财经法学 No. 5，2024
pp. 3 - 22

人工智能基础模型安全风险的平台治理

周　辉*

内容提要：人工智能基础模型的安全治理是人工智能法治化发展面临的重要命题。人工智能基础模型平台作为安全治理的主体，其发现、评估和缓解人工智能系统潜在风险的能力至关重要，具有调整和优化自身模型的作用。国内外人工智能基础模型平台积极布局安全风险治理，但仍然面临治理架构缺乏权威性和代表性、伦理规范过于抽象、测试算力不足、风险评估困难、平台存在利己偏好等挑战。有必要立足于中国人工智能基础模型平台安全治理的实际，完善压力驱动、动力保障、能力强化等机制，更好发挥人工智能基础模型平台对安全治理的特有优势和能动作用，支撑人工智能技术及其应用健康有序发展。

关键词：基础模型　基础模型平台　人工智能安全　人工智能风险　人工智能治理

一、问题的提出

人工智能是引领新一轮科技革命和产业变革的战略性技术。在以生成式人工智能为代表的新一代智能快速发展过程中，人工智能基础模型（以下简称“基础模型”）的能力也在快速提升。基于其巨量数据训练和自我监督的优势，基础模型可以适应和完成广泛的下游任务。作为新一代基础设施平台，基础模型将在推动社会进步、产业升级、科技创新等方面发挥重要作用，也将带来新的安全风险，并将一些风险进一步通过下游应用向社会不断传导。

针对基础模型，主要国家和地区在促进其创新发展的同时，也在积极探索安全风险的法律治理框架。早在 2018 年，习近平总书记主持十九届中央政治局第九次集体学习时就指出，要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可

* 周辉，中国社会科学院法学研究所副研究员。
本文获中国社会科学院学科建设“登峰战略”资助计划资助（DF2023XXJC07）。

靠、可控。〔1〕中国 2023 年 7 月出台的《生成式人工智能服务管理暂行办法》是世界范围出台的首部针对生成式人工智能综合治理的部门规章，明确要求在模型生成和优化的过程中防止歧视产生，服务提供者要使用具有合法来源的数据和基础模型。美国 2023 年 10 月在组织两轮由主要基础模型平台企业参与的自律承诺的基础上，出台了人工智能行政命令，围绕两用基础模型〔2〕设计了自我治理为主、配以行政指导和事后监管的治理框架。欧盟《人工智能法案》在 2023 年底的修改中，专门增加了针对通用人工智能系统的特别要求，明确其必须遵守保证模型透明度、进行模型评估和确保网络安全等基本原则。

基础模型发展速度之快和平台生态内治理之复杂，对法律治理的敏捷性、操作性、可预期性提出了新的更高要求。作为直接面对和实际了解基础模型安全风险的主体，基础模型平台主动建立有效的安全治理机制，有利于更好地及时实现安全治理的目标。目前，国内外基础模型平台已经通过对抗测试、评估审计等方式，积极探索平台内的安全风险治理，但仍存在伦理治理、技术治理、自我治理等各种局限。从理论和制度上破解这些困境，更好释放基础模型平台治理的潜力和动力，既是对法律治理的重要补充，也是法律制度建设的重要导向。

因此，本文将以把握基础模型安全风险的特点和相关平台的治理主体定位为起点，深入研究国内外代表性基础模型平台的安全治理实践，在总结其机制内容的基础上，分析制约其安全风险治理效能发挥的理论困境，进而让其有压力、有动力、有能力实施好自我治理制度设计，以期为基础模型安全风险的平台自我治理提供系统性方案和相应的法律制度设计参考。

二、基础模型安全风险的技术逻辑与平台治理的主体定位

2018 年，BERT 和 GPT-2 等基础模型问世，其采用 Transformer 架构并通过自监督预训练显著提升了处理多种任务的能力。〔3〕进入 2020 年以后，基础模型进一步向跨模态和多模态演进，任务处理能力进一步提升。〔4〕基础模型可以被理解为一种在大量原始数据上预训练的深度学习模型，可针对特定任务进行调整（适配），能够适应和完成广泛的下游任务。基础模型已经在自然语言处理（NLP）和计算机视觉领域显示出强大的效能，因此备受关注。

在预训练阶段，基础模型通常在大规模的数据分布上进行训练，使用预训练损失函数来衡量

〔1〕 参见习近平：《论科技自立自强》，中央文献出版社 2023 年版，第 215 页。

〔2〕 根据该行政命令的定义，两用基础模型是指根据广泛数据进行训练的人工智能基础模型，一般实行自我监督，包含至少数百亿个参数，适用于广泛的环境，并且在对安全、国家经济安全、国家公共卫生或安全，或这些问题的任何组合造成严重风险的任务中具有高水平的表现，或可以经过很简单的修改后具有高水平的表现。See Whitehouse, *Executive Order on the Safe, Secure and Trustworthy Development and Use of Artificial Intelligence*, available at https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/, last visited on May 10, 2024.

〔3〕 参见李舟军、范宇、吴贤杰：《面向自然语言处理的预训练技术研究综述》，载《计算机科学》2020 年第 3 期；Ashish Vaswani, Noam Shazeer, Niki Parmar, et al., *Attention Is All You Need*, in Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, et al. eds., Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Curran Associates Inc., 2017, pp. 6000 - 6010.

〔4〕 参见赵朝阳、朱贵波、王金桥：《ChatGPT 给语言大模型带来的启示和多模态大模型新的发展思路》，载《数据分析与知识发现》2023 年第 3 期。

模型在输入数据上的表现。这个阶段的目标是最小化预训练损失，生成一个能捕获广泛数据分布特征的模型。

基础模型的适配阶段，则是在特定下游任务的数据上调整预训练模型，通过优化特定的适配损失函数来实现。不同的适配方法可能会调整模型参数的不同子集，通过微调（fine-tuning）或提示调整（prompt-tuning）等技术，形成服务于专业领域的专用模型（task-specific models）来适用于准确率、专业度要求较高的场景。〔5〕

（一）基础模型的安全风险

据斯坦福大学统计，自 2023 年 9 月至 2024 年 4 月就有 120 多个新的基础模型面世，使已知的全球模型总量超过 330 个。〔6〕基础模型已在多个领域中展现了卓越的性能，包括但不限于语言理解、图像识别、自然语言生成等，成为企业和研究者开发高效、智能应用的基石。基础模型的构建和训练依赖于复杂的机器学习算法及大量数据集。然而，这些基础技术涉及多重潜在的风险因素，包括技术层面的风险、模型之间的相互关联性带来的风险以及模型泛化能力可能引起的风险传播问题。这些风险可能影响基础模型的安全性和可靠性。

第一，风险发现的困难性。技术风险包括数据偏差和不公平，如果用来训练模型的数据集存在偏见或者不平衡的问题，那么模型很可能会学到并且放大这些偏见，从而导致不公平的结果。例如，如果一个用于评估贷款申请的模型是基于历史数据训练出来的，而这些历史数据中包含了对某些群体的偏见，那么这个模型很可能会对所涉群体的贷款申请作出不公平的判定。此外，基础模型由于存在大量不可解释的内部参数，往往被认为是“黑盒”，其决策过程难以理解和解释。〔7〕这限制了其他主体对模型的理解，增加了误用和滥用的风险，同时也使得监管和审计变得更加困难。

第二，风险的易发性。基础模型面临的攻击面极其广泛，传统的网络安全攻击如网络钓鱼、供应链攻击、零日漏洞等，以及特定于基础模型的攻击如模型投毒、模型反转、对抗性攻击等都可以对基础模型造成较大威胁。攻击者可以通过微小的、经过精心设计的输入变化来欺骗模型，导致错误的输出。〔8〕这类安全问题在敏感应用中尤其危险，如自动驾驶车辆的安全系统。

第三，风险的关联性。基础模型的一个核心特性是它们能够在多个任务和领域中通用。然而，这种关联性也带来了风险。一是跨域误用，即一个为特定任务训练和优化的模型可能在其他领域产生不可预见的、有时是有害的结果。例如，原本设计用于回答日常生活问题的通用对话模型，如果被用于提供专业的医疗建议，可能会导致用户在自身健康问题上作出错误判断。二是知识泄露，即基于大规模数据集训练的模型可能无意中记忆并在后续与用户的交互中泄露敏感个人

〔5〕 See Yao Fu, Hao Peng, Litu Ou, et al., *Specializing Smaller Language Models Towards Multi-step Reasoning*, in Andreas Krause, Emma Brunskill, Kyunghyun Cho, et al. eds., Proceedings of the 40th International Conference on Machine Learning (ICML'23), JMLR.org, 2023, pp. 10421 - 10430.

〔6〕 See *AI Foundation Models Update Paper*, available at https://assets.publishing.service.gov.uk/media/661941a6c1d297c6ad1dfeed/Update_Paper_1_.pdf, last visited on May 10, 2024.

〔7〕 参见周辉：《算法权力及其规制》，载《法制与社会发展》2019 年第 6 期。

〔8〕 See Sella Nevo, et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, available at https://www.rand.org/pubs/research_reports/RRA2849-1.html, last visited on May 10, 2024.

信息或重要数据。

第四，泛化能力过强可能导致不当内容的输出。基础模型在未见过的数据或任务上表现出的性能，虽然是其强大的优势，但也存在风险。模型可能在某些情况下过度泛化，忽略细微差别，导致不准确或不适当的输出。例如，一个文本生成模型可能在生成涉及少数群体的内容时，使用刻板印象或过度泛化的语言。在某些情况下，模型可能需要在泛化能力和针对特定任务的优化之间取得平衡。过度优化可能导致模型在特定任务上表现出色，但在广泛应用时性能下降。〔9〕

由于基础模型强大的预训练模型能力，能够在多个领域内产生深远影响。这种能力也带来了误导性内容、隐私侵犯和歧视性偏见的风险。因此，需要对其进行监管以确保这些技术的发展和应用符合道德和法律标准，减少潜在的风险和危害。

（二）作为治理主体的基础模型平台

基础模型平台，是指将基础模型进行封装和优化，给广大终端用户、基础模型开发者、下游应用部署者提供使用方便的数字基础设施。〔10〕其应用消除了专业与通用的壁垒，极大降低了用户的使用门槛，使人工智能在真正意义上进入了普通人的生产和社会生活。〔11〕基础模型平台通过提供强大的预训练模型已成为推动人工智能技术发展的重要力量，被认为是“数字技术市场的‘看门人’”〔12〕。相较于政府的传统监管角色，基础模型平台在安全治理中展现出独特的重要性。基础模型的数据训练、结果生成、应用开发等直接影响了相关主体的权利义务分配，是需要独立监管的对象。〔13〕由于技术和应用场景不断进步和拓展，原先的监管重点和方式可能已不完全契合新的风险点。基础模型平台作为基础模型的开发者和提供者，也是基础模型应用生态体系的连接者和组织者，应作为安全治理的主体，减少基础模型应用潜在的风险和危害，保护公众利益，确保人工智能技术的健康发展和应用。

1. 基础模型平台作为安全治理主体的重要性

基础模型可能会存在知识幻觉、价值偏见、噪声污染等缺陷，带来包括网络安全、信息内容安全、算法安全、数据安全、知识产权保护等方面的风险。〔14〕此外，基础模型是一种以通用性

---

〔9〕 See European Commission, *Explanatory Memorandum*, available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN, last visited on May 10, 2024.

〔10〕 相较于一般应用软件平台或互联网服务平台，基础模型平台的服务对象和业务范围都更加广泛和开放，其可以将人、信息、服务、算法模型整合在一起，功能包括：为模型开发者、终端用户等提供使用和创新人工智能的工具；将海量终端用户与人工智能模型连接起来，让普通用户通过自然语言等交互方式便捷使用智能服务；连接不同的人工智能模型，实现多模态智能的融合应用等等。基础模型平台为人机协同、智能应用开发、跨界创新提供了广阔空间，其内涵已经超越了传统意义上的电商平台和信息内容平台。

〔11〕 参见朱嘉珺：《生成式人工智能虚假有害信息规制的挑战与应对——以 ChatGPT 的应用为引》，载《比较法研究》2023 年第 5 期。

〔12〕 於兴中、郑戈、丁晓东：《生成式人工智能与法律的六大议题：以 ChatGPT 为例》，载《中国法律评论》2023 年第 2 期，第 2 页。

〔13〕 参见张凌寒：《生成式人工智能的法律定位与分层治理》，载《现代法学》2023 年第 4 期。

〔14〕 See Rishi Bommasani, et al., *On the Opportunities and Risks of Foundation Models*, available at https://doi.org/10.48550/arXiv.2108.07258, last visited on May 10, 2024；刘艳红：《生成式人工智能的三大安全风险及法律规制——以 ChatGPT 为例》，载《东方法学》2023 年第 4 期；苏宇：《大型语言模型的法律风险与治理路径》，载《法律科学（西北政法大学学报）》2024 年第 1 期；郭春镇：《生成式 AI 的融贯性法律治理 ——以生成式预训练模型（GPT）为例》，载《现代法学》2023 年第 3 期等。

和多功能性为特征的算法模型，其所具有的缺陷和风险也会被所有下游应用所继承。〔15〕作为为广大用户提供基础模型服务和构建基础模型使用生态的主体，基础模型平台可以主动构建“模型即服务”的产业链条，〔16〕具备强大的技术能力，对外输出自然语言交互服务和下游应用服务，其运行的安全性和规范性也关乎众多参与主体的切身利益。这意味着，作为重要的治理主体，基础模型平台要实施有效的自我治理，也要接受政府监管、行业自律、社会监督、用户共治，从而引导基础模型规范发展。

基础模型已成为推动人工智能技术进步的关键因素，在图像识别、自然语言处理、语音识别等多个领域均展现出卓越的性能。这类模型通过在大规模数据集上进行预训练，掌握了丰富的、高层次的数据标识，从而具备了跨任务和跨领域的泛化能力。基础模型的核心优势在于其强大的数据效率和广泛的适用性，它们能够在多种不同的下游任务上实现快速适应和微调，极大地提高了模型的实用性和灵活性。但是，具有强大实际应用能力的基础模型在实际部署过程中可能存在许多风险，包括技术层面的风险、模型之间的相互关联性带来的风险等。基础模型的风险会带来数据歧视或偏见、数据隐私的安全性问题以及算力资源的消耗问题。在实现基础模型平台的安全治理过程中，基础模型是应重点把握的对象。

基础模型平台通过提供预训练的深度学习模型、制定行业标准、促进跨界合作以及提高公众意识等方式，确保了人工智能技术的安全、可靠和伦理发展。强化基础模型平台在安全治理中的主体作用，对于应对人工智能技术快速发展带来的挑战至关重要。

第一，基础模型平台对于发现、评估和缓解人工智能系统潜在风险至关重要。这些平台由于掌握着模型的设计、训练和部署的核心技术且直接参与模型设计和应用部署，能够从源头上识别并解决可能导致安全问题的技术缺陷。通过内部的技术审查和风险评估，平台能够有效地预防或降低因模型偏见、数据隐私泄露、对抗性攻击等问题引起的风险。〔17〕

第二，基础模型平台在制定与实施安全治理标准的实践中扮演着领导角色。它们不仅能够为人工智能领域的安全治理制定技术标准，还能通过行业内的合作和共识，推动这些标准的广泛采纳。此外，基础模型平台通过公开透明的方式分享安全最佳实践，为整个行业的安全治理提供指导和参考。

第三，基础模型平台在促进基础模型的合作和沟通中处于主导地位。基础模型平台拥有技术资源和专业知识，具备跨领域整合能力，能够有效地联结政府机构、学术界、行业协会以及公众等多方利益相关者。通过组织高层次技术论坛、发起跨界研究项目、建立开放的数据共享机制等多种方式，基础模型平台能够促进多方深入对话，进而合作应对人工智能发展带来的复杂挑战。

第四，基础模型平台在提升公众对人工智能安全的认识和理解方面发挥着重要作用。通过教

〔15〕 参见张璐：《通用人工智能风险治理与监管初探——ChatGPT 引发的问题与挑战》，载《电子政务》2023 年第 9 期。

〔16〕 参见张欣：《面向产业链的治理：人工智能生成内容的技术机理与治理逻辑》，载《行政法学研究》2023 年第 6 期。

〔17〕 See AI Now Institute, Amba Kak & Sarah Myers West, *Five Considerations to Guide the Regulation of “General Purpose AI” in the EU's AI Act*, available at https://ainowinstitute.org/wp-content/uploads/2023/04/GPAI-Policy-Brief.pdf, last visited on May 10, 2024.

育和宣传活动，平台能够帮助公众了解人工智能技术的潜在风险以及如何有效防范这些风险。〔18〕这不仅增强了公众对人工智能安全的意识，也为构建一个安全、可信的人工智能应用环境奠定了基础。

2. 基础模型平台作为安全治理主体的优势

与权威治理主体相比，基础模型平台在安全治理中扮演着更加多元化的角色。他们不仅是基础模型的提供者，而且是基础模型的治理者、组织者和参与者，能够自主进行规则制定，协调多方主体利益以及参与公共政策和标准的制定，承担法律和社会责任。基础模型平台与政府在安全治理中应形成互补，共同推动人工智能技术的健康发展和安全应用。

第一，基础模型平台有技术先知及快速响应的优势。基础模型平台作为技术的直接开发者和应用者，对自身产品的性能和潜在风险有着深刻的理解。与监管部门相比，基础模型平台能够更快速地识别技术漏洞，响应安全威胁，实现对风险的即时管理和控制。

第二，基础模型平台能够进行内部治理与自我调整。基础模型平台拥有调整和优化自身模型的能力，可以通过内部治理机制，如改进数据集、优化算法、增强模型安全性等措施，有效减少偏见、提高透明度和可解释性。这种自我调整能力使得平台能够在问题发生初期就进行干预，减轻可能的负面影响。

第三，基础模型平台可以从技术应用的上游实现风险的有效控制。基础模型平台开发和部署各类智能应用，处于基础模型产业链条的上游。由于基础模型平台对下游应用具有广泛而深远的影响，其在源头采取的安全治理措施对预防各类风险具有关键作用。〔19〕

三、基础模型平台安全治理的实践机制

在基础模型风险逐步显现的情况下，基础模型平台正致力于构建多维度的安全治理机制，以应对日益复杂的技术和伦理挑战，包括建立治理架构、明确伦理规范、进行对抗测试和开展评估审计等，旨在保障基础模型平台的安全性、可靠性和合规性。

（一）建立治理架构

基础模型平台的治理架构是一套全面的管理框架，它通过规则、政策和流程来引导人工智能技术的研发、应用和部署。如 OpenAI 发布内部自治性 Model Spec，通过设定目标、规则和默认值来推动基础模型平台的自我管理，确保基础模型行为符合 OpenAI 的道德和安全标准。它作为平台级别的权威性文件，为模型提供了一个行为准则的框架。〔20〕

---

〔18〕 See Isaac Triguero, Daniel Molina, Javier Poyatos, et al., *General Purpose Artificial Intelligence Systems (Gpais): Properties, Definition, Taxonomy, Societal Implications and Responsible Governance*, available at https://arxiv.org/abs/2307.14283, last visited on May 10, 2024.

〔19〕 See Liang Chen, Tony Tong, Shaoqin Tang & Nianchen Han, *Governance and Design of Digital Platforms: A Review and Future Research Directions on a Meta-Organization*, 48 Journal of Management 147 (2022).

〔20〕 See OpenAI, *Introducing the Model Spec*, available at https://openai.com/index/introducing-the-model-spec, last visited on May 10, 2024.

基础模型平台的治理架构致力于提升模型的安全性和可靠性，确保技术进步遵循伦理和社会责任，同时增强透明度并建立问责制度，有效管理人工智能相关风险。当前基础模型平台针对基础模型往往采取内外部相结合的治理机制，在内部构建专门负责模型安全的部门，在外部借助第三方技术或伦理委员会进行评估或邀请外部专家参与模型的评估。

内部治理模式指的是平台依靠其内部的管理团队、法律团队、技术团队等部门来负责制定和执行治理策略和规则。管理团队负责制定平台的整体战略和治理政策，确保治理结构符合平台的长期目标。法律团队确保平台的运营和治理结构遵守当地和国际的法律法规，处理与法律相关的各种问题。技术团队负责实施技术层面的安全措施，如数据加密、访问控制等，以及开发和部署安全相关的技术解决方案。

在这种模式下，决策过程、安全策略、技术标准的制定和执行大多在组织内部完成，虽然可能会听取外部意见，但核心治理结构和决策权主要集中在平台自身。例如，Google 拥有庞大的内部治理架构，包括安全工程团队、隐私法律团队和伦理委员会等，这些团队负责监管公司所有产品和服务的安全性和合规性。Google 已经成立了负责任的人工智能委员会，由高级管理人员代表组成，以评估模型开发和部署中的新型高风险问题，并将新模型的评估和质量保证整合到现有的信任、安全和企业风险管理协议中。〔21〕 Meta 则通过内部安全团队、工程团队和内容管理团队来处理用户隐私、数据安全和内容监管等问题，并且设立独立的监督委员会负责监督内部治理。

除了平台内部治理架构的完善，吸收外部、独立专家参与治理，也是基础模型平台安全治理架构的重要实践。例如百度的飞桨（Paddle）平台吸纳了多方面的独立专家参与到平台的治理过程中，他们通过开放的技术委员会和伦理委员会参与治理过程的评估，以确保技术的发展既能够符合伦理标准，也能够响应市场和社会的需求。OpenAI 除了依靠内部的治理机制外，还与外部的伦理和安全研究组织合作，包括联系外部专家帮助探测系统映射和评估风险，以及在对模型进行测试时定期进行模型安全性的评估。通过这种合作，OpenAI 能够从独立第三方获得关于其基础模型潜在风险的客观评价，并据此调整其治理策略。

（二）明确伦理规范

人工智能的伦理规范是引导人工智能技术研发与应用的原则性要求，强调在模型层面让人工智能理解人类的价值和伦理原则，确保其在促进社会福祉的同时减少负面影响。伦理规范关注透明度、公平性、责任性、隐私保护及安全性等方面内容，旨在平衡创新与保护的需要，增进公众对人工智能的信任和接纳，并为技术开发与运营各方提供负责任行动的指导框架。〔22〕 人工智能伦理规范内涵包含三方面：一是人类在开发和使用人工智能相关技术、产品及系统时的道德准则及行为规范，二是人工智能体本身所具有的符合伦理准则的道德编程或价值嵌入方法，三是人工

〔21〕 See Royal Hansen & Phil Venables, *Introducing Google's Secure AI Framework*, available at https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/, last visited on May 10, 2024.

〔22〕 See Anna Jobin, Marcello Ienca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 Nature Machine Intelligence 389 (2019).

智能体通过自我学习推理而形成的伦理规范。〔23〕

当前一些基础模型平台在用户服务条款内容中提及伦理规范原则或出台专门的伦理规范原则。多数平台在服务网页明显位置上提及一些伦理原则，例如以人为本、负责任的开发人工智能等。例如，谷歌 Deepmind 出台与人工智能伦理相关的专门性平台原则，以明确其在人工智能研发和应用中的伦理标准，进而评估人工智能系统，应对可能带来的风险。这些原则包括不会设计或部署违反人权的人工智能技术、坚持对社会有益、避免制造不公平的偏见、坚持以人为本、构建隐私设计等。为实施这些原则，谷歌 Deepmind 也建立责任与安全委员会（RSC）进行机制化的内部监督。OpenAI 通过专门的伦理原则和用户协议中的相关条款体现了其对人工智能伦理规范的承诺。这些独立于用户协议的伦理原则强调了技术的安全性、透明度、公平性和合作性，旨在促进人工智能技术的负责任使用和全球性的合作共享。OpenAI 强调不制造有害或歧视性的偏见以确保人工智能系统的安全性，并推动技术的公平与普遍利益。此外，OpenAI 的用户协议中明确了对这些伦理规范的法律承诺，确保用户在遵守这些原则的基础上使用其技术。Anthropic 用户协议涵盖了伦理使用的条款，确保其技术的使用者能在一个明确的伦理框架下操作。

（三）进行对抗测试

对抗测试（red teaming）是一种主动安全审计和风险评估手段，在安全治理中模拟攻击者的视角和行为，以发现和修复安全漏洞。这种方法起源于军事领域，用于测试策略的稳固性。〔24〕对抗测试在人工智能领域的应用展示了技术平台对自身安全策略的自律以及行业和政府组织对这些活动的支持和监管。当前基础模型平台会利用对抗测试对模型的各种性能进行测试，而且不同周期过程都会进行对抗测试，评估模型的危险能力或风险程度以采取不同等级的安全管控措施防范基础模型带来的风险。〔25〕不同平台，例如 Meta、Anthropic、Google、Hugging Face、Stability 和 OpenAI 等参与了如 DEFCON 这样的公开红基础模型平台队测试活动，这不仅体现了基础模型平台各自对模型安全性进行自我检测与提升的承诺，而且也反映了行业内对协作与透明度的重视。

美国国防部在其红队操作手册中强调，有效的对抗测试不仅需要技术专家的参与，还需法规与政策的支持，确保测试的全面性和深入性。这一点在人工智能领域尤为重要，因为人工智能系统的复杂性和潜在的社会影响要求更高层次的安全保障。国防部的文件还提到，红队测试应包括技术、操作和战略层面的全方位评估，这对人工智能平台来说意味着不仅要测试技术的鲁棒性，还要考虑伦理和社会责任。

从平台自律的角度看，像 OpenAI 和 Anthropic 这样的公司设立内部红队，定期进行安全性和伦理性的评估，显示出企业内部对持续改进和安全保障的重视。这些公司通常还会发布测试结果和改进措施，以增强外部利益相关者的信任。例如，Meta 进行红队测试旨在确保模型（Llama

〔23〕 参见国家人工智能标准化总体组、全国信标委人工智能分委会：《人工智能伦理治理标准化指南（2023 版）》，载 https://www.aipubservice.com/airesource/fs/人工智能伦理治理标准化指南.pdf，最后访问时间：2024 年 5 月 10 日。

〔24〕 See Deep Ganguli, Liane Lovitt, et al., *Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned*, available at https://doi.org/10.48550/arXiv.2209.07858, last visited on May 10, 2024.

〔25〕 See Meta, *Llama 2: Open Foundation and Fine-Tuned Chat Models*, available at https://lfs.aminer.cn/misc/Llama%202%20at%20KDD%20LLM%20Final.pdf, last visited on Mar. 2, 2024.

2、Llama 3）的安全性和可靠性。这项测试通过与内部员工、合同工和外部供应商的各种团体进行合作，深入探测模型可能面临的广泛风险类别和攻击媒介。

行业层面，行业协会或职业团体组织的红队测试活动也日益增多，这些活动不仅推动了标准的制定和最佳实践的分享，还促进了跨公司间的学习和协作。例如，在由全球人工智能安全联盟（GAISA）发起的年度人工智能安全大会上，成员公司如 Google、Microsoft 和 Amazon 共同参与红队测试。这种合作不仅促进了技术的跨公司评估，也推动了关于数据保护和隐私安全的最佳实践的共享。通过这些活动，各公司能够相互学习和改进，增强了整个行业的安全性和透明度。此外，技术论坛和研讨会也常常设有红队测试的专题讨论，这促进了从技术人员到管理层的全面参与和意识提升。

政府组织则通过制定相关政策和标准，监管这些测试活动，确保它们的执行既符合伦理标准，也能有效提升国内人工智能技术的整体安全水平。政府的介入有助于确保这些测试不仅限于自愿性质，还能成为行业标准的一部分，从而有利于提升整个行业的安全防护水平。例如，美国国家标准与技术研究院（NIST）制定了一系列关于人工智能安全测试和评估的指导原则。这些指导原则为政府部门和私营部门实施红队测试提供了框架和标准，确保测试的严谨性和全面性。政府还通过赞助研究项目和组织挑战赛来推动对抗测试技术的发展。如美国国防部国防高级研究计划局（DARPA）举办的 Cyber Grand Challenge，就是一个全自动的系统安全比赛，旨在通过对抗测试推动自动化防御系统的发展。这些政府举措不仅提升了技术的安全标准，也为人工智能的伦理使用和技术发展设定了明确的政策导向。

（四）开展评估审计

基础模型的风险评估是一种系统性方法，用于识别和分析在开发、部署和使用基础模型（如机器学习模型、深度学习模型等）过程中可能遇到的潜在风险。这种评估涉及识别可能对模型的性能、安全性、可靠性和伦理标准产生负面影响的因素，以及这些因素可能造成的影响程度和可能性。〔26〕基础模型平台开展风险评估的目的是优先级排序、制定缓解措施，以将风险降低到可接受的水平。这包括技术风险（如数据偏差、过度拟合等）、安全风险（如数据泄露、模型被恶意利用等）、法律和合规风险（如违反个人信息保护法等）以及伦理风险（如算法偏见、隐私侵犯等）。

例如，OpenAI 发布灾难风险的预备框架，详细介绍了其用于评估和减轻日益强大的基础模型所带来的灾难性风险的全面框架。该框架着重于跟踪和监控各类灾难风险水平，包括网络安全、化学、生物、辐射和核威胁，以及模型自主性。通过创建评估套件和监控解决方案，该框架旨在预测风险的未来发展并通过一个动态记分卡持续追踪风险水平。微软对于防范人工智能风险作出了自愿承诺，承诺实施 NIST AI 风险管理框架，为高风险模型治理提供安全性与可靠性的实践，并支持对高能力模型建立许可制度，以更好规范模型的安全开发与部署。〔27〕谷歌

〔26〕 See Toby Shevlane, Sebastian Farquhar, et al., *Model Evaluation for Extreme Risks*, available at https://doi.org/10.48550/arXiv.2305.15324, last visited on May 10, 2024.

〔27〕 See *Voluntary Commitments by Microsoft to Advance Responsible AI Innovaton*, available at https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2023/07/Microsoft-Voluntary-Commitments-July-21 - 2023.pdf, last visited on Mar.18, 2024.

DeepMind 在极端风险模型评估的报告中提出极端风险的模型评估过程包括“危险能力评估”和“对齐评估”。〔28〕阿里巴巴当前针对模型的全生命周期中不同角色的职责划分和不同阶段进行评估和风险治理。在模型训练阶段，采用定义风险和基准的方式提高评测能力。在服务上线阶段选用优质的安全评估模型进行安全核验，而且在这个阶段进行算法的评估与备案。在内容生成阶段，采取生成内容的审核阻断机制对生成内容进行分类分级处理。在内容传播阶段，通过标识和溯源的安全运营方式对风险内容进行及时处理。

基础模型审计是一个独立的检查过程，旨在评估模型的设计、开发和操作是否遵循了既定的标准和最佳实践。这包括对模型的数据处理、算法选择、性能评估、安全措施和伦理影响进行综合评价。审计的目的是确保模型的透明度、可解释性和公正性，同时识别并解决潜在的风险和弱点。基础模型平台对模型的审计过程可能涉及代码审查、性能测试、安全漏洞扫描、合规性检查以及伦理影响评估等多个方面。通过审计，组织可以提高利益相关者的信任，确保模型的负责任使用，并符合法律、行业标准和伦理要求。

第三方审计和监督在基础模型平台治理中起关键作用，确保这些平台的技术和操作符合行业标准和法律要求，同时可能引入操作和财务挑战。例如，微软会定期进行第三方安全评估和合规性审查，以确保其产品和服务不仅符合技术标准，还遵守如欧盟《一般数据保护条例》（GDPR）等数据保护法规。这样的审计既是公司内部安全策略的一部分，也是对外界承诺的一种证明。〔29〕Amazon 通过其云服务 AWS 可能会接受类似的第三方审计，特别是在数据中心安全和客户数据保护方面。考虑到 AWS 广泛的客户基础和服务范围，这种审计对于确保客户信任和满足监管要求尤为重要。〔30〕谷歌 DeepMind 明确参与伦理和安全性方面的第三方审计，以确保其研究和开发活动符合最高的伦理标准和行业规范。〔31〕

四、基础模型平台实施安全治理的理论困境

基础模型平台的安全治理对于保障基础模型及相关服务应用的安全性、可靠性和合规性至关重要。但是，基础模型平台的自我治理、伦理治理、技术治理仍存在内在缺陷，治理的权威性和代表性缺失、伦理规范过于抽象、安全测试的算力消耗成本较高、风险本身的复杂性都影响了其治理效能的发挥。问责和救济机制的有效性局限和平台自我治理的利己偏好，进一步影响了基础模型安全治理的成效。

---

〔28〕 See Rishi Bommasani, et al., *On the Opportunities and Risks of Foundation Models*, available at https://doi.org/10.48550/arXiv.2108.07258, last visited on May 10, 2024.

〔29〕 See Ram Shankar Siva Kumar, *Best Practices for AI Security Risk Management*, available at https://www.microsoft.com/en-us/security/blog/2021/12/09/best-practices-for-ai-security-risk-management, last visited on May 10, 2024.

〔30〕 See Amazon, *Enhancing Frontier AI Safety*, available at https://aws.amazon.com/cn/uki/cloud-services/uk-gov-ai-safety-summi, last visited on May 10, 2024.

〔31〕 See AI-Tech Park, *Responsible Use of AI: EqualAI, Google DeepMind, Microsoft & Others*, available at https://ai-techpark.com/responsibleuse-of-ai-equalai-google-deepmind-microsoft-others, last visited on May 10, 2024.

（一）治理架构的权威性和代表性缺失

基础模型平台作为私主体，其实施的治理并没有法律的授权，本就缺乏足够的权威性。实践中，基础模型平台的治理架构主要依赖于平台自身的管理团队、技术团队和法律团队等内部资源。无论是决策过程，还是安全策略、技术标准的制定和执行主要在组织内部完成。虽然这种模式有助于快速响应和调整，但也可能忽视或无法充分代表外部公众和用户的利益。

尽管一些平台尝试通过引入外部声音，例如设立独立监督委员会或吸纳独立专家参与治理过程，增加治理架构的多元性，但这些努力往往受限于实际操作和影响力的范围。外部专家和监督委员会可能缺乏足够的权力和资源来实质性地影响平台的决策和政策制定。在缺少外部利益相关者参与的情况，治理架构的权威性和代表性不足更加凸显。

权威性和代表性不足，导致基础模型平台制定和执行平台规则难以得到平台内外部成员广泛接受，甚至可能被忽视，从而无法有效地管理和减少基础模型安全风险。

（二）伦理规范过于抽象和复杂

在安全治理实践中，伦理规范过于抽象和复杂是一个普遍存在的问题。伦理规范往往表述为一般性的原则和标准，如公平、透明、负责任等，这些原则虽然在理论上获得广泛认可，但在具体应用到复杂多变的技术场景时，不同的利益相关者可能对同一伦理原则有不同的理解和预期，难以达成共识。伦理规范的抽象性导致不同平台在理解和执行这些规范时存在认知不一致。例如，关于“公平”的定义、如何量化“隐私保护”的有效性以及“责任”的具体承担方式等问题，不同平台可能有不同的解释和实施策略。

Google 通过建立负责任的人工智能原则和高级别的伦理审查委员会，强调在人工智能开发和部署过程中的伦理考量和透明度。〔32〕Google 的人工智能原则明确了遵循伦理标准的重要性，并设立了外部咨询委员会来增加决策的多元性和透明度。OpenAI 则注重与外部伦理和安全研究组织的合作，以第三方的视角评估基础模型的潜在风险和伦理问题。OpenAI 通过定期的安全审计和风险评估，试图在模型开发的各个阶段引入透明度和可解释性。百度的飞桨平台通过建立开放的技术委员会和伦理委员会吸纳多方面的独立专家参与到平台的治理过程中。这种做法旨在通过集合多方视角来提高平台的伦理标准和社会责任感。

在对齐问题中，尽管不同平台都在努力提高伦理规范的透明度和可解释性，但在如何对齐这些规范以及如何在全球范围内统一执行方面，仍然面临着重大挑战。不同地区的法律法规、文化背景和社会价值观对伦理规范的解读和要求各不相同，这给平台在全球范围内实施统一的伦理规范带来了额外的复杂性。此外，技术的快速发展也使得伦理规范需要不断更新和调整，以适应新的技术和应用场景，这进一步加大了伦理规范对齐和实施的难度。〔33〕

基础模型平台在实施安全治理时，需要在伦理规范的应然要求和具体实施之间找到平衡点。

〔32〕 See Google, *Secure, Empower, Advance How AI Can Reverse the Defender's Dilemma*, available at https://services.google.com/fh/files/misc/how-ai-can-reverse-defenders-dilemma.pdf, last visited on May 10, 2024.

〔33〕 See Ethan Perez, Sam Ringer, et al., *Discovering Language Model Behaviors with Model-Written Evaluations*, available at https://doi.org/10.48550/arXiv.2212.09251, last visited on May 10, 2024.

同时，这也要求平台之间、平台与社会各界之间建立更加有效的沟通和合作机制，共同推动伦理规范的统一理解、对齐和执行。

（三）安全测试的算力消耗缺乏可靠的保障机制

在基础模型平台的安全治理中，安全测试是一个至关重要的环节，特别是对抗测试和广泛的模型安全性测试。这些测试旨在发现和修复潜在的安全漏洞和弱点，确保模型的安全性、健壮性和可信度。然而，这些测试过程往往需要消耗大量的算力资源，给平台带来了显著的资源挑战。

对大规模的基础模型，尤其是深度学习模型，进行全面和深入的安全测试需要巨大的计算资源。这不仅包括单次测试的资源消耗，还包括持续的监控和更新所需的资源。平台需要投入显著的硬件和软件资源以及与之相关的财务成本。红蓝对抗测试、漏洞扫描、压力测试等多种测试方法都是评估基础模型安全性的重要手段。这些测试不仅要覆盖模型的各个方面，还要模拟各种潜在的攻击场景，进一步增加了测试的难度和算力需求。

除了算力资源的需求外，进行有效的安全测试还需要相应的技术能力，包括能够设计和实施复杂测试场景的能力，以及能够解析测试结果并据此优化模型的技术专长。不同平台在这方面的能力可能存在差异，这影响了测试的质量和效率。大型技术公司如 Google、OpenAI 等可能有足够的资源和技术能力进行广泛和深入的测试。然而，小型或初创公司可能难以承担相应的成本，导致在安全性验证方面的不足，缺少进行测试、评估所需的资源。〔34〕即使是资源充足的公司，对于如何进行最有效的安全测试也可能有不同的策略和偏好。测试的算力消耗和技术能力要求为基础模型平台的安全治理带来了显著的挑战。这要求平台不仅要投入必要的资源，还需要不断提升技术能力，同时寻求更高效的测试方法，以确保在资源和成本可承受的情况下实现模型的安全性和可靠性。此外，行业内的合作和知识共享也可能对提高测试效率和降低成本起到积极作用。优质的算法也可以提高算力应用的效率从而减少一定的算力资源消耗。

（四）风险评估的不完全性与评估难题

在基础模型平台实施安全治理的过程中，进行全面且准确的风险评估是一个重大挑战。这个问题主要体现在两个方面：一是风险难以完全预判，即在模型开发和部署的早期阶段难以识别所有潜在的风险点；二是评估过程中难以做到全面，即难以覆盖所有可能的使用场景和环境因素。

1. 风险难以完全预判

第一，技术不断演进与发展。随着人工智能技术的快速发展，新的算法和模型结构不断涌现。这种快速的技术迭代带来了新的安全挑战，使得在模型开发早期阶段识别所有潜在风险变得更加困难。例如，深度学习模型的复杂性使得它们的行为难以预测，可能在特定条件下表现出意料之外的行为，从而引入新的安全风险。

第二，数据驱动具有不确定性。当代基础模型，尤其是基于大数据的深度学习模型，性能高度依赖于训练数据。数据中的偏见、不准确性或敏感性可能未被充分识别，导致模型在实际应用

〔34〕 See Noam Kolt, et al., *Responsible Reporting for Frontier AI Development*, available at https://doi.org/10.48550/arXiv.2404.02675, last visited on May 10, 2024.

中产生不可预见的风险。例如，OpenAI 的 GPT 系列模型在生成文本时可能会反映或放大训练数据中的偏见，而这种风险在开发初期可能难以完全预测。〔35〕

第三，跨领域应用具有复杂性。基础模型被广泛应用于医疗、金融、法律等多个领域，每个领域的特定环境和要求都可能带来独特的风险。例如，使用机器学习模型进行医疗诊断时，即使是极小的错误率也可能导致严重的后果，而这些风险在跨领域应用时难以一一预测和评估。

2. 评估难以做到全面

第一，基础模型应用场景广泛。基础模型的应用场景极为广泛，每个场景都有其特定的风险因素。全面评估所有可能的使用场景和环境因素是一项巨大的挑战。如在图片识别、自然语言处理和自动驾驶不同领域都会有不同的安全要求和风险考量。由第三方安全公司、法律顾问或行业组织执行的外部审计监督，也会面临难以理解具体场景下基础模型运行逻辑的挑战，从而可能导致第三方审计过程变成一种纯粹的外部审批程序，这不仅不利于发现实际问题，而且可能增加企业的合规成本。

第二，基础模型技术和环境动态变化。技术和应用环境的快速变化意味着昨天的风险评估可能不再适用于今天。不断出现的新技术、新攻击手段以及用户行为的变化都可能引入新的风险。例如，随着社交媒体使用模式的变化，基于社交媒体数据的基础模型可能面临新的隐私和安全问题。

第三，风险评估需要考虑多方面的风险因素。完全的风险评估需要考虑技术、法律、伦理和社会等多个维度的因素，这不仅需要跨学科的知识，还需要不断更新的行业和社会标准。如基础模型平台在处理用户数据时不仅要考虑技术安全，还要应对不断变化的全球数据保护法规和公众对隐私的关切。

（五）问责与救济机制缺乏有效性

在基础模型平台实施安全治理过程中，问责机制的建立和救济机制的充分性是实现长期可持续发展的关键因素。然而，构建一个能够真正实现问责并提供有效救济的机制面临着诸多挑战。

第一，确定责任归属在复杂的人工智能系统中尤为困难。人工智能决策过程的不透明性，加之模型训练和部署涉及多方参与，当出现错误或争议时，很难追溯到具体的责任主体。这种情况在跨国公司运营的平台中尤为明显，不同国家和地区的法律法规差异使得制定统一的问责标准和救济机制更加复杂。〔36〕

第二，现有法律框架面对基础模型规范存在不足。评估人工智能决策导致的损失、确定赔偿范围和形式以及如何实施监管措施等问题，需要法律理论和实践经验的进一步探索。在这种情况

〔35〕 See OpenAI, *Frontier Risk and Preparedness*, available at https://openai.com/blog/frontier-risk-and-preparedness, last visited on May 10, 2024.

〔36〕 See Meta, *Building Generative AI Responsibly*, available at ai.meta.com/static-resource/building-generative-ai-responsibly/, last visited on May 10, 2024.

下，平台在建立问责和救济机制时可能会遇到一些法律和监管环境方面的不确定性，这可能会影响问责与救济机制的有效性和实施。

第三，救济机制的不足也是一个突出问题。在许多情况下，用户和受害者发现自己在面对基础模型引起的问题时缺乏有效的救济途径，相关的争议解决过程复杂且耗时，而结果也往往难以预测。基础模型平台如何确保基础模型的透明度和可追溯性，以及为受影响方提供有效救济，成为关键问题。

（六）平台自我治理存在利己偏好

基础模型平台本身作为私主体，自我治理的出发点首先是平台秩序基础上的私主体利益。基础模型平台作为安全治理的主体，其治理活动也会遵循相应的商业逻辑，体现市场化的思维方式。基础模型平台在提供服务、获取收益的同时，也承担着维护平台秩序、保障用户权益的责任。然而，平台在履行这些责任时，并非完全出于无私的公益考量，而是会受到自身商业利益的驱动和制约。这种利益驱动的治理动机，使得平台在面对不同风险时表现出明显的治理偏好。对于输出内容显然违法、造成个人隐私泄露等直接威胁其商业利益和声誉的风险，基础模型平台的治理力度可能较大，但面对属于细微偏见或者难以辨明是否真实的内容时，由于缺乏明确的法律约束或管理成本过高，基础模型平台的治理动力可能不足。

对于那些与其核心商业利益直接相关、对平台声誉和用户信任构成严重威胁的风险，如违法内容、隐私泄露等，平台往往有强烈的治理动力。这是因为这类风险一旦失控，不仅会引发用户流失和收益下滑，还可能使其面临监管处罚和公众质疑，对平台的可持续发展造成重大冲击。出于自身利益的考虑，平台对此类风险的治理意愿和投入程度通常较高。

相比之下，对于一些细微的偏见、模糊的界限、难以甄别的内容，平台的治理动力可能不足。这类问题往往缺乏明确的法律红线，且很难在短期内或者直接对平台产生显著的负面影响，对于企业而言治理成本高、收益低。在商业利益的权衡下，平台可能会降低对此类风险的关注度和资源投入，甚至选择性地“忽视”某些问题，以节约治理成本，维持表面的平台秩序。

平台自我治理中的动力偏好，客观反映了商业逻辑对其行为的影响。在利益驱动下，平台对不同风险的治理意愿和力度存在差异，这种差异可能导致其在某些领域的治理存在短板和漏洞，往往也是“头疼医头脚痛医脚”。对平台而言，只要模型的输出结果没有明显的歧视性，且没有引发用户的强烈不满和抗议，就可能认为问题不大，缺乏进一步排查和纠正基础模型本身偏差的动力，部分基础模型存在的问题也往往因此被隐藏了下来。

五、基础模型平台安全治理的完善路径

在全球人工智能相关立法已对基础模型开发应用提出原则性要求的情况下，如何构建适用于基础模型平台的安全治理机制，平衡创新与风险，是当前世界各国在人工智能监管和促进人工智能发展的实践中亟待解决的问题。本文从应对基础模型平台的治理挑战出发，基于现有的治理实

践，分析基础模型平台安全治理的重要性、治理困境及应对策略，旨在为我国基础模型平台安全治理提供参考。

美国政府曾两次召集行业领先的基础模型平台企业，推动其作出对模型开发应用进行更全面安全评估的自律承诺，在此基础上，美国政府还发布行政命令，针对具有双重用途风险的基础模型，要求 NIST 等机构编写安全测试指南，并要求模型开发者及时向政府汇报模型开发工作。欧盟《人工智能法》针对大模型这类通用人工智能系统的特殊风险，要求相关模型的提供者和部署者进行模型评估、评估和对抗测试。中国学者起草的《人工智能示范法 2.0（专家建议稿）》和《人工智能法（学者建议稿）》，分别从设定基础模型研发者的法定义务和明确基础模型相关主体法律责任的角度，提出基础模型安全治理的规范要求。国内外普遍认识到加强基础模型安全治理的必要性和紧迫性，并从法律规制、监督指导、企业自律等路径提出了治理思路。这为进一步推动基础模型平台安全治理机制的构建提供了参考。

相对于美欧对基础模型平台安全治理所作出的初步制度设计以及 OpenAI、Anthropic、Meta 等平台已实施的合规举措而言，中国的基础模型平台发展及其治理仍有一定差距，虽然学者起草的提案提出了规范性要求，但离具体实施落地尚有一定距离。前述治理困境在中国未来平台安全治理中可能逐步显现，甚至更加突出。基础模型平台安全治理机制的完善应在《生成式人工智能暂行规定》等一系列管理规定的基础上，采取针对性措施增强基础模型平台安全治理效能。

（一）压力驱动：完善监管机制，推进社会监督

基础模型平台在发展过程中面临着诸多安全和伦理挑战，来自监管部门和社会的外部压力能够倒逼平台加强管控水平，形成平台加强安全治理的外部约束和驱动力量，确保基础模型平台的安全治理规范有序实施。

1. 完善基础模型平台监管机制

在探索基础模型平台安全治理的完善路径时，应从抽象复杂的伦理规范中提炼切实可行的核心伦理原则，并将其转化为具体的法律规范和可操作的技术标准。通过法律规则将抽象自律的伦理规范或平台规则转为更加具体强制的法律规范，构建基础模型平台应承担的法定义务，从而确立基础模型平台的合规预期，为问责救济机制的实现提供制度支撑。

第一，增加有效的制度供给。“公共安全治理的推进有赖于国家/政府治理能力的提升。”〔37〕将人工智能领域的必要伦理规范转化为具体的法律规范和可操作的技术标准，这包括但不限于数据隐私保护、算法透明度和可解释性、公平性以及非歧视性原则。这一过程需要跨学科的合作，包括法律、伦理学、计算机科学等领域的专家需共同参与，确保制定的规范既科学合理，又具有可操作性。《人工智能示范法 2.0（专家建议稿）》针对基础模型研发者，要求其建立风险管理、模型管理和数据管理等方面的制度，确保其为安全治理投入必要资源，并采取有效手段处置平台生态内的违规活动等。〔38〕这些制度设计就可以作为未来法律制度设计的重要参考。

〔37〕 张春艳：《大数据时代的公共安全治理》，载《国家行政学院学报》2014 年第 5 期，第 104 页。

〔38〕 参见周辉等：《人工智能示范法 2.0（专家建议稿）》，载 https://doi.org/10.5281/zenodo.10974163，最后访问时间：2024 年 5 月 10 日。

第二，强化监管指引。除了制定规范之外，还需要相应的行政指导和监管框架来确保这些规范的有效实施。这包括建立监管机构、制定监管流程和机制以及明确各方面的责任和义务。监管机构应发挥其行政指导作用，引导平台遵守国家安全、数据保护等相关法律法规，并确保这些规范在技术实施上的可行性。如美国通过人工智能行政令对人工智能监管作出了整体部署，指示 50 多个联邦机构执行 100 多项关于人工智能监管的具体行动，促进各联邦机构对人工智能潜在风险以及现有法律、监管工具的适用性和延展性的统一理解；同时注重行业治理，通过监管与行业密切互动，推进达成共识的最佳实践、行业标准、测试环境等，以及制定严格的红队测试标准。中国在这方面可以借鉴美国经验，建立有效的信息共享和沟通机制，以及相应的安全评估和认证体系，保证人工智能系统的安全、可靠和值得信赖。

第三，完善法律责任归属规则。应当明确在人工智能系统造成损害时的责任归属以及受害者的救济途径。这需要在法律层面对人工智能造成的损害进行定义，明确责任主体、归责原则和问责路径，进而实现有效的法律救济。

2. 实现对基础模型平台的有效制衡

在面对安全治理困境时，基础模型平台需要进一步开放和扩大外部利益相关者的参与，通过外部利益相关者与基础模型平台间的相互制衡，形成治理的合力，控制前沿模型的滥用风险，敦促基础模型平台更加积极、主动地实施自我治理，切实履行安全治理责任。〔39〕

第一，建立和完善多元主体参与的社会治理机制。平台应确保不同利益相关者的声音能够被听到并纳入决策过程中。这不仅有助于增强治理架构的权威性和多元性，也是提升公众信任和接受度的关键，有利于激发不同主体参与治理的意愿和责任感。政府和行业组织可以共同发起建立由专家、学者、消费者代表等多方参与的社会监督机构。这些机构不仅能够对平台的安全治理进行定期检查和评估，还能作为公众与平台之间沟通的桥梁，收集和反馈公众对于人工智能安全的关切和建议，增加外部公众和用户在治理过程中的发言权和影响力。〔40〕

第二，优化第三方审计机制。其一，为了避免第三方审计异化为单纯的审查流程，审计过程中应注重实质性的安全检查与评估，关注平台在实际操作中的安全治理能力和效果。同时，鼓励采用创新的审计方法，如采用人工智能技术辅助审计，提高审计的效率和准确性。〔41〕其二，明确第三方审计目标和标准。在进行第三方审计时，应明确审计的目标和标准，避免泛泛而谈。具体而言，可以围绕数据保护、用户隐私、算法公平性等关键领域设定详细的审计标准和指标，确保审计过程的针对性和有效性。〔42〕其三，增强审计的透明度和公开性。通过公开审计报告、审计结果等信息，增强审计过程的透明度，让公众能够了解平台的安全治理状况和存在的问题，从

〔39〕 See Sella Nevo, et al., *Securing AI Model Weights: Preventing Theft and Misuse of Frontier Models*, available at https://www.rand.org/pubs/research_reports/RRA2849-1.html, last visited on May 10, 2024.

〔40〕 See Anthropic, *Make Safe AI Systems Deploy Them Reliably*, available at https://www.anthropic.com/research, last visited on May 10, 2024.

〔41〕 See Inioluwa Deborah Raji, Peggy Xu, et al., *Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance*, available at https://ar5iv.labs.arxiv.org/html/2206.04737, last visited on May 10, 2024.

〔42〕 See Al-Tech Park, *Responsible Use of AI: EqualAI, Google DeepMind, Microsoft & Others*, available at https://ai-techpark.com/responsibleuse-of-ai-equalai-google-deepmind-microsoft-others/, last visited on May 10, 2024.

而提升公众对平台的信任度。

（二）动力保障：提供投入激励，增强安全实效

基础模型平台安全治理需要投入大量的人力、物力和财力。确保基础模型平台能够按照监管要求和国家标准有效实施安全治理，除通过监管执法、第三方监督对平台企业施加治理压力外，也需要适当调动起企业主动加强自身安全保障能力的积极性，使提升服务可靠性、安全性的政策取向与平台企业以营利为目标的需求尽可能协调一致。

1. 引入税收抵免优惠

与传统网络安全风险相比，基础模型安全风险的变化更加多样，会因模型自身代码架构、技术逻辑的不同而有区别。同时，随着各基础模型迭代发展，先前已识别的安全风险可能带来新的威胁。2023 年，来自世界各国的近 20 位人工智能科学家和治理专家联署声明指出，研发者所采取的现有安全措施将因前沿人工智能的高性能被轻易破解。为了满足人工智能模型的安全防护需要，研发者需至少将三分之一的研发经费用于安全研究，同时政府机构需要以同等比例支持学术与非营利性的人工智能安全与治理研究，〔43〕以能够运用与模型自身复杂度和性能相匹配的安全治理工具。对于研发投入本就巨大的基础模型而言，为基础模型投入类似比例的安全支出，显然将为模型研发者及基础模型平台带来更沉重的成本压力。

采用税收优惠等方式提供激励，以鼓励私主体在特定领域或方向上进行投入，是世界各国普遍采用的政策工具。例如，《中华人民共和国企业所得税法实施条例》（2019 年修订）等法律法规就明确，购置并实际使用有关规定中列举的环境保护、节能节水、安全生产等专用设备时，设备投资额的一定比例可以从企业应纳税额中抵免。此类成本分摊和经济激励措施同样也可用于实施基础模型平台治理，鼓励平台投入资源主动研发、应用各类安全防护措施，提升自身系统安全性。

2. 激励机制设计

第一，明确激励的适用条件。激励机制应赋予企业足够的动力使之在安全等方面进行投入，但同时也应制定享受激励政策的标准，避免政策适用范围过广、浪费税收资源。可在相关立法中明确，基础模型研发者、提供者及基础模型平台在研发安全治理工具、应用防护技术等方面的投入，可以在计算应税额时予以一定比例的加计扣除或税额抵免。这不仅可以减轻平台在前期为提升安全性进行投入的经济负担，还能鼓励更多的平台主动行动，提升自身的安全治理能力。但是，相关安全投入及其他安全治理举措需严格遵循国家标准，并依法接受检查和监督。

第二，确保税收激励起到预期效果。在设计不同类型投入所获激励的额度时，需要尊重平台自我治理存在动力偏好这一客观现实，引导、促使平台发现和重视那些虽不直接显现但对模型安全和社会影响至关重要的问题。在激励机制的设计中，需要充分考虑基础模型平台既有的安全投

〔43〕 See Bengio Y, Hinton G, Yao A, et al., *Managing AI Risks in an Era of Rapid Progress*, available at https://arxiv.org/abs/2310.17688, last visited on Aug. 25, 2024; Center for Human-Compatible AI, *Prominent AI Scientists from China and the West Propose Joint Strategy to Mitigate Risks from AI*, available at https://humancompatible.ai/?p=4695, last visited on Aug. 25, 2024.

入做法，避免“一刀切”的激励措施对平台原有治理积极性产生负面影响。应当在问题导向的基础上，有的放矢地提供针对性激励，引导平台强化薄弱环节，保障基础模型开发所需关键投入的可得性，弥补治理短板。〔44〕针对基础模型平台普遍存在的问题进行有侧重的激励，可以在尊重平台既有安全投入的基础上，纠偏补缺，使激励的引导效果最大化。

第三，优化平台信息披露机制。平台的安全治理等信息披露情况可以作为是否给予平台激励或给予何种程度激励的重要参考指标。在设计税收抵免优惠等制度时应明确，基础模型平台应真实地进行重要安全信息的披露，尤其是平台为实施安全治理在哪些方面进行了投入、支出了何种成本。错报、漏报等行为应给予适当的处罚和公示，并以此作为警示。对于在基础模型安全保护方面作出积极贡献的平台，在已有的激励政策之外还可以通过奖励、优先政策等方式给予肯定，促进良好的行业自律风气。此外，考虑到信息披露可能对平台经营造成影响，应允许基础模型平台向政府和其他主体披露信息的详细程度有所不同，在确保政府监管需求的同时，尽可能降低商业敏感信息泄露的风险。

（三）能力强化：鼓励技术革新，培育用户素养

治理能力强化对于基础模型平台实施安全治理具有重要意义。基础模型平台的治理能力从根本上决定了治理的质量和成效，是其实现可持续、可信赖发展的内在要求和必由之路。通过鼓励技术创新，平台可以持续增强安全防护能力，完善治理手段和机制，从源头上防范和化解风险隐患。通过培育用户素养，提高模型开发者、部署者、使用者专业素养和责任意识，有助于帮助他们自觉担当、积极作为，切实提高基础模型的安全水平。

1. 鼓励安全治理技术层面的创新

加强技术创新，特别是在安全技术方面的创新，是确保人工智能系统安全性的核心。这要求基础模型平台持续进行研发投入，不仅需要在算法层面进行优化，也要在安全防护技术上进行突破。例如，美国国家标准与技术研究院（NIST）启动 GenAI 项目，旨在应对生成式人工智能技术带来的内容真实性挑战。NIST 发布一套标准化的测试任务、数据集和评估指标，引导基础模型平台创建“内容真实性”检测系统，鼓励其识别虚假或误导性人工智能生成信息的来源。〔45〕OpenAI 发布了“深度伪造检测器”用于打击虚假信息生成，OpenAI 声称，其新推出的检测器能够以 98.8% 的准确率识别出由旗下最新版图像生成器 DALL-E 3 所创建的图像。〔46〕由 OpenAI、Meta、Adobe，BBC，Intel，Microsoft 等组织联合形成的内容来源和真实性联盟（C2PA）正在通过制定技术标准来证明媒体内容的来源和历史（或来源），以解决在线误导性信息的普遍存在。〔47〕这些技术的发展不仅能够提升人工智能系统本身的安全性，同时也能够更好

〔44〕 See *AI Foundation Models Technical Update Report*, available at https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf, last visited on May 10, 2024.

〔45〕 See NIST, *Evaluating Generative AI Technologies*, available at https://ai-challenges.nist.gov/genai, last visited on May 10, 2024.

〔46〕 See New York Times, *OpenAI Releases “Deepfake” Detector to Disinformation Researchers*, available at https://www.nytimes.com/2024/05/07/technology/openai-deepfake-detector.html, last visited on May 10, 2024.

〔47〕 See C2PA, *The Coalition for Content Provenance and Authenticity*, available at https://c2pa.org/, last visited on May 10, 2024.

地降低安全测试成本，为用户提供更加安全、可信的服务体验。

第一，鼓励基础模型平台及外部产业发展安全治理的技术或专业的安全模型。可以推动基础模型平台开发专门针对人工智能系统的安全评估工具或治理模型，包括但不限于安全编码标准、漏洞评估体系以及风险管理框架。〔48〕工具不仅需要能够适应人工智能技术的特点，也应当能够满足不同行业和应用场景的具体需求。通过建立行业共识，形成统一或兼容的安全标准和最佳实践，可以有效提高整个行业的安全治理水平。〔49〕

第二，政府和行业组织应当指导产业有序发展，鼓励并支持红蓝对抗活动的开展，为参与者提供必要的资源和支持。通过发展基础模型安全治理技术或专业的安全治理模型，为对抗测试提供一定的技术支持。此外，应当建立红蓝对抗的标准化流程和评估体系，以确保演练活动的质量和效果，同时促进安全知识和经验的分享和传播。

2. 强化用户教育与鼓励用户参与

用户作为人工智能技术的直接受益者和风险承担者，在平台安全治理中的作用不可忽视。强化用户教育和鼓励用户参与，有助于在基础模型平台内形成更加安全、可靠和可持续的人工智能技术应用环境，同时提高基础模型平台安全治理的实效性，应对当前的安全治理困境，为平台的长期发展和技术进步提供可持续且全面的支持。

第一，用户教育应覆盖人工智能技术的基本原理、可能的风险以及安全使用的最佳实践。这不仅有助于用户在使用基础模型平台提供的服务时作出更明智的选择，还能使用户意识到自己在数据保护和维护网络安全中的角色。通过在线课程、公开讲座、互动问答等多种形式，可以有效提高用户教育的覆盖面和参与度。

第二，鼓励用户实际参与基础模型平台安全治理。上游模型开发者应向下游部署者提供充分、准确的信息，下游部署者应向最终用户提供作出知情选择所需的信息，让各类用户基于其角色承担安全责任。〔50〕一般用户也可以通过参与公开的安全评估、反馈系统漏洞、参与社区讨论等方式直接参与到平台的安全治理中来。例如，一些平台设立了漏洞奖励计划，鼓励用户报告潜在的安全问题。通过这种方式，平台不仅能够及时发现并修复安全漏洞，还能够建立与用户间的信任和合作关系。此外，推动社会共治是提升平台自我治理实效的关键。〔51〕通过建立开放透明的沟通机制，平台、用户、政府和第三方组织可以共同参与到安全治理中来，形成有效的监督和反馈系统。这种多方参与的共治模式强调“认同、共识与合意”，〔52〕有助于提高治理方案的全面性和有效性，同时也能够提升社会对人工智能技术发展的整体认知和接受度。

---

〔48〕 See Howandwhat, *Facebook Stakeholder Analysis*, available at https://www.howandwhat.net/stakeholders-facebook/, last visited on May 10, 2024.

〔49〕 See Anthropic, *Constitutional AI: Harmlessness from AI Feedback*, available at https://www.anthropic.com/news/constitutional-ai-harmlessness-from-ai-feedback, last visited on Mar. 10, 2024.

〔50〕 See *AI Foundation Models Technical Update Report*, available at https://assets.publishing.service.gov.uk/media/661e5a4c7469198185bd3d62/AI_Foundation_Models_technical_update_report.pdf, last visited on May 10, 2024.

〔51〕 See *Policy Updates*, available at https://www.aisafetysummit.gov.uk/policy-updates/#company-policies, last visited on May 10, 2024.

〔52〕 参见罗豪才、宋功德：《认真对待软法——公域软法的一般理论及其中国实践》，载《中国法学》2006 年第 2 期。

六、结　语

作为人工智能发展的基础设施，基础模型平台在支撑技术创新和产业发展的同时，也应承担起安全治理的主体责任，寻求实现平台自我治理与外部法律规制的合力，从而有效应对基础模型安全风险的复杂性和不确定性。目前欧美等国已经在尝试通过立法完善基础模型平台的外部监管，并通过国家级技术研究机构指引基础模型平台创新安全治理技术，在基础模型平台的安全治理上已经走在前列。

在已有部门立法的基础上，中国人工智能安全治理的法律框架正在加速构建，全国人大常委会 2024 年度立法工作计划将人工智能健康发展方面的立法项目列入预备审议项目。国务院 2024 年度立法工作计划中明确提出“将预备提请全国人大常委会审议人工智能法草案”。我国基础模型平台的安全治理也将会有更明确的法律规制和外部要求。未来还需要在理论和实践层面深化基础模型安全治理的研究，持续跟踪分析国外基础模型平台的代表性先进做法，将合理的平台规则不断提升为科学的法律制度，更好地以法治方式实现人工智能技术进步与安全发展的深度融合、动态平衡。

Abstract: The safety governance of AI foundation models is a crucial issue facing the legalized development of artificial intelligence. As the main subjects of safety governance, AI foundation model platforms play a vital role in their ability to identify, evaluate, and mitigate potential risks in AI systems, possessing the capability to adjust and optimize their own models. Domestic and international AI foundation model platforms are actively positioning themselves for safety risk governance, but still face challenges such as lack of authoritative and representative governance structures, overly abstract ethical standards, insufficient computing power for testing, difficulties in risk assessment, and platforms' self-interest bias. It is necessary to improve mechanisms such as pressure-driven, motivation-guaranteed, and capability-enhancing measures based on the practical realities of safety governance for Chinese AI foundation model platforms. This will better leverage the unique advantages and proactive roles of AI foundation model platforms in safety governance, supporting the healthy and orderly development of AI technology and applications.

Key Words: foundational model, foundational model platform, artificial intelligence safety, artificial intelligence risk, artificial intelligence governance

（责任编辑：张金平）