政策指南

2025 年 2 月

IDAIS

人工智能安全国际对话

关于我们

人工智能安全国际对话

人工智能安全国际对话（IDAIS）致力于汇聚全球顶尖科学家并开展合作以降低人工智能可能带来的风险。2023年10月，图灵奖得主Yoshua Bengio和姚期智（Andrew Yao），联合加州大学伯克利分校教授Stuart Russell及清华大学智能产业研究院院长张亚勤，共同举办首届人工智能安全国际对话。

人工智能安全国际论坛

人工智能安全国际对话举办获人工智能安全国际论坛支持。人工智能安全国际论坛是由Fynn Heide和Conor McGurk联合创立，不接受任何企业捐赠者资助。

如果您热衷于推进本文件提及的政策领域研究与行动，请通过邮件与我们联系：idais@saif.org

介绍

自2023年底以来，人工智能安全国际对话已召集全球顶尖科学家与学者，就前沿人工智能系统风险及应对风险所需的相关治理干预措施达成共识。基于这一共识，人工智能安全国际对话参与者签署公开声明，阐述人工智能安全和治理的一系列目标。

这本政策指南旨在将共识所声明的目标与政策制定者、慈善家、企业家及研究人员考虑采取的直接政策行动挂钩，以改善人工智能安全和治理状况。

该指南根据公开声明目标分为四个关键政策领域：

- 人工智能安全研究
- 测试与评估
- 国内治理
- 国际治理

每一政策领域都有一组自共识声明直接得出的建议目标。针对每一目标，该指南提供当前目标进展情况评估及为进一步行动以实现该目标可考虑采取的政策。

该指南为动态文件，会在每次对话后得到更新。指南也是分模块的，读者可通过阅读概述部分确定具有主要兴趣和相关的领域，以便进行更深入的研究。

读者应注意到，尽管声明中的高层目标得到所有签署方认同，但本指南提出的政策行动只是根据多种来源提出的建议，这些来源包括对话中的讨论，以及进一步的研究和分析。因此，具体的政策建议并不一定能反映所有签署者的观点。

概述

读者可以通过点击下表“共识声明中的目标”跳转到感兴趣的领域。

图例: - 进展有限的领域 - 进展适度的领域

类别	共识声明目标	潜在政策
安全研究	保证人工智能安全研发的必要资源	• 以来源广泛的资金资助人工智能安全组织或相关国家级安全组织。 • 为人工智能安全研究支出引入税收减免。 • 在公共采购指南中规定最低安全投资水平（或其他安全相关标准）。 • 通过提供财政和实物资源支持公立大学开展人工智能研究。 • 发布针对第三方安全研究的具体提案或资助申请。举办竞赛、企业孵化器及创新奖以加速安全研究。 • 成立全球合作基金以资助人工智能安全研发。
	安全国际合作	• 开展全球联合研究项目以应对最重大的安全挑战。 • 为人工智能安全组建研究网络并举办有影响力的会议。 • 为人才培养、交流和相互学习开辟新渠道。 • 开设联合授课大学课程、研讨会和模块。
	支持验证方法研究	• 设置一个治理相关的验证研究议程。 • 明确须进行国际合作的具体验证项目。 • 合作就新型硬件支持验证机制进行压力测试。 • 组织验证研究全球竞赛。

类别	共识声明目标	潜在政策
测试和评估	定义并评估 红线和预警阈值	• 发展具有安全许可的内部技术能力，以进行与武器开发和网络攻击红线相关的能力评估。 • 要求企业对前沿模型进行评估，并与相关监管机构共享评估结果，以增强国家能力和可见性。 • 建立有关红线和预警阈值的明确的端到端国内外信息披露流程。 • 开发方法以验证必要的缓解措施是否到位。 • 研究红线/阈值违规协议。 • 研究和开发可信赖地关闭已跨越或有跨越红线风险的人工智能系统的方法。
	建立第三方审计与评测	• 开发全面的评测框架以明确第三方评测人员的作用。 • 开发全面的审计框架以明确第三方审计人员的作用。 • 开发和共享评测基础设施。 • 提供算力或API额度等资源以支持第三方评测人员。 • 设立补助金和奖金以鼓励相关能力评测的研发工作。
	要求提供安全案例和/或保证	• 直接资助“安全案例”研究项目。 • 直接资助可能在设计上实现安全的可替代科学或工程范式。
国内治理	落实人工智能事故跟踪和报道	• 建立事故跟踪与报道的综合方法。 • 鼓励关于人工智能系统更为积极的记录，以便更好对事故进行分析。 • 研究并开发自动化元数据收集机制。
	监控大规模数据中心	• 将云计算供应商视为中介，以获得对于人工智能发展趋势更好的可见性，更有效地对模型开发商进行监管。 • 开发隐私保护工具以实现更大范围的治理活动。 • 开发稳健的算力工作负载分类方法。
	落实人工智能事故跟踪和报道	• 建立事故跟踪与报道的综合方法。 • 鼓励关于人工智能系统更为积极的记录，以便更好对事故进行分析。 • 研究并开发自动化元数据收集机制。

类别	共识声明目标	潜在政策
	采取风险管理	• 开展针对前沿人工智能实验室人工智能风险管理研究，借鉴其他安全关键行业。 • 为风险管理系统设定监管要求。 • 评估企业实践是否遵循既有政府指导和改进建议。
	要求进行部署后监控	• 要求企业执行并共享互联的部署后监控数据。 • 收集人工智能系统接入关键基础设施的有关数据。 • 投资技术治理手段，以增加智能体系统的可见性。
	为人工智能从业者制订专门职业道德准则	• 开发人工智能“希波克拉底誓词”并要求人工智能从业者进行宣誓。 • 规定“警告权”，推动人工智能从业者履行职业道德责任。
国际治理	将人工智能安全视为全球公共产品	• 确保人工智能安全国际合作免受更广泛的人工智能地缘政治竞争影响。 • 识别、开发并共享安全技术，如“人工智能允许行动联系机制”。 • 开展更进一步研究，以明确“人工智能安全作为全球公共产品”的概念。
	全球对齐红线标准并建立有条件市场准入	• 在既有协助框架下为红线标准制定确定合适的渠道。 • 建立监管市场，协调监管目标与市场准入的机制。 • 实施司法认证程序和多边出口管制制度以管理有条件市场准入。
	制订和协商国际协议和/或成立国际组织	• 围绕哪些前沿人工智能治理要素应被国际化建立共识。 • 就是否应通过现有组织或新程序建立关键国际前沿风险治理功能达成一致。 • 实施信任建立措施，确保具有必要的国际信任以构建更广泛的治理程序。 • 就关键新兴安全标准进行协商。

人工智能安全研究

保证人工智能安全研发的必要资源

目标

我们**呼吁领先的人工智能开发商将其人工智能研发的三分之一的最低支出承诺用于人工智能安全**，并呼吁政府机构以至少相同比例资助学术和非营利性人工智能安全和治理研究。 -人工智能安全国际对话-牛津，*2023*

需要额外的资金来支持该领域的发展：我们**呼吁人工智能开发商和政府资助者将至少三分之一的人工智能研发预算投入到安全领域。** -人工智能安全国际对话-北京，*2024*

各国需要为人工智能安全研究所提供足够的资源，并继续召开峰会，支持其他国际治理举措。 -人工智能安全国际对话-威尼斯，*2024*

国家、慈善机构、企业、和专家应设立一系列全球人工智能安全与验证基金。这些资金应当逐步增加，直至其在全球人工智能研发支出中占据重要比例，以充分支持并增强独立研究能力。 -人工智能安全国际对话-威尼斯，*2024*

进展

有限

人工智能安全的⅓研发投资将达到约400亿美元。但目前全球人工智能安全资助不太可能达到10亿美元，这表明人工智能安全研发支出不到总金额的0.5%。[1]

潜在政策

- **以来源广泛的资金资助人工智能安全研究所或相关国家级安全组织。**不同国家在该方面存在很大差异——英国为其人工智能安全研究所提供超1亿美元资助，而美国人工智能安全研究所仅

1 鉴于这个领域缺乏标准化数据，任何估算都将不够精确。此估算的目标是提供一个数量级上的大体判断，使用以下方法：大多数AI研发是私人投资，排除并购活动后，2024年斯坦福AI指数估计约为1100亿美元。我们假设政府投资约占私人投资的5%，因此AI总投资大约为1200亿美元，其中33%为400亿美元。目前，AI安全承诺主要由政府主导，金额略超过2亿美元。

获得英国总额约10%的资助。各国可考虑获得资助的多种渠道，如从政府部门获得暂时性或永久性资金、实施有针对性的税收和再分配政策以及/或从慈善家和企业获取自愿捐款等 。[23]

- **为人工智能安全研究支出引入税收减免。**中国社会科学院法学研究所周辉提出人工智能模型法草案，介绍了人工智能税收减免概念。税收减免额将至少是人工智能开发商和供应商研发或采购用于安全和治理目的装备费用的30%。[4]
- 在公共采购指南中规定最低安全投资水平（或其他安全相关标准）。各国可通过将一系列标准作为公共采购要求的一部分以塑造企业行为。例如，美国行政管理和预算局发布政策，要求联邦机构负责任地采购人工智能，这些政策包括要求充分的测试、保障措施和外部人工智能红队。[5]
- **通过提供财政和实物资源支持公立大学开展人工智能研究。** 在大学设立人工智能安全研究主席，将帮助建设人工智能安全这一学术领域，诸如美国国家人工智能研究资源试点项目等倡议能够帮助学者获得稀缺的、难以获得的资源（如算力）。[6]
- **发布针对第三方安全研究的具体提案或资助申请。**例如，英国人工智能安全研究所拨款1000万美元用于系统性人工智能安全，以支持社会层面缓解人工智能负面影响的研究。[7]
- **举办竞赛、企业孵化器及创新奖以加速安全研究。**在这方面有针对性的加速项目包括X奖项和Entrepreneur First的防御加速项目。[8]
- **成立全球合作基金以资助人工智能安全研发。**合作基金可汇聚多个组织资源，并允许提供大规模资金。这些资金可以通过公私合作的形式设立，允许政府资金更进一步吸收私人捐赠，用于支付政府拨款和/或直接投资人工智能安全项目。现有合作基金的例子包括TED设立的Audacious项目和气候工作基金会。[9]

实现人工智能安全国际合作

目标

全球研究界在人工智能和其他学科领域的共同努力至关重要；**我们需要一个由专门的人工智能安全研究和治理机构组成的全球网络。** -人工智能安全国际对话-牛津，*2023*

我们鼓励通过访问研究人员项目和组织深度人工智能安全会议与工作组，建立一个更强大的全球技术网络以加速人工智能安全研发和合作。 -人工智能安全国际对话-北京，*2024*

---

[2] 选择性税收用于资助一些公共项目，这些项目旨在应对某个行业带来的公共安全风险。例如，在加利福尼亚州，烟草税用于资助反吸烟运动。

[3] 在美国等国家，政府机构可以接受私人基金会的捐款。如需进一步信息，请见: https://exponentphilanthropy.org/qa/can-a-private-foundation-make-a-grant-to-a-government-agency/

[4] ZHOU Hui et al., 'The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version', 16 April 2024, https://doi.org/10.5281/ZENODO.10974162.zh.

[5] Shalanda Young, 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence' (Executive Office of the President, Office of Management and Budget, 1 November 2023), https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

[6] 'National Artificial Intelligence Research Resource Pilot', n.d., https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.

[7] 'Tech Secretary Unveils £8.5 Million Research Funding Set to Break New Grounds in AI Safety Testing', 22 May 2024, https://www.gov.uk/government/news/tech-secretary-unveils-85-million-research-funding-set-to-break-new-grounds-in-ai-safety-testing.

[8] 'XPrize', n.d., https://www.xprize.org/; Matt Clifford, 'Introducing Def/Acc at EF', 20 May 2024, https://www.joinef.com/posts/introducing-def-acc-at-ef/.

[9] 'The Audacious Project', n.d., https://www.audaciousproject.org/faq; 'ClimateWorks Foundation', n.d., https://www.climateworks.org/.

全面的验证最终可以通过多种方式进行，包括第三方治理（如独立审计）、软件（如审计跟踪）以及硬件（如人工智能芯片上的硬件支持治理机制）。**为确保全球信任，跨国联合开发验证方法，并对其进行压力测试将变得尤为重要。** -人工智能安全国际对话-威尼斯，*2024*

进展

适度

建立专门人工智能安全研究组织全球网络方面进展适度。人工智能安全研究所国际组织网络（AISIs）已被设立以协调国家间具有代表性的人工智能安全实体。另外，双边人才交换协议如英国和美国人工智能安全研究所国际组织网络谅解备忘录也已制订完成。

安全已成为重要人工智能会议的部分内容。例如，安全和对齐论坛已作为国际机器学习大会（ICML）、北京智源大会以及世界人工智能大会等会议的一部分而召开。

然而，目前尚没有专门针对人工智能安全的完整会议，全球人工智能实验室间的制度化关系仍相对有限。

潜在政策

- **开展全球联合研究项目以应对最重大的安全挑战。**来自其他领域的例子包括人类基因组计划，这是一种国际努力，在13年的时间里生成了人类基因组第一组序列，聚集了来自美国、英国、法国、德国、日本和中国20个组织的研究人员。[10]
- **为人工智能安全组建研究网络并举办有影响力的会议。**围绕现有机器学习生态建立网络和举办会议方面取得一些进展。然而，仍需建立更多专门用于人工智能安全研究的基础设施以加速其作为学术学科的发展。关于有大影响力的会议的潜在模型包括美国联邦储备委员会年度杰克逊·霍尔经济研讨会，该研讨会吸引大量全球学术专家和中央银行领导参与，且经常对货币政策和全球经济起重要影响。[11]
- **为人才培养、交流和相互学习开辟新渠道。**由全球顶尖人工智能安全科学家组织的具体人工智能安全交流项目、奖学金和伙伴计划，将为更协调的全球领域铺平道路。这些项目可借鉴一系列广泛的现有项目如施瓦茨曼学者项目、富布赖特或卢斯得奖学金。

[10] 'Human Genome Project', n.d., https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project.
[11] 'How Jackson Hole Became an Economic Obsession', *The New York Times*, 25 August 2023, https://www.nytimes.com/2023/08/24/business/economy/jackson-hole-economic-conference.html.

- **开设联合授课大学课程、研讨会和模块。**尽管一些大学开设了人工智能安全课程（如加州大学伯克利分校人工智能安全入门），但更多这样的努力——尤其如果是由全球顶尖学术专家共同组织的——将有助于快速培养对这个领域拥有共同理解的全球人才。

支持验证方法研究

目标

除了人工智能安全基础研究，**这些资金的其中一部分将专门用于隐私保护和安全验证方法的研究，为国内治理和国际合作提供支持。**这些验证方法将允许各国可信地核实人工智能开发者的评估结果，以及他们在安全报告中指定的任何缓解措施是否到位。在未来，这些方法还可能允 许各国验证其他国家提出的相关安全声明，包括对安全保障体系的遵守情况，以及重大训练运行的申报。 -人工智能安全国际对话-威尼斯，*2024*

进展

有限
允许某一主体检查另一主体提出的人工智能安全相关声明的验证方法方面研究进展有限。对一些主题如硬件支持机制已有些初步的工作，但许多验证问题仍未得到充分研究，而可信保证可能需要广泛的研究和压力测试。

潜在政策

- **推进现有研究议程中确定的高优先级研究。**有一系列正在进行或已发表的研究议程，确定了推进验证研究需要解决的关键研究问题。[12] [13] [14] [15] [16] [17]总的来说，这些研究议程向慈善家、各国和研究人员提供路线图，以指导在关键领域进一步利用资源取得进展。

[12] Mauricio Baker et al. (Forthcoming). RAND.
[13] Ben Harack et al. (Forthcoming). Oxford University.
[14] Scher & Thiergart, 'Mechanisms to Verify International Agreements About AI Development', November 2024, https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development
[15] Anka Reuel et al., 'Open Problems in Technical AI Governance' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.14981; Miles Brundage et al., 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims' (arXiv, 20 April 2020), http://arxiv.org/abs/2004.07213.
[16] Akash Wasil et al. (Forthcoming).
[17] James Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024, https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf.

- **明确须进行国际合作的具体验证项目**。一些验证项目要么更有力，要么只有在国际范围内进行时，才成为可能。1988年美苏联合验证试验是一个很好的例子，当时允许两国近距离测试另一方地下爆炸当量以验证1974年以来一直处于僵局的《禁止核试验阈值条约》。[18]
- **合作就新型硬件支持验证机制进行压力测试**。许多关于验证的建议包括允许对协议进行全球压力测试并建立信任的开源组件。来自世界各地的研究人员可以尝试进行压力测试并改进开源技术以建立协作验证机制。[19]
- **组织验证研究全球竞赛**。国家可以发起旨在开发安全验证方法的全球竞赛，类似于由美国国家标准和技术研究院发起的关于后量子密码学标准化的全球竞赛。[20]

[18] '30th Anniversary of JVE: Snezhinsk Scientists Reach out to Congratulate the US Colleagues', n.d., https://nonproliferation.org/lab-to-lab-joint-verification-experiment.
[19] Petrie et al., 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees'.
[20] 'Post-Quantum Cryptography', n.d., https://www.nist.gov/pqcrypto.

检测和评测

定义并评测红线和预警阈值

目标

我们还建议划定明确的红线，**一旦逾越，将强制要求通过快速、安全的关闭程序立即终止人工智能系统（包括所有副本）**。各国政府应合作实例化并保留这种能力。此外，在部署之前以及在最先进模型的培训期间，开发人员应向监管机构证明他们的系统不会跨越这些红线，以使监管机构满意。-人工智能安全国际对话-牛津，*2023*

我们**提出了人工智能发展作为国际协调机制的红线**……

- 自我复制或改进：如果没有明确的人类批准和帮助，任何人工智能系统都不应能够复制或改进自身。这包括自身的精确副本以及创建具有相似或更强能力的新人工智能系统。
- 寻求权力：任何人工智能系统都不应采取行动过度增加其权力和影响力。
- 协助武器开发：任何人工智能系统都不应大幅提高行为者设计大规模杀伤性武器或违反生物或化学武器公约的能力。
- 网络攻击：任何人工智能系统都不应能够自主执行网络攻击，从而导致严重的财务损失或同等伤害。
- 欺骗：任何人工智能系统都不应该能够持续导致其设计者或监管者误解其跨越任何前述红线的可能性或能力。 -人工智能安全国际对话-北京，*2024*

我们应建立预警阈值，即模型的能力水平表明该模型可能会越过或接近越过红线。该方法建立在现有的自愿承诺（如负责扩大政策）的基础上，对不同框架进行统一和协调。能力低于预警阈值的模型只需有限的测试和评估，而对于超出这些预警阈值的高级人工智能系统，我们则必须采用更严格的保障机制。 -人工智能安全国际对话-威尼斯，2024

进展

适度
在“红线”概念方面的直接工作有限，但相关工作仍在持续取得进展。例如，几家领先人工智能企业已经在人工智能首尔峰会签署《前沿人工智能安全承诺》，承诺明确某一模型或系统所带来的风险被认为无法容忍的阈值。人工智能企业和人工智能安全组织已经在设计和运行相关能力评估以测试是否模型是否满足预定阈值方面取得进展。

潜在政策

更广泛的旨在确保人工智能研发必要资源并通过国内监管塑造各国对于模型开发商行动的可见性，也可以为这个目标作出贡献。

- **为红线、预警阈值和其他关键阈值开发共同分类方法。**尽管人工智能安全国际对话——《北京人工智能安全国际共识》提供了一组基本的红线，还需要大量进一步工作以在更广泛的利益相关者中就红线的定义和预警阈值等相关概念建立共识。[21] 也需要进一步工作以明确这些概念如何相互联系的。经济合作与发展组织曾开展类似的多方利益相关者协商程序，通过人工智能事件专家组建立人工智能事件和危险级别分类。[22]
- **发展具有安全许可的内部技术能力，以进行与武器开发和网络攻击红线相关的能力评测。**国家安全机构在评估网络和化学、生物、辐射和核威胁（CBRN）方面具有专业知识。要深入了解这些领域的人工智能风险，需要访问机密信息和国家安全专家。各国政府在发展内部技术能力方面具有独特优势，可以设计并运行模型评测，同时与秘密和绝密信息来源及专家保持联系。[23]
- **要求企业对前沿模型进行评测，并与相关监管机构共享评测结果。**针对企业实施前沿模型评测并共享结果的广泛要求将帮助塑造国家能力和可见性，这对于未来制订红线相关评测具体要求是必需的。在这一领域已取得进展——几家美国企业已自愿承诺进行内部和外部测试及评估，并根据行政命令要求与美国政府共享超过特定计算阈值的模型的结果。[24] [25]《欧盟人工智能法》也要求构成系统性风险的通用人工智能模型开发商须开展模型评测。[26]
- **建立有关红线和预警阈值的明确的端到端国内外信息披露流程。**目前，围绕国家要求企业报告的与安全和保障测试相关的信息类型、政府内不同部门应该如何协调这些信息、以及政府哪些部门应对新出现的威胁进行回应，仍存在一定模糊性。各国应与企业和学术专家合作，明确哪些行为体（如承包方、第三方等）应向哪一政府部门报告，并具体说明提供什么类型的信息最有价值。[27] 该领域的进一步工作还可以明确什么类型的信息应在国际层面共享。
- **开发方法以验证必要的缓解措施是否到位。** 人工智能系统通过不同预警阈值并可能引起更大风险时，有必要采取缓解措施管理正不断上升的风险，包括保密措施（如强化网络空间安全）以预防模型权重被盗、安全措施（如红队、内容过滤器）以确保安全部署是必要的。目前，这样的缓解

[21] 当早期预警阈值被突破时，可能表示已达到关键模型能力或风险水平，应采取预定的行动。这些阈值是朝着红线发展，但其严重程度较低于红线，提供了提前警告，表明模型可能正在接近或超过红线。
[22] 'Defining AI Incidents and Related Terms', OECD Artificial Intelligence Papers, vol. 16, OECD Artificial Intelligence Papers, 6 May 2024, https://doi.org/10.1787/d1a8d965-en.
[23] Akash Wasil et al., 'AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.17347.
[24] 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', § 4 (2023).
[25] 'White House Voluntary AI Commitments' (White House, n.d.), https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.
[26] 'EU AI Act', § 55 (2024).
[27] Joe O'Brien et al., 'Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.01420.

措施是由人工智能企业自愿实施的，但未来，各国可能需要求企业实施并证明这些缓解措施已经部署到位。因此，需要更进一步的研究以开发和制定这样的验证方法。

- **研究红线/阈值违规协议。**针对模型评测表明红线已经或可能被跨越时，各国、企业和其他关键行为体如何采取行动的研究十分有限。需要更进一步的工作以理解国内和国际应对协议应该是什么样的。潜在协议的一个例子是协调暂停，这是一个当模型评估表明某一不被接受的风险阈值已经被跨越时，建议人工智能企业同意暂停在特定领域集体研发的协议；只有当某一安全标准被达到后才恢复这些领域的开发。[28]
- **研究和开发可信赖地关闭已跨越或有跨越红线风险的人工智能系统的方法。**该领域的研究还处于起步阶段，一旦发现人工智能系统跨过红线，发展安全关闭人工智能系统的方法可能是必要的。拟议解决方法包括在人工智能硬件（如半导体芯片）上安装一个“关闭按钮”。[29]

建立第三方审计与评估生态系统

目标

政府应该要求前沿人工智能模型开发商接受独立第三方审计，**评估其信息安全和模型安全性。** -人工智能安全国际对话-牛津，*2023*

国应要求开发者定期进行测试，判断模型是否具备带来潜在风险的能力，**并通过第三方独立的部署前审计保证透明度，确保这些第三方获得必要的权限，**包括开发者的员工、系统和记录等必要证据，以核实开发者的主张。此外，**对于超出早期预警阈值的模型，各国政府可要求开发者在进一步训练或部署这些模型前，必须获得独立专家对其安全报告的批准。**-人工智能安全国际对话-威尼斯，*2024*

[28] Jide Alaga and Jonas Schuett, ‘Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers’ (arXiv, 2023), https://doi.org/10.48550/ARXIV.2310.00374.
[29] ‘Considerations and Limitations for AI Hardware-Enabled Mechanisms’, 10 March 2024, https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/.

进展

适度
第三方审计和评测生态系统发展方面进展适度。

评测往往包括测试运行，以理解模型或系统所带来的风险。在一些国家，人工智能安全研究所、国家研究组织、非营利组织和营利企业已经成立并扮演第三方评估机构角色。然而，政府要求的第三方评测进展缓慢。一些人工智能企业正向第三方评测人员提供部署前访问权限；然而，这样的访问是自愿的且是根据人工智能企业规定条款提供。

审计是一类广泛的活动，可能包括企业治理体系、产品安全以及产品对下游用户的影响。政府要求审计方面进展很大。例如，到2026年，《欧盟人工智能法》将要求某类高风险人工智能系统供应商进行第三方合格性评估，这是审计方式的一种。

潜在政策

更广泛的旨在确保人工智能研发必要资源的行动也可以为此目标作出贡献。

- **开发全面的审计框架以明确第三方审计人员的作用。**第三方将在直接评测模型或审计验证由企业开展的某些评测中扮演一个日渐重要的角色。与研究人员和企业合作的各国，需要明确第三方测试在哪些领域是最为必要的，以及应对不同类型风险需要哪类第三方测试人员（如政府相关的或私人）。[30] 这也要求就部署前模型访问权限和可信的评测基础设施使用等议题达成共识。
- **开发全面的评测框架以明确第三方评测员的作用。**审计可能针对技术供应商治理系统（如人工智能开发商风险管理系统）与模型特性验证（如稳健性），审查模型缺陷文档（如模型卡），以及/或模型对下游用户的影响。[31] 公共部门和私营部门间的外部审计责任分配需要建立共识。一些审计，特别是涉及敏感信息的白盒子或灰盒子审计，可能需要公共主体的参与，就像核能安全和生命科学领域的情况一样。另一方面，涉及较少敏感信息的黑盒子模型审计和治理审计可能由受监管的私人审计人员市场更有效率地开展。[32]
- **开发和共享评测础设施。**评估基础设施如智能体框架、能力引导和分级工具将使第三方评估人员能够更有效率且更便宜地评估模型。[33] 例如，英国人工智能安全研究所拥有开源评估框架，而新加坡人工智能Verify基金会拥有开源测试平台Project Moonshot，允许企业更容易在其模型上运行既有测试和自动红队测试。[34] 应鼓励企业也做这样的事情。

[30] 'Third-Party Testing as a Key Ingredient of AI Policy', 25 March 2024, https://www.anthropic.com/news/third-party-testing.
[31] Jakob Mökander et al., 'Auditing Large Language Models: A Three-Layered Approach', *AI and Ethics*, 30 May 2023, https://doi.org/10.1007/s43681-023-00289-2.
[32] Merlin Stein et al., 'Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.20847.
[33] Gabriel Mukobi, 'Reasons to Doubt the Impact of AI Risk Evaluations' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2408.02565.
[34] 'Project Moonshot: An LLM Evaluation Toolkit', n.d., https://aiverifyfoundation.sg/project-moonshot/; 'Inspect: An Open-Source Framework for Large Language Model Evaluations', n.d., https://inspect.ai-safety-institute.org.uk/.

- **提供算力或API额度等资源以支持第三方评测人员。**国家，企业和慈善家可以建立共同的资源池（如国家研究云、API额度）以支持第三方评估人员。例如，英国人工智能研究资源，一个用于人工智能研究的高级计算机集群，正获得10亿美元用以支持研发。[35]
- **设立补助金和奖金以鼓励相关能力评测的研发工作。**一个例子是Anthropic资助第三方组织倡议，发展评估以有效衡量人工智能模型高级能力如应对网络安全与化学、生物、辐射和核威胁能力以及模型自主性。[36] 有前景的研究方向包括提升模型评估可靠性、可量测性、全面性以及确定监管智能系统最佳做法。[37]

要求提供安全案例和/或保证

目标

(政府）还应要求人工智能开发人员与相关机构**分享全面的风险评估、风险管理政策以及第三方评估和部署后系统行为的预测。** -人工智能安全国际对话-牛津，*2023*

开发人员有责任令人信服地证明红线不会被跨越，**例如通过严格的经验评估、定量保证或数学证明。** -人工智能安全国际对话-北京，*2024*

开发者应该**提交高置信度的安全案例，并以一种能够说服科学界相信其系统设计是安全的方式进行量化，**这也是其他安全关键工程学科的常见做法。此外，足够先进系统的安全报告应讨论开发者的组织流程，包括有利于安全的激励机制和问责结构。-人工智能安全国际对话-威尼斯，*2024*

进展

有限
要求开发商提供更全面的安全保证方面进展有限。然而，在安全案例、可证明的安全性和其他方法方面已经做了许多工作，可能可以为人工智能系统安全提供更大的确定性。

潜在政策

- **直接资助可能在设计上可通过设计实现安全的可替代科学或工程范式。** 今天，大多人工智能发展聚焦现有基于转化模型的深度学习范式。在这个范式下构建的系统可能永远不够安全，无

[35] 'AI Research Resource Funding Opportunity Launches', 24 January 2024, https://www.ukri.org/news/ai-research-resource-funding-opportunity-launches/.
[36] 'A New Initiative for Developing Third-Party Model Evaluations', 1 July 2024, https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations.
[37] Anka Reuel et al., 'Open Problems in Technical AI Governance' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.14981.

法在重要关键安全领域部署人工智能系统（如航空）。因此，各国和慈善家可以直接资助其他研究方法如“保证安全的人工智能”。[38]

- **直接资助“安全案例”研究项目。**安全案例是一个结构性的论证，说明人工智能系统在特定训练或部署背景下是安全的。对于关键安全领域的部署，或前沿人工智能表现出造成潜在灾难性风险迹象时，正面论证系统的安全性可能是必要的。一个安全案例可能包括足够强有力的控制措施，模型尽管有能力造成伤害但可信任的证明，或其他证明安全的方法。[39] 一些人工智能企业，如Anthropic，已经开始雇佣研究人员负责该项任务，而其他组织如Apollo研究中心已经开始构建特定领域的论证方法，以详细构建人工智能在“阴谋”这一领域的安全案例。[40] [41] [42]

[38] David ‘davidad’ Dalrymple et al., ‘Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems’ (arXiv, 2024), https://doi.org/10.48550/ARXIV.2405.06624.
[39] Joshua Clymer et al., ‘Safety Cases: How to Justify the Safety of Advanced AI Systems’ (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.10462.
[40] ‘Safety Case Specialist (Safety Mitigations) (London)’, n.d., https://www.linkedin.com/jobs/view/safety-case-specialist-safety-mitigations-london-at-anthropic-3956180495/?originalSubdomain=uk.sa.
[41] Mikta Baseni et al., ‘Towards Evaluations-Based Safety Cases for AI Scheming’ (Apollo Research, 1 November 2024), https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6724bd285d993323e03b89d6/1730460969317/Toward_evaluations_based_safety_cases_for_AI_scheming.pdf.
[42] Marie Davidsen Buhl et al., ‘Safety Cases for Frontier AI’ (arXiv, 28 October 2024), http://arxiv.org/abs/2410.21572.

国内治理

本章节国内治理行动不包括测，后者单独列出论述。

实施国内模型登记

目标

> 在国内监管中，我们**建议对超过一定能力阈值的模型（包括开源副本和衍生品）的创建、销售或使用进行强制注册，**以使政府能够获得对新兴风险的关键且目前缺失的可见性。-人工智能国际安全对话-牛津，*2023*
>
> 我们应该立即**对超过一定计算或能力阈值的人工智能模型和训练运行进行国内注册**。注册应确保政府能够了解其境内最先进的人工智能以及阻止危险模型传播和运行的手段。-人工智能安全国际对话-北京，*2024*

进展

适度

模型和训练运行登记方面进展适度。在欧盟和美国，模型开发商被要求或者即将被要求告知政府如果他们开发超过特定计算阈值的人工智能模型，因为计算阈值是有关能力和风险的重要指标。在中国，模型开发商被要求在中国国家互联网信息办公室登记所有面向社会公开的模型。计算阈值在两部由中国顶尖法律学者和专家撰写的人工智能法律提案草案中被提及。

潜在政策

- **确保模型登记政策对新兴人工智能风险具有更广泛、更深层次的可见性。**各国已经采取初步措施，要求开发商就特定类型的人工智能模型进行登记备案。[43] 除了现有登记要求外，各国可能还要考虑表明未来模型登记所需要的更广泛的文档和数据，为目前数据收集创造激励机制。这可能包括已经通过模型卡片公开披露之外的数据，如模型训练过程具体信息，包括所使用数据库、数据清洗实践、相关硬软件组件和这些组件已知漏洞。[44]

[43] 例如，拜登政府关于人工智能的行政命令要求双用途基础模型的开发者在开发或计划开发超过计算阈值的模型时，通知政府。中国的模型注册规定适用于所有面向公众的生成性人工智能模型，但不适用于用于研发的模型。在欧盟，分类为高风险的人工智能系统需要进行注册，且那些可能带来系统性风险的通用人工智能模型的开发者也需要通知欧盟人工智能办公室。

[44] Noam Kolt et al., 'Responsible Reporting for Frontier AI Development' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2404.02675.

- **就应引起更大力度监管审查的精确的算力、能力和风险阈值开展研究并达成共识。**欧盟和美国现有模型登记要求广泛使用计算阈值作为风险替代指标。[45]这是一种较为生硬的做法，仅为监管机构提供较为初步的、不确定的线索。关于阈值、相关评估和测试方法的进一步细化，将确保对于人工智能开发商更严格的审查与有关能力和风险的具体指标挂钩。[46]

监控大规模数据中心

目标

> **政府应监控大规模数据中心**..... -人工智能国际安全对话-牛津, *2023*

进展

有限
监控大规模数据中心方面进展有限。美国政府已发布拟议定规则， 要求云计算供应商（他们经常运行大规模数据中心）验证外国用户身份，并报告与外国用户的、可能导致所训练人工智能大模型被用于网络攻击的交易。

虽然其他司法管辖区可能拥有适用于前沿人工智能相关数据中心监控的现有法律，但这方面工作的探索仍然有限。

潜在政策

- **将云计算供应商视为中介，以获得对于人工智能发展趋势更好的可见性，更有效地对模型开发商进行监管。**鉴于云计算供应商市场的集聚度，监管云计算供应商将为政府提供更高的治理杠杆。[47] 例如，对于达到特定基础设施安全要求、同意验证模型开发商使用云计算服务是否完全符合国内要求的国内云计算供应商，政府可以为其颁发许可。[48]更广泛地说，计算供应商可能是一个很有发展前景的治理中介，通过：
    - 为人工智能模型提供额外的物理安全和网络空间安全。
    - 坚持记录高层次信息，如用户算力的使用情况。
    - 证实关键属性，如用户身份和人工智能系统属性。
    - 针对不合规用户实施算力访问限制。

[45] Sara Hooker, 'On the Limitations of Compute Thresholds as a Governance Strategy' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2407.05694.
[46] Leonie Koessler, Jonas Schuett, and Markus Anderljung, 'Risk Thresholds for Frontier AI' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2406.14713.
[47] Lennart Heim et al., 'Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation' (arXiv, 2024), https://doi.org/10.48550/ARXIV.2403.08501.
[48] 'Global Governance: Goals and Lessons for AI' (Microsoft, n.d.), https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0.

- **开发隐私保护工具以实现更大范围的治理活动。**隐私保护工具允许云计算供应商验证更多详细信息，如与治理相关的具体工作负载。目前，具体工作负载验证仍然是不可能的，因为计算供应商需直接访问模型或代码以验证特定治理相关属性（如模型评测是否已运行），这违反他们与用户签订的保密协议与隐私协议。关于机密计算和其他隐私保护方法的进一步研究将在不暴露敏感数据的情况下允许更具体的工作负载验证。[49]
- **开发稳健的算力工作负载分类方法。**计算工作负载分类方法允许云计算供应商决定由用户执行的运算类型是否作为一部分纳入监督机制（如大型人工智能训练运行）。需要进一步研究，创建能够抵御对抗性操作的计算工作负载分类方法。目前，对抗性行为体可改变其工作负载的计算模式，以躲避检测。[50]

落实人工智能事故跟踪和报道

目标

政府应该……跟踪人工智能事件 -人工智能安全国际对话-牛津, *2023*

进展

适度
事故跟踪方面进展适度。人工智能伙伴关系（Partnership on AI）和经济合作与发展组织（OECD）都维护不同人工智能事件数据库。《欧盟人工智能法》要求高风险人工智能系统和对系统构成威胁的通用人工智能系统开发商要向国家市场监督管理部门或欧盟人工智能办公室（EU AI Office）报告严重事故。除这些例子以外，全面的人工智能事故跟踪仍然较为有限。

潜在政策

- **建立事故跟踪与报道的综合方法。**这样的系统可以将重大事故的强制报告、轻微故件的自愿报告与公共门户网站的公众报告结合。该系统可由一个标准化数据收集框架支撑以确保来自不同来源搜集数据的可比性，由一个调查团队深入探究重大事故起因。[51]
- **鼓励对人工智能系统更为积极的记录，以便更好对事故进行分析。**目前人工智能事故处理往往是被动型的，即系统信息和事故在事后记录。更多人工智能事前系统主动型记录已经出现，使

[49] Heim et al., 'Governing Through the Cloud'.
[50] Heim et al.
[51] Ren Bin Lee Dixon and Heather Frase, 'An Argument for Hybrid AI Incident Reporting' (Center for Security and Emerging Technology, March 2024), https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

得事件发生时更易于追溯人工智能系统全生命周期。这也将帮助调查人员确定系统失灵早期信号大概是什么样。[52]

- **研究并开发自动化元数据收集机制。**飞行记录仪等机制——也被称为“黑匣子”——常被用于提供事件的关键背景性和技术性元数据。元数据在日常操作中不会被检查，但是在调查中非常有帮助。人工智能系统类似机制尚未被开发，但可能包括定期检查的模型技术数据“快照”（如模型权重、输入和输出）。[53]

采取风险管理和评测实践

目标

还应**要求人工智能开发人员与相关机构分享全面的风险评估、风险管理政策。**-人工智能国际安全对话-牛津，*2023*

进展

适度
风险评测和风险管理方面进展适度。几家领先人工智能企业在人工智能首尔峰会签署《前沿人工智能安全承诺》，要求企业开展风险评测和缓解计划。一些企业已经发布类似计划，多是负责任扩展策略、前沿安全和准备政策，并成立专业团队聚焦现在和未来系统的安全。
一些司法管辖区，如欧盟，已经开始要求开发商实施风险管理系统，而其他司法管辖区如美国，也提出若干自愿风险管理框架供企业选择采用。
然而，仅有限证据证明更全面的风险管理实践已经被执行，如三道防线风险防控法，可能包括企业内部实施多层风险监控和独立审计。

潜在政策

- **开展针对前沿人工智能实验室的人工智能风险管理研究。**风险管理是一个贯穿模型全生命周期的持续性迭代过程，由风险识别、分析和评估构成。关键工具包括用于构建风险分类法、用于收集专家对不同风险情境可能性预测的德菲尔技术，以及基于可能性和潜在严重性帮助决策者确定何时须解决具体风险的风险矩阵。尽管已有初步研究开发了一些方法，但仍需要通过进一步工作来确立其他影响风险管理的关键因素，如风险偏好和风险承受能力。[54]

[52] Violet Turri and Rachel Dzombak, ‘Why We Need to Know More: Exploring the State of AI Incident Documentation Practices’, in *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society* (AIES ’23: AAAI/ACM Conference on AI, Ethics, and Society, Canada: ACM, 2023), 576–83, https://doi.org/10.1145/3600211.3604700.
[53] Dixon and Frase, ‘An Argument for Hybrid AI Incident Reporting’.
[54] Leonie Koessler and Jonas Schuett, ‘Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries’ (arXiv, 2023), https://doi.org/10.48550/ARXIV.2307.08823.

- **为风险管理系统设定监管要求**。各国可在提供清晰指导方针方面发挥作用，明确企业如何执行风险评估以及应向监管者或第三方审计人员报告什么材料。例如，《欧盟人工智能法》第9条明确规定，风险管理系统应包括风险类型分析、企业必须进行的识别和评估（如通过上市后监管系统评估部署可能引起的风险）等具体信息。
- **评估企业实践是否遵循既有政府指导和改进建议**。研究者和监管者可以评估企业遵守现有指导并履行自愿承诺的程度。人工智能政策和策略研究所（Institute for AI Policy and Strategy）的报告就是一个这样的例子，它评估了Anthropic负责任扩展策略和英国政府发布的政策指南。[55]

要求进行部署后监控

目标

> 高级人工智能系统可能会逐渐增加与其他人工智能系统和用户进行的复杂多智能体交互，而这可能导致难以预测的潜在风险。部署后的监控是整个保障体系的关键部分，它可以包括**对模型行为的持续自动评估、人工智能事故追踪的集中数据库，以及人工智能在关键系统中的应用报告**。进一步的保障还可以通过自动化运行时验证来实现，**例如确保安全报告中的假设条件依然成立**，并在模型运行到超出预期范围的环境时安全地关闭系统。-人工智能安全国际对话-威尼斯，*2024*

进展

适度

一些形式的部署后监控取得了适度进展。OpenAI等人工智能开发商监控使用模式以检测其人工智能系统被滥用的情况，如被用于网络攻击。[56]部署模型时，腾讯和阿里巴巴等中国企业运行全面的输入输出过滤器，以确保恶意提示和输出被过滤。[57] 另外，事故报告提及的一些进展也有助于稳健的部署后监控。最后，《欧盟人工智能法》要求高风险人工智能系统开发商实施上市后监控，以搜集人工智能系统相关表现数据并监督《欧盟人工智能法》其他要求遵守情况。

人工智能企业和人工智能安全研究所也正通过雇佣研究人员专注于评估人工智能系统的社会影响，以增强对人工智能社会层面影响的理解能力。[58]

然而，许多部署后监控已由模型开发商完成，向政府提供对关键安全行业等的可见性有限。

---

[55] Bill Anderson-Samways et al., 'Responsible Scaling: Comparing Government Guidance and Company Policy' (Institute for AI Policy and Strategy, 11 March 2024), https://www.iaps.ai/research/responsible-scaling.
[56] 'Influence and Cyber Operations: An Update' (OpenAI, 9 October 2024), https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.
[57] '大模型安全与伦理报告 / Tencent Large Model Safety-Security and Governance Report' (Tencent Research Institute, 24 January 2024), https://mp.weixin.qq.com/s/KCWw9gBwUnzywyNW_K8-4A.l.
[58] 'Psychological and Social Risks (Societal Impacts) Workstream Lead (UK AISI)', n.d., https://boards.eu.greenhouse.io/aisi/jobs/4399639101; 'Research Engineer, Societal Impacts (Anthropic)', n.d., https://boards.greenhouse.io/anthropic/jobs/4251453008.

注：事件跟踪在独立的章节中有具体讨论。

潜在政策

- **要求企业收集并共享互联的部署后监控数据。**目前，政府已限制针对部署后数据的可见性。联合部署后监控是指不同类型的部署后信息（如模型集成和使用，应用程序使用，事故信息）相互联结，以便更好进行全面风险评估，并将部署后信息与具体风险缓解措施相联系。 这一方法在其他行业已取得一定影响力。例如，美国食品和药物管理局（FDA）监控并将药品对人群的影响与医生的观察相联系，并告知是否应用新预警标签等决定。[59]
- **监控应触发自动关闭的令人担忧的模型特征的激活** 。一些人工智能安全研究方向的进展，诸如机制可解释性和表征工程，正使得人工智能开发商能够隔离不良或有害行为（如欺骗）所对应的特征（如高级概念）。[60] 在最近的一篇探索性博客文章中，Anthropic主张运行时自动检测出应该立刻关闭的特征是可能的。[61]
- **收集人工智能系统接入关键基础设施的有关数据。**拜登政府2023年人工智能行政命令（2023 AI Executive Order）呼吁风险管理部门评估在关键基础设施领域使用人工智能的潜在风险。[62]这些机构可以且应该在协助关键基础设施运营商负责任地使用人工智能方面发挥作用，应从人工智能开发商和/或关键基础设施运营商中获取人工智能系统使用数据。[63]
- **投资技术治理手段，以增加智能体系统的可见性。**可见性，特别是针对日益增加自主性的人工智能系统活动的部署后可见性，目前是极端受限的。各国可以利用技术治理手段以提升可见性，如代理标识（水印、身份）、实时监控以及活动日志记录以确保使用者和政府能够拥有对于智能体社会使用情况的可见性。[64]

为人工智能从业者制订专门职业道德准则

目标

> 各国可以帮助**建立人工智能工程的伦理规范**，例如要求工程师承担类似于医疗或法律专业人士的个人责任，保护公众利益。 -人工智能国际安全对话-威尼斯，*2024*

进展

[59] Merlin Stein, Jamie Bernardi, and Connor Dunlop, 'The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI' (arXiv, 7 October 2024), http://arxiv.org/abs/2410.04931.
[60] Andy Zou et al., 'Representation Engineering: A Top-Down Approach to AI Transparency' (arXiv, 10 October 2023), http://arxiv.org/abs/2310.01405.
[61] Roger Grosse, 'Three Sketches of ASL-4 Safety Case Components', 5 November 2024, https://alignment.anthropic.com/2024/safety-cases/.
[62] 'Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence', § 4 (2023).
[63] Kyle Crichton et al., 'Securing Critical Infrastructure in the Age of AI' (Center for Security and Emerging Technology, 15 October 2024), https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.
[64] Alan Chan et al., 'Visibility into AI Agents' (arXiv, 17 May 2024), http://arxiv.org/abs/2401.13138.

有限

尽管一些行业协会如英国计算机协会声称从事高风险计算机技术行业的技术专家，特别是人工智能领域的，应是符合道德实践独立标准的注册专业专家，但各国似乎并没有考虑制订这样的规范。[65]

潜在政策

- **开发人工智能“希波克拉底誓词”并要求人工智能从业者进行宣誓。**希波克拉底誓词作为医生个人预防性原则，宣誓“第一，不伤害自己或他人”，将病人的利益放在首位，并坚持高标准职业道德操守。人工智能从业者，特别是那些在可能将灾难性风险带给人类的前沿人工智能系统工作的人，需要进行这样的宣誓。尽管违反誓词的处罚难以执行，誓词中的原则可以在人工智能从业者被要求遵守的具体职业标准中反映出来。
- **规定“警告权”，推动人工智能从业者履行职业道德责任。**与医生一样，职业道德责任可能与企业激励行为相冲突。[66]然而，许多人工智能潜在风险尚未被现有法律所涵盖。 由几位前沿人工智能实验室前职员共同签署的公开信呼吁根据“警告权”加强对吹哨人的保护，其他人工智能从业者也在参议院听证会上关注该议题。[67]

[65] ‘Living with AI and Emerging Technologies: Meeting Ethical Challenges through Professional Standards’ (BCS, Chartered Institute for IT, 15 February 2024), https://www.bcs.org/media/jgmfqo2i/living-with-ai-and-emerging-technologies.pdf.
[66] Sharon A. Clark, ‘The Impact of the Hippocratic Oath in 2018: The Conflict of the Ideal of the Physician, the Knowledgeable Humanitarian, Versus the Corporate Medical Allegiance to Financial Models Contributes to Burnout’, *Cureus* 10, no. 7 (30 July 2018): e3076, https://doi.org/10.7759/cureus.3076.
[67] ‘A Right to Warn about Advanced Artificial Intelligence’, 4 June 2024, https://righttowarn.ai/; Sophie Luskin, ‘Need for Whistleblower Protections in Artificial Intelligence Industry Discussed in Senate Judiciary Hearing’, *Whistleblower Network News*, 24 September 2024, https://whistleblowersblog.org/corporate-whistleblowers/need-for-whistleblower-protections-in-artificial-intelligence-industry-discussed-in-senate-judiciary-hearing/.

国际治理

国内治理行动、关于人工智能安全资助和合作以及测试和评估活动，将在各相关章节中有所反映。这一节将特别关注特定目标的国际层面。

将人工智能安全视为全球公共产品

目标

> **人工智能安全是一项全球公共产品**，应得到公共和私人投资的支持，并广泛分享安全方面的进步。 -人工智能安全国际对话-牛津，*2023*
>
> 由于人工智能带来的风险具有全球性，**我们必须将人工智能安全视为全球公共产品**，并为实现这些风险的全球治理而努力……**我们呼吁各国将人工智能安全视为一个独立于人工智能能力地缘战略竞争的合作领域，**专注于国际学术与技术合作。-人工智能安全国际对话-威尼斯，*2024*

进展

适度
人工智能峰会（AI Summits）聚集了来自不同文化背景和政治体制的关键人工智能力量以讨论人工智能安全议题。人工智能安全和风险也持续被作为中美政府间国际对话的关键议题。同时，许多中国顶尖专家开始传播将人工智能安全视为全球公共产品的理念。

尽管如此，现今情况是脆弱的，人工智能安全与围绕人工智能竞赛的更广泛地缘政治紧张相联系，存在真正的风险。

潜在政策

- **确保人工智能安全国际合作免受更广泛的人工智能地缘政治竞争影响。**国家可以采取措施以确保限制性政策如出口管制不会在无意间影响人工智能安全合作。反例是——2019年美国对于华为的出口管制导致华为被无意间排除在一些标准制定组织外。诸如为服务商能源效率设定标准并通过维护软件工具包以确定服务商能源效率水平的标准性能评估公司（SPEC）等组织， 无法与华为互动，导致该议题全球标准制定出现意见分歧。出口管制最终在两年之后得到调整，但迄今为止尚远不能修复上一次对于服务商能源效率标准国际协作的伤害。[68]

[68] Nigel Cory, 'The U.S.-China Tech Conflict Fractures Global Technical Standards: The Example of Server and Datacenter Energy Efficiency', 22 August 2023, https://itif.org/publications/2023/08/22/the-us-china-tech-conflict-fractures-global-technical-standards-the-example-of-server-and-datacenter-energy-efficiency/.

- **识别、开发并共享安全技术，如“人工智能允许行动联系机制”。**允许行动联系机制（PALs）是电子机械锁的一种，在冷战时期通过预防核武器未经授权的意外使用提升了核安全。该技术的核心要素由美国开发并与苏联共享，旨在促进全球核安全。美国国家安全关键人物如杰森·马西尼，前任曾任国家安全委员会技术事务协调员和现任兰德公司CEO，曾经呼吁开发人工智能允许行动联系机制并和竞争者共享。[69]人工智能允许行动联系机制的例子包括严格增强安全性的安全技术和方法，如异常检测机制以确定人工智能系统何时表现出潜在危险行为。[70]
- **开展更进一步研究，以明确“人工智能安全作为全球公共产品”的概念。**国际货币基金组织（International Monetary Fund）将全球公共产品定义为个人免费使用、对个人好处小且好处将在较远的未来实现的产品。[71]定义的部分内容及其他类似的全球共同产品（如环境）可能适用于人工智能安全，但要确定人工智能安全作为全球公共产品应是什么样的可能需要新的知识进步。尽管中国人工智能治理学者在该方向有早期工作基础，更进一步的工作仍然是必要的。[72]

对齐全球红线标准并建立有条件市场准入

目标

国内监管机构应采用全球统一的要求，以防止跨越这些红线。进入全球市场的条件应以符合国际审计确定的这些全球标准的国内法规为条件，从而有效防止违反红线的系统的开发和部署。 -人工智能安全国际对话-北京，*2024*

进展

有限
迄今为止，在通过对齐全球红线标准建立有条件市场准入方面没有取得任何进步。需要通过更进一步的工作详细说明并实施红线相关标准，将标准与有条件市场准入相联系。

潜在政策

- **在既有协助框架下为红线标准制定确定合适的渠道。**一系列现有程序——人工智能安全研究所、人工智能峰会、国际标准化组织（ISO）/国际电子委员会（IEC）——在红线被划定后可以为国

[69] Jeffrey Ding, ‘Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies’, *European Journal of International Relations*, 27 April 2024, 13540661241246622, https://doi.org/10.1177/13540661241246622.
[70] Oliver Guest and Zoe Williams, ‘Topics for Track IIs: What Can Be Discussed in Dialogues about Advanced AI Risks without Leaking Sensitive Information?’ (Institute for AI Policy and Strategy, 2 May 2024), https://www.iaps.ai/research/dialogue-topics.
[71] Moya Chin, ‘What Are Global Public Goods?’, December 2021, https://www.imf.org/en/Publications/fandd/issues/2021/12/Global-Public-Goods-Chin-basics.
[72] 王迎春 et al., ‘人工智能安全作为 全球公共产品 研究报告 / AI Safety as Global Public Goods Working Report’, 5 July 2024, https://www.sipa.sjtu.edu.cn/Kindeditor/Upload/file/20240704/%E7%A0%94%E7%A9%B6%E6%8A%A5%E5%91%8A%E6%89%8B%E5%86%8C-04.pdf.

家对齐国内红线标准提供渠道。鉴于只有少数国家有能力开发前沿模型，首先统一关键人工智能力量的标准可能更有效。

- **建立监管市场，作为协调监管目标与市场准入的机制。**技术发展的步伐和技术转型的范围可能超过拥有庞大但效率低的官僚机构的国家的直接适应和应对能力。相反地，国家可以考虑将治理方面的技术外包给持证的、合法的私人监管者。[73]在这一模式下，各国将为私人监管者规定需要实现的高层次目标。世界各国可一同建设最低限度的全球监管目标，使人工智能开发商在进入市场时须获得私人监管者相应的认证。
- **实施司法认证程序和多边出口管制制度以管理有条件市场准入。**例如，可建立一个国际人工智能组织（International AI Organization），大致类似于国家原子能机构（IAEA）。该组织将负责制订标准、证明各国有充足能力监控和管理前沿人工智能开发商、协商执行等。特别是，它可以与国家合作，通过多边出口管制机制，拒绝不遵守规定的国家获得相关人工智能原材料（如半导体）和向关键市场出口人工智能产品（如模型）的能力。[74]

制订和协商国际协议和/或成立国际组织

目标

世界各国政府——尤其是领先的人工智能国家——有责任制定措施，防止恶意或粗心行为者造成最坏的结果，并遏制不计后果的竞争。**国际社会应以此为导向，努力建立先进人工智能的国际协调流程。** -人工智能安全国际对话-牛津，*2023*

我们应该采取措施防止最危险技术的扩散，同时确保广泛获得人工智能技术的好处。为了实现这一目标，我们应该建立多边机构和协议，通过执行**机制来安全、包容地管理通用人工智能的发展，以确保不跨越红线并广泛分享利益。** -人工智能安全国际对话-北京，*2024*

就应对先进人工智能系统所需的技术和制度措施，各国应达成一致，无论这些系统的开发时间线如何。为促进这些协议的达成，我们**需要建立一个国际机构，将各国人工智能安全监管部门聚集在一起**，在制定和审核人工智能安全法规方面，推动不同司法管辖区的对话与合作。**该机构将确保各国采纳并实施一套基本的安全准备措施**，包括模型注册、信息披露与预警机制……随时间推移，这一机构还将制订标准并承诺使用验证方法实施国内安全保障框架。这些方法可以通过激励和处罚机制互相强制执行，如将市场准入与遵循全球标准相挂钩。-人工智能安全国际对话-威尼斯，*2024*

[73] Gillian K. Hadfield and Jack Clark, 'Regulatory Markets: The Future of AI Governance' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2304.04914.
[74] Robert Trager et al., 'International Governance of Civilian AI: A Jurisdictional Certification Approach' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2308.15514.

进展

有限

尽管国际治理努力激增，但建立各种与红线相关、聚焦高级人工智能风险的国际协作程序方面进展仍然有限。

几个国际协作程序已被建立，包括联合国高级别咨询机构（HLAB）、人工智能峰会程序、G7广岛程序（G7 Hiroshima Process）、高级人工智能安全国际科学报告和中国全球人工智能治理倡议。关于国际组织，联合国高级别咨询机构发布的中期报告指出几个须由全球人工智能治理实行的关键职能。多个专家也提出几种聚焦人工智能治理的多边组织新模式。

然而，关于人工智能灾难性风险的国际协作非常有限，且较少关注关于红线的实施机制。

潜在政策

- **围绕哪些前沿人工智能治理要素应被国际化建立共识。**当单靠国内治理无法充分解决风险时，需要国际化的治理。除其他原因外，如果风险是跨国性质的（如极端恐怖主义等来自某一司法管辖区的风险可能影响其他区域），以及国家不太想采取代价高昂的单边行动管理风险的情况下（如气候承诺），可能也需要通过国际化解决问题。从完全由国际组织制定的标准，到仅确保国内标准互认，国际化是一个具有广泛范围的概念。建立一个国际风险治理体系，需要就什么需要被国际化、为了什么原因以及国际化要达到什么程度达成共识。[75]
- **就是否应通过现有组织或新程序建立关键国际前沿风险治理功能达成一致。**新的程序和框架允许在没有原有制度负担情况下更清晰地聚焦，但也较少嵌入更广泛的框架中，面临潜在的资源和合法性问题。各国应在前沿人工智能治理职能上达成一致，决定是否设立一个新程序（如人工智能峰会程序）或依赖现有渠道（如联合国人工智能高级别咨询机构）进行。在某些情况下，通过几个机构分解一项职能以服务不同但相关的目标是必要的。例如，可能有必要将就人工智能风险达成科学共识的职能分解为两部分：由联合国领导的、涉及成员国深度参与的更广泛程序，类似于政府间气候变化专门委员会（IPCC）；聚焦高级人工智能安全的独立报告，继续英国委托编写“高级人工智能安全国际科学报告”所启动的程序。[76]
- **实施信任建立措施，确保具有必要的国际信任以构建更广泛的治理程序。**信任建立措施（CBMs）是旨在减少因不明确和不透明而导致的不确定性和风险升级的程序和措施，同时也有助于化解危机局势。信任建立措施具有三个关键功能：提升透明度；提供沟通和协商渠道；为合作、协作和整合提供途径。在人工智能大背景下，信任建立措施可能包括风险热线、事故共享、

[75] For a taxonomy of international governance functions, see Matthijs M. Maas and José Jaime Villalobos Ruiz, ‘International AI Institutions: A Literature Review of Models, Examples, and Proposals’, *SSRN Electronic Journal*, 2023, https://doi.org/10.2139/ssrn.4579773.
[76] Claire Dennis et al., ‘The Future of International Scientific Assessments of AI’s Risks’ (Carnegie Endowment for International Peace, 27 August 2024), https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks?lang=en.

模型卡片、内容来源、协作红队训练、桌面演练和数据集及评测共享。[77] 就关键新兴安全标准进行协商。通过双边工作组、现有标准制定机构或新兴机构如安全研究所网络，国家可以为关键领域制定标准，如识别关键风险，并决定哪一类型模型或训练运行登记信息须在国际共享。双边风险映射演练的一个例子是将新加坡风险框架与美国风险框架相联系的交叉对照表。[78]

[77] Sarah Shoker et al., 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings' (arXiv, 2023), https://doi.org/10.48550/ARXIV.2308.00862.
[78] 'Joint Mapping Exercise between Singapore IMDA and the US NIST', 13 October 2023, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise.

参考文献

'30th Anniversary of JVE: Snezhinsk Scientists Reach out to Congratulate the US Colleagues', n.d. https://nonproliferation.org/lab-to-lab-joint-verification-experiment.

'A New Initiative for Developing Third-Party Model Evaluations'. Anthropic, 1 July 2024. https://www.anthropic.com/news/a-new-initiative-for-developing-third-party-model-evaluations.

'A Right to Warn about Advanced Artificial Intelligence', 4 June 2024. https://righttowarn.ai/.

'AI Research Resource Funding Opportunity Launches', 24 January 2024. https://www.ukri.org/news/ai-research-resource-funding-opportunity-launches/.

Alaga, Jide, and Jonas Schuett. 'Coordinated Pausing: An Evaluation-Based Coordination Scheme for Frontier AI Developers'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2310.00374.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, Jade Leung, Cullen O'Keefe, Jess Whittlestone, Shahar Avin, et al. 'Frontier AI Regulation: Managing Emerging Risks to Public Safety'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2307.03718.

Anderson-Samways, Bill, Shaun Ee, Joe O'Brien, Marie Buhl, and Zoe Williams. 'Responsible Scaling: Comparing Government Guidance and Company Policy'. Institute for AI Policy and Strategy, 11 March 2024. https://www.iaps.ai/research/responsible-scaling.

Baseni, Mikta, Marius Hobbhahn, David Lindner, Alex Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, et al. 'Towards Evaluations-Based Safety Cases for AI Scheming'. Apollo Research, 1 November 2024. https://static1.squarespace.com/static/6593e7097565990e65c886fd/t/6724bd285d993323e03b89d6/1730460969317/Toward_evaluations_based_safety_cases_for_AI_scheming.pdf.

Buhl, Marie Davidsen, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. 'Safety Cases for Frontier AI'. arXiv, 28 October 2024. http://arxiv.org/abs/2410.21572.

Brundage, Miles, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, et al. 'Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims'. arXiv, 20 April 2020. http://arxiv.org/abs/2004.07213.

Chan, Alan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, et al. 'Visibility into AI Agents'. arXiv, 17 May 2024. http://arxiv.org/abs/2401.13138.

Chin, Moya. 'What Are Global Public Goods?', December 2021. https://www.imf.org/en/Publications/fandd/issues/2021/12/Global-Public-Goods-Chin-basics.

Clark, Sharon A. 'The Impact of the Hippocratic Oath in 2018: The Conflict of the Ideal of the Physician, the Knowledgeable Humanitarian, Versus the Corporate Medical Allegiance to Financial Models Contributes to Burnout'. *Cureus* 10, no. 7 (30 July 2018): e3076. https://doi.org/10.7759/cureus.3076.

Clifford, Matt. 'Introducing Def/Acc at EF', 20 May 2024. https://www.joinef.com/posts/introducing-

def-acc-at-ef/.

'ClimateWorks Foundation', n.d. https://www.climateworks.org/.

Clymer, Joshua, Nick Gabrieli, David Krueger, and Thomas Larsen. 'Safety Cases: How to Justify the Safety of Advanced AI Systems'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2403.10462.

'Considerations and Limitations for AI Hardware-Enabled Mechanisms', 10 March 2024. https://blog.heim.xyz/considerations-and-limitations-for-ai-hardware-enabled-mechanisms/.

Cory, Nigel. 'The U.S.-China Tech Conflict Fractures Global Technical Standards: The Example of Server and Datacenter Energy Efficiency', 22 August 2023. https://itif.org/publications/2023/08/22/the-us-china-tech-conflict-fractures-global-technical-standards-the-example-of-server-and-datacenter-energy-efficiency/.

Crichton, Kyle, Jessica Ji, Kyle Miller, and John Bansemer. 'Securing Critical Infrastructure in the Age of AI'. Center for Security and Emerging Technology, 15 October 2024. https://cset.georgetown.edu/publication/securing-critical-infrastructure-in-the-age-of-ai/.

Dalrymple, David 'davidad', Joar Skalse, Yoshua Bengio, Stuart Russell, Max Tegmark, Sanjit Seshia, Steve Omohundro, et al. 'Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2405.06624.

'Defining AI Incidents and Related Terms'. OECD Artificial Intelligence Papers. Vol. 16. OECD Artificial Intelligence Papers, 6 May 2024. https://doi.org/10.1787/d1a8d965-en.

Dennis, Claire, Hadrien Pouget, Robert Trager, Jon Bateman, Renan Araujo, Belinda Cleeland, Malou Estier, et al. 'The Future of International Scientific Assessments of AI's Risks'. Carnegie Endowment for International Peace, 27 August 2024. https://carnegieendowment.org/research/2024/08/the-future-of-international-scientific-assessments-of-ais-risks?lang=en.

Dennis et al., 'What Should be Internationalised in AI Governance', November 2024, http://oxfordmartin.ox.ac.uk/publications/what-should-be-internationalised-in-ai-governance

Ding, Jeffrey. 'Keep Your Enemies Safer: Technical Cooperation and Transferring Nuclear Safety and Security Technologies'. *European Journal of International Relations*, 27 April 2024, 13540661241246622. https://doi.org/10.1177/13540661241246622.

Dixon, Ren Bin Lee, and Heather Frase. 'An Argument for Hybrid AI Incident Reporting'. Center for Security and Emerging Technology, March 2024. https://cset.georgetown.edu/publication/an-argument-for-hybrid-ai-incident-reporting/.

EU AI Act, § 55 (2024).

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, § 4 (2023).

'Global Governance: Goals and Lessons for AI'. Microsoft, n.d. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW1lhQ0.

Guest, Oliver, and Zoe Williams. 'Topics for Track IIs: What Can Be Discussed in Dialogues about Advanced AI Risks without Leaking Sensitive Information?' Institute for AI Policy and Strategy, 2

May 2024. https://www.iaps.ai/research/dialogue-topics.

Grosse, Roger. 'Three Sketches of ASL-4 Safety Case Components', 5 November 2024. https://alignment.anthropic.com/2024/safety-cases/.

Hadfield, Gillian K., and Jack Clark. 'Regulatory Markets: The Future of AI Governance'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2304.04914.

Heim, Lennart, Tim Fist, Janet Egan, Sihao Huang, Stephen Zekany, Robert Trager, Michael A Osborne, and Noa Zilberman. 'Governing Through the Cloud: The Intermediary Role of Compute Providers in AI Regulation'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2403.08501.

Hooker, Sara. 'On the Limitations of Compute Thresholds as a Governance Strategy'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.05694.

'Human Genome Project', n.d. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project.

'Influence and Cyber Operations: An Update'. OpenAI, 9 October 2024. https://cdn.openai.com/threat-intelligence-reports/influence-and-cyber-operations-an-update_October-2024.pdf.

'Inspect: An Open-Source Framework for Large Language Model Evaluations', n.d. https://inspect.ai-safety-institute.org.uk/.

'Joint Mapping Exercise between Singapore IMDA and the US NIST', 13 October 2023. https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2023/nist-imda-joint-mapping-exercise.

Kim, Jerome H, Peter Hotez, Carolina Batista, Onder Ergonul, J Peter Figueroa, Sarah Gilbert, Mayda Gursel, et al. 'Operation Warp Speed: Implications for Global Vaccine Security'. *The Lancet Global Health* 9, no. 7 (July 2021): e1017–21. https://doi.org/10.1016/S2214-109X(21)00140-6.

Koessler, Leonie, and Jonas Schuett. 'Risk Assessment at AGI Companies: A Review of Popular Risk Assessment Techniques from Other Safety-Critical Industries'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2307.08823.

Koessler, Leonie, Jonas Schuett, and Markus Anderljung. 'Risk Thresholds for Frontier AI'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2406.14713.

Kolt, Noam, Markus Anderljung, Joslyn Barnhart, Asher Brass, Kevin Esvelt, Gillian K. Hadfield, Lennart Heim, Mikel Rodriguez, Jonas B. Sandbrink, and Thomas Woodside. 'Responsible Reporting for Frontier AI Development'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2404.02675.

'Living with AI and Emerging Technologies: Meeting Ethical Challenges through Professional Standards'. BCS, Chartered Institute for IT, 15 February 2024. https://www.bcs.org/media/jgmfqo2i/living-with-ai-and-emerging-technologies.pdf.

Luskin, Sophie. 'Need for Whistleblower Protections in Artificial Intelligence Industry Discussed in Senate Judiciary Hearing'. *Whistleblower Network News*, 24 September 2024. https://whistleblowersblog.org/corporate-whistleblowers/need-for-whistleblower-protections-in-artificial-intelligence-industry-discussed-in-senate-judiciary-hearing/.

Maas, Matthijs M., and José Jaime Villalobos Ruiz. 'International AI Institutions: A Literature Review of Models, Examples, and Proposals'. *SSRN Electronic Journal*, 2023. https://doi.org/10.2139/ssrn.4579773.

Mökander, Jakob, Jonas Schuett, Hannah Rose Kirk, and Luciano Floridi. 'Auditing Large Language Models: A Three-Layered Approach'. *AI and Ethics*, 30 May 2023. https://doi.org/10.1007/s43681-023-00289-2.

Mukobi, Gabriel. 'Reasons to Doubt the Impact of AI Risk Evaluations'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2408.02565.

'National Artificial Intelligence Research Resource Pilot', n.d. https://new.nsf.gov/focus-areas/artificial-intelligence/nairr.

'NTU Serving up New Undergrad Course on Meat Alternatives', 25 June 2021. https://www.edb.gov.sg/en/business-insights/insights/ntu-serving-up-new-undergrad-course-on-meat-alternatives.html.

O'Brien, Joe, Shaun Ee, Jam Kraprayoon, Bill Anderson-Samways, Oscar Delaney, and Zoe Williams. 'Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.01420.

Petrie, James, Onni Aarne, Nora Ammann, and David 'davidad' Dalrymple. 'Interim Report: Mechanisms for Flexible Hardware-Enabled Guarantees', 23 August 2024. https://yoshuabengio.org/wp-content/uploads/2024/09/FlexHEG-Interim-Report_2024.pdf.

'Post-Quantum Cryptography', n.d. https://www.nist.gov/pqcrypto.

'Project Moonshot: An LLM Evaluation Toolkit', n.d. https://aiverifyfoundation.sg/project-moonshot/.

'Psychological and Social Risks (Societal Impacts) Workstream Lead (UK AISI)', n.d. https://boards.eu.greenhouse.io/aisi/jobs/4399639101.

'Research Engineer, Societal Impacts (Anthropic)', n.d. https://boards.greenhouse.io/anthropic/jobs/4251453008.

Reuel, Anka, Ben Bucknall, Stephen Casper, Tim Fist, Lisa Soder, Onni Aarne, Lewis Hammond, et al. 'Open Problems in Technical AI Governance'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.14981.

'Safety Case Specialist (Safety Mitigations) (London)', n.d. https://www.linkedin.com/jobs/view/safety-case-specialist-safety-mitigations-london-at-anthropic-3956180495/?originalSubdomain=uk.

Scher & Thiergart, 'Mechanisms to Verify International Agreements About AI Development', November 2024, https://techgov.intelligence.org/research/mechanisms-to-verify-international-agreements-about-ai-development

Shoker, Sarah, Andrew Reddie, Sarah Barrington, Ruby Booth, Miles Brundage, Husanjot Chahal, Michael Depp, et al. 'Confidence-Building Measures for Artificial Intelligence: Workshop Proceedings'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2308.00862.

Stein, Merlin, Jamie Bernardi, and Connor Dunlop. 'The Role of Governments in Increasing Interconnected Post-Deployment Monitoring of AI'. arXiv, 7 October 2024. http://arxiv.org/abs/

2410.04931.

Stein, Merlin, and Connor Dunlop. 'Safe beyond Sale: Post-Deployment Monitoring of AI'. Ada Lovelace Institute, 28 June 2024. https://www.adalovelaceinstitute.org/blog/post-deployment-monitoring-of-ai/.

Stein, Merlin, Milan Gandhi, Theresa Kriecherbauer, Amin Oueslati, and Robert Trager. 'Public vs Private Bodies: Who Should Run Advanced AI Evaluations and Audits? A Three-Step Logic Based on Case Studies of High-Risk Industries'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.20847.

'Tech Secretary Unveils £8.5 Million Research Funding Set to Break New Grounds in AI Safety Testing'. UK Government, 22 May 2024. https://www.gov.uk/government/news/tech-secretary-unveils-85-million-research-funding-set-to-break-new-grounds-in-ai-safety-testing.

'The Audacious Project', n.d. https://www.audaciousproject.org/faq.

Smialek, Jeanna. 'How Jackson Hole Became an Economic Obsession'. *The New York Times*, 25 August 2023. https://www.nytimes.com/2023/08/24/business/economy/jackson-hole-economic-conference.html.

'Third-Party Testing as a Key Ingredient of AI Policy'. Anthropic, 25 March 2024. https://www.anthropic.com/news/third-party-testing.

Trager, Robert, Ben Harack, Anka Reuel, Allison Carnegie, Lennart Heim, Lewis Ho, Sarah Kreps, et al. 'International Governance of Civilian AI: A Jurisdictional Certification Approach'. arXiv, 2023. https://doi.org/10.48550/ARXIV.2308.15514.

Turri, Violet, and Rachel Dzombak. 'Why We Need to Know More: Exploring the State of AI Incident Documentation Practices'. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 576–83. Canada: ACM, 2023. https://doi.org/10.1145/3600211.3604700.

Wasil, Akash, Tom Reed, Jack Miller, and Peter Barnett. 'Verification Methods for International AI Agreements', 2024. https://doi.org/10.2139/ssrn.4938419.

Wasil, Akash, Everett Smith, Corin Katzke, and Justin Bullock. 'AI Emergency Preparedness: Examining the Federal Government's Ability to Detect and Respond to AI-Related National Security Threats'. arXiv, 2024. https://doi.org/10.48550/ARXIV.2407.17347.

'White House Voluntary AI Commitments'. White House, September 2023. https://www.whitehouse.gov/wp-content/uploads/2023/09/Voluntary-AI-Commitments-September-2023.pdf.

'XPrize', n.d. https://www.xprize.org/.

Young, Shalanda. 'Advancing Governance, Innovation, and Risk Management for Agency Use of Artificial Intelligence'. Executive Office of the President, Office of Management and Budget, 1 November 2023. https://www.whitehouse.gov/wp-content/uploads/2023/11/AI-in-Government-Memo-draft-for-public-review.pdf.

Zou, Andy, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, et al. 'Representation Engineering: A Top-Down Approach to AI Transparency'. arXiv, 10 October 2023. http://arxiv.org/abs/2310.01405.

ZHOU Hui, ZHU Yue, ZHU Lingfeng, SU Yu, YAO Zhiwei, WANG Jun, CHEN Tianhao, et al. 'The Model Artificial Intelligence Law (MAIL) v.2.0 - Multilingual Version', 16 April 2024. https://doi.org/10.5281/ZENODO.10974162.

'大模型安全与伦理报告 / Tencent Large Model Safety-Security and Governance Report'. Tencent Research Institute, 24 January 2024. https://mp.weixin.qq.com/s/KCWw9gBwUnzywyNW_K8-4A.

王迎春, 贾 开, 陈 玲, 赵 静, 秦川申, 袁 媛, 傅宏宇, and 梁兴洲. '人工智能安全作为 全球公共产品 研究报告 / AI Safety as Global Public Goods Working Report', 5 July 2024. https://tinyurl.com/AIGPGWP