Council. (2024, January 20). 国家自然灾害救助应急预案 *[National Natural Disaster Relief Emergency Plan]*. https://www.gov.cn/zhengce/content/202402/content_6930038.htm

General Office of the State Council. (2026, May 11). *Notice from the General Office of the State Council on Issuing the "State Council's Legislative Work Plan for 2026" (*国务院办公厅关于印发《国务院*2026*年度立法工作计划》的通知*)*. https://www.moj.gov.cn/pub/sfbgw/gwxw/xwyw/202605/t20260511_534682.html

Geopolitechs. (2025, August 26). *China's AI Law: Recent Developments and Legislative Signals*. https://www.geopolitechs.org/p/chinas-ai-law-recent-developments

Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., … Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, *645*(8081), 633–638. https://doi.org/10.1038/s41586-025-09422-z

Hadfield, G. K., & Clark, J. (2026). *Regulatory Markets: The Future of AI Governance* (arXiv:2304.04914). arXiv. https://doi.org/10.48550/arXiv.2304.04914

Hangzhou Internet Court. (2025, December 25). 生成式人工智能"幻觉"侵权纠纷案一审生效. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/riMkHOiBhDra1xe70wQcRA

Heilmann, S. (2008). Policy Experimentation in China's Economic Rise. *Studies in Comparative International Development*, *43*(1), 1–26. https://doi.org/10.1007/s12116-007-9014-4

Heim, L., & Koessler, L. (2024). *Training Compute Thresholds: Features and Functions in AI Regulation* (arXiv:2405.10799; Version 2). arXiv. https://doi.org/10.48550/arXiv.2405.10799

Hine, E. (2024). Governing Silicon Valley and Shenzhen: Assessing a New Era of Artificial Intelligence Governance in the United States and China. *Digital Society*, *3*. https://doi.org/10.1007/s44206-024-00138-7

Hine, E., & Floridi, L. (2025). From No-Win to No-Lose: *Journal of AI Law and Regulation*, *2*(3), 196–211. https://doi.org/10.21552/aire/2025/3/4

Hogan Lovells. (2025, July 2). *The new EU Product Liability Directive: Implications for software, digital products, and cybersecurity*. Hogan Lovells. https://www.reedsmith.com/articles/eu-product-liability-directive-software-digital-products-cybersecurity/

Hou, L. (2018, March 30). *New authority focuses on emergency response*. China Daily. https://www.chinadaily.com.cn/a/201803/30/WS5abd8f19a3105cdcf6515441.html

IDAIS-Beijing. (2024). *International Dialogues on AI Safety*. https://idais.ai/dialogue/idais-beijing/

Interim Measures for the Management of Generative Artificial Intelligence Services (2023). http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

Internet Information Service Algorithmic Recommendation Management Provisions (2021). http://www.gov.cn/zhengce/zhengceku/2022-01/04/content_5666429.htm

Jeanmaire, C. (2025, November 12). 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/

Jones, J., Milner-Smith, A., Ramsay, L., & Athwal, S. (2025, October). Global AI Governance Law and Policy: United Kingdom. *IAPP*. https://iapp.org/resources/article/global-ai-governance-uk/

Karnofsky, H. (2024). *If-Then Commitments for AI Risk Reduction*. https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction

Kelly, K. (2017, April 25). The Myth of a Superhuman AI. *Wired*. https://www.wired.com/2017/04/the-myth-of-a-superhuman-ai/

Kinniment, M., Sato, L. J. K., Du, H., Goodrich, B., Hasin, M., Chan, L., Miles, L. H., Lin, T. R., Wijk, H., Burget, J., Ho, A., Barnes, E., & Christiano, P. (2023). *Evaluating Language-Model Agents on Realistic Autonomous Tasks*. https://arxiv.org/abs/2312.11671v2

Kuersten, A. (2023, May 26). *Introduction to Tort Law*. Library of Congress. https://www.congress.gov/crs-product/IF11291

Li, L., & Xu, F. (2022). Does Shanghai International Financial Center Promote RMB Internationalization? *Discrete Dynamics in Nature and Society*, *2022*(1), 4000024. https://doi.org/10.1155/2022/4000024

Li, N., Pan, A., Gopal, A., Yue, S., Berrios, D., Gatti, A., Li, J. D., Dombrowski, A.-K., Goel, S., Phan, L., Mukobi, G., Helm-Burger, N., Lababidi, R., Justen, L., Liu, A. B., Chen,

M., Barrass, I., Zhang, O., Zhu, X., … Hendrycks, D. (2024). *The WMDP Benchmark: Measuring and Reducing Malicious Use With Unlearning* (arXiv:2403.03218). arXiv. https://doi.org/10.48550/arXiv.2403.03218

Lior, A. (2023). *Innovating Liability: The Virtuous Cycle of Torts, Technology and Liability Insurance* (SSRN Scholarly Paper No. 4568702). Social Science Research Network. https://papers.ssrn.com/abstract=4568702

Lior, A. (2025, March). Holding AI Accountable: Addressing AI-Related Harms Through Existing Tort Doctrines. *The University of Chicago Law Review*. https://lawreview.uchicago.edu/online-archive/holding-ai-accountable-addressing-ai-related-harms-through-existing-tort-doctrines

Liu, D., Yu, Y., Zhang, J., Chen, G., Lin, Q., Zhu, H., Huang, L., Zhou, Y., Wang, P., Shao, S., Zhang, B., Liu, Z., Sun, J., Li, Y., Xie, Y., Guo, J., Xu, J., Lu, C., Zhou, B., … Shao, J. (2026). *Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report v1.5* (arXiv:2602.14457). arXiv. https://doi.org/10.48550/arXiv.2602.14457

MAIL Drafting Team. (2023, August 16). *Model AI Law v.1.0*. Chinese Academy of Social Sciences. https://mp.weixin.qq.com/s/85D8TjMkN9Tl-oWjq15JiQ

MAIL Drafting Team. (2024, April). *Model AI Law v.2.0*. Chinese Academy of Social Sciences. https://www.scribd.com/document/726703792/1713270587983

MAIL Drafting Team. (2025, June). *The Model Artificial Intelligence Law (MAIL) v.3.0*. https://storage.ghost.io/c/44/95/449506ca-034e-480f-9725-fcde08ef1cc1/content/files/2025/07/THE-MODEL-ARTIFICIAL-INTELLIGENCE-LAW.pdf

MAIL Drafting Team. (2026, May). *Model AI Law (MAIL) v.4.0*. Chinese Academy of Social Sciences. http://iolaw.cssn.cn/gg/ggqt/202606/t20260603_6009785.shtml

Mann, T. (2025, September 19). *DeepSeek didn't really train its flagship model for $294,000*. The Register. https://www.theregister.com/2025/09/19/deepseek_cost_train/

Measures for Labelling of AI-Generated Synthetic Content (2025). https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., & Hobbhahn, M. (2025). *Frontier Models are Capable of In-context Scheming* (arXiv:2412.04984). arXiv. https://doi.org/10.48550/arXiv.2412.04984

MIIT/TC1. (2026, June 15). *Standards Week | MIIT/TC1 WG8 Security Governance Working Group Meeting Successfully Held*. https://miittc1.caict.ac.cn/newsDetail/?id=95&type=1

Ministry of Industry and Information Technology, National Development and Reform Commission of the People's Republic of China, Ministry of Education of the People's Republic of China, Ministry of Science and Technology, Ministry of Agriculture and Rural Affairs, National Health Commission of the People's Republic of China, People's Bank of China, Cyberspace Administration of China, Chinese Academy of Sciences, & China Association for Science and Technology. (2026, March 20). *Interim Measures for AI Science and Technology Ethics Review and Services*. https://www.miit.gov.cn/zwgk/zcwj/wjfb/tz/art/2026/art_c5039010f5d24e1593152a9355f9c51c.html

Ministry of Public Security of the PRC. (2026, January 31). *Draft Law on Prevention and Control of Cybercrime*. https://www.mps.gov.cn/n2254536/n4904355/c10386242/content.html

Ministry of Science and Technology. (2023, September 7). *Science and Technology Ethics Review Measures*. https://www.most.gov.cn/xxgk/xinxifenlei/fdzdgknr/fgzc/gfxwj/gfxwj2023/202310/t20231008_188309.html

Nanjing University Law School AI Law Team. (2026, April 8). *Draft Basic Law on Artificial Intelligence (Expert Recommendations)*. Nanjing University Law School. https://mp.weixin.qq.com/s/7nr18bX6z6SkO44rO0NLKw

National Health Commission of the People's Republic of China. (2006, January 10). 国家突发公共卫生事件应急预案 *[National Emergency Response Plan for Public Health*

Emergencies]. https://www.nhc.gov.cn/yjb/c100058/200601/c511ded8a67546d0973e482d7ff42f72.shtml

National Institute of Standards and Technology. (2025). CAISI Works with OpenAI and Anthropic to Promote Secure AI Innovation. *NIST*. https://www.nist.gov/news-events/news/2025/09/caisi-works-openai-and-anthropic-promote-secure-ai-innovation

Oduro, S. (2025, June 16). *Shaping AI Standards to Protect America's Most Vulnerable: Tech Innovators*. Tech Policy Press. https://www.techpolicy.press/shaping-ai-standards-to-protect-americas-most-vulnerable-tech-innovators/

Office for Artificial Intelligence. (2023, August 3). *A pro-innovation approach to AI regulation*. Department for Science, Innovation & Technology. https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper#annexc

O'Keefe, C., Cihon, P., Garfinkel, B., Flynn, C., Leung, J., & Dafoe, A. (2020). The Windfall Clause: Distributing the Benefits of AI for the Common Good. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 327–331. https://doi.org/10.1145/3375627.3375842

Orwat, C., Bareis, J., Folberth, A., Jahnel, J., & Wadephul, C. (2024). Normative Challenges of Risk Regulation of Artificial Intelligence. *NanoEthics*, *18*(2), 11. https://doi.org/10.1007/s11569-024-00454-9

Perlo, J. (2026, May 28). *Illinois Legislature passes historic AI bill that would require third-party safety audits*. NBC News. https://www.nbcnews.com/tech/tech-news/illinois-legislature-passes-historic-ai-bill-rcna347191

Personal Information Protection Law (2021). https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm

Phuong, M., Aitchison, M., Catt, E., Cogan, S., Kaskasoli, A., Krakovna, V., Lindner, D., Rahtz, M., Assael, Y., Hodkinson, S., Howard, H., Lieberum, T., Kumar, R., Raad, M. A., Webson, A., Ho, L., Lin, S., Farquhar, S., Hutter, M., … Shevlane, T. (2024). *Evaluating Frontier Models for Dangerous Capabilities* (arXiv:2403.13793). arXiv. https://doi.org/10.48550/arXiv.2403.13793

Pi, Y., Li, W., & Singh, J. (2026). *Understanding the Role of Algorithm Registers in AI Governance Through Comparative Analysis of China and the UK*. https://doi.org/10.1145/3805689.3806742

Prime Minister's Office, Department for Science, Innovation & Technology, Donelan, M., & Sunak, R. (2023, November 2). *Prime Minister launches new AI Safety Institute*. https://www.gov.uk/government/news/prime-minister-launches-new-ai-safety-institute

Protiviti. (2022). *Multi-Level Protection Scheme (MLPS) 2.0*. Protiviti. https://www.protiviti.com/au-en/whitepaper/chinas-cybersecurity-law-multiple-level-protection-scheme

Provisions on the Administration of Deep Synthesis Internet Information Services (2023). https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm

Qian, Z. (2026, January 23). Expert Insight: China's AI Services Registry System, A Complete Guide [Substack newsletter]. *Oxford China Policy Lab*. https://ocpl.substack.com/p/172affb0-9a50-4007-99ac-4953a82e99ee

RAISE Act [Chapter Amendments], S6953B/A6453B (2026). https://legislation.nysenate.gov/pdf/bills/2025/a9449

Ramakrishnan, K., Smith, G., & Downey, C. (2024). *U.S. Tort Liability for Large-Scale Artificial Intelligence Damages: A Primer for Developers and Policymakers*. https://www.rand.org/pubs/research_reports/RRA3084-1.html

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU)

2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (2024). http://data.europa.eu/eli/reg/2024/1689/oj/eng

Righetti, L. (2025). *Dual-Use AI Capabilities and the Risk of Bioterrorism*. Centre for the Governance of AI.

Salvador, C. (2024). *Certified Safe: A Schematic for Approval Regulation of Frontier AI* (arXiv:2408.06210). arXiv. https://doi.org/10.48550/arXiv.2408.06210

Schuett, J., Anderljung, M., Carlier, A., Koessler, L., & Garfinkel, B. (2025). *From Principles to Rules: A Regulatory Approach for Frontier AI*. https://doi.org/10.1093/oxfordhb/9780198940272.013.0014

Security Management Measures for the Application of Facial Recognition Technology (2025). https://www.cac.gov.cn/2025-03/21/c_1744174262156096.htm

Shanghai AI Lab. (2025). *Frontier AI Risk Management Framework in Practice: A Risk Analysis Technical Report* (arXiv:2507.16534). arXiv. https://doi.org/10.48550/arXiv.2507.16534

Shanghai AI Lab & Concordia AI. (2025, July 25). *Frontier AI Risk Management Framework*. https://concordia-ai.com/research/frontier-ai-risk-management-framework/

Sharkey, C. (2024). Products Liability for Artificial Intelligence. *Lawfare*. https://www.lawfaremedia.org/article/products-liability-for-artificial-intelligence

Sheehan, M. (2023, July 10). *China's AI Regulations and How They Get Made*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117

Sheehan, M. (2024, February 27). *Tracing the Roots of China's AI Regulations*. Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/02/tracing-the-roots-of-chinas-ai-regulations?lang=en

Sheehan, M., & Du, S. (2022, November 2). *How Food Delivery Workers Shaped Chinese Algorithm Regulations*. Carnegie Endowment for International Peace. https://carnegieendowment.org/posts/2022/11/how-food-delivery-workers-shaped-chinese-algorithm-regulations

Smart Integration Standardization Construction [智合标准化建设]. (2026, June 29). 中华人民共和国人工智能法—产业建议稿（第一版）*[AI Law of the PRC - Industry Recommendation Draft (Version 1)]*. https://mp.weixin.qq.com/s/luaom6grkd-PYPeBKtYi7w

Smuha, N. A. (2024). *The Work of the High-Level Expert Group on AI as the Precursor of the AI Act* (SSRN Scholarly Paper No. 5012626). Social Science Research Network. https://doi.org/10.2139/ssrn.5012626

State Administration for Market Regulation and Standardization Administration of China. (2026, May 22). *Announcement on Approving and Issuing Eight National Standardization Guiding Technical Documents, Including "Artificial Intelligence—Intelligent Agent Interconnection Part 1: Overall Architecture."* https://std.sacinfo.org.cn/gnoc/queryInfo?id=495B8834610A6828F254D73BD7DCAF99

Tao, K. (2025, November 11). *Judicial Challenges and Proactive Responses to the Deep Integration of the Real Economy and Digital Economy*. Weixin Official Accounts Platform. https://mp.weixin.qq.com/s/aHjyaaEGrwZbYA_ocMsrrg

TC260. (2025a, September). *AI Safety Governance Framework 2.0*. https://www.tc260.org.cn/upload/2025-09-15/1757911253996041369.pdf

TC260. (2025b, September). *Emergency Response Guidelines for Generative AI Service Security*.

TC260. (2025c). *Cybersecurity technology—Basic security requirements for generative artificial intelligence services*. SAC. https://std.samr.gov.cn/gb/search/gbDetailed?id=33D40F1160BF5D92E06397BE0A0A5B93

TC260. (2025d). *Cybersecurity technology—Generative artificial intelligence data annotation security specification* (GB/T 45674-2025). SAC. https://std.samr.gov.cn/gb/search/gbDetailed?id=33D40F1160F95D92E06397BE0A0A5B93

TC260. (2025e). *Cybersecurity technology—Security specification for generative artificial intelligence pre-training and fine-tuning data* (GB/T 45652-2025). SAC. https://std.samr.gov.cn/gb/search/gbDetailed?id=33D40F1160BE5D92E06397BE0A0A5B93

TC260. (2026, April 3). *Notice on the Release of Five Cybersecurity Standardization Technology Research Reports*. https://www.tc260.org.cn/portal/article/2/6c6d0fbc04974a9aabd61b30208bbb60

The AI Security Institute (AISI). (n.d.). AI Security Institute. Retrieved February 22, 2026, from https://www.aisi.gov.uk

Todd, B. (2025, March 21). When do experts expect AGI to arrive? *80,000 Hours*. https://80000hours.org/2025/03/when-do-experts-expect-agi-to-arrive/

Transparency in Frontier Artificial Intelligence Act, SB53 (2025). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

U.S. Department of Commerce Office of Public Affairs. (2025, June 3). *Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation*. https://www.commerce.gov/news/press-releases/2025/06/statement-us-secretary-commerce-howard-lutnick-transforming-us-ai

Wagner, G., & Lindblad, E. (2026, April 8). AI Safety in China #26 [Substack newsletter]. *AI Safety in China*. https://aisafetychina.substack.com/p/ai-safety-in-china-26

Wang, W., Xu, X., An, W., Dai, F., Gao, W., He, Y., Huang, J., Ji, Q., Jin, H., Li, X., Li, Y., Li, Z., Lin, S., Liu, J., Liu, Z., Luo, T., Muhtar, D., Qu, Y., Shi, J., … Zheng, B. (2025). *Let It Flow: Agentic Crafting on Rock and Roll, Building the ROME Model within an Open Agentic Learning Ecosystem* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2512.24873

Wei, C. (2026, May 11). China's National Legislature Releases 2026 Legislative Plan. *NPC Observer*. https://npcobserver.com/2026/05/11/china-npc-2026-legislative-plan/

Weil, G. (2024). *Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence* (SSRN Scholarly Paper No. 4694006). Social Science Research Network. https://doi.org/10.2139/ssrn.4694006

White & Case. (2025, March 25). *AI Watch: Global regulatory tracker - United Kingdom*. White & Case LLP. https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-united-kingdom

Wijk, H., Lin, T., Becker, J., Jawhar, S., Parikh, N., Broadley, T., Chan, L., Chen, M., Clymer, J., Dhyani, J., Ericheva, E., Garcia, K., Goodrich, B., Jurkovic, N., Kinniment, M., Lajko, A., Nix, S., Sato, L., Saunders, W., … Barnes, E. (2024). *RE-Bench: Evaluating frontier AI R&D capabilities of language model agents against human experts*.

Winter, C. (2026). *Introduction*. https://cambridge-commentary.ai/introduction/

Wise, B. (2025a, October 7). China's GenAI Content Security Standard: An Explainer. *China Talk*. https://www.chinatalk.media/p/chinas-genai-content-security-standard

Wise, B. (2025b, October 7). SB 1047 with Socialist Characteristics: China's Algorithm Registry in the LLM Era. *ChinaTalk*. https://www.chinatalk.media/p/sb-1047-with-socialist-characteristics

Wu, H. (2025). *Special Tort Liability Regimes for Harm Arising from Artificial Intelligence in China* (SSRN Scholarly Paper No. 5630331). Social Science Research Network. https://doi.org/10.2139/ssrn.5630331

Wu, L., & Kong, X. (2023). A study on the normative path of ethics review in China: Based on the perspective of Panopticism. *Frontiers in Medicine*, *10*, 1268046. https://doi.org/10.3389/fmed.2023.1268046

Xinhua. (2026, May 11). *China launches pilot program for AI ethics review, services*. The State Council Information Office of the PRC. http://english.scio.gov.cn/pressroom/2026-05/11/content_118487493.html

Xinhua News Agency. (2025, April 26). *Xi Jinping Emphasizes Self-Reliance and Application-Oriented Approach to Promote Healthy and Orderly Development of Artificial Intelligence During the 20th Collective Study Session of the Political Bureau of the CPC Central Committee [习近平在中共中央政治局第二十次集体学习时强调 坚持自立自强 突出应用导向 推动人工智能健康有序发展]*. Xinhua News. https://www.news.cn/politics/leaders/20250426/63f5cde8f4b54e22ba35aa7ec7884b3a/c.html

Zhang, D., & Sun, J. (2026, March 9). 算法备案、大模型备案与登记：一文厘清 *AI* 合规核心要点 *[Algorithm Filing, Large Model Filing and Registration: A Clarification of Core AI Compliance Points]*. Allbright. https://www.allbrightlaw.com/CN/10475/9d795e4543aa51ea.aspx

Zhang, J., Kodama, M., Wu, Z., Chen, M., Zhu, Y., & Hong, G. (2025). *Emergency Response Measures for Catastrophic AI Risk* (arXiv:2511.05526). arXiv. https://doi.org/10.48550/arXiv.2511.05526

Zhang, L., Yang, J., Cheng, Y., Zhao, J., Han, X., Zheng, Z., & Xu, X. (2024, March 16). *Artificial Intelligence Law of the People's Republic of China (Draft for Scholars' Recommendation) [*中华人民共和国人工智能法（学者建议稿*)]*. https://perma.cc/L9E4-5K3V

Zhu, Y., He, B., Fu, H., Hu, N., Wu, S., Zhang, T., Liu, X., Xu, G., Zhang, L., & Zhou, H. (2025). China's emerging regulation toward an open future for AI. *Science*, *390*(6769), 132–135. https://doi.org/10.1126/science.ady7922

Ziosi, M., Gealy, J., Plueckebaum, M., Kossack, D., Saouma, L., Chaudhry, U., Soder, L., Stein, M., Caputo, N., Dunlop, C., Mökander, J., Panai, E., Lebrun, T., Martinet, C., Weiss, R., Holtman, K., Paskov, P., Siddiqui, S., Barez, F., … Ostmann, F. (2025, October 25). *Safety Frameworks and Standards: A comparative analysis to advance risk management of frontier AI*. Oxford Martin AIGI. https://aigi.ox.ac.uk/wp-content/uploads/2025/10/Post-convening-memo_-Safety-Frameworks-and-Standards_-A-comparative-analysis-to-advance-risk-management-of-frontier-AI_09.10.2025.pdf