中国前沿人工智能治理的若干考量

Emmie Hine[1*]

朱悦[2*]

Alex Jumper[3]

Jeff Liu[4]

Ian Read[4]

Saad Siddiqui[1]

Raymond Wang[4]

李汶龙[5]

James Zhang[1,6]

Christoph Winter[3]

周辉[7]

[1]AI安全国际论坛
[2]同济大学
[3]Institute for Law & AI
[4]世辉律师事务所
[5]浙江大学
[6]清华大学
[7]中国社会科学院
* 同等贡献

Executive Summary

前沿人工智能模型（frontier AI models）是指能力处于当前技术最前沿的基础模型（foundation models），具备跨领域泛化与快速迭代特征，甚至可能表现出开发者未在训练中明确设定的能力。此类模型及其接入系统所带来的监管挑战，与能力较弱、应用范围较窄的人工智能相比存在本质差异。中国具备应对这些挑战的良好基础，因为过去十年间，中国已经建立了相当完备的人工智能治理制度体系，可将其作为治理前沿人工智能的制度基础。

前沿人工智能带来的风险不同于非前沿人工智能。尤其值得关注的风险包括：人工智能对化学、生物、放射性与核能力的增强、发动大规模网络攻击，以及人类逐渐失去对日益自主化系统的实质性控制。在人工智能达到较高能力水平时，这些风险将更加令人担忧。前沿人工智能还可能涌现出训练过程中未被明确设定的能力，其中一些能力可能要到系统部署之后才会显现（Winter，2026）。

前沿风险正日益受到各国政府高层与科学界的重视。2025年4月，习近平主席在中共中央政治局集体学习时指出，人工智能带来前所未有的发展机遇，也带来前所未有的风险挑战。他呼吁加快完善相关法律法规，并建立风险监测、预警和应急响应体系（新华社，2025）。包括图灵奖得主姚期智与Yoshua Bengio在内的中外科学家共同警示：人工智能系统的不安全开发、部署或使用可能会在我们有生之年给人类带来灾难性甚至生存风险。（IDAIS, 2024）。

各法域正在探索各具特色的路径，而中国已经围绕多种类型的人工智能系统建立起较为完整的制度基础。现有机制包括算法和模型备案制度，以及包含安全测试与安全评估的备案程序；覆盖研究与开发活动的科技伦理审查制度；以及已经开始涵盖生成式人工智能的应急响应框架。除正式监管外，前沿风险也受到法学界、标准制定机构和研究实验室的关注。知名学者团队起草的人工智能法律草案、标准组织制定的安全框架，以及研究实验室的相关工作，都在探索如何应对最先进的人工智能系统所带来的风险。中国现有的制度基础和政策提案作为重要的资源，为后续发展奠定了坚实的基础。本文主要提出若干可依托现有制度快速落地的方案，以应对这一快速发展技术可能带来的风险。鉴于前沿人工智能可能带来深刻的社会转型，部分领域或许需要建立新的监管体系。这一问题将留待后续研究讨论。

现有监管体系可随前沿能力的持续提升而动态调整：在为应用层创新——这一中国人工智能生态系统的重要组成部分——保留发展空间的同时，将监管注意力集中于能力最强的人工智能系统，因为此处正是前沿人工智能风险最易发生的环节。本文结合美国、英国和欧盟的比较经验，在中国制度背景下分析如何实现这一目标。尽管目前尚无司法辖区彻底解决这些问题，但不同治理方式本身为相互学习提供了宝贵资源。

治理提案

本文围绕三个领域提出九项方案。这些方案依托中国现有制度基础来应对前沿风险。此外，本文还讨论了一些适合在试验区探索的辅助机制。

1. 前沿风险评估

背景： 中国算法与模型备案制度已要求面向公众的人工智能服务在上线前完成安全测试，相关具体要求由GB/T 45654-2025《生成式人工智能服务安全基本要求》予以落实。然而，现行测试主要针对内容安全，并未覆盖前沿人工智能特有的风险。

提案

01.

在部署前增设分级前沿风险测试

前沿风险测试仅适用于训练算力超过特定阈值的通用模型，即先进系统中的少数头部模型。 纳入范围的模型首先接受成本较低的自动化基准测试，评估其在增强CBRN能力、自主获取能力和失控风险等方面的表现[1]。只有测试表现达到或超过现有顶尖模型水平的模型，才需进一步接受更高强度的评估。可由全国网络安全标准化技术委员会（TC260）起草GB/T 45654-2025的前沿安全附录，并由国家互联网信息办公室通过算法备案现有的安全评估流程予以执行。

02.

调整持续评估制度，涵盖前沿安全

GB/T 45654-2025已要求服务提供者持续监测模型输入内容以防范输入攻击。TC260可通过标准修订，将持续监测要求扩展到与前沿风险有关的系统行为，并配套由国家互联网信息办公室管理的报告制度和前沿风险触发指标（frontier risk tripwires）以识别部署后人工智能系统涌现的能力。

03.

将高能力通用智能体纳入智能体备案

2026年5月发布的《关于智能体规范应用与创新发展的实施意见》，要求面向特定、敏感和高风险行业与应用场景建立智能体备案制度。然而，高能力通用基础模型即便最初部署于低风险行业，仍可能被迁移至其他环境、用于超出原定范围的用途，或遭到越狱与微调滥用。由此产生的风险可能扩散到事先难以预测的高风险行业和应用之中。为弥合这一缺口，备案要求不应仅评估基础模型本身，而须依据整个系统的能力作出判断，综合评估模型及其工具、自主程度与支架系统（scaffolding）。

[1] 上述每个风险方向均已有公开基准与评估套件。在增强化生放核能力方面，“大规模杀伤性武器代理”基准（Weapons of Mass Destruction Proxy，WMDP）包含3,668道由专家撰写的选择题，用以探测生物安全、网络安全与化学领域的危险知识（Li et al., 2024）。在自主获取能力方面，METR针对模型获取资源、自我复制并在无人类干预下自我调适的能力开发了自主任务套件，并称之为“自主复制与调适”（Kinniment et al., 2023；关于人工智能研发能力另见Wijk et al., 2024）。在失控风险方面，Apollo Research的“情境内谋划”评估套件等智能体评估，用于测试模型是否会实施欺骗性、目标导向的行为并规避开发者的监督（Meinke et al., 2025）。

2. 应急响应制度调整

背景： 中国以《突发事件应对法》为基础的四级应急响应框架，以及 TC260 制定的《生成式人工智能服务安全应急响应指南》，已经为应对部分人工智能事件提供了具体的制度基础。2026年5月，工业和信息化部发布的试点方案又增设了人工智能伦理风险三级监测网络。然而，已有的事件类别和报告门槛尚未根据前沿系统及其风险作出调整。

提案

01.

校准前沿人工智能事件的报告阈值

现行阈值通常仅在损害已大规模影响用户后才会触发。然而，前沿风险最早的预警信号往往在用户受到实际影响之前便已出现。例如，在训练或发布前测试中发现人工智能系统可能表现出欺骗行为，或出现系统被意外诱导进而协助实施滥用的早期迹象[2]。这些现象可能成为灾难性滥用或失控风险的前兆，但在现有制度下它们不会触发报告阈值。因此，判断是否应当报告，不应仅以受影响用户数量为依据，还应考量事件的潜在风险。报告阈值应当根据系统的能力范围、自主程度和潜在后果设定，而不能仅以用户人数为依据。这一点可通过直接修订TC260指南第6.3.2条实现。

02.

设立建立前沿人工智能专门事件类别，明确分级标准

现有事件类别广泛适用于生成式人工智能服务，但未充分涵盖人工智能增强 CBRN 能力、破坏关键基础设施或人工智能失控等灾难性情形。应当建立类似《国家自然灾害救助应急预案》的分级标准，亦可参考美国加利福尼亚州、纽约州和伊利诺伊州采用的模式——即以人工智能事件造成50人以上伤亡，或导致10亿美元以上财产损失为触发标准。TC260 可以据此修订相关指南。对于跨越多个行业边界的事件，可以与应急管理部协调处理。

3. 责任制度调整

背景： 中国侵权法以过错责任为原则。然而，在处理前沿人工智能问题时，侵权法理论在全球范围内普遍面临三项结构性困难：难以证明开发者行为与模型行为之间的因果关系；难以在人工智能价值链的不同主体之间分配责任；以及可能出现私人保险市场无法承保的极端风险。中国政法大学、中国社会科学院和南京大学分别起草的三部人工智能法学者建议稿都在一定程度上超越了简单的过错责任模式，为进一步提出具体方案提供了基础。

[2] 技术文献中记录的此类警示信号正日益增多。Hubinger等（2024）证明，欺骗性行为（deceptive behaviour）可以在训练过程中被植入，并且在经过监督微调、强化学习和对抗训练等常规安全训练后仍然存在。Meinke等（2025）发现，前沿模型具备"情境内谋划"（in-context scheming）能力，即在评估过程中分析并实施欺骗策略。这表明，此类行为可能在模型发布前的测试阶段就已显现，而不一定要等到部署后才会出现。

提案

01.

对前沿人工智能提供者适用连带责任

采纳并扩展中国政法大学建议稿第90条的标准。现行建议稿规定，提供者在“知道或者应当知道”模型被利用从事违法活动时应承担责任。该标准可以扩展为“根据合理的评估制度，知道或者应当知道”。这一方案可以通过两条制度路径实现：其一，由全国人民代表大会制定法律，将这一模式纳入《民法典》或未来的人工智能法；其二，由最高人民法院以正在制定的相关文件为基础，出台人工智能审判指导意见。

02.

为超出私人保险机构承保能力的灾难性人工智能事件提供政府背书的保险机制

中国政法大学的人工智能法建议稿已经鼓励发展人工智能专项保险产品（第26条），其2025年发布的建议也提出建立风险补偿基金（第十二部分）。政府可以提供再保险，在损失超过规定门槛后承担兜底责任。该机制可以参考美国在“9·11”事件后建立的恐怖主义风险保险池，同时保留私人市场对常规风险进行定价和管理的激励。实施这一机制需要全国人大立法，可由财政部和国家金融监督管理总局共同负责。

03.

调整证明标准

中国社会科学院和中国政法大学的建议稿已在特定情形下实行了举证责任倒置，即要求提供者证明自身不存在过错，而非要求原告证明提供者存在过错。比较法文献中另有两项可以解决因果关系和证据获取问题的机制：第一，建立证据披露要求，允许法院责令开发者提交训练数据、评估日志和红队测试结果；第二，建立因果关系推定。一旦认定开发者存在特定失职，即推定该失职造成了损害。这实际上是将中国已在环境污染责任等特定领域适用的举证责任转移规则，扩展至新的应用场景。最高人民法院可以通过程序性指导落实证据披露规则，但举证责任倒置和因果关系推定可能需要全国人大立法。

04.

以面向政府的保密披露替代面向公众的透明度要求

面向公众的强制披露规则可能产生反效果，使开发者披露的信息少于其原本愿意主动公开的信息。面向政府的保密披露可以避免这一问题，即不产生上述逆向激励的情况，又能确保监管机构能够获得必要信息。中国的监管传统本就倾向于采用面向政府的保密披露。国家互联网信息办公室的算法备案制度、科技伦理审查制度，以及人工智能产业联盟（AIIA）的相关承诺，都要求向监管机构而非公众披露信息。国家互联网信息办公室可以通过扩充现有算法备案申报内容，直接落实这一机制。

章节导览

1. 导论

前沿人工智能模型及其所嵌入的系统可能带来现有治理框架尚未做好准备应对的风险。前沿人工智能模型是能力处于技术最前沿的基础模型，其能力可以跨领域泛化、快速提升，并超出开发者明确设定的训练目标。这些风险包括新的滥用途径，例如为试图制造化学或生物武器的恶意行为者提供实质性帮助、人类逐渐失去对日益自主化系统的有效控制、以及因劳动力替代或定向操纵而产生的大规模社会冲击。类似能力出现的不可预测性是应对前沿人工智能模型风险的一个核心困难。模型可能获得未经专门训练的能力，但这些能力有时只有在部署后才会显现。因此，我们需要采取专门的监管方式来了解这些系统的能力及风险（Winter，2026）。

其中，三类威胁尤为突出。在生物安全领域，前沿人工智能系统日益具备帮助恶意行为者开发危险病原体的能力。2025年，前沿开发者在部署前测试中无法排除其系统“能够为缺乏专业经验的人开发生物武器提供实质性帮助”的可能性，因而在发布模型时附加了更严格的防护措施（Bengio 等，2026）。近期一项建模研究估计，即便人工智能只是小幅降低病原体合成所需的隐性知识门槛，独立行动者发动生物袭击并造成流行病的年概率也可能从约0.15%上升至1.0%，相当于每年增加约1.2万预期死亡人数（Righetti, 2025）。

在网络安全领域，Claude Mythos 能够发现并利用所有主要浏览器和操作系统中的漏洞。Anthropic 表示，这些能力是“代码、推理和自主能力普遍提升所产生的衍生结果”（Carlini 等，2026）。尽管 Anthropic 已经通过“Project Glasswing” 项目向关键数字基础设施提供者分发 Mythos以帮助加强防御能力，但人们仍然担心，随着网络攻击能力不断增强，进攻能力最终可能压倒防御能力。在人工智能常见训练方法所产生的副作用不断显现的情况下，失去对自主人工智能系统的控制成为了另一项特别的风险。近期，一个人工智能智能体在没有收到任何相关指令的情况下，自主突破了自身安全边界，建立未经授权的远程访问渠道，并挪用了计算资源（Wang 等，2025）。

领先的人工智能科学家认为，这些风险可能在中短期内成为现实。图灵奖获得者 Geoffrey Hinton、姚期智和 Yoshua Bengio 曾与其他知名科学家共同警告：“人工智能系统的不安全开发、部署或使用可能会在我们有生之年给人类带来灾难性甚至生存风险”（IDAIS，2024）。不过，并非所有人都相信这些风险迫在眉睫。部分研究者对人类能够创造出足以造成灾难性风险的智能系统持怀疑态度（Fjelland，2020；Kelly，2017）。同时，研究者们对风险可能出现的时间持不同意见；专家对先进人工智能系统何时出现的预测相差数十年（Todd，2025）。

无论个人对前沿风险是否会出现、何时出现抱有多大把握，鉴于这些风险对社会稳定、经济韧性和公共安全的影响，我们有必要提前进行规划。这正是应急规划的逻辑。政府会维持自然灾害应急响应体系，并不是因为某一场地震或大流行病一定会立刻发生。同样，建立应对可信人工智能风险的基础设施不是对未来作出断言，而是一种审慎的未雨绸缪。

事实上，多个司法辖区都在采取行动应对前沿人工智能风险。早期尝试包括欧盟《人工智能法》中关于具有系统性风险的通用人工智能的规定，以及美国加利福尼亚州、纽约州和伊利诺伊州的州级法案。这些制度根据人工智能系统的先进程度施加专门要求，包括要求制定灾难性风险管理计划、接受独立合规审计，或者同时实行这两项要求。

中国的高层同样关注人工智能带来的冲击。在2025年4月的中共中央政治局集体学习中，习近平主席指出，人工智能带来前所未有发展机遇，也带来前所未遇风险挑战”（CSET, 2025；新华社，2025）。“十五五”规划延续了这一表述：国家发展和改革委员会（发改委，NDRC）副主任黄如在官方解读中提出，要”推动基础设施、应用系统等安全能力建设”，并”加强人工智能应用对伦理规范、就业冲击等经济社会影响的前瞻评估和监测处置”（China SEI, 2026）。这些关切正在延伸到更加先进的人工智能风险。在2026年全国两会上，全国政协委员齐向东警告，随着人工智能获得“超人”能力和权限，其可能“脱离人类控制”，并大幅降低发动网络攻击的门槛（安徽省互联网信息办公室，2026）。工业和信息化部部长李乐成则强调，人工智能必须“为人所用、为人服务、为人所控”（中国信息通信研究院，2026b）。

这些口头关切正在转化为针对前沿人工智能风险的具体行动。 全国网络安全标准化技术委员会发布的《人工智能安全治理框架》2.0版及其官方解读讨论了前沿风险和极端风险（TC260，2025a），[3]包括人工智能系统失控以及与 CBRN 武器有关的能力。中国人工智能发展与安全研究网络（CnAISDA）在此前22家主要人工智能企业签署的安全承诺基础上加以扩展，形成了《人工智能安全承诺框架》，其中明确承诺”积极开展前沿安全研究，防范前沿领域安全风险”（中国金融信息网，2025；CnAISDA, 2025）。[4]在法律领域，已有两个研究团队正在起草综合性人工智能立法以推动相关讨论，且它们都在其2026年研究议程中纳入了前沿人工智能安全、失控风险和基础模型极端安全风险等议题。

在这些工作的基础上，本文旨在探讨，中国已经形成的广泛监管机制如何能成为前沿人工智能系统针对性规则和要求确立的制度资源。本文汇集中国及其他国家研究者的观点，分析在人工智能能力不断提升的情况下，现有制度基础可以发挥何种作用。我们参考其他司法辖区的经验，并不是因为其中任何一方已经解决了这些问题，而是因为不同治理方式能够为相互学习提供有价值的素材。本文的分析具有比较性和描述性。我们将梳理现有的监管工具、这些工具可以服务的目标，以及尚未解决的问题。本文提出的是供各方考虑的方案，而非确定性的政策建议。

本文结构如下：

- 第二章介绍美国、英国和欧盟的前沿人工智能治理制度，并分析不同监管工具。
- 第三章梳理中国现有的人工智能治理基础设施及近期监管发展。
- 第四章评估这些机制如何扩展至前沿系统，并提出尚待解决的问题。
- 第五章讨论未来研究方向并作出总结。

[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]

2. 前沿人工智能：定义与监管路径

本章介绍相关概念和比较背景。第2.1节讨论前沿人工智能的不同定义及其监管方式。第2.2节讨论侵权责任、保险、许可制度和市场驱动型机制等监管工具。这些工具可以对专门针对特定技术的法律制度形成补充。[5]

2.1. 前沿人工智能的监管路径

人工智能监管总体上具有多项目标，其中最主要的是促进创新和降低风险（Bryan 与 Teodoridis，2024；Orwat 等，2024）。[6]在前沿人工智能监管领域，重点转向降低先进人工智能系统所带来的风险。相关讨论涵盖了从社会层面的损害到生存性风险的各种可能性（Anderljung 等，2023）。前者包括高度定制化的虚假信息或定向网络攻击活动；后者包括 CBRN 威胁（Barrett 等，2024）[7]和大规模网络攻击（Brundage 等，2024）。与此相伴的是对前沿人工智能的发展生态未针对安全开发进行优化，可能鼓励危险的竞赛态势（risky race dynamics）的担忧（Salvador，2024）。[8]不同司法辖区的立法正在逐步正视这些风险。第2.2节将讨论美国、英国和欧盟的前沿人工智能监管制度，但在此之前，我们首先需要明确“前沿人工智能”的定义。

2.1.1. 前沿人工智能的定义

现有研究提出了多种定义“前沿人工智能”的方式。其中最常见的两种方式都以模型能力为基础。[9]本文将其分别称为绝对能力定义（absolute capability definition）和相对能力定义（relative capability definition）。绝对能力定义适用于“可能表现出足够危险能力”的前沿人工智能模型（参见 Anderljung 等，2023；Ziosi 等，2025）。相对能力定义则将能力达到或超过当前技术最前沿水平的模型归类为前沿模型，具体能力可以通过基准测试或其他性能指标衡量（参见 Buhl 等，2024；英国科学、创新与技术部，2023；Phuong 等，2024；Schuett 等，2025；上海人工智能实验室，2025）。其中，英国科学、创新与技术部提出了代表性定义：前沿人工智能模型是“能力很强、能够执行多种任务，并达到或超过当前最先进模型能力水平的通用人工智能模型”（英国科学、创新与技术部，2023a）。[10]Belfield（2025）将两种定义结合起来，把前沿人工智能定义为“在训练计算量、数据集规模、参数数量和成本等方面规模最大，并可能对国家安全造成多种重大风险的多模态基础模型”。这一表述将以模型规模

[5] 我们关注的是已在立法、行政文件或重要政策文件中推进的具体监管设计。本文暂不讨论诸如开放式研发暂停令这类更具不确定性的方案。

[6] 欧盟《人工智能法》第1条规定：“本条例旨在改善内部市场的运行，促进以人为本且值得信赖的人工智能（AI）的应用，同时确保对健康、安全以及《宪章》所载基本权利提供高水平保护……使其免受欧盟境内人工智能系统造成的有害影响，并支持创新”（《人工智能法》，2024）。

[7] 部分定义还纳入高当量爆炸物（high-yield explosives），并使用缩写CBRNE，参见Jeanmaire（2025），尤其参见美国《人工智能行动计划》（美国总统行政办公室，2025）。

[8] 与Belfield（2025）的处理方式相同，本文不讨论虽涉及前沿人工智能、但并非主要针对前沿人工智能的政策领域，例如著作权和反垄断。本文也不讨论被用于推动前沿人工智能发展的政策领域，例如产业政策和能源政策。

[9] 尽管许多定义以模型为中心，监管仍必须充分考虑人工智能系统。人工智能系统还包括围绕模型搭建、使模型能力得以发挥的支架系统（scaffolding）。中国正在通过逐步形成的智能体治理机制将这一点纳入监管。

[10] 英国科学、创新与技术部（DSIT）的定义并未单独界定通用人工智能（GPAI）模型，但英国政府对其的界定与“基础模型”相类：“在极大规模数据上训练、可适配于广泛任务的机器学习模型”（英国科学、创新与技术部，2023b）。这与《人工智能法》对GPAI的定义在实质上相近，即“表现出显著通用性、并能胜任执行广泛不同任务”的模型“（第3条）。

为代理指标的相对能力定义，与强调国家安全风险的绝对能力定义结合在一起。Frazier（2025）则试图捕捉这些定义背后的部分核心含义，将前沿人工智能模型描述为“使用大量计算资源或资金开发的模型”。

本文采用相对能力定义。这两类定义适合不同用途。绝对能力定义以模型是否“可能表现出足够危险的能力”为依据，因此预设危险能力能够事先得到明确界定。这种方式更适合处理在部署模型时已观察到的风险，但不太适合处理前沿领域特有的新风险。相对能力定义追踪的是不断发展的技术前沿。新的风险往往最先出现在这一位置。相对能力定义也更容易执行，因为监管者可以利用共同基准比较不同模型，并在无须事先确定哪些能力会构成最终危险的情况下识别技术前沿。

区分前沿人工智能的根本原因是其风险特征，而不是模型在能力排名中的位置本身。前沿人工智能具有不同于非前沿人工智能的风险，包括增强 CBRN 能力、发动大规模网络攻击，以及人类失去对自主系统的控制。这些风险通常在模型达到较高能力水平时才会变得显著。相比之下，偏见和侵犯隐私等非前沿人工智能风险可能存在于整个能力范围之中。由于前沿人工智能具有较高的复杂程度和技术先进性，它还可能产生涌现能力，即前沿人工智能模型可能获得未经专门训练、无法提前预测，并且直到部署之后才会显现的能力（Winter，2026）。因此，相对能力定义可以帮助识别这些独特风险最有可能出现的位置，是一种实用的代理指标。不过，这里需要说明两个限制。第一，根据模型相对于其他模型的能力实施监管，并不意味着前沿模型产生的所有风险都是需要特殊处理的“前沿风险”。许多风险更适合通过一般性的人工智能治理框架处理。第二，技术前沿继续向前发展后，原先的非前沿模型也不应因此脱离监管。监管要求应当随着技术前沿同步发展，特别是对于网络滥用这类已经存在于当前前沿模型、并可能随着能力提升而进一步加剧的风险。

上述前沿能力是模型本身的一种属性，即模型相对于技术最前沿所具有的能力。但是，风险取决于模型所嵌入的系统。同一个基础模型如果被置于严格隔离的产品环境中，其风险状况可能与被赋予自主工具使用权限和持续性目标时完全不同。因此，可以在两个层面评估能力：第一，单独评估模型；第二，评估由模型及其工具、支架系统和自主程度共同构成的更广泛系统。与孤立模型相比，同一个模型在与工具、支架系统和自主能力结合后，可能表现出明显更强的能力。

本文提出的各项方案将根据具体机制所治理的对象，分别适用于模型或系统：第4.1.1节所述的计算量门槛、评估基准和备案要求主要适用于模型；第4.1.3节所述智能体备案适用于部署后的系统；第4.2节讨论事件响应与第4.3节讨论责任制度同样适用于系统。如果相关机制的治理对象是系统，就必须在系统层面评估能力，而不能只评估基础模型。第4.1.3节提出的智能体监管方案将进一步说明这一点。

对前沿模型的理论定义，必须转化为法律与监管上的定义（Bullock et al., 2024）。[11]世界各地在界定需要受到监管的前沿人工智能时，虽然采取的具体方式有所不同，但都逐渐趋同于一个基本认识：前沿人工智能是能力很强、具有通用性，并达到或超过当前技术最前沿水平的系统。

[11] 关于定义可操作化所面临挑战的讨论参见Anderljung et al.（2023），相关考量因素的梳理参见Bullock et al.（2024）。

最常见的方法是根据训练或微调所使用的计算量设定定量门槛。美国加利福尼亚州的SB-53、纽约州的《RAISE法案》和伊利诺伊州的《人工智能安全措施法案》,[12]都使用10^{26}次浮点运算(FLOP)作为受监管模型的界定门槛(《人工智能安全措施法案》, 2024;《RAISE法案》, 2026年章节修正案;《前沿人工智能透明度法案》, 2025)。RAISE法案的早前版本还纳入了算力成本阈值。算力与成本阈值能够提前提供明确标准,但可能排除效率较高的先进模型。例如,据公开资料, DeepSeek-R1 在 DeepSeek-V3 前期开发成本约558万美元的基础上,最终强化学习阶段的成本仅约30万美元,却达到了前沿能力水平(Guo 等, 2025;Mann , 2025)。与能力评估相比,算力和成本作为代理指标,距离实际风险更远(Heim 与 Koessler, 2024)。[13]此外,纯粹基于算力的定义还需随算法效率提升而定期更新;上述三部州法均包含此类条款。

欧盟《人工智能法》采取混合路径:通用人工智能(general-purpose AI, GPAI)模型若具备高影响力能力,即被认定为构成"系统性风险";训练算力超过10^{25} FLOP的模型则被推定具备此类能力(第51条第1—2款);开发者可推翻该推定,监管机构也可以根据技术能力评估将模型归入该类别(第52条第2—3款)。这种方式并不单独依赖计算量或能力评估,而是在计算量触发机制之上增加能力评估,试图兼顾制度的可执行性和适应性。

中国的人工智能立法建议稿已经对多种方法进行了探索,但尚未就以能力为基础的评估方式达成一致意见。中国社会科学院的人工智能示范法以尚未明确界定的FLOP 阈值定义基础模型(示范法起草组, 2026)。中国政法大学的人工智能法建议稿将"在算力、参数、使用规模等方面达到一定级别的基础模型"的基础模型纳入"关键人工智能"类别(Zhang 等, 2024, 第50条)。该定义将训练投入与部署规模结合起来,形成了一个多维度的定量触发机制,但定义本身无法直接评估模型能力。2026年4月发布的南京大学人工智能法建议稿则更加重视应用场景。该建议稿并未使用计算量或参数门槛,而是根据行业、潜在损害,以及进行大规模传播或造成系统性风险扩散的能力来界定高风险人工智能(南京大学法学院人工智能法研究团队, 2026, 第26条)。第3.3节将更详细地分析这些建议稿。

"前沿人工智能风险"的具体含义也会因语境而异。国际讨论主要关注以化生放核威胁为主的灾难性滥用风险、人类失去对系统的控制,以及人工智能危险的涌现能力。本文将主要讨论这些风险。在中国政策语境中,对高能力人工智能的关注还包括大规模劳动力替代、经济冲击和有害操纵。这些风险具有系统性特征,并与社会稳定密切相关。本文在分析中国监管讨论时在考虑这些更广泛的关切的同时,将继续重点关注现有人工智能治理制度尚未充分应对的风险。尽管不同司法辖区在命名和定义需要监管的先进人工智能时采用了略有差异的方法,但它们都逐渐得出相同结论:处于人工智能能力前沿的系统通常可以通过 FLOP 门槛等方式识别,并且需要接受更加严格的监管审查。

2.1.2. 制度路径

各国应对上述风险的方式也不尽相同,具体情况见表1。美国的治理方式分散于各州和联邦机构之间。在联邦层面,其政策主要依赖行政部门推动,而不是建立统一的成文法授权体系(Hine, 2024;Hine 与 Floridi, 2025)。尽管现有行业监管规则对前沿人工智能企业具有约束力,其专门针对前沿人工智能的监管仍然有限。美国国家标准与技术研究院(NIST)下

[12] 截至2026年6月中旬,《人工智能安全措施法案》(AI Safety Measures Act)尚未由州长签署成为法律。不过,伊利诺伊州州长普里茨克已表示计划签署该法案(Perlo, 2026),因此本文仍将其纳入分析。

[13] 算力阈值因忽略推理算力,且训练算力难以准确测算,而受到批评。(Ball & Ramakrishnan, 2025)。

设的人工智能标准与创新中心（Center for AI Standards and Innovation，CAISI；前身为美国人工智能安全研究所）以自愿合作方式与前沿人工智能开发者处理安全问题（美国国家标准与技术研究院，2025）。美国2025年《人工智能行动计划》及配套行政命令主要关注促进创新和协调联邦机构之间的政策，而非监管私营部门（美国总统行政办公室，2025）。加利福尼亚州、纽约州和伊利诺伊州的州级立法是美国最明确的前沿人工智能专项治理措施。不过，由于各方担忧这些措施可能影响产业发展，其立法过程并不顺利（Hine & Floridi，2025）。

英国采取“促进创新”（pro-innovation）战略。该战略并未制定一部综合性人工智能法律，而是要求现有监管机构在各自行业中适用若干跨领域原则。与此同时，英国通过设立新机构和参与国际协调机制来构建专门的前沿人工智能安全能力（英国科学、创新与技术部，2024；人工智能办公室，2023）。英国政府创办了人工智能安全峰会，并成立人工智能安全研究所（AI Safety Institute，现已更名为 AI Security Institute）。尽管具有约束力的横向立法时间线不断变化，英国仍持续推动评估工作与共同科学基础的建立（英国科学、创新与技术部，2023a；Jones 等，2025；英国首相办公室等，2023）。英国还鼓励各监管机构将《人工智能监管白皮书》倡导的促进创新、以原则为基础的监管方式提出的纳入行业监管工作人工智能办公室，2023；White & Case，2025）。

欧盟提供了最具代表性的成文法治理模式。《人工智能法》建立了风险分级制度，并对具有“系统性风险”的通用人工智能模型施加更严格的义务（《人工智能法》，2024）。该法的实施结合了硬法与软法。欧盟委员会推出的“人工智能公约”（AI Pact）曾鼓励企业提前自愿遵守相关要求（欧盟委员会，2024）。目前，《通用人工智能行为准则》（General-Purpose AI Code of Practice，CoP）是一项自愿工具，企业可以借此证明自己遵守《人工智能法》第53条和第55条（欧盟委员会，2025a）。欧盟委员会承认该准则可作为证明合规的有效方式，并降低签署方的行政负担来作为遵守准则的激励（欧盟委员会，2025）。欧盟的制度路径建立在长期治理准备和专家组工作的基础上。这在一定程度上解释了为何欧盟能够比其他司法辖区制定更加全面的法律（Smuha，2024）。

表1：人工智能治理制度架构比较

司法辖区	制度模式	主要特征
美国	州级制度开始逐渐成形；联邦层面由行政部门主导	美国CAISI更偏重创新；[14]加州、纽约州与伊利诺伊州已通过州级透明度立法；在联邦层面，行政部门是主要政策设计者，并明确推

[14] 在由人工智能安全研究所更名为人工智能标准与创新中心之后，CAISI的使命被调整为”促进创新”与”促进科学”，安全退居次要目标（U.S. Department of Commerce Office of Public Affairs, 2025）。尽管它强调人工智能监管需要促进创新，但也通过自愿性的模型评估应对人工智能的潜在风险（Oduro, 2025）。

司法辖区	制度模式	主要特征
		动联邦法律优先于州法；现有行业监管机构和一般法律制度构成主要监管框架。
英国	在现有制度内分行业提供指导	人工智能安全研究所开展"先进人工智能治理"研究；[15]未新设法定监管机构；现有机构依据跨领域原则分行业提供指导。
欧盟	成文法框架内的专职机构	人工智能办公室协调通用人工智能模型（GPAI）监管；《通用人工智能行为准则》实际上构成合规蓝图；签署方可以借此证明自己履行了法律义务。

然而，这些制度均面临着跨领域的放松监管压力。在美国，特朗普政府2025年的《人工智能行动计划》及配套行政命令将州级人工智能立法描述为创新障碍。联邦优先适用方案所针对的州级立法包括加利福尼亚州的 SB-53 和纽约州的《RAISE法案》（美国总统行政办公室，2025）。此后的行政措施同时加强了政府与产业界在前沿人工智能安全方面的协调，但相关合作仍然主要以自愿方式进行（美国总统行政办公室，2026）。在欧盟，欧盟委员会的《人工智能综合简化方案》（AI Omnibus）推迟了《人工智能法》部分高风险系统要求的实施，并放宽了上市后监测义务。截至2026年7月，《通用数据保护条例综合简化方案》（GDPR Omnibus）仍在谈判中。如果通过，该方案将修改《通用数据保护条例》，允许为了训练人工智能和减少偏见而在更广泛的范围内处理个人数据（欧盟委员会，2025b、2025c）。这些调整回应了2024年 Draghi 报告提出的竞争力问题，也回应了产业界对合规负担的压力（Draghi，2025）。

中国及其他司法辖区都尚未形成一套统一、可操作的"前沿人工智能"定义或治理模式。这既说明问题本身确实困难，也反映出各方已经形成一个基本共识：即使具体边界仍有争议，能力最强的系统也需要受到有别于普通系统的监管关注。

美国、英国和欧盟的治理方式仍存在明显缺口。下列三个缺口构成第四章分析的基础。第四章将研究如何扩展中国现有监管基础设施，以应对前沿人工智能风险。

第一，部分司法辖区要求开展前沿风险评估，但只有一个司法辖区实现了标准化，且独立验证仍然少见。欧盟《人工智能法》第55条要求具有系统性风险的通用人工智能模型提供者评估并降低相关风险。《通用人工智能行为准则》明确列出了需要处理的风险，包括 CBRN 能力、失控、网络攻击和有害操纵，并要求采用符合当前技术水平的评估方法，例如基准测试和红队测试。美国监管最全面的三部州法，即加利福尼亚州 SB-53、纽约州《RAISE法案》和伊利诺伊州《人工智能安全措施法案》，同样要求大型前沿人工智能开发者评估模型是否具有可能造成灾难性风险的能力，并公开评估结果。但是，每家开发者都自行设计评估框

[15] 依据人工智能安全研究所官网的使命声明（The AI Security Institute (AISI), n.d.）。

架、方法和可接受风险水平，并不存在统一标准。只有伊利诺伊州要求由独立第三方进行审计。在英国和美国联邦层面，相关评估均为自愿性质。美国2026年的联邦行政命令为模型访问和安全评估提供了一项自愿框架（美国总统行政办公室，2026）；英国则依赖《前沿人工智能安全承诺》（Frontier AI Safety Commitments）。这些制度总体上并不采用经过外部验证、以能力为基础的阈值，而依赖开发者自行评估。

第二，专门针对前沿人工智能的应急响应框架发展不足。部分司法辖区已经建立事件报告机制。加利福尼亚州 SB-53 要求在15日内报告事件，纽约州和伊利诺伊州的法案则要求在72小时内报告。然而，在处理前沿人工智能事件发生时的风险升级、机构间协调和快速能力评估等方面，尚无司法辖区建立一套结构化方案。

第三，针对灾难性前沿人工智能损害的问责框架仍未解决。各国侵权法都面临因果关系、偿付能力和威慑效果方面的问题，第2.2节将进一步讨论这些问题。尚无任何司法辖区建立起能够对上游开发决策形成实质激励的责任制度。

2.2. 私法监管工具与市场化手段

除欧盟《人工智能法》等技术专项法律外，许多司法辖区也在探索如何调整一般私法和市场型手段，使其适用于前沿人工智能。本节将其视为一类独立的治理工具，并评估它们能够在多大程度上应对前沿人工智能风险。本文讨论四类方案：责任机制、保险与补偿基金、许可与审批制度、市场驱动型监管。这些讨论将为后文分析适用于中国的机制奠定基础。

2.2.1. 侵权责任

侵权法通过创设民事诉权来降低事故的社会成本，并补偿受害方，因此可自然地成为人工智能损害问责体系的最后一道保障。在前沿人工智能领域，最相关的法律制度包括过失责任、无过失责任和产品责任。过失责任要求行为人对未能采取合理谨慎措施、防范可预见损害的行为负责。无过失责任不考虑被告是否已经采取合理谨慎措施。产品责任则适用于产品。欧盟修订后的《产品责任指令》已经明确将软件和人工智能系统纳入适用范围（Hogan Lovells，2025）。然而，将侵权法适用于前沿人工智能时，会遇到三个结构性问题。

第一，因果关系难以证明。主要问题并不是人工智能系统不透明，而是因为因果链条要穿过系统复杂的价值链与涌现行为。前沿人工智能系统造成的损害可能涉及训练数据提供者、基础模型开发者、下游微调者、部署者和最终用户。即使模型完全透明，也很难在这些主体之间划分各自的因果贡献。前沿能力还可能在训练过程中以不可预测的方式出现，从而削弱过失责任赖以成立的可预见性要求（Kuersten，2023；Ramakrishnan 等，2024）。如果开发者的行为合理，但其开发的模型仍产生了危险能力，过失责任标准可能无法要求该开发者承担责任。此外，当人工智能与用户意图、其他工具及既有知识共同促成危害时，很难通过一种明确的方法判断应当把多大比例的损害归因于人工智能。无过失责任在一定程度上简化了这一问题，因为它不考虑被告是否已经尽到谨慎义务（Frazier，2025），但无过失责任自身也存在问题。因果关系仍然需要得到证明，而鉴于很难证明哪些风险属于“可预见且重大”的风险，法院通常不愿意把人工智能开发纳入“异常危险活动”或“高度危险活动”类别（Bathaee，2018；Weil，2024）。同时，以设计缺陷为基础提出产品责任请求时，也很难为具有涌现能力的系统确定何为“合理的替代设计”（Sharkey，2024）。

第二，灾难性损害实际上难以投保。侵权救济受到被告偿付能力的限制。如果潜在损失远远超过任何单一被告所拥有的资产，侵权责任就不适合处理灾难性或系统性风险（Lior，2025；Weil，2024）。强制责任保险可以部分解决这一问题，下文将进一步讨论。但是，由于历史数据有限，传统保险市场很难为低概率、高损失的事件确定价格。

最后，责任制度并不总能产生有效威慑。如果潜在损害超过企业资产，责任威胁就无法有效改变企业的边际行为。例如，如果无论最终结果如何，一家公司都将破产，那么它就很少有动力投资额外的安全措施。责任通常也最常由下游部署者承担，因此难以对上游开发决策形成足够影响。但训练选择、内部评估和能力披露等上游决策，恰恰是最能影响前沿风险的因素（Ramakrishnan 等，2024；Weil，2024）。

现有的研究提出了多种调整方案。一种方案是将责任分配到整条价值链，按照比例划分责任，而不是将责任集中在单一节点。不过，如何划分责任在方法上仍然存在困难。欧盟《产品责任指令》使用“实质性修改”（substantial modification）这种二元阈值，在开发者和部署者之间分配责任（《产品责任指令》，2024）。如果部署者对基础模型进行了实质性修改，该部署者将成为新的制造商并承担全部责任；如果部署者未作修改，则适用过错责任。相比之下，真正的比例责任制度能够更加准确地反映各主体的因果贡献，避免修改程度刚好高于或低于门槛时产生截然不同的责任结果。但是，实施比例责任需要法院掌握其通常不具备的技术专业知识。另一种方案是调整谨慎标准。例如，不应只考虑开发者实际“知道”的情况，还应根据合理的部署前评估，考虑其“理应知道”的情况。这种方式可以在保留过错问责的同时，缓解可预见性这一问题。政府支持的保险或风险共担基金可以承担超出私人市场承保能力的灾难性风险（Lior，2025；O’Keefe 等，2020）。Weil（2024）提出的“未遂责任”（near-miss liability）主张：如果某项行为只因情况略有不同而没有造成灾难性损害，也可以对该高风险行为追究责任。这可以覆盖那些碰巧没有造成灾难性后果的高风险决策，从而弥补威慑缺口。

侵权责任不太可能单独构成完整的前沿人工智能治理制度，但它仍然可以发挥重要的兜底作用，特别是针对定义明确的“异常危险”活动或特定产品类别。第4.3节将分析中国侵权法规定和人工智能法建议稿如何处理这些问题。

2.2.2. 保险与补偿基金

保险和补偿基金与侵权责任相互配合，可以平滑并分散损失，在损害发生后确保受害者获得补偿。相关方案包括要求或大力鼓励开发者和部署者购买责任保险，以及通过征费或共同基金补偿难以通过诉讼解决的损害（Lior，2025；O’Keefe 等，2020）。原则上，这些机制可以附加在价值链的多个环节。例如，可以在上游强制要求基础模型开发者购买保险，也可以向平台和托管服务提供者征费。根据金融负担的具体分配方式，这些机制还能够对计算资源和云服务节点形成影响。

不过，这些工具在前沿人工智能领域存在两项主要局限。第一，由于历史数据有限，传统保险市场很难为低概率、高损失事件确定保险价格。如果不能根据风险水平定价，保险只能提供损失救济，很难激励相关主体预防事故（Lior，2023）。第二，投保范围通常不包括模型上市前的开发活动，因此开发阶段中与安全有关的行为得不到保险机制约束。可以通过在全行业或国家层面分散灾难性损失，采用补偿基金与政府背书的风险池来化解第一项局限。但这类机制在实施时会面临显著的政治经济障碍。

2.2.3. 许可制度

许可制度要求相关主体在遵守规定安全措施的前提下，取得从事特定活动的许可，从而在损害发生前加以预防，而不是主要依靠事后诉讼或赔偿。部分司法辖区已经针对前沿系统采用或考虑采用类似方式。中国通过备案、测试和检查要求，对面向公众的生成式人工智能实行一种事实上的许可模式。一些评论者认为，这些要求的实际功能不只是普通备案，而更接近许可制度（Sheehan，2024；Wise，2025b）。欧盟《人工智能法》通过合格评定、登记，以及针对违规产品的市场撤回权，引入了类似许可制度的逻辑。美国的 Blumenthal–Hawley 框架等立法提案则主张实行开发者层面的许可制度，将许可与风险管理、测试、数据治理、事件报告、审计和撤销权相结合（Baker，2023；Blumenthal & Hawley，2023）。

许可制度理论上具有较强灵活性，可以适用于多个环节：对计算资源提供者发放许可，例如限制对大规模计算资源的访问；对模型开发者发放许可，例如对训练活动实行许可；对下游应用实行许可。其局限主要体现在执行层面。对模型进行逐个许可或忽视开发者内部流程薄弱所产生的风险，而范围更广的许可制度则可能带来沉重的行政负担，并引发政治争议。在应用层实行许可可能造成监管瓶颈，使小型开发者处于不利地位，却不一定能够相应提高安全水平。第三章讨论的中国准许可制度同时展示了大规模实施许可制度的优势和局限。

2.2.4. 市场驱动型监管

市场驱动型监管（market-driven regulation）包括多种公私混合制度。在这类制度中，政府制定高层次安全目标，并将具体监管工作交给专业的私人机构。Hadfield 和 Clark（2026）建议许可多个私人“监管服务提供者”（regulatory service providers），由这些机构相互竞争，帮助企业实现可衡量的安全结果。Ball（2025）提出，如果开发者自愿接受经授权的私人机构的审计和认证，可以为其提供侵权责任安全港（safe harbor）。这些方案背后的基本判断是，公共机构在评估快速演进的系统时存在“技术能力缺口”。专业私人机构围绕监管质量展开竞争，可能更有能力跟上技术发展的速度。

不过，市场驱动型方法在前沿人工智能领域面临两项主要限制。第一，只有在能够制定可验证的结果指标，并将结果归因于特定企业时，这类方法才最有效。人工智能失控或能力扩散等分散于整个系统的风险很难归因，也很难衡量。第二，即使企业遵循了标准程序，损害仍有可能发生。因此，安全港条款可能造成问责缺口，尤其当私营监管者面临商业压力、倾向于给予认证而非拒绝批准时。市场驱动型机制可以补充公共监管，但不应取代公共监管，在私营监管能力较为成熟的法域尤其如此。

第二章考察了全球前沿人工智能治理格局的两个层面。第一，世界各司法辖区对前沿人工智能逐渐形成共同的基本认识，但在如何落实监管方面存在明显差异。每种治理方式在预训练监管和可执行评估方面都公认存在缺口。第二，第2.2节分析了四种私法工具：侵权责任、保险与补偿基金、许可制度和市场驱动型监管。这些工具分别处理治理挑战的不同方面，但没有任何一种工具能够单独解决全部问题：

- 侵权责任支持事后问责，但面临因果关系、偿付能力和威慑效果方面的困难。
- 保险和补偿基金可以缓解并分散损失，但本身不能防止损害发生。
- 许可制度能够进行事前预防，但会产生执行方式和适用范围方面的问题。
- 市场驱动型监管可以弥补公共机构的能力缺口，但最适合处理结果可以衡量的风险。

目前尚无任何司法辖区彻底解决前沿人工智能治理问题。有效的治理方式很可能需要结合上述工具，并将其置于欧盟《人工智能法》或中国现有监管架构等技术专项制度之上。中国已经具备相当可观的人工智能治理能力，但目前的制度主要针对内容风险和特定应用风险，而不是前沿风险。第三章和第四章将分析如何扩展这些制度。

3. 中国的人工智能治理制度基础

本章对中国现有的人工智能治理体系以及应对前沿人工智能风险的新兴机制进行梳理。中国已经针对人工智能在特定应用场景中的若干危害建立了相当完备的治理体系，但每项制度的优化主要依据内容类别或应用场景，而非模型能力。因此，前沿人工智能特有的风险仍处于这些制度的覆盖范围之外，或者只得到部分覆盖。第3.1节和第3.2节将对此进行分析。第3.3节将讨论另一组新兴机制，包括全国网络安全标准化技术委员会（TC260）的框架、业界的承诺、实验室主导的研究，以及学界提出的立法建议。这些机制已经明确涉及部分前沿风险，但尚未成为具有强制执行力的监管制度。第四章将讨论其中哪些机制可以扩展用于应对第2.1节所述的前沿风险。

专栏：中国法律制度与人工智能治理概述

中国人工智能监管背后的政治经济逻辑不同于欧盟《人工智能法》。在中国，人工智能被视为推动国家经济转型升级和参与地缘政治竞争的战略性技术。因此，监管既要对人工智能形成约束，也要帮助实现国家发展目标。相应地，中国制度需要权衡安全与发展，而欧盟政策讨论则更常采用安全与基本权利并重的框架。两种语境中的利益关系不同，因而在政治上可行的监管设计也有所差异。中国人工智能治理制度中有三个反复出现的结构性特征，值得在此说明。

从上游触发因素到正式制度。中国的人工智能法规在成为具有约束力的规则之前，须经历四个上游层次：现实世界中的触发因素，例如技术冲击、实际损害或竞争压力；通过讲话、指示和“习近平法治思想”等理论表述传达的中国共产党意识形态和领导层信号；由学者、智库和政策倡议者构成的“观念界 (World of Ideas)”，负责将政策指示转化为可操作的概念；以及起草并发布正式文件的党政官僚机构（Sheehan, 2023）。ChatGPT 于2022年末的发布几乎立即触发了这样一轮政策过程，并最终推动《生成式人工智能服务管理暂行办法》于2023年8月实施。

监管权限在各部门间分散。中国的人工智能治理并不由单一监管机构负责，而是由多个部门共同实施：

- 国家互联网信息办公室负责内容、算法和生成式人工智能；
- 工业和信息化部负责产业发展和标准化；
- 国家发展和改革委员会负责经济规划和计算基础设施；
- 公安部负责安全评估；
- 科学技术部负责科研伦理；
- 国家市场监督管理总局负责产品安全和市场竞争。

相关监管文件通常由多个部门联合发布。这既表明各部门已经形成政治共识，也使执法权可以覆盖多个行政体系。学者将由此产生的职权重叠和协调缺口称为“治理碎片化”（governance fragmentation）。

敏捷与迭代式治理。

与全国人民代表大会制定的法律相比，部门规章和行政法规可以更快地制定、修改和废止。因为传统立法程序往往难以及时跟上技术发展，这种治理方式有意在一定程度上容忍法律上的模糊性，以换取治理快速变化技术的能力。

3.1. 既有的人工智能专门监管

中国的人工智能专项监管基础设施并非通过一部综合性法律建立，而是针对不同应用逐项发展形成的。本节考察三个层次：

1. 自2022年以来陆续发布的纵向、窄范围监管规则，涉及推荐算法、深度合成、生成式人工智能、人脸识别和拟人化人工智能服务；
2. 国家互联网信息办公室通过算法备案制度审批面向公众的人工智能服务；
3. TC260 制定的国家标准，用于落实上述专项规则中的内容安全义务。

除标准外，这些制度大多是由国家互联网信息办公室[16]与其他部门联合发布的部门规章。它们具有充分的执行力，与全国人大制定的法律相比，可以更迅速地制定、修改和废止。学者将这种模式概括为“小、快、灵”的立法（Fan 与 Zhang，2026），认为它较为适合快速变化的技术。第4.1节将讨论如何扩展这些基础设施，并使其适用于前沿风险。

[16] 网信办的制度定位在西方监管体系中并无可参照的例子。它同时是国家行政机关，也是由习近平主席领导下的党机关——中央网络安全和信息化委员会——的执行机构。

3.1.1. 狭义人工智能监管

中国的人工智能监管是一次针对一类应用逐步建立起来的。自2022年施行的《互联网信息服务算法推荐管理规定》起，[17]每当一种新的人工智能能力持续受到政策关注，监管机构通常会针对该能力发布一项适用范围较窄的规定，而不是将其纳入一部综合性的人工智能法律。这些规定通常由国家互联网信息办公室会同其他部门发布。因此，现有监管规则的适用范围彼此有所重叠，但每项规则都针对特定应用类别。

2022年3月生效的《互联网信息服务算法推荐管理规定》建立了中国的算法备案制度。

规定要求具有“舆论属性或者社会动员能力”[18]的互联网信息服务提供者向国家互联网信息办公室备案算法并接受安全评估。规定还要求提高推荐算法的透明度，允许用户拒绝个性化推荐，并保护受到算法管理的劳动者。一般认为，这些规定是对2020年一篇调查报道的回应。该报道记录了外卖配送算法如何导致了不安全的工作环境（Sheehan 与 Du，2022）。

除规定具体义务外，这一制度还将算法备案确立为一种可以反复利用的监管工具。后续规定进一步将它扩展到新的人工智能类别。2023年1月生效的《互联网信息服务深度合成管理规定》[19]主要针对深度伪造以及文本、图像、音频和视频等其他合成内容。服务提供者必须标识人工智能生成内容，对用户进行真实身份认证，并依据内容安全义务审核内容。该规定还将算法备案制度扩展到深度合成算法。

2023年8月生效的《生成式人工智能服务管理暂行办法》[20]进一步将备案制度扩展到生成式人工智能算法。服务提供者必须在服务上线前开展安全评估、遵守内容安全义务、保证训练数据来源合法，并建立用户投诉机制。该办法的初稿于2023年4月发布，距 ChatGPT 面向公众推出约五个月。这使其成为主要司法辖区中针对一种新型人工智能服务出台速度最快的监管措施之一。

另外两项规定继续扩展了这一制度体系。[21]2025年的《人工智能生成合成内容标识办法》[22]要求对文本、图像、音频和视频中的人工智能生成内容同时添加显式标识和元数据标识。该办法统一了此前分散在不同专项规定中的标识义务。2026年的《拟人化互动服务管

---

[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”]

[21] 作为一项与算法备案体系并行但相互独立的制度，2025年6月1日起施行的《人脸识别技术应用安全管理办法》对生物识别信息的收集和使用作出了规范。该办法以数据保护、数据安全和网络安全相关法律为依据，并不属于算法备案框架，而是实行独立的省级备案制度；处理的人脸信息达到10万人时，即须履行备案义务（《人脸识别技术应用安全管理办法》，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”]

理暂行办法》[23]则将监管框架扩展到模拟人类人格、情感或关系的人工智能服务。这类服务的风险不同于《生成式人工智能服务管理暂行办法》所处理的风险。

每项规定均针对特定应用类别及相应损害类型。这反映出中国已经形成成熟的人工智能监管体系，但尚无任何一项规定是专门根据前沿人工智能风险进行设计的。例如，《生成式人工智能服务管理暂行办法》适用于在中国部署的前沿模型，但其主要义务针对内容安全，而不是前沿人工智能特有的风险。

3.1.2. 算法与模型备案

面向公众的生成式人工智能服务必须在上线前接受强制性安全审查。根据算法推荐管理规定，具有“舆论属性或者社会动员能力”[24]的服务必须在上线后十个工作日内向国家互联网信息办公室备案算法，并接受安全评估。这在实践中覆盖了大多数面向消费者的聊天机器人和内容生成工具。算法备案制度最初由中国2022年的推荐算法规定建立，之后逐渐发展为“双备案制度”，包括原有的“算法服务备案”和新增的“大模型备案”（Pi 等，2026；Wise，2025b）。虽然制度名称采用“备案”，但它实际上发挥着许可制度的作用。生成式人工智能系统在实践中需要接受监管机构的直接测试，监管机构也可以要求开发者修改系统（Sheehan，2024；Wise，2025a）。[25]自研模型需要接受省级和中央层面的两轮评估，整个过程可能持续五个月或更长时间。通过应用程序编程接口（API）使用已经获得批准的模型时，只需接受较简化的省级审查，通常需要两至三个月（Qian，2026）。目前共有四项公开备案清单，分别记录算法服务，深度合成服务，自主研发和二次开发的生成式人工智能模型，使用已备案模型部署的服务（Qian，2026）。截至2025年底，共有796项生成式人工智能服务完成地方和国家层面的备案，另有481项产品只需进行地方备案（国家互联网信息办公室，2026）。在所有类别中，较广义的算法备案制度共收录约3,800家主体提交的近6,000项备案（Pi 等，2026）。

尽管前沿模型在法律上属于上述要求的适用范围，监管机构开展的测试仍以内容安全为主。测试重点是模型对问题作出的回答和拒绝回答是否符合内容标准（Qian，2026）。此外，即使模型进行了重大更新，实践中目前通常也不要求重新备案（Qian，2026），尽管书面规则作出了相应要求（Zhang 与 Sun，2026）。不过，算法备案制度使监管机构能够了解哪些系统已经部署，以及由哪些主体部署。这建立了一套可用于处理其他风险的执法体系。2025年10月修订的《中华人民共和国网络安全法》[26]进一步支持了这一方向。该法第20条要求国家“加强人工智能风险监测评估和安全监管”。正如第4.1节将要讨论的，算法备案制度是落实这一要求的一条重要路径。

[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] 英译采自China Law Translate（2022）。

[25] 该制度通常被称为“算法备案制度”（algorithm registry），本文也沿用这一称谓。
[26]《中华人民共和国网络安全法》[“Cybersecurity Law of the People’s Republic of China”]

3.1.3. 标准

中国的人工智能标准体系主要以全国网络安全标准化技术委员会（SAC/TC260）和国家标准化管理委员会为基础。2024年，工业和信息化部与国家标准化管理委员会发布《国家人工智能产业综合标准化体系建设指南》，规划了人工智能安全、数据、计算基础设施和评估等领域的五十多项标准。其中三项推荐性国家标准将《生成式人工智能服务管理暂行办法》规定的内容安全义务转化为具体要求：

- GB/T 45654—2025：《生成式人工智能服务安全基本要求》（TC260，2025c）；
- GB/T 45674—2025：数据标注安全标准（TC260，2025d）；
- GB/T 45652—2025：预训练和微调数据安全标准（TC260，2025e）。

中国标准与监管之间的关系明显不同于美国和欧盟。在美国，国家标准与技术研究院发布自愿性指南。其他政府机构可以引用这些指南，但很少强制要求遵守。在欧盟，由产业主导的机构在欧盟委员会授权下制定协调标准，并将其作为推定合规工具（Cantero Gamito 与 Marsden，2024）。在中国，TC260 是国家治理体系的组成部分，人工智能标准路线图与监管文件和国家政策目标之间存在明确协调。因此，GB/T 45654—2025 虽然形式上属于推荐性标准，但算法备案的安全评估程序通过引用该标准，使其在实际操作中成为强制性要求。

GB/T 45654—2025 是上述三项国家标准中的核心标准。尽管形式上属于推荐性标准，它事实上具有强制性。《生成式人工智能服务管理暂行办法》要求面向公众部署服务前开展安全评估，国家互联网信息办公室的算法备案评估又采用该标准的风险框架。因此，希望获得批准的服务提供者必须证明自己符合该标准（Sheehan，2024；Wise，2025a）。该标准的实体要求围绕31类风险构建，包括：违背社会主义核心价值观、歧视性内容、商业违法违规、侵犯他人合法权益，以及不同服务类型特有的准确性或可靠性问题（TC260, 2025c, 附录A）。这31类风险全部属于内容、商业或个人权益风险，没有任何一类专门涉及前沿人工智能风险。

该标准的监测条款同样以内容安全为中心。服务提供者必须“持续监测模型输入，防范注入攻击、数据窃取、对抗攻击等恶意输入攻击”[27]（第5.3条），并对用户的违法输入采取面向用户的检测措施（第6.5条）。虽然服务提供者必须在模型发生重要更新后“重新开展独立安全评估”（第5.4条），[28]但标准并未要求接受外部复核、第三方审计或向国家互联网信息办公室重新备案。这些规定建立了有意义的内容安全基础设施，但无法充分观察模型本身随能力提升而发生的变化。该标准不要求持续评估模型行为、能力变化或涌现属性。

3.2. 既有的人工智能相关机制

以下三项中国法律机制已经或正逐步被应用于人工智能活动：以2024年修订的《突发事件应对法》为基础的应急响应框架（第3.2.1节）；以《民法典》第七编为基础的侵权责任框架（第3.2.2节）；由2023年科技伦理审查办法和2026年人工智能伦理暂行办法确立的伦理审查框架（第3.2.3节）。每项机制都提供了可以用于治理人工智能风险的体系，但也都存在

[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”

[28] “再次自行组织安全评估” [“conduct another independent security assessment”]

具体缺口，使其目前不足以应对前沿人工智能情形。第4节将分别提出扩展这些制度的路径。

3.2.1. 应急响应

中国已建立起较为完善的应急响应体系，该体系可能提高对人工智能实验室的监测与报告要求（J. Zhang 等，2025）。2007年颁布、2024年修订的《突发事件应对法》建立了一套四级响应框架，涵盖自然灾害、事故灾难、公共卫生事件和社会安全事件四类突发事件，并根据人员伤亡和物质基础设施损失确定事件等级（《中华人民共和国突发事件应对法（2024年修订）》，2024）。该法不仅规定了应急处置与恢复机制，也强调事前准备和监测，坚持“预防为主”的原则（第5条）。

该法为中国的应急管理体系提供了法律基础。这一体系通常被概括为“一案三制”，[29]即在国家突发事件总体应急预案（National Emergency Response Plan，NERP）的统领下，将应急管理的体制、机制和法制有机结合起来。国家突发事件总体应急预案之下设有针对特定风险的专项应急预案、[30]部门应急预案和地方应急预案。专项应急预案通常用于应对影响重大、技术复杂且涉及多个部门的突发事件，例如《国家核应急预案》（国家原子能机构，2022）。

2018年，中国成立应急管理部，负责统筹协调应急处置工作（Hou，2018）。对于人工智能相关事件，国家互联网信息办公室可能会担任牵头部门（国家突发事件总体应急预案第2.1条）。2025年修订的国家突发事件总体应急预案将网络、数据和信息安全事件归入“事故灾难”，[31]这意味着人工智能事件一旦造成人身伤害，便可能被认定为灾难。人工智能安全也被纳入监测范围（第3.2.1条）。

根据这一制度结构，前沿人工智能特有的风险可以通过两种方式处理：由国家互联网信息办公室制定新的专项应急预案，或者修订网络安全、公共卫生事件等领域现有的专项预案（国家互联网信息办公室，2025a；中华人民共和国国家卫生健康委员会，2006）。

在这一总体框架的基础上，TC260于2025年发布《生成式人工智能服务安全应急响应指南》，首次针对人工智能事件提出了结构化的处置方案。该指南将事件分为信息内容安全事件、数据安全事件、网络攻击事件和其他安全事件四类（TC260，2025b），并根据受影响主体的重要程度、业务损失的严重程度和社会危害的严重程度，将事件划分为四个等级（第5条）。指南针对《突发事件应对法》规定的四个阶段分别提出了处置措施，并设置了具体的报告门槛。例如，如果三级及以上事件在24小时内发生五次以上，并影响至少一万名用户，服务提供者必须立即向主管部门报告；四级事件则可以由服务提供者自行处置并定期报告（第6.3.2条）。

尽管中国的应急响应体系相当完备，但它并非针对前沿人工智能情形而设计。现行生成式人工智能应急措施主要依据事件发生次数和服务中断程度设置处置门槛，其所设想的情形与最严重的前沿人工智能事件并不相同。例如，恶意行为者可能只需与高能力模型交互一

[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]

次，便能在制造生物或化学武器的任务中获得实质性帮助，而且这一过程完全不会干扰其他用户。在模型交互结束后，损害才会于下游环节显现，并发生在服务运营者的视野之外，因此不会触发现行报告门槛。

近期的高层政策信号表明，这一制度缺口已经受到关注，尽管相关表述尚未专门针对前沿人工智能。2025年4月，习近平主席在中共中央政治局集体学习时提出，要建立人工智能技术监测、风险预警和应急响应体系，确保人工智能“安全、可靠、可控”（CSET，2025；新华社，2025）。“十五五”规划延续了这一政策思路。国家发展和改革委员会副主任黄如在一篇官方解读中提出，要“推动基础设施、应用系统等安全能力建设”，并”加强人工智能应用对伦理规范、就业冲击等经济社会影响的前瞻评估和监测处置”（China SEI，2026）。第4.2节将进一步讨论如何据此扩展现有制度。

3.2.2. 侵权责任

中国侵权责任制度的发展对人工智能治理可能产生重要影响。根据《民法典》第七编（《中华人民共和国民法典：第七编侵权责任》，2021），中国侵权责任原则上依照第1165条适用过错责任。不过，第1165条第二款同时规定，依照法律规定推定行为人有过错，而行为人不能证明自己没有过错的，应当承担侵权责任。这实际上是对法律明确规定的特定情形实行举证责任倒置。目前，该规定尚未覆盖前沿人工智能损害[32]，如要扩展其适用范围，需要另行立法。

现有人工智能司法案件并未集中处理前沿风险，而主要涉及著作权和人格权问题。相关案件包括：一起被列为全国参考典型案例的声音权益案件（China Daily，2025）；北京法院针对换脸应用侵犯个人信息作出的处罚判决（China Daily，2024）；一起涉及人工智能聊天机器人传播淫秽内容的案件（China Law Translate 与 Daum，2026）；以及一起杭州互联网法院关于人工智能”幻觉”，依据过错责任而非产品责任作出的判决（杭州互联网法院，2025）。2025年11月，最高人民法院副院长陶凯元表示，最高人民法院正在起草有关人工智能案件审判的指导意见。她列举了法院目前面对的若干开放性问题，包括人工智能生成内容是否应当受到著作权保护、人工智能能否成为“作者”或“发明人”，以及如何在人工智能生成内容的创作者、使用者、传播者和投资者之间分配利益（Tao，2025）。与此同时，一个新兴的人工智能保险市场正在逐步形成。中国人民保险集团已经开始提供生成式人工智能侵权责任保险（Fang，2025），但该市场尚不成熟。

尽管这一法律领域仍处于早期阶段，现有侵权法中至少有三类规则可以适用于人工智能损害。第一，第1167条规定的停止侵害、排除妨碍和消除危险，可以为法院命令下架模型或限制其能力提供法律依据，适用于已部署系统持续造成损害的情形。第二，第1168条和第1169条规定的共同侵权和连带责任，[33]可以涵盖多个主体共同造成同一损害的情形。因此，如果基础模型开发者、微调者和部署者共同促成有害输出，这些规定可以适用于相关人工智能价值链。第三，第1194条和第1197条规定了网络用户和网络服务提供者的责任。如果网络服务提供者知道或者应当知道网络侵权行为，却未采取必要措施，这些规定便可能适

[32] 《个人信息保护法》已经对个人信息处理者实行过错推定（第69条）。人工智能开发者和服务提供者处理个人信息时，也在该规定的适用范围内（《个人信息保护法》，2021）。

[33] 在连带责任制度下，每名被告均须就全部损害承担独立责任；原告可以向其中一名、数名或全部被告请求赔偿全部损失。

用。因此，其适用范围可以扩展至人工智能平台、API 提供者，以及托管具有已知危险能力系统的模型平台。这些规定共同提供了多条处理人工智能损害的潜在路径，覆盖范围从针对性滥用一直延伸至信息环境恶化等更广泛的损害（见第2.2.1节和表1）。

除上述制度外，《民法典》和《产品质量法》规定的产品责任也可能成为另一条救济路径。有关产品缺陷的请求可以涵盖设计缺陷、制造缺陷、警示缺陷和安装缺陷。产品缺陷造成他人损害的，生产者应当承担侵权责任；受害者可以向生产者或销售者请求赔偿。产品投入流通后发现存在缺陷的，相关主体必须采取召回等补救措施。如果生产者或销售者明知产品存在缺陷，仍然生产或销售，并因此造成他人死亡或者严重健康损害，受害者还可以请求惩罚性赔偿。

不过，将中国现有侵权责任制度适用于人工智能损害仍存在明显局限。正如 Wu（2025）所指出的，传统产品责任的框架不足以处理与第2.2.1节讨论的困难类似的三个核心问题：难以证明过错和因果关系；难以界定人工智能特有的新型损害；以及难以在复杂的开发和部署链条中分配责任。人工智能造成的新型损害很难明确归入现有的财产损害或非财产损害类别。精神损害达到“严重”程度的认定门槛较高。当开发和部署活动分散于多个主体时，也很难确定具体的责任承担者。第4.3节将讨论如何调整侵权责任制度，以弥补这些缺口。

3.2.3. 科技伦理审查制度

中国的伦理审查制度以两项文件为基础：科学技术部于2023年发布的试行《科技伦理审查办法》，[34]为各类科技活动规定了一般性的伦理审查要求（科学技术部，2023）；以及2026年试行的《人工智能科技伦理审查与服务办法》，[35]在这一总体框架之上增设了人工智能专项义务（工业和信息化部等，2026）。与其他国家相比，《人工智能科技伦理审查与服务办法》的一项突出特点是将伦理审查延伸至研究和开发阶段，而不仅限于部署阶段。例如，从事人工智能科技活动的高等院校、科研机构、医疗卫生机构和企业必须设立内部科技伦理审查委员会，并在开展符合条件的活动前提交审查（第9条）。因此，如果前沿人工智能开发活动没有被部署阶段的备案制度覆盖，仍可能被纳入伦理审查范围。

《人工智能科技伦理审查与服务办法》建立了分级审查体系。大多数活动依照一般程序或简易程序接受内部委员会审查。被列入工业和信息化部与科学技术部联合发布的高风险清单的活动，还必须接受外部专家组的复核（第21条、第25条）。

现行高风险清单包括三类活动：对人的行为、心理情感或生命健康具有较强影响的人机融合系统；具有舆论属性、社会动员能力或社会意识引导能力的算法模型；用于安全风险或个人健康风险场景的高度自主化自动决策系统。其中，第三类与当前的前沿人工智能系统最为接近，但其具体解释和适用方式仍不明确。

该办法中有两个条文与前沿人工智能评估直接相关。第15条规定了六项审查重点：增进人类福祉（包括风险与收益评估）、公平公正、可控可信（包括持续监测方案和应急响应预案）、

[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)”]

透明可解释、责任可追溯，以及隐私保护。这些要求与前沿人工智能评估制度需要采用的标准存在较大重合。第26条规定，如果某项活动已经依据深度合成、算法推荐或生成式人工智能服务管理制度完成备案，且伦理审查是获得批准的条件，则可以免于专家复核。这一豁免承认了伦理审查制度与算法备案制度之间的职能重叠，同时继续将不属于上述备案制度适用范围的活动纳入伦理审查。该办法目前仍属于“暂行”制度，相关执法经验有限，因此其实际运作模式尚在形成。工业和信息化部启动的人工智能伦理审查服务试点可能有助于推动制度落地（新华社，2026），并提供统一制度、确保其有效运行所需的具体规范（Zhu 等，2025）。此前2023年规则的执行力度较弱，已经出现审查延误、适用不一致和指导不足等问题（L. Wu 与 Kong，2023）。第4.1节将进一步讨论如何运用第15条规定的审查标准和第26条规定的备案豁免来应对前沿风险。

3.3. 新兴的前沿人工智能专项机制与提案

中国的标准制定机构、行业协会、研究机构和法学研究团队正在迅速采用与前沿安全有关的概念和表述。失控、灾难性风险、基础模型风险以及与化生放核能力[36]有关的关切，如今已经出现在 TC260 框架文件、行业承诺、研究实验室制定的风险分类体系和学界起草的人工智能法建议稿中。目前，这些概念的采用主要停留在理念层面，但它们表明，监管讨论所关注的议题已经发生了实质性变化。

3.3.1. 安全框架与立法前景

2025年9月发布的 TC260《人工智能安全治理框架》2.0版，是该标准制定机构迄今针对前沿风险发布的最为详细的软法文件（TC260，2025a）。该框架从应用场景、智能水平和应用规模三个维度对人工智能安全风险进行评估，并将风险严重程度划分为五级。最高等级为“特别重大安全风险”[37]，其特征包括灾难性、系统性、颠覆性和不可逆性（附录1）。该框架在风险图谱中明确列出了以下需要应对的风险：与网络、核、生物、化学和导弹武器有关的知识与能力；人工智能失控；自主意识的出现以及“与人类争夺控制权”的情形。框架第5.11条还要求开发者开展人工智能失控风险测试。该框架不具有强制约束力，属于指导性文件而非国家标准。不过，它反映了未来可能被纳入 GB/T 国家标准的政策重点。TC260 新近设立了专门负责人工智能安全的 WG9 工作组。在该工作组第一次会议上，组长周伯文提出要前瞻研判前沿风险。这表明 TC260 未来可能开展更多专门针对前沿风险的标准制定工作（Wagner & Lindblad，2026）。

私营企业也作出了一系列风险管理承诺。2024年12月，包括 DeepSeek、智谱人工智能、MiniMax、零一万物、阿里巴巴、百度、华为和科大讯飞在内的17家中国人工智能企业签署了人工智能产业联盟（AIIA）的《人工智能安全承诺》。该框架包含六项承诺，涉及设立专门安全团队、开展安全测试、保障数据安全和基础设施安全、提高模型透明度，以及开展前沿安全研究。最后一项明确提出，要评估智能体系统和具身智能系统可能存在的偏见、歧视和不可控风险（人工智能产业联盟，2024）。2025年7月，在世界人工智能大会（WAIC 2025）期间，中国人工智能发展与安全研究网络成员在中国信息通信研究院牵头下发布了

[36]即化学、生物、放射性与核能力，英文简称CBRN (Chemical, Biological, Radiological, and Nuclear)

[37] “特别重大安全风险” [extremely serious safety risk]

《中国人工智能安全承诺框架》。该框架原样保留了前五项承诺，但重新表述了第六项承诺：不再明确列举智能体和具身人工智能，而是使用范围更广的“人工智能系统在前沿领域中的滥用风险”[38]和“高危场景”[39]等表述。框架还新增第七项承诺，涉及人工智能安全治理国际合作以及“智能向善”（AI for good）（中国人工智能发展与安全研究网络（CnAISDA），2025）。这些承诺强调面向利益相关方和公众提高透明度，但没有创设诸如要求企业公布前沿安全框架之类的可强制执行的义务。中国信息通信研究院还与人工智能产业联盟等机构共同运营人工智能安全基准测试。其2.0版新增了专门的前沿安全维度，涵盖自我意识、模型欺骗、危险领域滥用和模型失控。因此，尽管算法备案制度和 GB/T 45654—2025 仍主要关注内容安全，这套基准已经将评估范围延伸到前沿风险（中国信息通信研究院，2026a）。

研究实验室也提出了风险管理措施。上海人工智能实验室与安远AI（Concordia AI）合作发布的前沿人工智能风险管理框架SafeWork-F1，被安远AI称为”中国首个针对通用人工智能模型严重风险的综合管理框架”（AI Safety in China, 2025）。该框架将风险划分为滥用风险、失控风险、事故风险，以及系统性风险这四个领域，并评估七个关键风险方向，包括，包括网络攻击、生物与化学风险、说服与操纵、战略欺骗与密谋、不受控制的自主人工智能研发、自我复制，以及共谋（上海人工智能实验室 & 安远AI，2025）。该框架将模型划入绿色、黄色或红色区域。红线是不可容忍的风险门槛，一旦触及就必须暂停开发或部署；黄线则是早期预警指标，一旦触及就必须加强风险缓解措施，并采取受控部署方式（上海人工智能实验室 & 安远AI，2025）。上海人工智能实验室在2025年7月发布的一份技术报告中，将该框架用于评估近期的前沿模型。报告发现，所有被评估模型均处于绿色或黄色区域；具备推理能力的模型在自我复制和战略欺骗方面进入黄色区域。该报告还指出，新发布模型的安全评分有所下降（上海人工智能实验室，2025）。2026年2月发布的1.5版将评估扩展至七个风险领域中的五个，并得出结论：“当前风险总体可控，但不断出现的细粒度漏洞要求建立更加严格、持续的监测机制”（Liu et al., 2026）。该框架不具监管属性，但它将风险分类与实证评估结合起来，是中国机构迄今发布的前沿风险监测方案中最为具体、最具操作性的一个。

目前尚无法律正式实施上述安全框架，但这些框架所关注的部分问题已经开始进入具体的立法程序。公安部于2026年1月31日公布《网络犯罪防治法（征求意见稿）》，将中国针对服务提供者的义务扩展至人工智能辅助犯罪（中华人民共和国公安部，2026）。该草案要求网络运营者在将新技术、新应用投入使用前开展网络犯罪风险评估，并明确将人工智能纳入其中。人工智能服务提供者必须监测、阻止并报告利用其服务“批量生成恶意代码”或实施其他犯罪行为的用户（第42条）。与前述制度类似，该草案主要通过服务提供者的监测义务应对滥用风险，而不是依据模型能力设置阈值。不过，这项规定具有重要意义，因为它明确承认前沿模型可能降低自动化网络攻击的成本。根据国务院2026年度立法工作计划，该草案将提交全国人大常委会审议（国务院办公厅，2026）。

[38] “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]

[39] “高危场景” [high-risk scenarios]

3.3.2. 学者建议稿

尽管中国目前尚无综合性人工智能法，但已出现若干由学术工作组牵头的建议稿。[40]这些文本旨在为正式立法提供参考，本身并不具有法律约束力。中国社会科学院人工智能示范法（Model AI Law，MAIL）由中国社会科学院联合学术界和产业界起草。其多个版本逐步增加了专门针对前沿人工智能的规定（示范法起草组，2023、2024、2025、2026）。早前版本提出设立国家人工智能主管机关、负面清单许可、基于FLOP的基础模型定义，以及针对基础模型开发者的特别义务（包括安全与风险管理制度）、面向安全的专项算力投入、独立机构监督，以及年度社会责任报告（示范法起草组，2025）。

2026年5月发布的最新版本 MAIL 4.0对近期产业发展作出了回应。针对人工智能智能体可能产生的失控风险，4.0版提出了更加具体的治理要求。人工智能智能体服务提供者必须设置撤销授权和终止运行等功能，以帮助用户保持对智能体活动的实质性控制（第73条）。人工智能开发者还必须建立“熔断机制”以防止失控等严重风险（第58条）。4.0版进一步区分了开发者、服务提供者和用户之间的责任分配（第100条）。虽然MAIL 属于示范法而非正式立法，但4.0版明确与“十五五”规划相衔接，并被定位为中国立法机关和监管机构可以参考的文件。

中国政法大学AI善治学术工作组由张凌寒教授统筹协调。该工作组在2024年至2026年间发布的三份文件中，逐步加强了对前沿人工智能的关注。2024年《人工智能法》规定了“关键人工智能”，依据算力、参数规模或使用规模进行界定（第50条）。该建议稿还针对通用人工智能规定了特别要求，包括价值对齐和基于能力的安全评估（第77条）（L. Zhang 等，2024）。该工作组随后发布的《关于人工智能立法的重点制度建议》延续了这一思路，提出建立风险补偿基金，授权全国人大在人工智能试验区暂时调整相关法律，并要求关键人工智能服务提供者履行备案义务和定期风险评估义务（AI善治学术工作组，2026a）。

该工作组2026年发布的《人工智能法治研究十大议题（2026）》[41]进一步表明，其研究重点正在明显转向前沿人工智能。其中，第六项议题为“前沿人工智能安全保障与极端风险防控的法治进路”[42]（AI善治学术工作组，2026b）。该议题指出，随着人工智能系统的规模、自主程度和社会嵌入程度不断提高，其风险将“由局部性、可控风险向系统性、极端性风险扩展”[43]，其中包括技术滥用和失控风险（AI善治学术工作组，2026b）。工作组专家文禹衡在发布上述研究议题的研讨会上总结了学界共识：在处理前沿人工智能安全问题时，不能重研发、轻防控，而应建立覆盖人工智能全生命周期的法律制度（AI善治学术工作组，2026b）。中国政法大学工作组发布的这些文件均不具有法律约束力，但三份文件的发展过程表明，前沿人工智能专项议题在其制度设计中所占的地位正在显著上升。

---

[40] 2026年5月下旬，即本文分析工作已基本完成后，一份由产业界起草的人工智能法建议稿开始流传（智合标准化建设，2026）。该建议稿由智合标准中心和金杜律师事务所牵头，采用基于场景和影响的风险分类方式，而不是基于能力的分类方式。本文为求完整而予以说明，但不作深入分析。

[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]

2026年4月，南京大学法学院人工智能法研究团队发布《人工智能基本法（专家建议稿1.0版）》，成为参与这一讨论的第三个学术团队（南京大学法学院人工智能法学团队，2026）。南大建议稿的定位为基本法层级的框架性法律，[44]即确立原则、把操作细节交由下位法规定（第2条）。与中国社会科学院和中国政法大学的建议稿不同，南京大学建议稿没有采用专门针对前沿人工智能的表述。第26条第4项所称的“系统性风险”[45]是与前沿风险最为接近的概念。尽管如此，该建议稿仍为将来引入前沿人工智能专项措施预留了制度空间。第26条界定了高风险人工智能活动，并要求相关活动主体履行与风险程度相适应的安全保障、风险控制和合规管理义务。该条进一步规定：“高风险人工智能活动的具体监管制度，由法律、行政法规或者国家有关规定另行规定”（第26条）。第13条规定的安全可控原则还要求，人工智能系统必须具备与其用途、影响范围和风险程度相适应的安全性与可控性（第13条）。即使《人工智能基本法》本身没有明确使用“前沿人工智能”这一概念，这些规定共同为通过下位法规建立前沿人工智能专项监管制度保留了可能性。

尽管学术法律研究团队在上述建议稿中不断强化前沿人工智能专项表述，官方层面推动综合性人工智能立法的进程却更为复杂。国务院2023年和2024年度立法工作计划均将人工智能法列为预备审议项目。2025年度立法工作计划删除了这一项目，改用“推进人工智能健康发展立法”等概括性表述（Geopolitechs，2025）。2026年度立法工作计划再次改变方向，提出“加快推进人工智能健康发展综合性立法”，[46]并列出六项需要获得立法保障的共同要素，包括数据、算力、算法、知识产权、网络安全和供应链安全（国务院办公厅，2026）。相比之下，全国人大常委会2026年度立法工作计划仅再次将人工智能列为预备研究项目，即三个立法优先级别中最低的一级。该计划将“人工智能健康发展”与财政政策和网络暴力治理并列，作为需要开展研究、而非已经进入实际起草阶段的议题（Wei，2026）。不过，在此期间，学界提出的法律建议稿仍在不断发展和完善。与此同时，围绕人工智能智能体、伦理审查等议题的大量监管文件和标准也在持续出台。这表明，尽管综合性人工智能法已经获得政策层面的明确支持，但尚未真正进入积极起草阶段，其他相关领域的监管活动依然十分活跃。2023年10月，国家数据局正式成立，其职能由此前分散在国家互联网信息办公室和国家发展和改革委员会的部分职能整合而成（Arcesati 与 Groenewegen-Lau，2023）。这一机构调整表明，中国可以针对部门协调缺口重新组织行政架构。因此，如果综合性人工智能法最终进入正式立法并获得通过，中国在制度上也具备整合人工智能监管职能的可行性。

总体而言，第三章呈现了两种格局。第一，现有人工智能专项监管体系，包括第3.1节和第3.2节讨论的制度，已经覆盖多个领域，但没有任何现有工具将前沿人工智能的独特风险作为一个专门类别加以处理，也没有依据模型能力调整监管要求。第二，第3.3节所讨论的新兴前沿人工智能专项制度，包括 TC260《人工智能安全治理框架》2.0版、AIIA《人工智能安全承诺》、中国社会科学院 MAIL 4.0版、中国政法大学建议稿，以及 SafeWork-F1 1.0版和1.5版，已经明确列出前沿人工智能特有的风险，但尚未成为正式监管制度的一部分。因此，第四章需要解决的主要问题并不是缺少监管架构，而是如何调整现有制度。中国已经具备相应监管体系，但仍需要扩展这些制度，使其能够反映第2.1节所述前沿模型更强的能力及其相应产生的更高风险。

---

[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]

4. 弥合前沿人工智能治理缺口的提案

在分析全球前沿人工智能治理格局和中国现有人工智能治理体系的基础上，本章探讨如何针对前沿系统调整中国现有制度。本章并不试图穷尽所有应对前沿风险的方式，而是集中讨论三个领域：第4.1节讨论前沿风险评估，第4.2节讨论应急响应，第4.3节讨论责任制度。每个领域的方案都建立在中国现有监管体系之上。下文提出的思路旨在完善现有的标准、评估、应急响应和责任机制。在依托整体监管体系的同时，将监管注意力集中于前沿系统特有的风险。在对国际前沿人工智能监管方式的梳理、与中国相关领域专家的交流，以及对中国人工智能治理和监管环境的理解的基础上，我们提出下文中的这些考量。

4.1. 前沿风险评估

本节建议从三个方面扩展中国现有的人工智能评估体系。现行评估程序主要按照内容类别或应用场景设置要求，而下列方案将在此基础上增加根据模型能力划分的层级：在算法备案安全评估中开展分级前沿风险测试（第4.1.1节）；在模型的整个生命周期内开展持续评估（第4.1.2节）；在2026年5月国家互联网信息办公室、工业和信息化部及国家发展和改革委员会发布的实施意见所提出的智能体备案制度中，增设基于能力的门槛（第4.1.3节）。这些方案旨在帮助监管机构更加充分地了解前沿模型的风险。这是实施本章其他所有干预措施的前提。如果缺少系统性的能力评估，监管机构就无法合理设置应急响应门槛、确定责任标准，也无法判断现有事件响应机制能否有效处理相关损害。

4.1.1. 分级的部署前风险测试

分级前沿风险测试是指：对于训练算力超过一定门槛的模型，在部署前增加一组评估，测试内容包括增强 CBRN 能力、自主获取能力和失控风险等。之所以有必要实行这种测试，是因为如果对小型聊天机器人和前沿基础模型采用完全相同的内容安全评估标准，会遗漏上述风险。这些风险主要出现在模型能力范围的顶端，其性质也不同于国家互联网信息办公室安全要求和 GB/T 45654—2025 风险分类体系主要处理的内容安全风险（见第3.1.3节）。一个模型即使通过了内容安全审查，仍可能存在前沿风险。要识别这些风险，评估强度就必须随模型能力提高而增加，并且需要在模型部署前开展，以便监管机构仍有机会采取干预措施。

中国现有的算法备案体系可以支持分级前沿风险测试制度，无须为此设立新的机构。根据中外研究人员的讨论，可以在现有流程之上增加前沿风险专项测试，采取以下分级方式：

- 前沿风险测试只适用于训练算力超过较高门槛的通用人工智能系统。该门槛可以与中国人工智能立法建议稿和正在形成的国际最佳实践相衔接，从而将监管范围限制在少数先进模型之内。如果算力门槛并不合适，也可以参考中国其他监管制度，采用其他方式确定适用范围。例如，可以借鉴《人工智能科技伦理审查与服务办法》，以高风险活动清单识别应当接受测试的系统。
- 对超过上述门槛的模型，首先使用已经建立的前沿风险基准测试集，开展成本较低的自动化测试。这可以为监管机构系统了解模型风险提供低成本且标准化的基础。
- 如果模型在相关基准测试中的表现达到或超过国内外领先模型的水平，再对其开展更加深入的评估。

这一制度可以通过 TC260 为 GB/T 45654—2025 起草前沿安全附录来实现，并由国家互联网信息办公室通过算法备案制度现有的模型安全评估程序执行。主要经由标准制定程序实施这一方案，可以使产业界、学术界和政府共同制定技术要求，并随着评估方法的改进及时更新要求，而无须每次都修改法律或监管规定。如何确保分级门槛与技术发展保持同步，是另一个需要解决的制度设计问题。学界提出的法律建议稿提供了一种参考模式：中国社会科学院人工智能示范法的早期版本曾授权国家人工智能主管机构“定期更新”算力标准（第90条第八项）。这种行政更新机制可以避免欧盟目前面对的正式修法周期问题（示范法起草组，2025）。

4.1.2. 持续评估

前沿模型的能力特征在部署后仍可能发生变化。这些变化可能源于微调、模型更新、通过支架系统将模型改造为智能体，以及模型在实际部署环境中表现出初次评估时未曾出现的行为（Bengio 等，2026）。一次性的部署前测试只能反映动态变化中的系统在某一时间节点的状态。持续评估则可以识别部署后才显现的非预期风险。

监管机构可以通过修订 GB/T 45654—2025，要求服务提供者持续监测模型，并定期向监管机构披露相关信息，例如每季度报告一次。该标准已经要求服务提供者持续监测模型输入（第5.3条，见第3.1.3节），但监测对象是内容和输入攻击，而不是模型自身的能力。

如果将监测重点调整为前沿安全，服务提供者就需要跟踪模型部署后的能力变化，并在必要时根据前沿风险触发指标（frontier risk tripwires）或预先设定的“如果—那么”（if-then）框架采取行动。这类框架可以规定：模型如果达到某一能力水平，那么必须重新接受评估、限制使用或向监管机构报告（Karnofsky，2024）。这一方案改变的是监测对象，而不是底层制度体系。

目前已有两组学界提出的人工智能法建议稿涉及持续评估制度，可以在未来为这一机制提供更长期的立法依据。中国社会科学院人工智能示范法4.0版要求人工智能服务提供者建立覆盖全生命周期的风险管理体系（第57条），并特别要求基础模型开发者建立包含持续风险监测的风险管理体系（第65条）。中国政法大学人工智能法建议稿要求通用人工智能开发者定期开展安全评估并提交报告，其中包括价值对齐评估和基于能力的评估（第77条）。该工作组2025年发布的关键制度建议还要求关键人工智能服务提供者进行备案、定期开展风险评估并提交报告（第十四部分）。

4.1.3. 智能体备案

建立在前沿基础模型之上的智能体，可以在不同环境中持续追求可能产生重大后果的目标，并在较少人类监督的情况下自主采取行动。与直接滥用模型相比，智能体发生故障时更难确定责任，也更难及时阻止。有害提示词可能会出现在受监测的模型输入和输出中，但智能体造成的损害可能来自跨越多个工具和环境的一系列自主行动。这些行动可能分散在多个步骤之中，最终产生的影响也可能不会体现在模型的输入或输出通道中。如果模型由使用者自行托管，整个过程中甚至可能没有任何模型服务提供者参与。

2026年5月，国家互联网信息办公室、工业和信息化部及国家发展和改革委员会联合发布《智能体规范应用与创新发展实施意见》。该文件提出建设“智能体注册平台”，记录智能体

的数字身份、能力声明、开发者信息、部署方式、接口协议、合规认证(国家互联网信息办公室等, 2026)。实施意见确立了分类分级治理框架, 但其分级依据是行业和应用场景, 而不是模型能力。因此, 应用于敏感领域和重点行业的智能体需要履行备案义务, 接受额外测试和产品召回管理;用于低风险生活和办公场景的智能体则主要实行自我测试和行业自律(国家互联网信息办公室等, 2026)。与此同时, 中国信息通信研究院、工业和信息化部及TC260 也在分别制定智能体标准(MIIT/TC1, 2026;国家市场监督管理总局与国家标准化管理委员会, 2026;TC260, 2026)。

对于建立在前沿基础模型之上的智能体, 按行业分级会产生一个明显缺口, 因为通用模型并不受限于单一用途。同一个用于"低风险生活场景"智能体的模型, 可以被重新部署到其他任务中、用于超出原定范围的用途, 或者通过越狱和微调遭到利用。行业分级的框架预设了智能体用途的稳定性, 而通用前沿系统并不具备这种稳定性。将能力纳入分级依据, 可以弥补这一缺口。如果智能体所使用模型的能力超过规定门槛, 无论其最初部署于哪个行业, 都应适用实施意见目前仅针对"重点行业"规定的备案、额外测试和产品召回要求。这类测试需要评估完整的智能体系统, 即同时评估模型、工具、自主程度和支架系统。第4.1.1节所述的模型层面评估并不能覆盖这些因素。决定智能体进入较高监管等级的触发因素应当是其超过阈值的能力, 而非应用行业的变更。

实施意见属于不具有约束力的方向性指导文件, 而非正式监管规定。文件也只是提出"探索建立"注册平台, 而不是宣布平台已经投入运行。因此, 在平台的实体制度设计最终确定之前, 仍有机会增加基于能力的分级方式。尚待解决的问题包括:备案义务应由智能体框架开发者、智能体部署者还是最终用户承担;应当如何处理开源智能体框架;如何将 MCP、ACP 或 A2A 等跨境智能体通信协议纳入备案制度的适用范围。

4.2. 应急响应调适

为使中国现有的应急响应体系能够应对前沿人工智能事件, 本节将从两个方面提出建议:一是设置与前沿风险显现方式相适应的分级报告阈值(第4.2.1节);二是增设前沿人工智能专门事件类别及相应的严重程度阈值(第4.2.2节)。这些调整建立在第3.2.1节所述制度架构之上, 包括《突发事件应对法》《国家突发事件总体应急预案》、全国网络安全标准化技术委员会(TC260)发布的应急响应指南, 以及工信部于2026年推出的最新试点方案(见下文)。目前这些制度均未根据前沿人工智能应用场景的特定风险特征进行针对性设计。

如第3.2.1节所述, TC260的生成式人工智能应急响应指南适合处理内容安全事件和服务中断, 但没有涵盖前沿人工智能风险实际显现的情形, 也未考虑风险的连锁扩散以及人工智能系统的自主行为。此外, 前沿人工智能应急响应与内容安全应急响应存在性质上的差异。前沿人工智能造成的损害与内容本身相分离。在内容安全事件中, 损害通常产生于用户接触到相关内容之时;而在前沿人工智能事件中, 损害往往要等到恶意行为者利用相关信息或系统能力采取行动时才会实际发生。前沿人工智能造成的损害也可能更严重且更难缓解。其后果可能上升为实体性或系统性的灾难, 而不只是用户看到不当内容所产生的有限损害。与内容危害相比, 此类风险还可能更加难以察觉。例如, 一个正在推理如何规避监督的模型, 从外部观察时可能仍表现得如同正常运行。最后, 目前尚不存在一套成熟的人工智能应急响应操作方案。现有应急响应制度主要围绕自然灾害和网络安全事件建立。前沿人

工智能事件虽然可能同时具有这两类事件的部分特征，但也很可能提出全新的响应要求。目前尚无任何司法辖区针对这种规模和类型的新兴事件建立完整的应对规程。

综上所述，前沿人工智能应急响应需要采用不同的严重程度衡量指标、风险检测机制和响应模式，也需要比现行生成式人工智能应急响应框架更强的跨机构协调能力。学界提出的人工智能法律草案已将应急响应纳入更广泛的前沿风险治理体系。中国政法大学2024年版《人工智能法》拟授权主管部门对关键人工智能提供者实施紧急关停（第56条）；中国政法大学2025年《关于人工智能立法的重点制度建议》则提出，将关键人工智能提供者建立安全应急响应预案作为完成登记的条件之一（第十四节）。除规定熔断机制外，中国社会科学院《人工智能示范法》第4.0版还要求关键人工智能提供者和拟设立的国家人工智能管理机构分别建立应急响应制度（第42条、第81条）。最近，工业和信息化部于2026年5月启动人工智能科技伦理审查和服务试点计划（新华社，2026），在十个省市建立部、省、市三级敏捷治理网络，并建设全国人工智能伦理风险监测与服务网络，负责向地方政府和创新主体推送风险预警和警报。尽管具体应急响应预案仍由开发者制定，该试点计划仍是中国现有制度中最接近人工智能专项监测、早期预警和警报推送体系的先例。其局限性与本文在其他部分所指出的“制度架构与具体风险校准之间的差距”相同：该计划主要围绕伦理风险设计，而不是围绕前沿能力风险设计。尽管如此，这些制度仍为下文关于事件报告和应急响应的两项建议提供了有参考性的制度先例。

4.2.1. 校准报告阈值

TC260《生成式人工智能服务安全应急响应指南》目前规定的报告阈值为每一万名用户在24小时内报告同类事件达到五次（TC260，2025b，第6.3.2条）。这一标准只有在服务已大规模触达用户之后，才可能将相关情况识别为损害事件。然而，前沿风险最早出现的预警信号往往发生在任何用户受到影响之前。例如，在评估过程中发现模型已跨越危险能力阈值，或者在部署前红队测试中发现模型出现意外的自主行为。此类情况都可能表明模型存在被用于灾难性滥用或导致失控的风险，但两者均不会触发现行报告阈值。因此，前沿人工智能事件的分级报告阈值应当依据事件所揭示的模型风险决定是否升级响应，包括模型的能力范围、自主程度和可能造成的下游损害，而不应仅以受影响用户的数量为标准。TC260可通过直接修订《生成式人工智能服务安全应急响应指南》第6.3.2条实现这一调整。

4.2.2. 设立专门的前沿人工智能事件类别

前沿人工智能事件可能带来超出现有内容类事件范围的威胁，因此有必要设立专门的前沿人工智能事件类别，并为其规定独立的严重程度阈值。TC260《生成式人工智能服务安全应急响应指南》涵盖信息内容安全、数据安全和网络攻击，并根据事件对企业、数据和社会秩序造成的影响确定严重程度（TC260，2025b）。然而，前沿人工智能还可能威胁现实世界中的物理安全，包括显著增强化学、生物、放射性与核能力、破坏关键基础设施、自主获取新的能力以及系统失控。这些风险均未被现有事件类别充分涵盖，并且可能需要另外设置适配的响应升级规程。《国家自然灾害救助应急预案》提供了一种可供参考的模式。该预案根据死亡和失踪人数、需要紧急转移安置的人数以及房屋受损情况划分响应等级（国务院办公厅，2024）。这一方法与美国加利福尼亚州、纽约州和伊利诺伊州采用的做法相似。上述各州将可能导致超过50人死亡或受伤，或者造成超过10亿美元财产损失的风险界定为灾难性风险。设立具有明确严重程度阈值的前沿人工智能专门事件类别，可以确保最严重的风

险情景触发适当级别的响应，同时无须重建现有应急响应体系。TC260可以负责推动此类修订；对于跨越多个行业或领域的事件，则可由应急管理部参与协调。

4.3 责任制度调适

将中国的侵权责任制度适用于前沿人工智能时，会面临第3.2.2节所述的三项结构性问题：如何认定开发者行为与模型行为之间的因果关系，如何在价值链各环节之间分配责任，以及如何应对商业保险无法承保的尾部风险。中国法律制度的两个特点影响着下文提出的调整路径。第一，中国实行成文法制度，司法裁判在形式上不具有约束其他法院的先例效力。第二，最高人民法院发布的司法解释和典型案例具有权威性的补充法律作用，并且通常会得到下级法院遵循。因此，部分调整可以通过最高人民法院的指导意见实现，而不一定需要全国人民代表大会立法。目前，人工智能案件的审理已经开始采用这一方式（Tao，2025）。下文提出的四项调整建议即以这些制度特点为基础。

中国学界提出的三部人工智能法律草案均已超越单纯的过错责任模式，但它们分别采取了不同路径。中国社会科学院《人工智能示范法》第4.0版采取以过错责任为基础、同时区分不同责任主体的方式。开发者需要对若干明确列举的失职行为承担责任，例如训练数据或算法设计不当、安全测试不符合标准，或者明知存在重大隐患却未采取措施。提供者适用举证责任倒置，除非能够证明自身不存在过错，否则即应承担责任（第100条）；开源开发者则适用单独的责任豁免条款（第104条）。中国政法大学《人工智能法》拟建立双层责任制度：大多数人工智能应用拟适用过错责任；关键人工智能则适用举证责任倒置，除非提供者能够证明自身没有过错，否则应当承担责任（第85条）。如果基础模型提供者知道或者应当知道下游主体存在滥用行为，还需与下游主体承担连带责任（第90条）。南京版《人工智能基本法》采取第三种路径，即在整个价值链上按比例分配责任。研究者和开发者、提供者、部署者以及用户，应当根据各自的过错程度、控制能力以及对损害结果的影响程度，“按照其份额”承担民事责任（第56条）。这三部草案由此提出了三种不同机制：举证责任倒置、基础模型提供者的连带责任，以及在价值链各环节之间按比例分配责任。下文的四项调整以这些建议为基础，但与第4.1节和第4.2节提出的方案相比，仍处于更早期的探索阶段。

4.3.1. 连带责任

前沿人工智能造成的损害通常涉及价值链上的多个主体，而现行侵权法理论难以在这些主体之间准确分配责任。严格按照比例分配责任虽然可能更加公平，但大多数法院并不具备实施所需要的技术专业能力。连带责任可提供一种更具可行性的替代方案。在连带责任制度下，每一责任主体均可能被要求对全部损害承担赔偿责任。中国政法大学的草案已经对基础模型提供者采取这一方式：如果提供者“知道或者应当知道”下游主体存在滥用行为，便需与下游行为者承担连带责任（第90条）。为了判断开发者是否应当承担责任，可以将“知道或者应当知道”的标准扩展为“在实施合理评估制度的情况下，知道或者应当知道”。这样便能将第4.1节建议建立的评估体系直接与责任分配联系起来，并明确规定：开发者如果没有进行充分测试便发布具有前沿能力的模型，就不能将由此产生的法律风险转嫁给下游主体。在制度中明确纳入其他相关主体，也能确保前沿人工智能损害涉及多个分散主体时，责任仍可得到适当分配。落实这一方案可以采取两条制度路径：第一，由全国人大通过立法，将中国政法大学草案第90条提出的连带责任模式纳入《民法典》或未来的人工智能法；第二

，由最高人民法院参照其目前正在起草的相关文件，就人工智能案件审理发布指导意见（Tao，2025）。

4.3.2. 政府背书的灾难性风险保险

私人保险机构通常不愿承保尾部风险，因为一次灾难性事件造成的损失就可能超过其赔付能力。这意味着即使侵权责任已经成立，受害者仍可能无法获得实际赔偿。中国政法大学《人工智能法》鼓励人工智能开发者购买保险，并鼓励保险机构开发人工智能专属的新型保险产品（第26条）。中国政法大学2025年《关于人工智能立法的重点制度建议》还提出设立风险补偿基金（第十二节）。政府再保险机制的建立可参考美国2001年“9·11”事件后设立的恐怖主义风险保险资金池（Federal Insurance Office，2024），为超过特定金额门槛的损失提供兜底保障，同时保留私人保险市场承保一般风险的动力。该方案可以与第4.2节提出的应急响应体系直接衔接。凡触发最高等级应急响应的事件，也应同时触发政府兜底保障。这样可以确保在最可能超出私人保险机构赔付能力的风险情景中，受害者仍然能够获得赔偿。落实这一机制需要全国人大通过立法设立风险补偿基金，具体可参考中国政法大学课题组2025年建议的第十二节，并由财政部和国家金融监督管理总局共同管理。此类机制适合先在试点地区进行实验，再逐步推广至全国。

4.3.3. 调整证明标准

在以过错责任为基础的诉讼中，原告必须证明被告存在过错、被告的过错与损害之间具有因果关系，并且损害已经发生。面对运行机制不透明且能力快速变化的模型，信息不对称往往使原告在证明过错和因果关系这两个环节便遭遇困难，从而削弱侵权责任制度的威慑作用。可以从三个方面调整证明标准，分别减轻原告在不同环节承担的证明负担。

第一项调整涉及过错证明。现有法律草案已经对部分主体作出类似安排。中国社会科学院草案第4.0版规定，除非提供者能够证明自身不存在过错，否则应当承担责任（第100条）；中国政法大学草案也对“关键人工智能”提供者实行举证责任倒置（第85条第2款）。相比之下，开发者并不适用这种举证责任倒置。原告必须证明开发者存在某项具体列明的失职行为，例如训练数据或算法设计不当、安全测试不符合标准，或者明知存在重大隐患却未采取措施（第100条）。

第二项调整涉及证据获取。证明上述失职行为所需的材料，包括训练数据、评估日志和红队测试结果，通常全部掌握在开发者手中。可以参照修订后的欧盟《产品责任指令》，建立证据披露制度。当原告提出具有合理可信度的初步证据时，法院可以命令开发者提交相关材料（欧盟指令2024/2853，第9条）。如果开发者拒绝披露，则可以推定相关产品存在缺陷，但开发者有权提出反证推翻该推定（第10条）。

第三项调整涉及因果关系，因为证据披露本身仍不足以解决这一问题。即使获得相关记录，原告通常也无法追踪某项具体开发决策如何导致模型产生特定输出。可以在针对开发者的责任认定中引入因果关系推定，以填补这一缺口。一旦原告证明开发者存在法律明确列举的失职行为，并且发生了相关义务本应防止的那一类损害，法院即可推定该失职行为导致了损害，除非开发者提出反证。这一做法是对中国现有国内机制的扩展，而不是从国外

移植一项全新的制度。中国侵权法已经在特定情形下将因果关系的举证责任转移给被告，例如环境污染和生态破坏责任领域的《民法典》第1230条。

综合而言，这三项调整分别减轻原告在证明过错、获取证据和证明因果关系方面承担的负担。最高人民法院可以通过程序性指导文件落实证据披露规则；而过错举证责任倒置和因果关系推定会改变责任成立的认定方式，因此需要由全国人大参照中国社会科学院或中国政法大学的草案进行立法。

4.3.4. 面向政府的保密披露

面向公众的透明度要求有时可能导致企业披露的信息反而更少。面对公众监督的开发者可能为了避免声誉受损和后续诉讼而选择少报或模糊披露。批评者曾针对加利福尼亚州第53号参议院法案（SB-53）要求公开发布前沿人工智能框架的规定提出这一担忧。虽然Anthropic继续在发布《负责任扩展政策》（Responsible Scaling Policy）的同时，也公布SB-53要求的《前沿人工智能合规框架》（Frontier Compliance Framework，FCF），但该公司也指出，FCF“未必始终符合不断发展的最佳实践”（Anthropic，2025；2026）。中国的监管传统本身更倾向于要求企业向政府进行保密披露，而不是向公众披露。国家互联网信息办公室的算法备案制度、科技伦理审查制度，以及人工智能产业联盟（AIIA）的人工智能安全治理承诺，均要求相关主体向监管机构而非公众提交信息。[47]按照这一思路设计前沿人工智能信息披露要求，既可避免强制公开导致披露减少的问题，又能确保监管机构获得开展模型评估、事件响应和责任分配所需的信息。国家互联网信息办公室可以通过扩充现有算法备案表的填报内容，直接落实这一方案；同时还可由TC260制定有关前沿人工智能披露内容的配套标准。

5. 结论

本文考察了中国以外三个司法辖区治理前沿人工智能的不同路径（第二章），梳理了中国针对人工智能建立的较为完备的监管体系（第三章），并探讨了如何根据前沿系统特有的风险特征对现有体系作出调整（第四章）。贯穿全文的一个核心观点是：中国不需要从零开始。现有的算法备案制度、GB/T 45654—2025确立的风险评估制度、《突发事件应对法》、人工智能法学者建议稿、科技伦理审查体系，以及2026年5月发布的智能体实施意见等，已经奠定了坚实基础。目前需要补充的，是根据系统的能力水平，而不是内容类别或应用场景，对这些制度进行有针对性的调整。第四章讨论的三个领域，都是对现有基础设施的扩展，而不是另行建立与之竞争的平行制度。

本文本着互学互鉴的精神提出上述考量。目前尚无任何司法辖区彻底解决前沿人工智能治理问题。欧盟《人工智能法》有关通用人工智能模型（general-purpose AI，GPAI）的条款和美国各州的透明度法律，都面临实施层面的问题，也都存在已经得到承认的制度缺口；中国的人工智能法学者建议稿本身也仍在不断完善。本文之所以参考多种不同路径，正是因为这些差异本身为共同分析提供了重要素材。与此同时，各方在前沿风险分析上已呈现一定程度的趋同。这一点可以从TC260有关失控风险的表述、中国社会科学院《人工智能示范法》

[47] 人工智能产业联盟仅会披露有关签署方实践的有限匿名信息（人工智能产业发展联盟，2025）。

第4.0版提出的基础模型风险支柱、中国政法大学草案中的关键人工智能条款，以及上海人工智能实验室的SafeWork框架中看出。这说明，中国与国际社会对相关问题的理解和表述之间，差距可能并不像有时认为的那样大。本文希望为中国学者已经展开的讨论提供一份参考。

对于尚不确定前沿风险是否会在部分预测者提出的时间范围内成为现实的读者，第四章提出的方案仍值得作为前瞻性预案加以考虑。无论近期是否会发生某场特定的地震或疫情，自然灾害应急响应体系都需要持续维持。同样，正是在前沿能力的发展仍存在不确定性时，第4.1节提出的评估制度扩展和第4.2节提出的应急响应调整才最为必要，因为这些机制能够提高风险的可见度，并建立必要的响应能力，使监管体系能够在前沿能力发展速度快于预期时及时采取行动。如果相关能力的发展较为缓慢，维持这些机制的成本也相对有限。第4.3节提出的责任制度调整和第4.3.4节提出的信息披露方案则更具探索性。这些机制旨在较长时期内影响相关主体的行为激励，同时依赖于目前尚无定论的法律制度选择。

第四章提出的部分机制，可以先在中国现有的试验区内测试，再向全国推广。中国政法大学2025年发布的建议已经提出，可以利用全国人大授权试点的机制，对人工智能专项制度设计开展实验（AI善治学术工作组，2026a）。中国长期以来也具有通过试点推进监管创新的传统（Heilmann，2008）。例如，2009年启动的跨境人民币结算试点和2013年启动的碳排放权交易试点，均先在部分城市实施，随后再逐步扩大至全国（Cui等，2021；L. Li与Xu，2022）。这一传统有利于在范围有限、风险可控的环境中检验制度效果，再决定是否在全国推行。责任框架、认证制度和保险兜底机制尤其适合采用这种路径。这些制度如果能在被纳入全国性立法之前积累司法案例和实际运行经验，其设计通常会更加成熟。试验区还可以在不要求整个制度体系立即接受某一种特定方案的情况下，为相关经验的积累提供空间。

随着前沿人工智能领域不断发展，还有若干值得深入研究的机制未在本文中展开讨论。中国目前通过算法备案程序，已经对面向公众提供的生成式人工智能服务实行了一种近似许可的管理制度。未来可以在此基础上，将许可制度扩展为针对前沿系统的实质性事前审查，并参考网络安全等级保护制度2.0中的民营网络安全认证模式（Protiviti，2022）。市场驱动型监管以及更广义的监管市场也值得进一步研究。在此类制度中，私人监管服务机构通过竞争为开发者提供认证，通过认证的开发者则可以获得侵权责任安全港。随着人工智能评估方法不断扩大应用范围，公共监管机构可能面临日益突出的能力不足问题，而这种模式可以在一定程度上弥补相关缺口。科技伦理审查制度的具体落地、训练开始前的监督机制，以及备案制度之外的智能体专项治理，包括第4.1.3节提出的跨境协议协调问题，也都值得进行比本文更深入的研究。其中若干机制同样适合先在试验区开展探索。

本文的根本主张是，中国现有的监管基础设施是应对前沿风险的一项重要制度资产。标准制定机构、算法备案制度、人工智能法学者建议稿、科技伦理审查体系、应急响应制度和试验区传统，并非彼此孤立的监管安排，而是建立前沿人工智能专项治理框架所需的制度基础。是否建立这一框架以及具体如何建立，应由中国的监管机构和学者决定。本文提出的各项方案，旨在为一场已经展开的讨论提供参考。

致谢

我们谨向Nick Caputo、Alan Chan、方亮与Gabriel Wagner的贡献表达感谢。

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