2024 年第 1 期（总第 11 期）

人工智能与国际安全研究动态

ARTIFICIAL INTELLIGENCE AND INTERNATIONAL SECURITY STUDIES REVIEW

人工智能安全研究所国际网络会议前瞻

清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

战略与安全
SECURITY & STRATEGY
清华大学战略与安全研究中心
CENTER FOR INTERNATIONAL SECURITY AND STRATEGY TSINGHUA UNIVERSITY

人工智能安全研究所国际网络会议前瞻

编者按：为推进人工智能与国际安全领域的相关研究，清华大学战略与安全研究中心（CISS）组织研究团队定期跟踪最新国际研究动态，重点关注人工智能应用对国际安全带来的风险挑战，并针对人工智能安全领域国际动态、智库报告、学术论文等资料进行分析。本文是CISS推出的人工智能与国际安全研究动态第11期，主要分析美国举办的人工智能安全研究所国际网络会议。

2024 年 11 月 21-22 日，美国将举办人工智能安全研究所国际网络（International Network of AI Safety Institutes）第一次会议，汇聚全球各地政府代表、企业高管和学术界人士，推进全球合作，促进人工智能安全、可靠和值得信赖的发展。在人工智能技术快速发展背景下，各国高度重视人工智能安全和治理问题。本次峰会将在设定议题、达成共识和形成预期成果方面引领全球人工智能治理新方向。

一、峰会核心议题和动向

1.推进前沿人工智能模型的风险评估与安全测试

前沿人工智能模型，即指在规模、性能和创新性上达到或接近当前技术极限，具备广泛适用性和高计算能力的模型。这一模型展现出惊人潜力，但其强大的能力也带来了潜在安全隐患。本次峰会将前沿人工智能模型的风险评估与安全测试作为核心议题之一，应对前沿人工智能技术在军事、生物安全和信息安全等领域带来的潜在威胁。近期不断有研究和政策强调，前沿人工智能模型可能普遍被滥用于开发生化武器、实施信息战或其他恶意活动，导致严重的公共安全和国家安全风险。[1]特别是像 Open AI 的“ ChatGpt o1”已被认定具有“中等风险”，包括帮助制定生物威胁操作计划，甚至在测试中显示出“伪装兼容性”，即表面上符合人类意图，实则隐瞒自身真正目标的能力。[2]

与会各方将在英国、韩国两次“人工智能安全峰会”（AI Safety Summit）会议基础上，深入探讨如何通过安全测试和风险评估机制确保这些前沿人工智能模型可控性和安全性。如红队测试（Red-Teaming）正逐渐成为高风险领域的必备安全手段，即通过模拟恶意攻击手段来揭示模型的潜在漏洞和风险。[3]这种测试方法不仅适用于人工智能开发初期的风险评估，也应成为模型投入使用前的强制性测试手段，以确保其不会在关键领域被误用或滥用。美国和欧盟等国已将红队测试列为高风险人工智能系统的关键步骤，应对日益增长

的人工智能安全威胁，并为未来全球技术标准制定奠定基础。[4]

峰会预计还将就人工智能模型在军事和生物安全等高风险领域的应用制定更严格的技术安全标准。为此，政府、技术研发机构及国际标准组织可能会合作建立一套全面的安全评估框架，涵盖风险量化、透明度标准和数据管理等关键方面。此框架旨在指导各国在推动技术创新的同时，有效防控可能出现的风险，以达成技术发展与安全保障的平衡。

2.探讨构建人工智能安全治理的国际框架

人工智能技术的跨国应用特性导致单一国家的治理手段难以全面应对潜在风险。为此，建立国际一致的人工智能治理框架已成为应对人工智能安全挑战的当务之急[5]。本次峰会将不仅停留在抽象原则的讨论层面，而是着重讨论具体、可操作的治理机制，推动跨国合作。

主要参与方将在峰会上提出各自治理主张，其中一个重要议题是如何协调这些主张以实现一致标准。例如，会议或将推进人工智能生成内容的统一标识标准，使得人工智能生成内容能在发布时被以统一规格明确标记，便于用户辨别其来源，从而提高透明度和信任度。[6]此外，针对风险较高的人工智能模型，会议可能探讨设置风险评估和测试强制性标准，确保这些模型在部署前经过充分的安全验证。

该框架将不仅限于原则性的透明度、问责制和数据保护要求，而是力图纳入详细的技术性规范，如要求人工智能公司在开发和应用阶段提供可核查的模型行为报告，并开展强制性“红队测试”（red teaming）等安全评估活动，以识别潜在风险。峰会预期将讨论具有一定约束力的协议或意向书，涵盖包括透明度、责任追究、可解释性等关键要素，为未来的全球治理奠定基础。这种有约束力的框架不仅推动人工智能在成员国范围内的安全与负责任使用，也为技术创新提供更具确定性的法律环境，确保人工智能开发能够在伦理和社会安全框架下健康发展。

3.促进透明度与公众信任机制

透明度和信任机制构建是本次峰会重要议题之一。公众对人工智能的接受度和信任直接影响技术在各领域广泛应用，尤其是在医疗、教育和金融等涉及个人隐私与公共安全的高敏感领域。预计峰会将深入探讨如何通过透明度措施和信任机制提升公众对人工智能系统信任，为人工智能在全球范围内的安全部署奠定基础。

透明度措施可能包括采用可解释性技术，使人工智能决策过程更加透明，从而帮助用户区分人工智能生成内容与人工创作内容，减少错误信息。此外，为增加公众对人工智能系统决策过程的理解，峰会可能会倡导制定更严格的大模型

可解释性要求。这将涉及明确人工智能模型的工作逻辑、决策依据以及结果的可解释性和可追溯性，从而确保人工智能系统的行为符合伦理和法规标准。例如，在人工智能用于自动化诊断和金融交易决策时，可解释性能够有效避免“黑箱”操作，使监管机构和用户能够更清晰了解算法的实际运行方式。[7]

峰会还可能就数据保护和隐私管理提出新的国际标准和指南，特别是在处理用户个人数据方面，确保人工智能系统开发和应用过程中的隐私合规性。为此，数据管理标准将从数据收集、存储、传输到处理全链条实施严格的保护措施，并确保个人数据可追溯性。例如，通过引入类似于欧盟《通用数据保护条例》（GDPR）的隐私框架，强化对数据使用的监管，减少数据滥用风险。在国际合作层面，峰会也可能推动跨国数据共享和隐私保护的平衡发展，确保在大规模数据处理和人工智能系统训练时，能够保护用户隐私，避免潜在滥用风险。这些严格的标准和实践将为全球人工智能行业建立更高透明度和问责性，从而提升公众对人工智能技术的信任感，并进一步支持人工智能在全球范围内的负责任应用。

二、峰会预期成果

1.构建人工智能风险协同应对机制

基于英国和韩国峰会达成的共识成果，本次峰会将努力达成多项多边协议，进一步建立全球在人工智能安全风险防控方面的合作机制。协议内容将可能涵盖多层次的跨国协作机制，具体可能包括建立跨国情报共享和应急响应体系，以在人工智能相关风险或突发事件发生时提高各国集体应对能力。例如，协议可能明确要求签署国在面临大规模跨国虚假信息等人工智能安全事件时，通过快速反应、通报机制协同处理，有效遏制风险的扩散。[8]此类措施将为全球人工智能治理提供更强有力的支持，确保各国在应对和管控人工智能潜在威胁时能协调一致。

协议预计将建立更加系统的政策框架，以应对人工智能滥用可能带来的国家安全风险。这一框架可能包括具体的指导原则，规定各国如何在国家安全、认知风险、关键基础设施保护等敏感领域快速协调行动，防止人工智能被用于潜在军事、情报等高风险用途。通过此协议，签署国可以通过共享的分析资源和政策框架形成良好互操作性，在应对复杂人工智能安全事件时联合采取行动。[9]比如，协议可能促使各国在面临高风险人工智能模型或前沿人工智能技术可能带来的安全隐患时，采用相互认证的测试和评估标准，实现更高程度的风险识别和管理一致性。

2.扩大人工智能安全研究所国际合作网络

峰会另一个重要预期成果是推动国际人工智能安全研究所网络的发展。该网络旨在整合全球人工智能安全研究资源，将包括美国、欧盟、英国、韩国、日本等在内的人工智能安全研究机构力量汇聚起来，共同应对迅速增长的人工智能风险。自 2023 年英国人工智能安全峰会以来，该领域的跨国合作已取得初步成效，例如美国和英国于 2024 年 4 月签署谅解备忘录，将共同努力开发最先进人工智能模型测试。在 2024 年 5 月韩国首尔峰会上，各国签署《首尔意向声明》，承诺加强国际合作，建立一个跨国人工智能安全研究所联盟。

这一倡议的最终目标是通过建立一个全球性协作网络，推动人工智能安全技术的标准化和技术共享，为各国人工智能安全研究所提供资源和支持。未来，人工智能安全研究所国际合作网络将致力于进一步深化多国技术协作，重点推进高风险人工智能模型测试、评估和验证标准的统一化，以应对全球日益严峻的人工智能安全挑战。该网络将优先关注跨国合作的实践落地，计划在各成员国之间建立共享资源库，涵盖专业技术、数据集、基础设施和关键研究成果，从而在全球范围内提供更广泛支持。[10]此外，人工智能安全研究所国际合作网络还将积极参与并主导国际标准制定，与国际标准化组织（ISO）和国际电工委员会（IEC）等标准机构合作，推动人工智能安全技术标准化，确保不同国家和地区在

采用前沿人工智能技术时能够遵循一致的安全标准。在具体手段上，网络将采用红队测试、前部署评估验证（TEVV）等技术方法，并通过共享最佳实践来提升成员国人工智能安全应对能力。为巩固其影响力，该网络还将加强与私营部门、技术社区和非政府组织的联系，确保人工智能安全技术在各个层级应用中得到广泛认同和实施，最终实现具备高度响应力和适应性的全球人工智能安全治理框架。

3.发布全球人工智能安全治理声明

峰会将发布一份全球人工智能安全治理声明，界定人工智能安全治理的核心原则和关键执行方案。该声明不仅将框定各国在人工智能安全治理方面的共同原则与价值观，还将试图为今后各国在人工智能安全、伦理和透明度标准上的协调行动提供实质性政策指引。具体而言，这份声明可能会包括多个重要的透明度和问责制要求，特别是在风险评估、内容标识和责任分担等方面。文件或将呼吁各国建立系统的人工智能透明化机制，确保人工智能的决策过程、数据使用和模型行为的可解释性等。此外，声明还可能明确提出各国应对人工智能生成内容进行清晰标识，以避免虚假信息扩散，同时便于用户和监管机构对人工智能应用合法性和潜在风险进行识别和监督。[11]

在行业和企业层面，声明可能要求人工智能开发者采用

共识化的风险评估方法，包括通过红队测试、聘请专门人工智能安全官等方式识别和控制人工智能潜在安全漏洞和伦理风险。[12]这一政策性框架旨在推动国际社会就人工智能安全达成共同技术和管理标准，确保不同国家和地区的人工智能系统能够互相兼容，防止风险外溢，减少不必要的监管重叠和资源浪费。

声明还可能强调人工智能安全治理的国际合作，呼吁建立跨国人工智能安全研究机构网络，实现全球范围内的数据共享、人才培养和技术协作。特别是在标准制定方面，声明可能鼓励各国积极参与国际标准化组织的工作，推动人工智能安全规范的全球化和系统化。在全球范围内形成统一的人工智能安全治理模式，将对构建可信赖的人工智能发展环境起到至关重要的作用，为可能发生的重大安全事件提供应对方案，提升各国在人工智能发展方面的透明度和公信力[13]。这份声明最终将为全球人工智能治理提供方向性指导，确保人工智能的应用符合国际社会的共同利益和长远发展需求。

三、峰会影响分析

1. 增强全球人工智能治理的协调性

本次峰会的预期成果将显著提升各国在人工智能安全政策上的协调性。鉴于人工智能在医疗、金融、交通等多个关键领域的快速应用，各国对其监管的需求不断增加。然而，

由于政策差异，许多国家在应对人工智能潜在风险方面存在不协调，导致不必要的技术壁垒和重复管理。[14]通过此次峰会，参与国有望在一些核心人工智能安全原则上取得一致，包括透明度、可解释性、数据隐私与安全、以及人类监督等关键点。这些共识将有助于推动统一、可操作的风险评估标准，特别是在人工智能模型开发和测试阶段，例如在敏感领域推行更严格的红队测试和事前部署的测试、评估流程。

鉴于推动统一的全球人工智能治理标准难以实现，本次峰会更可能推动建立分层协同的治理框架，即在核心原则和高风险领域实施一致性治理，但在较低风险领域允许各国逐步推进适度的本地化治理。[15]这种框架将首先针对高风险的人工智能应用，如军事安全、公共基础设施保护等领域，以防范重大安全隐患。而在较低风险的商业和消费领域，各国则可根据峰会提供的指导原则进行相对独立的规范。对于跨国企业而言，这种治理协调将减少在不同国家部署人工智能技术时的合规成本，尤其是涉及模型透明度和数据共享的要求。通过建立人工智能安全机构间的合作网络，进一步共享测试数据、专家资源和基础设施，各国将更高效地应对人工智能在各个领域带来的复杂挑战，避免因重复管理导致的资源浪费。

2.加剧全球人工智能治理的阵营化趋势

本次峰会预计将加剧全球人工智能安全领域的“阵营化”趋势，推动国际治理分裂为西方主导的技术联盟和其他国家间对立态势。以七国集团（G7）为核心的西方国家在英国布莱切利和韩国首尔峰会中已形成初步合作框架，当前峰会可能进一步加固这一联盟关系，尤其在人工智能安全的测试、评估和治理标准方面。通过统一的监管标准、跨国的人工智能安全实验室和信息共享机制，西方国家可能会构建一个具备较强排他性的技术网络，以期在核心技术和高风险人工智能领域中保持主导地位。这一趋势意味着全球在技术治理和人工智能安全规范上的分歧将更显著，西方技术标准将逐渐成为该联盟内部统一准则，同时向外释放出一种“技术屏障”，对非联盟国家形成技术壁垒。

峰会很可能推动跨国安全治理协议，通过“核心圈层”和“外围圈层”模式强化联盟内部技术流通与信息保护。这意味着在高风险人工智能应用上，西方国家间合作将更加密切，而外围国家在技术准入上则会受到更严格审查和限制。例如，峰会可能推动建立人工智能风险分级管理系统，对不同等级的人工智能模型设置差异化的安全和透明性要求。这种差异化机制对西方国家而言，有助于提高人工智能模型的透明度、合规性并树立严格监管的正面形象。而对于非西方国家来说，这种分级管理体系将增设技术壁垒，使得它们在

开发和应用前沿人工智能技术时面临更繁琐的流程。[16]这种阵营化趋势带来的直接影响在于，人工智能治理的全球标准化进程将遭遇阻力，非西方国家在人工智能政策上面临选择压力。随着西方国家不断巩固其人工智能治理的安全标准，其他国家可能被迫在联盟规则和自主发展之间做出取舍，从而加剧全球人工智能治理的分裂，削弱全球人工智能治理的合作基础。

3.加剧人工智能国际安全领域“治理内卷”

本次峰会可能进一步加剧全球人工智能治理的“内卷化”现象。尽管各国难以迅速将监管规则落地，但为了吸引国际注意、展示政绩，各国可能继续投入资源举办类似的高层会议和多边倡议。这种象征性的“峰会效应”可能导致各国在人工智能治理中逐渐陷入形式化竞争，但实际治理成效可能有限。首先，在本次峰会激励下，各国预计会竞相推出新的人工智能监管工具和政策，以展示在人工智能安全治理方面的积极姿态。[17]然而，这些政策可能大多响应峰会提出的通用标准，缺乏具体的执行措施，因此往往仅停留在表面宣示阶段，未能结合各国实际需求和资源条件。[18]尤其是资源不足或技术相对薄弱的国家，可能会采取高度一致的政策框架，而缺乏落实执行的实际支撑，导致人工智能治理政策趋同。这样的同质化趋势形成“治理内卷”，各国在不

断投入精力推出新政策、塑造宏观框架，但在实际操作中，由于缺乏执行力和本地化调整，难以对人工智能安全治理产生实质助益。

峰会的全球性关注也加剧各国话语权竞争。为在国际平台上显示领导力，并避免深入细则导致的协调成本，许多国家更倾向于发布象征性声明，比如道德准则和安全红线，以彰显其对人工智能治理的重视。然而，这类声明大多缺乏强制执行的配套机制或具体监管手段，实际效用有限，更多出于政治宣传考量。这种治理表演的激增不仅削弱全球治理体系的实质推进，也分散各国在人工智能治理中的资源，使人工智能治理进展在大量象征性政策中停滞，难以推动有效的落地实施。

为在人工智能治理中强化自身影响力，各国或区域组织可能会争相组织独立的人工智能治理峰会，导致“竞争性峰会”数量激增，议题和目标的高度重叠使得国际资源分散、政策进一步碎片化。[19]缺乏协调的多边会议使得各国在不同的治理标准间徘徊，增加了政策执行复杂性，也造成宝贵资源浪费。总之，本次峰会存在引发政策“过度叠加”的现象，形成治理框架表面丰富、实际无效局面的风险。

撰稿：高隆绪

编辑：郑乐锋

审核：肖茜、董汀、孙成昊

参考文献：

[1] Withers, C., & Drexel, B. (2024). *Catalyzing Crisis*. CNAS. https://www.cnas.org/publications/reports/catalyzing-crisis

[2] Singer, S. (2024). *How the UK Should Engage China at AI's Frontier*. Carnegie Endowment for International Peace. https://carnegieendowment.org/posts/2024/10/lammy-china-ai-safety-cooperation?lang=en

[3] Goodfellow, I., Papernot, N., Huang, S., Duan, Y., Abbeel, P., & Clark, J. (2017). *Attacking machine learning with adversarial examples*. OpenAI Blog, 24, 1.

[4] Lang, C. (2024). *Advancing AI safety requires international collaboration. Here's what should happen next*. Atlantic Council. https://www.atlanticcouncil.org/blogs/new-atlanticist/advancing-ai-safety-requires-international-collaboration-heres-what-should-happen-next/

[5] Dafoe, A. (2018). *AI governance: a research agenda*. Governance of AI Program, Future of Humanity Institute, University of Oxford: Oxford, UK, 1442, 1443.

[6] Allen, G. C., & Adamson, G. (2024). *The AI Safety Institute International Network: Next Steps and Recommendations*. CSIS. https://www.csis.org/analysis/ai-safety-institute-international-network-next-steps-and-recommendations

[7] Kelly, E. (2024). *The U.S. Vision for AI Safety: A Conversation with Elizabeth Kelly, Director of the U.S. AI Safety Institute*. Csis.org. https://www.csis.org/analysis/us-vision-ai-safety-conversation-elizabeth-kelly-director-us-ai-safety-institute

[8] Meltzer, J. P., & Triolo, P. (2024). *The Bletchley Park process could be a building block for global cooperation on AI safety*. Brookings. https://www.brookings.edu/articles/the-bletchley-park-process-could-be-a-building-block-for-global-cooperation-on-ai-safety/

[9] Webster, G., & Hass, R. (2024). *A roadmap for a US-China AI dialogue*. Brookings. https://www.brookings.edu/articles/a-roadmap-for-a-us-china-ai-dialogue/

[10] Petroupoulos, A. (2024, September 10). *The AI Safety Institute Network: Who, What and How?* - ICFG. https://icfg.eu/the-ai-safety-institute-network-who-what-and-how/

[11] Allen, G. C., & Goldston, I. (2024). T*he Biden Administration's National Security Memorandum on AI Explained*. CSIS.org. https://www.csis.org/analysis/biden-administrations-national-security-memorandum-ai-explained

[12] Csernatoni, R. (2024). T*he AI Governance Arms Race: From Summit Pageantry to Progress?* Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2024/10/the-ai-governance-arms-race-from-

summit-pageantry-to-progress?lang=en
[13] Wyckoff, A. (2024). *A new institution for governing AI? Lessons from GPAI*. Brookings. https://www.brookings.edu/articles/a-new-institution-for-governing-ai-lessons-from-gpai/
[14] Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: barriers and pathways forward. *International Affairs*, 100(3), 1275-1286.
[15] Schmitt, L. (2022). Mapping global AI governance: a nascent regime in a fragmented landscape. *AI and Ethics*, *2*(2), 303-314.
[16] Paul, R. (2023). European artificial intelligence "trusted throughout the world": Risk-based regulation and the fashioning of a competitive common AI market. *Regulation & Governance*. https://doi.org/10.1111/rego.12563
[17] Pouget, H. (2024). *France's AI Summit Is a Chance to Reshape Global Narratives on AI*. Carnegie Endowment for International Peace. https://carnegieendowment.org/posts/2024/07/france-ai-summit-reshape-global-narrative?lang=en
[18] Acharya, A. (2004). How ideas spread: Whose norms matter? Norm localization and institutional change in Asian regionalism. *International organization*, 58(2), 239-275.
[19] Keohane, R. O., & Victor, D. G. (2011). The regime complex for climate change. *Perspectives on politics*, 9(1), 7-23.