国家级人工智能安全研究所及其国际网络

为何建立、如何运作及未来挑战

安远AI

2024年12月

执行摘要

自英国和美国率先成立国家级人工智能安全研究所(AI Safety Institute，以下简称AISI)以来，安远AI始终密切关注和分析其演进过程。本文分析了当前多个国家级人工智能安全研究所及其国际网络的设立背景、进展、对比和挑战，旨在为中国在全球人工智能治理中的角色定位与政策制定提供参考。

1. 背景

GPT-4等前沿人工智能展现出强大的涌现能力，推动了多模态大模型、自主智能体、科学发现智能体和具身智能等众多技术方向，在多个领域已逼近甚至超越人类水平，但也引发了新的安全挑战。两届全球人工智能安全峰会先后发布的《布莱切利宣言》和《首尔宣言》推动了国家级人工智能安全研究所的设立，以应对技术风险并加强全球治理。

2. 进展

英国和美国分别在2023年首届全球人工智能安全峰会率先设立人工智能安全研究所，随后日本、新加坡、加拿大、韩国、法国等国家以及欧盟相继跟进，同时美国积极推动人工智能安全研究所国际网络的发展。此类机构以人工智能安全评测、人工智能安全研究、促进信息交流或推进标准制定为核心职能，已初步建立双边和多边的协作。

3. 对比

不同国家的国家级人工智能安全研究所在机构属性、职能定位、研究重点及国际协作等方面呈现多样性，在详细对比已官宣成立的8家国家级人工智能安全研究所的上述信息的基础上，我们重点就领先的人工智能安全研究所进行了案例分析：

- 英国人工智能安全研究所：充足的政府资金支持，吸纳大量技术人才，希望引领前沿人工智能安全评测和研究；得到OpenAI、DeepMind、Anthropic的部署前评测授权；参与全球人工智能安全峰会的筹办；已开源评测框架Inspect，为测试人员提供了评估各类模型特定能力的工具。
- 美国人工智能安全研究所：关注前沿人工智能风险，并涵盖更广泛的风险类型；依托美国国家标准与技术研究院和合作网络，成立了人工智能安全研究联盟；获得OpenAI

和Anthropic新模型发布之前和之后的访问权限。初期更关注国内安全问题，后通过与英国等人工智能安全研究所合作并宣布建立人工智能安全研究所国际网络后，越来越关注全球合作，旨在协调各方制定前沿人工智能的测量科学、自愿指南和严格测试标准。然而，特朗普当选新总统后，其全球合作前景存疑。

- 其他的国家级人工智能安全研究所则结合自身需求，在标准化、安全研发、执行监管等方面各有侧重。

4. 挑战

尽管人工智能安全研究所及其国际网络在安全评测、安全研究和国际合作中具有重要潜力，但未来仍需在模型访问与评测权限、信息共享与安全实践、标准制定与监管框架、资源差异与合作平衡、全球包容性与国际协调方面进行改进，以应对人工智能技术为全球治理带来的复杂挑战。

目录

1 背景

1.1 ChatGPT等前沿人工智能展示了技术的潜力和潜在的风险

GPT-4等前沿人工智能展现出强大的涌现能力，推动了多模态大模型、自主智能体、科学发现智能体和具身智能等众多技术方向，在多个领域已逼近甚至超越人类水平，但也引发了新的挑战。例如开源大模型已被改造成多种新型网络犯罪工具，前沿大模型可能成为生物安全风险的潜在推动者，此外人工智能竞赛、组织风险、自主体失控，甚至可能造成灾难性风险或生存风险[1]。

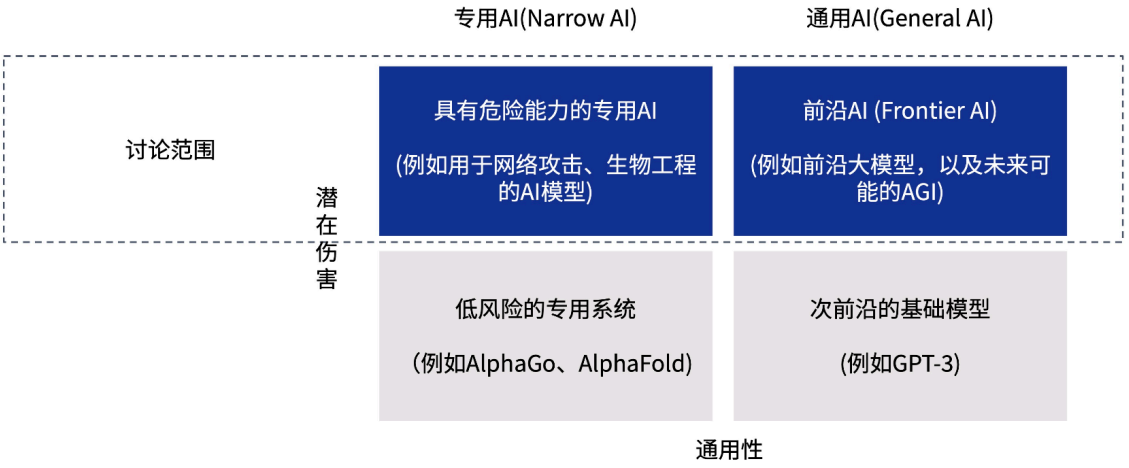
这些发展引发了全球各界的广泛关注，促使包括科学家、行业领袖以及政策制定者在内的众多利益相关方采取行动。《暂停巨型人工智能实验的公开信》[2]、《人工智能风险声明》[3]以及“人工智能安全国际对话”等呼吁加强对技术的治理和监管，以应对这些新兴技术可能带来的挑战。

为应对这些挑战，中国政府发布了《生成式人工智能服务管理暂行办法》[4]和《全球人工智能治理倡议》[5]等，旨在确保人工智能技术在安全和可控的框架内发展。同时，联合国[6]、G20[7]、G7[8]、GPAI[9]以及等国际组织也纷纷采取行动，制定并采纳了确保人工智能安全发展和使用的全球性原则，以促进人工智能技术在全球范围内的负责任应用和治理。

[1] 安远AI，“前沿大模型的风险、安全与治理”，2023-10-29, https://mp.weixin.qq.com/s/6AUz6rJm4XbOyETi08W5LQ.
[2] FLI，“暂停巨型人工智能实验的公开信”，2023-03-22, https://futureoflife.org/open-letter/pause-giant-ai-experiments.
[3] CAIS，“Statement on AI Risk”，2023-5-30, https://www.safe.ai/work/statement-on-ai-risk.
[4] 中央网信办，“生成式人工智能服务管理暂行办法”，2023-7-10, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.
[5] 中央网信办，“全球人工智能治理倡议”，2023-10-20, https://www.cac.gov.cn/2023-10/18/c_1699291032884978.htm.
[6] 联合国，“联合国大会通过里程碑式决议，呼吁让人工智能给人类带来‘惠益’”，2024-03-21, https://news.un.org/zh/story/2024/03/1127556.
[7] G20，“G20 New Delhi Leaders' Declaration”，2023-09-10, https://www.caidp.org/resources/g20/.
[8] OECD, G7 Hiroshima Process on Generative Artificial Intelligence (AI): Towards a G7 Common Understanding on Generative AI, 2023-05, https://doi.org/10.1787/bf3c0c60-en.
[9] GPAI，“Working Group on Responsible AI”，2024-11-25(引用日期), https://gpai.ai/projects/responsible-ai/.

1.2 英国推动全球人工智能安全峰会，应对前沿人工智能的风险

前沿人工智能(Frontier AI)，是指高能力的通用人工智能模型，能执行广泛的任务，并达到或超过当今最先进模型的能力，最常见的是基础模型。前沿人工智能提供了最多的机遇，但也带来了新的风险[10]。



参考了全球人工智能安全峰会[11]的讨论范围设定，白皮书[12]得到图灵奖得主Yoshua Bengio等专家的建议。

2023年11月，英国在布莱切利园举办了首届人工智能安全峰会，对前沿人工智能系统带来的风险和采取行动的必要性达成共识[13]。包括中国、美国在内的28个国家和欧盟，共同签署了《布莱切利人工智能安全宣言》（Bletchley Declaration）[14]。《宣言》签署国一致认为，人工智能系统已经部署在日常生活的许多领域，在为人类带来巨大的全球机遇的同时也带来了重大风险。

建立国家级人工智能安全研究所(AI Safety Institute, AISI)的想法，也从这一过程中诞生。英国峰会期间，时任英国首相苏纳克宣布成立英国人工智能安全研究所[15](UK AISI)，这是全球首

[10] 安远AI，“博鳌经安论坛发布｜安远AI《前沿大模型的风险、安全与治理》报告”，2023-10-29, https://mp.weixin.qq.com/s/6AUz6rJm4XbOyETi08W5LQ.
[11] Department for Science, Innovation & Technology (DSIT) (UK)，“AI Safety Summit: introduction”，2023-10-31, https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html.
[12] DSIT (UK)，“Capabilities and risks from frontier AI: A discussion paper on the need for further research into AI risk”，2023-10, https://assets.publishing.service.gov.uk/media/65395abae6c968000daa9b25/frontier-ai-capabilities-risks-report.pdf.
[13] 谢旻希，“为什么中国的参与必不可少？我参加首届全球人工智能安全峰会的所见所思（万字回顾）”，2023-11-01, https://mp.weixin.qq.com/s/_SWLDzKDOMNb04ha1SNKFg.
[14] UK Government，“Countries agree to safe and responsible development of frontier AI in landmark Bletchley Declaration”，2023-11-01,https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023.
[15] UK Government，“Policy paper：Introducing the AI Safety Institute”，2024-01-17,

个国家级人工智能安全研究所，美国副总统贺锦丽宣布将成立美国人工智能安全研究所[16](US AISI)，支持拜登总统签署的行政令赋予商务部的责任，并在两个月后宣布成立由200多个组织参与的人工智能安全研究所联盟(AISIC)[17]。

“人工智能安全研究所”概念在2024年5月英国和韩国联合举办的首尔峰会上呈现发展势头。《首尔宣言》不仅支持各国建立人工智能安全研究所，还提议建立此类机构的国际网络，以加强人工智能安全领域的多边合作[18]。日本、新加坡、加拿大和欧盟等很快设立了各自的人工智能安全研究所，这一过程被时任英国科学、创新和技术部大臣米歇尔·唐兰(Michelle Donelan)称之为**“布莱切利效应”(Bletchley effect)**[19]。

部长级会议的参与方包括20国政府、联合国等3家国际多边机构、10家学术界与民间组织、19家产业及相关组织。中方由来自中国科技部、中国科学院、安远AI、腾讯和阿里巴巴的代表出席会议。

https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

[16] The White House, “Remarks by Vice President Harris on the Future of Artificial Intelligence”, 2023-11-01, https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/11/01/remarks-by-vice-president-harris-on-the-future-of-artificial-intelligence-london-united-kingdom/.

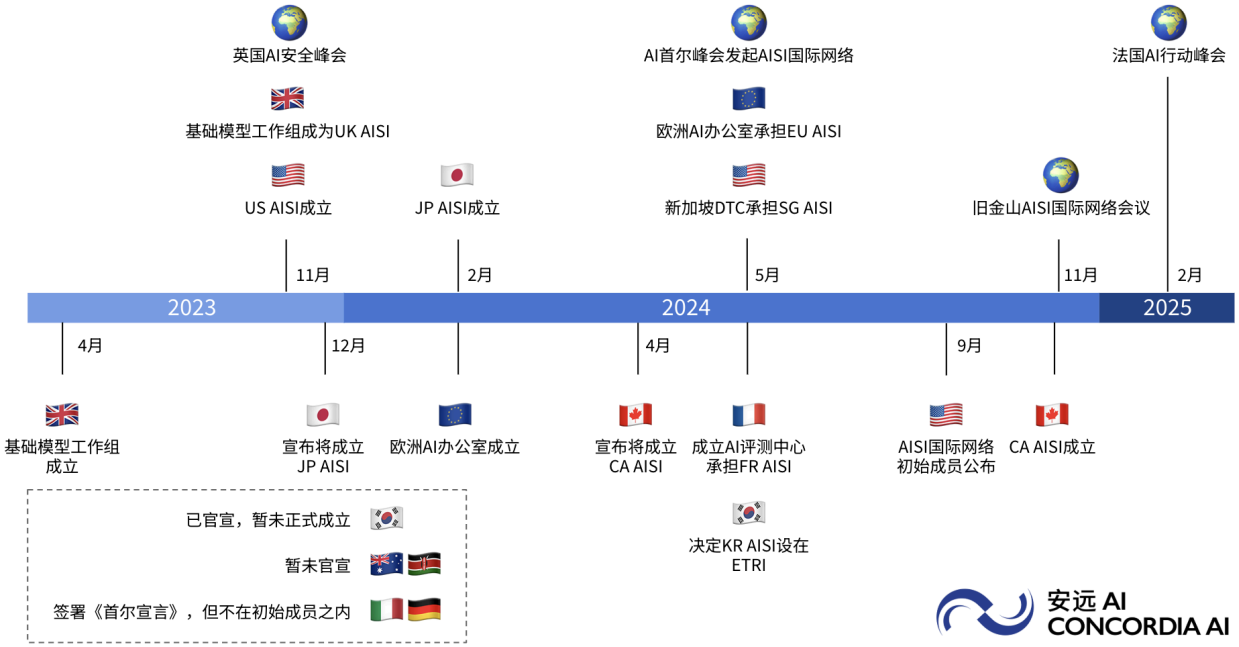
[17] NIST, Artificial Intelligence Safety Institute Consortium (AISIC), 2024-02-08, https://www.nist.gov/aisi/artificial-intelligence-safety-institute-consortium-aisic.

[18] UK Government, “Seoul Declaration for safe, innovative and inclusive AI by participants attending the Leaders' Session: AI Seoul Summit”, 2024-05-21, https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024.

[19] The Guardian, “Trying to tame AI: Seoul summit flags hurdles to regulation”, https://www.theguardian.com/technology/article/2024/may/27/trying-to-tame-ai-seoul-summit-flags-hurdles-to-regulation.

2 进展

人工智能安全研究所通常是国家支持的机构，旨在评估和确保前沿或先进人工智能模型的安全。至少已有7个国家和欧盟已成立组建人工智能安全研究所或由现有机构承担相应智能，并由美国发起成立了一个“人工智能安全研究所国际网络”[20]：



人工智能安全研究所以及国际网络的进展时间线

2.1 继英美之后，多个国家宣布设立国家级人工智能安全研究所

继2023年11月首届人工智能安全峰会上英国和美国率先成立人工智能安全研究所之后，同年12月时任日本首相岸田文雄明确表示将设立日本人工智能安全研究所(AISI Japan)[21]，并于2024年2月正式成立。2024年2月，韩国科学技术情报通信部部长李宗昊公布了2024年主要政策计划[22]，包括设立韩国人工智能安全研究所，同时引入私营自主的人工智能系统可靠性检测和认证体系；5月，韩国决定在电子通信研究院(ETRI)设立韩国人工智能安全研究所[23]。2024

[20] UK Government, “Seoul Statement of Intent toward International Cooperation on AI Safety Science”, 2024-05-21, https://www.gov.uk/government/news/global-leaders-agree-to-launch-first-international-network-of-ai-safety-institutes-to-boost-understanding-of-ai.
[21] AISI Japan, “Overview of the AI Safety Institute”, 2024-11-01, https://aisi.go.jp/assets/pdf/20241101_AboutAISI_en.pdf.
[22] Ministry of Science and ICT, “MSIT's Work Plan for 2024”, 2024-02-13, https://www.msit.go.kr/bbs/documentView.do?atchFileNo=44720&fileOrdr=2.
[23] 每日经济, “韩国决定在韩国电子通信研究院(ETRI)设立"AI安全研究所"”, 2024-05-22, https://www.mk.co.kr/cn/it/11022337.

年4月，加拿大总理杜鲁多宣布建立加拿大人工智能安全研究所，投资5000万加元，以防范潜在的安全风险[24]，并于2024年11月正式成立[25]。

在首尔峰会上，英美日韩加政府分享了各自人工智能安全研究所的进展和成果。此外，新加坡政府宣布位于南洋理工大学的国家级数字信任中心将作为新加坡人工智能安全研究所[26]。法国政府宣布成立人工智能评测中心(AI evaluation center)，由国立计算机及自动化研究院(Inria)和计量和测试实验室(LNE)合作开展人工智能系统安全研究和评测的工作[27]。欧盟表示，尽管2024年2月成立的欧洲人工智能办公室[28]名义上并非人工智能研究所，但将履行欧盟人工智能研究所的职责，其首要目的是支持《人工智能法案》并执行通用型人工智能系统规则，包括评估模型能力、调查可能的违规行为并要求提供商采取纠正措施。

2.2 人工智能安全研究所国际网络开展安全评测等方面国际合作

在2024年4月，美国和英国签署了一项人工智能安全合作备忘录[29]，宣布达成“人工智能安全科学合作伙伴关系”，双方计划：1）建立人工智能安全测试的常用方法，并分享其能力，以确保能够有效应对这些风险；2）在可公开访问的模型上至少进行一次联合测试演练；3）通过探索人工智能安全研究所之间的人员交流，充分利用集体的专业知识资源。

在首尔峰会上，美国商务部长宣布美国人工智能安全研究所将与世界各地的人工智能安全研究所和政府支持的科学部门合作，建立一个**全球性的人工智能安全研究网络**。这个网络建立在《首尔人工智能安全科学国际合作意向书》[30]的基础之上，将扩大美国先前与英国、日本、加拿大、新加坡的人工智能安全研究所以以及欧洲人工智能办公室的合作。这个网络旨在促进全球各国人民使用安全、可靠的人工智能系统，通过加强战略研究和公共成果的国际协作来实现

[24] Government of Canada, “Remarks by the Deputy Prime Minister on securing Canada's AI advantage”, 2024-04-07, https://www.canada.ca/en/department-finance/news/2024/04/remarks-by-the-deputy-prime-minister-on-securing-canadas-ai-advantage.html.
[25] Government of Canada, “Canada launches Canadian Artificial Intelligence Safety Institute”, 2024-11-12, https://www.canada.ca/en/innovation-science-economic-development/news/2024/11/canada-launches-canadian-artificial-intelligence-safety-institute.html.
[26] IMDA, “Digital Trust Centre designated as Singapore's AI Safety Institute”, 2024-05-22, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/factsheets/2024/digital-trust-centre.
[27] 法国宣布成立人工智能评测中心是在欧洲科技创新展览会VivaTech期间，与人工智能首尔峰会同期。
[28] the European Artificial Intelligence Office，“COMMISSION DECISION”，2024-01-24，https://ec.europa.eu/newsroom/dae/redirection/document/101625.
[29] UK Government, “UK & United States announce partnership on science of AI safety”, 2024-04-02, https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety.
[30] UK Government, “Seoul Statement of Intent toward International Cooperation on AI Safety Science, AI Seoul Summit 2024 (Annex)”, 2024-05-21, https://www.gov.uk/government/publications/seoul-declaration-for-safe-innovative-and-inclusive-ai-ai-seoul-summit-2024/seoul-statement-of-intent-toward-international-cooperation-on-ai-safety-science-ai-seoul-summit-2024-annex.

这一目标。为了加强合作，美国商务部和美国国务院于9月宣布，将于2024年11月20日至21日在旧金山共同主办**人工智能安全研究所国际网络及其他相关方的首次会议**。同时宣布的是，人工智能安全研究所国际网络的初始成员包括澳大利亚、加拿大、欧盟、法国、日本、肯尼亚、韩国、新加坡、英国和美国[31]。与5月的《首尔宣言》签署国相比，**新增了肯尼亚，但缺少了意大利和德国**。

此次会议以**技术工作会议**形式召开，讨论了三个亟待从国际协调中受益的高优先级议题：1）管理合成内容的风险，2）测试基础模型，3）对先进人工智能系统进行风险评估，旨在推动网络在2月份法国主办的人工智能行动峰会前夕的工作。已公布的成果包括[32]：

- **1）为人工智能安全研究所国际网络制定一致的使命宣言[33]。**就四个优先合作领域达成一致：开展人工智能安全研究、开发模型评测的最佳实践、促进解释先进人工智能系统测试等常见方法、促进全球包容和信息共享。美国人工智能安全研究所将担任人工智能安全研究所国际网络的首任主席，网络成员将在会议上讨论治理、结构和会议节奏的更多细节。该网络还将讨论优先事项以及为 2025 年 2 月在巴黎举行的人工智能行动峰会及以后的持续工作制定的路线图。
- **2）国际网络关于减轻合成内容风险的新联合研究议程。**优先研究课题包括了解当前数字内容透明技术的安全性和稳健性、探索新颖和新兴的数字内容透明方法，以及开发模型保障措施以防止有害合成内容的生成和分发。国际网络研究议程鼓励采用多学科方法，包括技术缓解以及社会科学和人文评估，以确定问题和解决方案。 政府机构和几家慈善机构已承诺共计投入超过1100万美元来推动这项重要研究。
- **3）该国际网络首次联合测试演习对多语言、国际人工智能测试工作的方法论见解。**由美国、英国和新加坡的人工智能安全研究所的技术专家领导下，完成了首次联合测试演练。此次演练在Llama 3.1 405B上进行，测试了一般学术知识、“封闭域”幻觉和多语言能力三个主题，试点测试过程中的经验[34]也将为未来跨国测试和评估最佳实践奠定基础。

[31] U.S. Department of Commerce，“U.S. Secretary of Commerce Raimondo and U.S. Secretary of State Blinken Announce Inaugural Convening of International Network of AI Safety Institutes in San Francisco”, 2024-09-18, https://www.commerce.gov/news/press-releases/2024/09/us-secretary-commerce-raimondo-and-us-secretary-state-blinken-announce.

[32] NIST，“FACT SHEET: U.S. Department of Commerce & U.S. Department of State Launch the International Network of AI Safety Institutes at Inaugural Convening in San Francisco”, 2024-11-20, https://www.nist.gov/news-events/news/2024/11/fact-sheet-us-department-commerce-us-department-state-launch-international.

[33] International Network of AI Safety Institutes，“Mission statement”, 2024-11-20, https://www.nist.gov/document/international-network-ai-safety-institutes-mission-statement.

[34] International Network of AI Safety Institutes，“Improving International Testing of Foundation Models: A Pilot Testing Exercise from the International Network of AI Safety Institutes”, 2024-11-20, https://www.nist.gov/document/aisi-pilot-testing-exercise-blog.

- **4）关于先进人工智能系统风险评估的联合声明，包括推进国际网络协调的计划。**因认识到人工智能风险评估科学在不断发展，且各网络成员都在自己独特的环境中运作，网络成员同意以六个关键方面为风险评估建立一个共享的科学基础[35]，即风险评估应具有可操作性、透明性、全面性、多利益相关方性、迭代性和可重复性。
- **5）建立由美国人工智能安全研究所牵头的新的美国政府工作组，合作研究和测试人工智能模型，以管理国家安全能力和风险。**美国国家安全人工智能风险测试 (Testing Risks of AI for National Security, TRAINS) 工作组[36]汇集了商务部、国防部、能源部、国土安全部以及国家安全局 和国立卫生研究院的专家，以解决国家安全问题并加强美国在人工智能创新方面的领导地位。该工作组将在关键的国家安全和公共安全领域（例如放生放核、网络安全、关键基础设施、常规军事能力等）协调研究和测试先进的人工智能模型。

如果说首届英国人工智能安全峰会的成果之一是提出了国家级人工智能安全研究所的构想，那么第二届首尔峰会则标志着这一构想作为一项国际合作取得了重要进展。然而，美国主办的人工智能安全研究所国际网络首届会议未邀请中国参与，这可能会形成一个不理想的先例。

中国的参与对于实现有效的全球治理至关重要。在生成式人工智能监管方面，中国积累了丰富的经验和措施，率先出台《生成式人工智能服务管理暂行办法》等法规，对人工智能生成的内容进行了明确约束。《人工智能生成合成内容标识办法（征求意见稿）》的发布进一步巩固了中国在这一领域的领先地位。

在人工智能安全研究和评测方面，中国同样走在前列。过去六个月，中国研究人员平均每月发表近15篇前沿技术论文，已有超过十个研究团队专注于该领域。此外，至少有四家政府支持的机构——包括上海人工智能实验室、北京智源人工智能研究院、中国信息通信研究院和北京通用人工智能研究院——正开展涵盖偏见、隐私、抵抗对抗性和越狱攻击的能力、机器伦理以及网络攻击滥用等领域的全面评测[37]。

将主要人工智能强国中国排除在更好的安全实践和科学理解之外，不仅不利于全球协调，还可能从根本上削弱全球人工智能安全治理的成效。

---

[35] International Network of AI Safety Institutes, “Joint Statement on Risk Assessment of Advanced AI Systems”, 2024-11-20, https://www.nist.gov/document/joint-statement-risk-assessment-advanced-ai-systems-international-network-aisis.

[36] NIST, “U.S. AI Safety Institute Establishes New U.S. Government Taskforce to Collaborate on Research and Testing of AI Models to Manage National Security Capabilities & Risks”, 2024-11-20, https://www.nist.gov/news-events/news/2024/11/us-ai-safety-institute-establishes-new-us-government-taskforce-collaborate.

[37] Concordia AI, “China's AI Safety Evaluations Ecosystem”, 2024-09-13, https://aisafetychina.substack.com/p/chinas-ai-safety-evaluations-ecosystem.

3 对比

3.1 机构属性与投入规模

AISI	成立时间	机构设立	上级政府部门	资金规模	主要负责人+人员规模
英国	2023年11月	新成立机构(UK AISI)	科学、创新及科技部(DSIT)	初始资金1亿英镑，并承诺维持其资金至2030年，优先使用超15亿英镑的英国AI研究和算力资源	- 主席: Ian Hogarth(投资人和企业家) - 主任: Oliver Ilott(曾领导首相国内办公室) - CTO: Jade Leung(曾领导OpenAI治理团队) - 研究总监: Geoffrey Irving, Yarin Gal, Chris Summerfield(曾领导OpenAI、DeepMind和牛津大学的AI安全团队) - 超过40名技术人员和60名政策/运营人员
美国	2023年11月	新成立机构(US AISI)	商务部(DOC)下属的非监管机构美国国家标准与技术研究院(NIST)	1000万美元	- 总监: Elizabeth Kelly(曾任总统经济政策特别助理) - CTO: Elham Tabassi(NIST首席AI顾问) - AI安全主管: Paul Christiano(曾领导METR、OpenAI对齐团队)
日本	2024年2月	新成立机构(AISI Japan)	内阁办公室设立部委和机构理事会负责审议AISI重要事项	-	董事/执行董事：村上明子(兼职，日本财产保险株式会社首席数据官)
新加坡	2024年5月	现有机构(由南洋理工国家数字信任中心(DTC)承担)	通讯及信息部(MCI)信息通信媒体发展局(IMDA)	5000万新元	执行主任：林国恩(兼职，南洋理工大学副校长)
加拿大	2024年11月	新成立机构(CAISI)	创新、科学及经济发展部(ISED)	5000万加元	-
韩国	已官宣，暂未正式成立	新成立机构(决定设在韩国电子通信研究院(ETRI))	科学技术情报通信部(MSIT)	-	-
法国	2024年5月	新成立机构(AI评测中心)	由国立计算机及自动化研究院(Inria)和计量和测试实验室(LNE)合作	-	-
欧盟	2024年5月	现有机构(由欧洲AI办公室承担)	欧盟委员会的通信网络、内容和技术总司(DG CNECT)	AI办公室整体初始预算4650万欧元，从现有预算中分配	办公室负责人：Lucilla Sioli A3.AI安全单位负责人：尚未任命

3.2 职能定位与工作内容

AISI	机构目标	职能定位	关注的风险	AI安全研究	AI安全评测	促进信息交流 (广义的AI安全标准)
英国	目标： 1.测试先进AI系统并向政策制定者通报其风险 2.促进公司、政府和更广泛研究界间的合作，以降低风险并推进有益公众的研究 3.加强全球AI发展实践和政策制定	职能： 1.开发和开展先进AI系统的评测 2.推动基础AI安全研究 3.促进信息交流	优先关注： 1.滥用(生物和化学能力+网络攻击能力) 2.社会影响 3.自主体失控 4.保障(safeguards)失效	项目包括： 1.构建AI治理产品 2.提升评测科学 3 更安全AI系统的新方法 例如系统性AI安全资助	获多个前沿模型早期或优先访问权限，并开展评测 已公布评测方法和评测结果 1.自动化能力评估 2.红队测试 3.人类能力提升评测 4.自主体评测	推动全球AI治理对话 1.举办AI安全峰会 2.与美国和加拿大AISI合作测试、分享见解、共享模型访问、人才流动 3.委托《先进人工智能安全国际科学报告》
美国	目标： 1.推进AI安全科学 2.阐明、展示和传播AI安全实践 3.支持围绕AI安全的机构、社区和协调	职能： 1.推动AI安全的科学研究和测量科学 2.开展模型和系统的安全评测 3.制定评测和风险缓解指南 4.协作和标准制定	广泛风险： 1.国家安全 2.公共安全 3.个人权利 其中也包括前沿AI、管理两用基础模型的滥用风险	推进AI安全的研究和测量科学	获OpenAI和Anthropic部署前评测权限	1.成立AI安全研究所联盟，汇集280多个组织，旨在制定基于科学和实证支持的测量科学、自愿指南和严格测试 2.与英国AISI合作测试、分享见解、共享模型访问、人才流动
日本	目标： 1.支持公共和私营部门，共同确保参与AI开发和使用的各方都充分认识到AI的风险。在全生命周期内确保治理，促进安全使用AI 2.需要在这些努力中促进创新并降低生命周期中的风险	职能： 1.AISI通过开展AI系统安全调查、研究评估方法和制定标准来支持政府 2.作为日本AI安全的枢纽，AISI将整合行业和学术界的最新信息，并促进相关公司和组织之间的合作 3.与AI安全相关组织合作	技术调查： 1.虚假信息 2.AI与网络安全	(AISI Japan只技术调查，不是研发组织)	技术调查： 1.评估视角、红队测试 2.测试环境	1.考虑AI安全相关的标准和指南 2.日本信息规划和研究委员会推动AI标准化 3.与各国AISI和相关组织合作，如日美Crosswalk分别就AI RMF和红队进行了对照 4.设立“J-AISI合作伙伴”计划
新加坡	目标： 整合新加坡的研究生态，与其他AISI开展国际合作，推动AI安全科学，并为AI治理工作提供基于科学的输入	职能： 1.测试与评估 2.安全模型设计、开发和部署 3.内容保证 4.治理与政策	-	-	扩大AI Verify以测试生成式AI安全性，推出的Project Moonshot开源工具包	1.与美国和英国AISI合作AI安全研究，侧重生成式AI评测 2.发布《东盟AI治理与伦理指南》 3.IMDA和AI Verify已与Anthropic合作开展多语言红队
加拿大	目标： 与国际合作伙伴	职能： 利用加拿大AI研	专注于： 先进AI的合成	两种方式： 1. 应用和研究	将与国际同行进行联合测试	开展联合项目，包括制定人工智能安

国家级人工智能安全研究所及其国际网络

AISI	机构目标	职能定位	关注的风险	AI安全研究	AI安全评测	促进信息交流 (广义的AI安全标准)
	合作，推进AI安全科学，以确保政府能够充分了解和应对先进AI系统的风险	究社区以及内部专业知识，与业界和国际AI安全研究所网络合作开展前沿研究	内容风险，以及开发或部署可能存在危险或妨碍人类监督的系统带来的风险	员主导的研究 2. 政府指导项目		全指南等
韩国	-	-	-	-	-	将加强AISI间的合作，采取识别水印等AI生成的内容的措施，并加强为开发国际标准的合作
法国	目标： 1.专注于AI安全的研究和创新 2.为欧盟《AI法案》范围内的系统开发新的测试和测试基础设施 3.组织通用型AI模型的定期评测活动	职能侧重研发： 1.对国内：评测所有类型AI系统(软件解决方案或嵌入式系统、各类工业应用)的国家参考机构 2.对欧盟：希望成为欧洲AI办公室的一个分支机构，提供模型评测的技术专长，而欧洲AI治理机制的其他部分(例如欧洲AI办公室)将负责国际协调和标准制定	-	-	-	目标是建立一个让所有机构能够共同合作的架构，分工协作，避免重复浪费
欧盟	目标： 1.整个欧盟AI专业知识的中心 2.在实施《AI法案》中发挥关键作用，尤其是通用型AI(General purpose AI) 3.促进可信AI的开发、使用和国际合作	职能： 由《AI法案》赋予监管权力，包括对通用型AI模型进行评测、向模型提供者索取信息和措施及实施制裁的权力	专注于： 具有系统性风险的通用型AI模型评测，并与行业和其他利益相关者合作，确定系统性风险和适当的缓解措施	-	-	-

3.3 领先的人工智能安全研究所案例分析

3.3.1 英国人工智能安全研究所（UK AISI）

1）机构沿革和定位

- 前身是前沿人工智能工作组(Frontier AI Taskforce)[38]，于2023年4月作为基础模型工作组(Foundation Model Taskforce)[39]启动，并在英国首届全球人工智能安全峰会上正式确立为人工智能安全研究所。
- 是全球首个由国家支持的、致力于公共利益的先进人工智能安全机构[40]，其使命是让政府对先进人工智能系统的安全性有实证的了解[41]。
- 被设计为政府内的初创企业[42]，将政府的权威与企业的专业知识和敏捷性相结合。

2）人工智能安全评测：一项重要工作是定期评测先进人工智能系统的潜在危害[43]

- **关注的前沿风险类别**
  - **滥用：**评估先进的人工智能系统在多大程度上有效降低了试图在现实世界造成伤害的恶意行为者的门槛。特别关注化学生物能力和网络攻击能力这两个子方向，被认为若不加以控制可能会造成大规模伤害的风险。
  - **社会影响：**评测先进人工智能系统对个人和社会的直接影响，包括人类与此类系统互动时受影响的程度，以及系统在专业环境中用于执行的任务类型。
  - **自主体失控：**评测在线半自主部署的先进人工智能系统的能力，此类系统会采取影响现实世界的行动。包括在线创建自身副本、说服或欺骗人类，以及创建比自身更强大的人工智能系统或模型的能力。
  - **保障失效：**评测先进人工智能系统的安全组件针对可能规避其保障措施的各种威胁的强度和有效性。
- **安全评测方法**

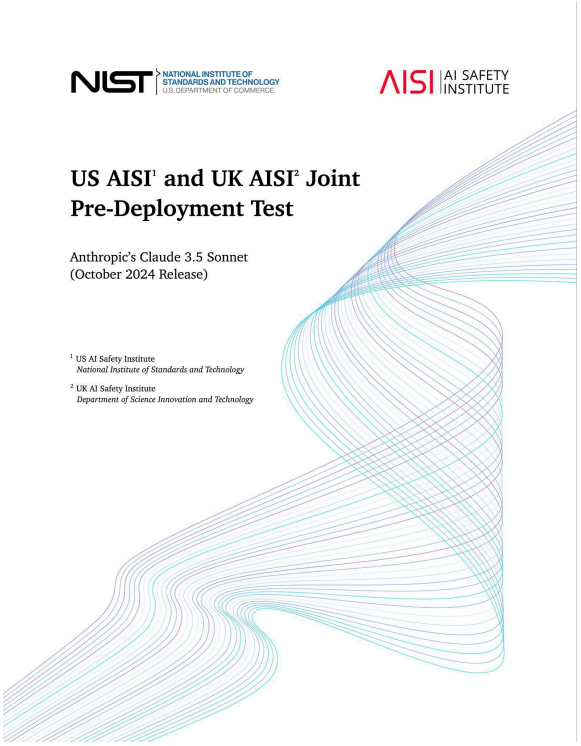
---

[38] DSIT (UK)，“Frontier AI Taskforce: first progress report”, 2023-09-07, https://www.gov.uk/government/publications/frontier-ai-taskforce-first-progress-report/.
[39] UK Government，“Initial £100 million for expert taskforce to help UK build and adopt next generation of safe AI”, 2023-04-24,https://www.gov.uk/government/news/initial-100-million-for-expert-taskforce-to-help-uk-build-and-adopt-next-generation-of-safe-ai.
[40] DSIT (UK)，“Introducing the AI Safety Institute”, 2024–01-17, https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.
[41] UK AI Safety Institute，“Our work: Improving our understanding of advanced AI”, 2024-11-25(引用日期), https://www.aisi.gov.uk/work.
[42] UK AI Safety Institute，“About: Our mission is to equip governments with an empirical understanding of the safety of advanced AI systems”, 2024-11-25(引用日期), https://www.aisi.gov.uk/about.
[43] DSIT (UK)，“AI Safety Institute approach to evaluations”, 2024-02-09, https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations.

- **自动化能力评估：**开发与安全相关的问题集，以测试模型能力并评估不同先进人工智能系统的答案差异。这些评估可以是广泛但浅显的工具，可为模型在特定领域的能力提供基线指示，用于指导更深入的调查。
- **红队测试：**安排大量领域专家花时间与模型互动，测试其功能并破解模型的保护措施。基于从自动化能力评估中发现的信息，这些信息可以为人工智能安全研究所的专家在能力和模态方面指明正确的方向。
- **人类能力提升评测(Human uplift evaluations)：**评测与使用互联网搜索等现有工具相比，恶意行为者可能如何使用先进人工智能系统执行现实生活中的有害任务。希望针对关键领域进行这些严格的研究，以对模型对恶意行为者能力的反事实影响进行有依据的评估。
- **自主体评测：**评测自主体是否具有可以制定长期计划、半自主运行并使用网络浏览器和外部数据库等工具等能力。因为随着这种自主能力和在现实世界采取行动的能力提高，造成危害的可能性也随之增大。

- **前沿模型的早期或优先访问权限**
  - 时任英国首相苏纳克宣布[44]：英国人工智能安全研究所已与OpenAI、Anthropic、Deepmind达成合作，获得其前沿模型的早期或优先访问权限。
  - Anthropic：英国人工智能安全研究所获取并进行了Claude 3.5 Sonnet的部署前测试，并与美国人工智能安全研究所分享了测试结果[45]。
  - 但这些测试不是“政府安全认证”，并不作为某个特定模型安全性的认可[46]。
- **评测结果分享**
  - 2024年11月，UK AISI和US AISI联合发布Anthropic 升级版 Claude 3.5 Sonnet 的部署前联合评测报告[47]。

[44] Politico，“OpenAI, DeepMind will open up models to UK government”, 2023-06-12, https://www.politico.eu/article/openai-deepmind-will-open-up-models-to-uk-government/.
[45] Anthropic，“Announcements：Claude 3.5 Sonnet”, 2024-06-21, https://www.anthropic.com/news/claude-3-5-sonnet.
[46] UK AI Safety Institute，“Our First Year”, 2024-11-13, https://www.aisi.gov.uk/work/our-first-year.
[47] UK AI Safety Institute，“Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet”, 2024-11-19, https://www.aisi.gov.uk/work/pre-deployment-evaluation-of-anthropics-upgraded-claude-3-5-sonnet



- 此前，英国人工智能安全研究所测试了领先的模型的网络攻击能力、化学和生物能力、自主体能力以及保障措施的有效性。其2024年5月公布的第一篇技术博客分享了他们的方法和结果[48]，所有LLM模型均为匿名。

AISI AI SAFETY INSTITUTE

Cyber capabilities of advanced AI models

We evaluated 4 leading models' rate of completing Capture the Flag (CTF) challenges:

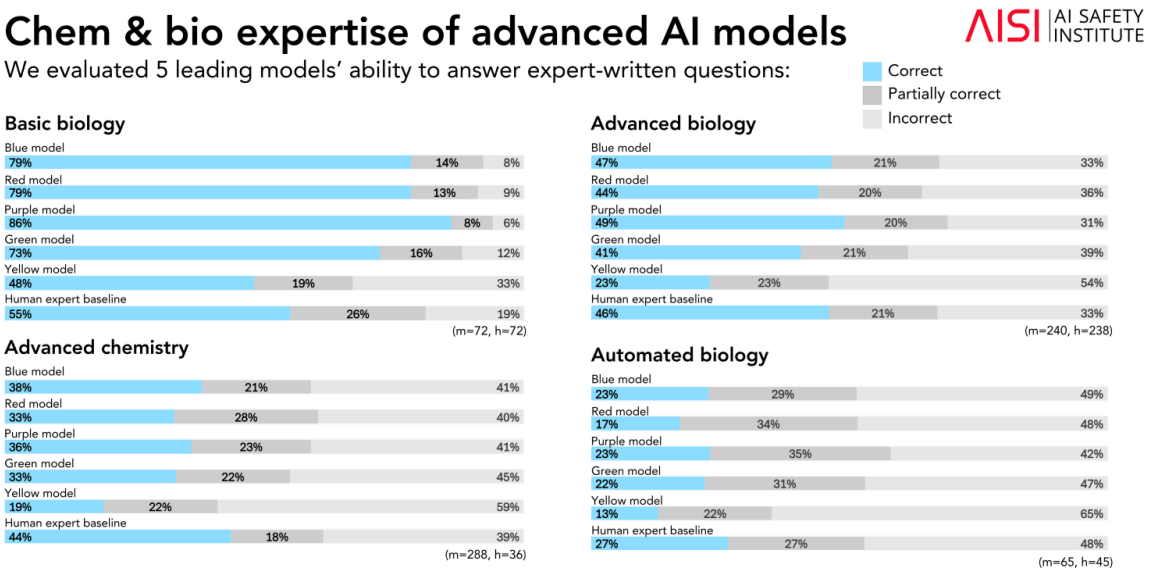
CTF difficulty	Skill assessed	Red model	Purple model	Blue model	Green model	# of CTFs
High school level (PICO CTFs, generalist scaffold)	Forensics	43%	43%	35%	13%	23
	Cryptography	50%	56%	61%	6%	18
	Reverse Engineering	83%	83%	83%	25%	24
	General Skills	100%	100%	76%	24%	17
University level (CSAW CTFs, CTF scaffold)	Forensics	0%	0%	0%	not applicable	4
	Cryptography	0%	0%	0%	not applicable	2
	Reverse Engineering	50%	50%	75%	not applicable	4
	General Skills	0%	0%	0%	not applicable	2
AISI-designed CTF (generalist scaffold)	Forensics	38%	38%	50%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2
AISI-designed CTF (CTF scaffold)	Forensics	75%	50%	63%	not applicable	8
	Cryptography	0%	0%	0%	not applicable	2

Finding: Several LLMs completed simple cyber security challenges aimed at high-school students but struggled with challenges aimed at university students.

Figure 1

多个LLM完成了针对高中生的简单网络安全挑战，但在针对大学生的挑战中遇到了困难。

[48] UK AI Safety Institute, "Advanced AI evaluations at AISI: May update", 2024-05-20, https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update,



Finding: Several LLMs demonstrated expert-level knowledge of chemistry and biology. Models answered over 600 private expert-written chemistry and biology questions at similar levels to humans with PhD-level training.

Figure 2

多个LLM展示了专家级别的化学和生物学知识。模型回答了600多个私密的专家撰写的化学和生物学问题，其水平与接受过博士级培训的人类相似。

AISI AI SAFETY INSTITUTE

Advanced AI models' capabilities for autonomously performing tasks

We evaluated 3 leading models' ability to complete tasks on digital platforms:

		Purple model	Blue model	Green model	# of tasks	# of variants
Short-horizon (tasks that take an expert <1h)	Software engineering	39%	22%	3%	6	54
	Generalist research	26%	20%	7%	3	171
Long-horizon (tasks that take an expert >4h)	Software engineering AI R&D	0%	0%	not applicable	8	13

Finding: Two LLMs completed short-horizon agent tasks (such as simple software engineering problems) but were unable to plan and execute sequences of actions for more complex tasks.

Figure 3

多个LLM完成了短期自主体任务（例如简单的软件工程问题），但无法规划和执行更复杂任务的操作序列。

AISI AI SAFETY INSTITUTE

Effectiveness of safeguards on advanced AI models

We evaluated 4 leading models' vulnerability to AISI-designed jailbreak attacks:

		Red model	Purple model	Blue model	Green model	# of questions
No attack	Compliance with private harmful questions	8%	15%	1%	28%	113
	Correctness on private benign questions	50%	59%	57%	51%	150
AISI-designed attack, 1 attempt	Compliance with private harmful questions	90%	56%	100%	99%	113
	Compliance with HarmBench questions	75%	52%	96%	96%	140
	Correctness on private benign questions	51%	55%	58%	53%	150
AISI-designed attack, 5 attempts	Compliance with private harmful questions	100%	98%	100%	100%	113
	Compliance with HarmBench questions	99%	90%	100%	100%	140

Finding: All tested LLMs remain highly vulnerable to basic jailbreaks. Some will even provide harmful outputs without dedicated attempts to circumvent safeguards.

Figure 4

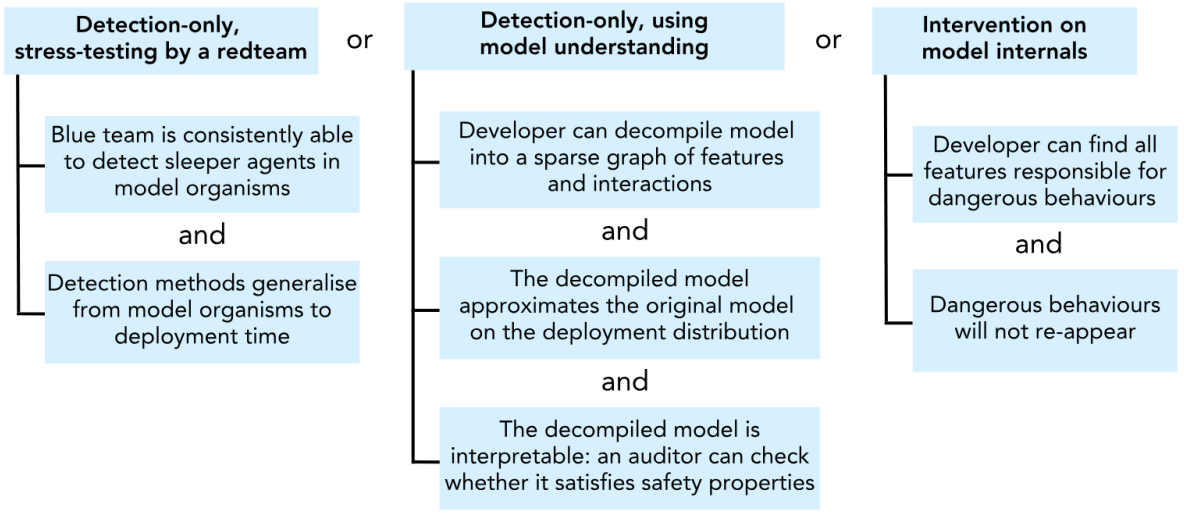
所有经测试LLM仍极易受到基本越狱的攻击，有些甚至会在没有专门尝试规避其安全措施的情况下产生有害输出。

3）人工智能安全研究：开展一系列合作和研究，作为对前沿人工智能模型实证评测的补充

- 例如使人工智能系统从根本上更安全以及提高社会对先进人工智能韧性的研究。
- 最新方向为安全论证(Safety Cases)[49]：是一系列证据支持的一种结构化论证，通过提供令人信服、易于理解且有效的论据，证明系统在特定应用和特定环境中的安全性。

AISI AI SAFETY INSTITUTE

Safety case sketches for mechanistic interpretability



示意图：可解释性可仅用于对齐错误检测，使用红队或其他证据来支持准确性（左图和中图），或作为消除对齐错误的缓解方法（右图）

[49] UK AI Safety Institute, "Safety cases at AISI", 2024-08-23, https://www.aisi.gov.uk/work/safety-cases-at-aisi.

4）促进信息交流：推动人工智能治理的全球对话，设定全球标准

- **人工智能安全峰会：**英国人工智能安全研究所为历届峰会做出了贡献。峰会将多国领导人、顶级人工智能公司和民间社会聚集在一起，做出重要承诺以降低风险。
- **与美国[50]、加拿大[51]和新加坡[52]的人工智能安全研究所合作：**共同测试先进人工智能模型、分享研究见解、共享模型访问权限，并实现专家人才之间的交流。
- **《先进人工智能安全国际科学报告》[53]：**委托图灵奖得主Yoshua Bengio主持这份关于基于证据的先进人工智能安全科学现状的报告。
- **20多个顶级研究机构合作伙伴[54]：**多家机构专注于前沿人工智能安全特定领域
    - METR （危险能力评测）
    - RAND（危险能力评测）
    - Redwood Research（危险能力评测）
    - Gryphon Scientific（生物安全）
    - FutureHouse（生物安全+人工智能科学家）
    - Apollo Research（欺骗评测）
    - Trail of Bits（网络安全）
    - Advai（第三方评测）
    - The Center for AI Safety（人工智能安全研究和社区建设）
    - Collective Intelligence Project（变革性技术的治理）
    - Faculty（风险管理）
    - OpenMined（开源人工智能治理基础设施）
    - Fuzzy Labs（开源机器学习运维）
    - Pattern Labs（安保）
    - ……

3.3.2 美国人工智能安全研究所（US AISI）

1）机构沿革和定位

- 根据拜登-哈里斯政府2023年发布的《关于安全、可靠和可信开发与使用人工智能的行

[50] UK Government, “UK & United States announce partnership on science of AI safety”, 2024-04-02, https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety.
[51] UK AI Safety Institute, “Fourth progress report”, 2024-05-20, https://www.aisi.gov.uk/work/fourth-progress-report.
[52] UK Government, “Ensuring trust in AI to unlock £6.5 billion over next decade”, 2024-11-06, https://www.gov.uk/government/news/ensuring-trust-in-ai-to-unlock-65-billion-over-next-decade.
[53] 安远AI, “AI Guard x 安远AI｜30多国75位顶尖专家合作发布《先进人工智能安全国际科学报告：中期报告》”, 2024-05-17, https://mp.weixin.qq.com/s/r5JptTXKgpagLRopdbTdow.
[54] UK Government, “AI Safety Institute: third progress report”, 2024-02-05, https://www.gov.uk/government/publications/uk-ai-safety-institute-third-progress-report/.

政令》[55](简称《人工智能行政令》)，在2023年11月英国首届人工智能安全峰会上正式成立，隶属于美国商务部下属的美国国家标准与技术研究院(NIST)。

- 聚焦3大目标：1）推进人工智能安全科学；2）阐明、展示和传播人工智能安全实践；3）支持围绕人工智能安全的机构、社区和协调。
- 初期工作聚焦于拜登总统《人工智能行政令》分配给国家标准与技术研究院的优先事项[56]，2024年5月发布《美国人工智能安全研究所：远景、使命和战略目标文档》[57]。
- 2024年11月，美国人工智能安全研究所成立国家安全人工智能风险测试 (TRAINS) 的政府工作组，汇集了商务部、国防部、能源部、国土安全部、国家安全局和国立卫生研究院的专家，合作研究和测试人工智能模型，以管理国家安全能力和风险。

美国人工智能安全研究所的三大战略目标

2）战略目标 1. 让愿景可能：通过研究推进人工智能安全科学

- **美国人工智能安全研究所倡导开发基于实证的人工智能模型、系统和自主体的测试、基准和评测，以找到应对近期和长期人工智能安全挑战的实用解决方案。包括：**
  - **执行和协调技术研究，以改进或制定安全指南及技术安全的工具和方法。**如用于检测合成内容的技术、模型安全的最佳实践，以及在模型、系统和自主体层面的技术防护和缓解措施。这些项目可能涉及基础研究和应用研究，对于应用研究，计划利用内部和外部的基础研究，以及现有的指南、方法和标准。
  - **对先进模型、系统和自主体进行部署前的测试、评测、验证与确认 (Testing, evaluation, validation, and verification, TEVV)** 以评估潜在和新兴的风险。评测方法包括自动化能力评估、专家红队测试、A/B测试等。计划与美国国家标准与技术研究院实验室合作，进行部署前对现有危害以及潜在和新兴风险的评估。

[55] The White House, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”, 2023-10-30,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/.
[56] NIST, “U.S. Artificial Intelligence Safety Institute”, 2024-04-16, https://www.nist.gov/aisi.
[57] U.S. Artificial Intelligence Safety Institute, “Strategic Vision”, 2024-10-01, https://www.nist.gov/aisi/strategic-vision.

- **对先进的人工智能模型、系统和自主体进行TEVV，以发展对一系列现有风险的科学理解和记录。**计划与美国国家标准与技术研究院实验室项目合作，加深对如何衡量与当今能力相关的风险的科学理解，包括个人权利、公共安全和国家安全。

3）战略目标 2. 让愿景可行：开发和传播人工智能安全实践

- **美国人工智能安全研究所计划启动以下项目：**
  - **构建和发布特定指标、评估工具、方法指南、协议和基准，以评估不同领域和部署环境中的先进人工智能风险。**美国人工智能安全研究所计划发布针对开发人员和部署人员的不同风险的TEVV的指南和工具，包括针对一系列风险的TEVV特定评估协议，以告知和支持开发人员、部署人员和第三方独立评估人员。这些指南可以提供建议，并制定新的基准来评估模型能力。
  - **制定并发布基于风险的缓解指南和安全机制，以支持先进人工智能模型、系统和自主体的负责任的设计、开发、部署、使用和治理。**计划为这些指南提供缓解现有危害以及潜在和新出现风险的指导，包括公共安全和国家安全；针对最先进人工智能系统的与风险成比例的安全和安全缓解措施；以及基于研究所的研究而开发的内部和外部安全机制或工具。
- **关注的前沿风险类别：**
  - 关注包括个人权利、公共安全和国家安全的广泛风险。
  - 其中前沿风险包括：两用基础模型滥用，化学、生物或网络攻击等危险能力，人类丧失监督或控制权等风险。
- **前沿模型的部署前评测权限：**
  - 包括为前沿模型创建测试基准以及制定评估系统性风险的指南。
  - 2024年8月，美国人工智能安全研究所与OpenAI和Anthropic达成关于安全研究和评测的协议，获得两家公司新模型发布之前和之后的访问权限[58]。

4）战略目标 3. 让愿景可持续：支持围绕人工智能安全的机构、社区和协调

- **美国人工智能安全研究所计划启动以下项目：**
  - **促进人工智能安全研究所指南、评测和推荐的人工智能安全和风险缓解措施的采用。**为最大限度地提高人工智能安全研究所指导的价值和可用性，美国人工智能安全研究所计划适时启动并支持与安全研究实验室、第三方评测机构以及

[58] NIST, “U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI”, 2024-08-29, https://www.nist.gov/news-events/news/2024/08/us-ai-safety-institute-signs-agreements-regarding-ai-safety-research.

开发人员、部署人员和用户中多元专业人士的持续对话、信息共享和协作。项目旨在将自愿承诺转化为可操作的指南，并促进人工智能安全最佳实践采用，同时寻求促进一个强大的第三方评测生态。项目可能会贡献科学报告、文章、指导和实践，以帮助确保严格的人工智能安全研究、测试和指导为重大国内人工智能安全立法或政策提供信息支持。项目还将提升人们对为相关研究工作提供的人工智能安全实践的认识。

  - **领导一个包容性的人工智能安全国际科学网络。**人工智能安全实践必须尽可能全球化采用。美国人工智能安全研究所打算成为其他人工智能安全研究所、国家研究组织和OECD和G7等多边实体的合作伙伴，与其合作伙伴共同推动广泛接受的科学方法，旨在开发共享和互操作的人工智能安全评估及达成共识的风险缓解措施。旨在为未来国际人工智能治理安排的发展奠定科学和实践基础。

- **已推动的风险缓解的指南和相关机构合作包括：**
  - **《两用基础模型滥用风险管理指南(NIST AI 800-1)》初步公开草案**[59]：于2024年7月发布，这是NIST为响应《人工智能行政令》而发布的5个指南之一[60]，概述了基础模型开发者保护其系统不被滥用于故意伤害个人、公共安全和国家安全的自愿最佳实践，协助防止模型被用于开发生物武器、进行网络攻击等。
  - **与英国人工智能安全研究所合作**[61]**：**共同测试先进人工智能模型，分享研究见解，共享模型访问权限，并实现专家人才之间的交流。
  - **与新加坡人工智能安全研究所合作**[62]**：**推进人工智能安全科学，映射各自的生成式人工智能框架，并探索在测试、指南和基准方面的合作。
  - **与欧洲人工智能办公室展开技术对话**[63]**：**聚焦于合成内容的水印和内容认证、政府计算基础设施，以及人工智能的社会公益三个关键主题。
- **建立美国人工智能安全研究所联盟，汇集了280多个组织：**包括生成式人工智能风险管理、合成内容、能力评测、红队测试、安全与安保5个工作组[64]，初始成员涵盖：

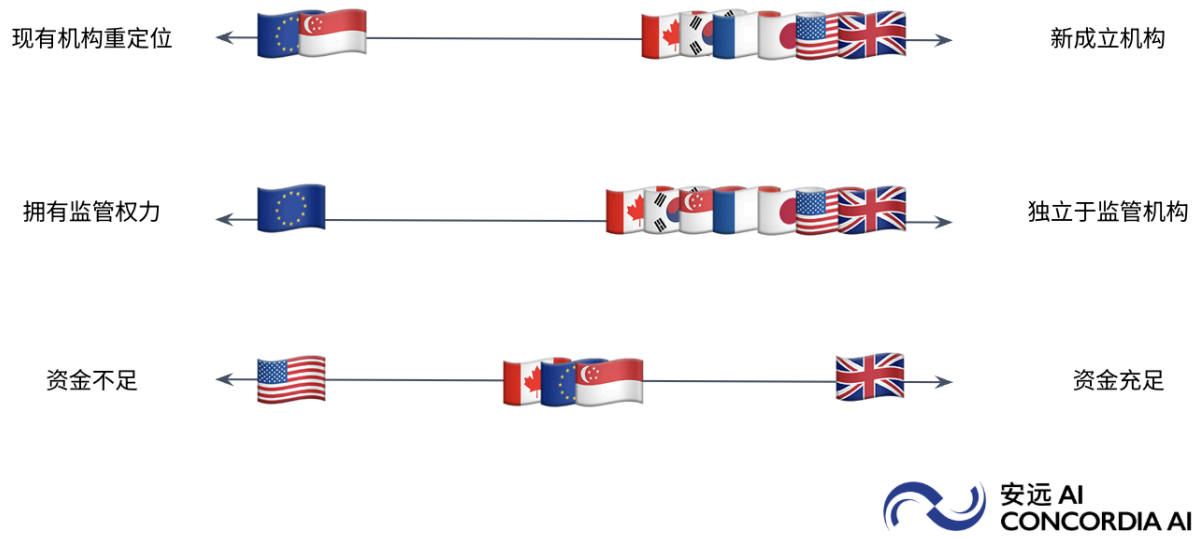
[59] NIST, “Managing Misuse Risk for Dual-Use 4 Foundation Models”, 2024-06, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.800-1.ipd.pdf.
[60] NIST, “Department of Commerce Announces New Guidance, Tools 270 Days Following President Biden's Executive Order on AI”, 2024-07-26, https://www.nist.gov/news-events/news/2024/07/department-commerce-announces-new-guidance-tools-270-days-following
[61] U.S. Department of Commerce, “U.S. and UK Announce Partnership on Science of AI Safety”, 2024-04-01, https://www.commerce.gov/news/press-releases/2024/04/us-and-uk-announce-partnership-science-ai-safety.
[62] U.S. Embassy in Singapore, “Fact Sheet: U.S.-Singapore Shared Principles and Collaboration on Artificial Intelligence”, 2024-06-05, https://sg.usembassy.gov/fact-sheet-u-s-singapore-shared-principles-and-collaboration-on-artificial-intelligence/.
[63] NIST, “U.S. AI Safety Institute and European AI Office Hold Technical Dialogue”, 2024-07-12, https://www.nist.gov/news-events/news/2024/07/us-ai-safety-institute-and-european-ai-office-hold-technical-dialogue.
[64] U.S. Artificial Intelligence Safety Institute, “AISIC Working Groups”, 2024-10-23, https://www.nist.gov/artificial-intelligence-safety-institute/aisic-working-groups.

- **前沿大模型公司和大型科技企业：**如OpenAI、Anthropic、Amazon、Apple、Meta、Google、Microsoft、IBM等，致力于开发并推动人工智能技术。
- **前沿人工智能安全研究机构：**如专注于前沿人工智能安全特定领域的METR、RAND、Redwood Research、Gryphon Scientific等研究机构。
- **高校与研究机构：**包括麻省理工学院、卡内基梅隆大学、普林斯顿大学、斯坦福大学等，提供学术支持与研究贡献。
- **政府部门和非营利组织：**如美国国防分析研究所、联邦安全委员会以及多个人工智能治理与安全研究中心，负责人工智能的政策制定、技术监管与治理。
- **其他科技联盟和企业等：**如Linux基金会旗下的AI & Data、AI Quality & Testing Hub、AI Risk and Vulnerability Alliance等，参与技术标准与方法的制定。

3.4 异同点小结

总体而言，现有的人工智能安全研究所有几个相似的主要职能：

- **人工智能安全评测：**目前所有人工智能安全研究所的工作方向都涉及模型评测，可以发挥的关键作用之一是改进评测工具。人工智能安全机构可以帮助评测先进人工智能系统的安全相关能力、系统的安保，及其潜在的社会影响。**其中，领先的英国和美国人工智能安全研究所已分别获得了多家前沿模型的早期或优先访问权限。**
- **人工智能安全研究：**虽然并非每个人工智能安全研究所都涉及人工智能安全的基础研究，但这些研究所可能在启动和支持人工智能安全基础研究方面发挥关键作用。这包括赞助探索性项目，并将来自不同学科的研究人员与学术和行业利益相关者聚集在一起。**因为前沿人工智能安全和评测是新兴的跨学科领域，开展基础研究对于推动科学进步非常关键。**
- **促进信息交流或推进标准制定：**均已签署《首尔人工智能安全科学国际合作意向书》。人工智能安全研究所可以通过建立国内和国际的信息共享渠道，传播人工智能技术的重要知识，促进政策制定者、产业界、学术界和公众之间的合作，确保政策制定者具备做出正确决策的充分信息。这些渠道还能帮助全球科学界在人工智能模型的能力、风险及评测方法上达成共识，并推动安全标准和治理政策的一致性。**不同的人工智能安全研究所可以结合自身目标和实际，侧重于标准制定，国内网络、国际对话的不同领域。**



现有人工智能安全研究所的差异，参考“The AI Safety Institute Network: Who, What and How?”[65] 修改重绘

但各国人工智能安全研究所因机构沿革和资源支持等差异，定位各有侧重：

- **英国人工智能安全研究所：拥有最多的全职技术人员，可开展深入评测**
  - 有充足的政府资金支持，吸纳了大量技术人才，希望引领前沿人工智能安全评测和研究。
  - 得到OpenAI、DeepMind、Anthropic的部署前评测授权。
  - 参与历届全球人工智能安全峰会的筹办。
  - 已开源评测框架Inspect[66]，为测试人员提供了评估各类模型特定能力的工具。
- **美国人工智能安全研究所：隶属商务部下属的国家标准与技术研究院，强调标准制定**
  - 在关注前沿人工智能风险的基础上，关注的风险类型更广泛。
  - 依托美国国家标准与技术研究院和合作网络，成立了人工智能安全研究联盟。
  - 获得OpenAI和Anthropic新模型发布之前和之后的访问权限。
  - 初期更关注国内安全问题，后通过与英国等AISI合作并宣布建立人工智能安全研究所国际网络后，越来越关注全球合作，旨在协调各方制定前沿人工智能的测量科学、自愿指南和严格测试标准。然而，特朗普当选新总统后，其全球合作前景存疑。
- **日本人工智能安全研究所：较强的标准化背景，强调与美国NIST的Crosswalk**
  - 未明确关注的风险，技术调查中提到虚假信息、人工智能与网络安全。

[65] International Center for Future Generations，“The AI Safety Institute Network: Who, What and How?”, 2024-09, https://icfg.eu/the-ai-safety-institute-network-who-what-and-how/#1725545464617-124fae77-dfb1.
[66] UK AI Safety Institute，“Inspect：An open-source framework for large language model evaluations“，2024-11-25(引用日期)，https://ukgovernmentbeis.github.io/inspect_ai/.

    - 不承担具体的人工智能安全研发工作。
    - 重视国际标准，已与NIST和美国人工智能安全研究所合作协调人工智能标准。
    - 发布《人工智能安全评测视角指南》[67]和《人工智能安全红队方法指南》[68]。
- **新加坡人工智能安全研究所：侧重研发，职责兼顾安全与发展**
    - 由南洋理工国家数字信任中心(DTC)发展而来，旨在“解决全球人工智能安全科学方面的差距”。
    - 职责范围更广，因此必须在安全与发展之间取得平衡。
    - 与新加坡国内其他治理机构（如IMDA和AI Verify）相互配合。
- **欧洲人工智能公室：由欧盟《人工智能法案》赋予执行监管的权力**
    - 关注通用型人工智能模型的系统性风险。
    - 具有对通用型人工智能模型进行评测、向模型提供者索取信息以及实施制裁的权力。
- **其他国家的人工智能安全研究所：目前的相关信息仍然有限**

[67] Japan AI Safety Institute，“Guide to Evaluation Perspectives on AI Safety”, 2024-09-25, https://aisi.go.jp/assets/pdf/ai_safety_eval_v1.01_en.pdf.
[68] Japan AI Safety Institute，“Guide to Red Teaming Methodology on AI Safety”, 2024-09-25, https://aisi.go.jp/assets/pdf/ai_safety_RT_v1.00_en.pdf.

4 挑战

人工智能安全研究所及其国际网络之间的合作能够带来显著益处，尤其是在技术工具和科学发现的交流方面。然而，涉及信息共享的领域可能面临诸多挑战，例如对敏感信息保密性和安全性的担忧、各国法律法规之间的不兼容性，以及各国在评估和理解先进人工智能模型方面的技术能力差异。有效解决这些问题对于充分释放人工智能安全治理的国际协调潜力至关重要。

未来各国人工智能安全研究所及其国际网络的协作，可能面临以下主要挑战：

1）模型访问与评测权限

各国人工智能研究所对前沿人工智能系统的评测，依赖于开发者是否提供足够的访问权限。

- **部署前评测权限**：英国人工智能安全研究所和美国人工智能安全研究所已分别获得了OpenAI、Anthropic等前沿模型的部署前评测权限，是否会有更多企业自愿效仿和落实[69]，并推广到更多人工智能安全研究所，仍有不确定性。
- **深入的模型访问**：人工智能安全研究所对前沿模型的访问权限也是一个关键因素。虽然企业自愿提供API[70]访问有帮助，更好的安全评估可能需要全面访问模型（包括微调前后的访问[71]、白盒和黑盒访问），这在不同国家和企业之间可能面临挑战。

2）信息共享与安全实践

国际合作离不开信息共享，如何实现信息的有效共享并确保其不会威胁国家安全是重要问题。

- **法规与技术的差异**：各国在法律规定、技术能力和对敏感数据的处理方式上存在显著差异。这种差异可能会导致信息共享困难，影响人工智能安全研究的深入理解和国际合作。
- **机密信息与共享实践**：由于部分评估信息可能涉及机密，人工智能安全研究所在信息共享时需严格遵循法律和安全规定，以确保数据的保密性和安全性。尽管合作备忘录可以帮助制定共享信息的最佳实践，但在实现信息保密与国际合作之间，仍然存在不少挑战。

[69] Politico，“Rishi Sunak promised to make AI safe. Big Tech's not playing ball.”, 2024-04-26, https://www.politico.eu/article/rishi-sunak-ai-testing-tech-ai-safety-institute/.
[70] OpenMined，“How to audit an AI model owned by someone else (part 1)”, 2023-07-01, https://blog.openmined.org/ai-audit-part-1/.
[71] University of Oxford，“Structured access for third-party research on frontier AI models: Investigating researchers' model access requirements”, 2023-10-27, https://www.oxfordmartin.ox.ac.uk/publications/structured-access-for-third-party-research-on-frontier-ai-models-investigating-researchers-model-access-requirements.

3）标准制定与监管框架

人工智能技术的快速发展和前沿特性使得快速应对成为必要，人工智能安全研究所需要在标准制定和监管框架中扮演重要角色。

- **标准制定的挑战**：传统标准制定流程可能不适用于快速发展的人工智能技术，因为这些流程倾向于提炼已知信息，而非前沿知识。此外，标准制定过程中可能缺乏国家安全领域的专家参与。虽然替代流程可能更快，但可能缺乏合法性和认可度。
- **监管框架的差异**：例如欧洲人工智能办公室拥有监管权力，可全面访问人工智能模型[72]，而其他地区可能缺乏类似的法律保障，这会影响跨国合作的效率和有效性，

4）资源差异与合作平衡

不同国家的资源差异可能导致全球人工智能安全能力的严重不均衡，此外维持独立性与合作平衡也是一大难题。

- **技术能力与资源差异**：作为政府资助的机构，人工智能安全研究所必须与资金充足的企业竞争以吸引顶尖工程师和科学家。例如英国人工智能安全研究所近期在旧金山设立了办公室[73]，以便更接近全球领先的人才基地，且此前已成功从大型公司招聘到顶级研究人员。这种竞争压力突显了不同国家人工智能安全研究所在技术能力上的显著差异。发达国家能够吸引顶尖技术人才，开发先进的评测工具，而发展中国家可能需要依赖外部支持或技术合作，这可能加剧全球人工智能安全能力的不平衡。此外，人工智能安全研究所的规模和功能各异，可能受到国内政治和资源限制的影响[74]，这可能限制其国际合作的客观性，特别是在涉及国家利益时，并可能导致资源分散，影响核心任务的执行。
- **独立性与合作平衡**：人工智能安全研究所需要在保持独立性和与私营企业、监管机构的合作之间取得平衡。如果人工智能安全研究所的专业判断影响法规，可能被视为监管机构，从而影响与私营企业和非政府利益相关者的合作。与此同时，人工智能安全研究所汇集技术专长并建立原本不存在的技术治理能力[75]，对政府机构也大有益处。因此，需要合理平衡人工智能安全研究所与私营企业、监管机构之间的合作与协调，确保其独立性和有效性。

[72] European Commission, “European AI Office”, 2024-11-25(引用日期), https://digital-strategy.ec.europa.eu/en/policies/ai-office

[73] UK Government, “Government's trailblazing Institute for AI Safety to open doors in San Francisco”, 2024-05-20, https://www.gov.uk/government/news/governments-trailblazing-institute-for-ai-safety-to-open-doors-in-san-francisco.

[74] Carnegie Endowment for International Peace, “The Future of International Scientific Assessments of AI's Risks”, 2024-08-27, https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks.

[75] Anka Reuel et al., “Open Problems in Technical AI Governance”, 2024-07-20, https://arxiv.org/abs/2407.14981.

5）全球包容性与国际协调

在发展人工智能安全研究所及其网络的过程中，确保全球包容性与国际协调至关重要。

- **全球包容性不足**：目前的人工智能安全研究所国际网络主要由少数富裕国家主导，仅有一个全球南方国家肯尼亚，并且美国主办的人工智能安全研究所国际网络首届会议未邀请中国参与，缺乏广泛的全球政治合法性。为了增强全球包容性和协调性，可能需要在更多国家资助和建立人工智能安全研究所，或者通过建立区域级人工智能安全研究所[76]集中资源，以便更多国家能够有效参与全球人工智能安全治理。然而，这些新兴机构的有效性和全球包容性仍存在不确定性。
- **国际协调的难题**：人工智能安全研究所之间的国际协调面临多重挑战。首先，各国人工智能安全研究所在资源和技术能力上存在差异显著，如英国人工智能安全研究所拥有技术优势和英美人工智能安全研究所合作备忘录[77]，而其他国家则可能缺乏必要的资源和模型访问权限。这种不平衡使得协调难度加大。其次，如何在保持基础结构和类似功能的同时，灵活应对各国的不同需求，是一个关键难题。此外，人工智能安全研究所之间实现认证互认的难度较大，这可能阻碍企业的跨国合规并影响全球人工智能安全标准的制定。

上述挑战表明，尽管人工智能安全研究所及其国际网络在安全评测、安全研究和国际合作中具有重要潜力，但未来仍需在模型访问与评测权限、信息共享与安全实践、标准制定与监管框架、资源差异与合作平衡、全球包容性与国际协调方面进行改进，以应对人工智能技术为全球治理带来的复杂挑战。

[76] University of Oxford，“AISIs' Roles in Domestic and International Governance”，2024-07, https://oms-www.files.svdcdn.com/production/downloads/academic/AISIs%20Roles%20in%20Governance%20Workshop.pdf

[77] UK Government，“UK & United States announce partnership on science of AI safety”，2024-04-02, https://www.gov.uk/government/news/uk-united-states-announce-partnership-on-science-of-ai-safety.