Formal Response

Recommendations for an AI Action Plan

Response to OSTP's Request for Information

Author
Center for Security and Emerging Technology

CSET CENTER *for* SECURITY *and* EMERGING TECHNOLOGY

March 2025

引言

乔治敦大学安全与新兴技术中心（CSET）根据2025年1月23日总统行政命令的要求，提交了以下关于人工智能（AI）行动计划发展的建议。这些建议来自CSET在评估和评估、生物安全、网络安全和AI、军事AI应用、技术监控、美中竞争分析以及AI劳动力研究等领域广泛的研究。针对特朗普政府的建议主要分为三个类别：1）美国可以采取的步骤来推进和确保其在开发尖端AI能力方面的领导地位，2）与中国的AI竞争的倡议，以及3）美国政府可以采取的措施以实现AI的益处同时减轻其风险。

我们首先提出了通过在研究、创新和人才方面的投资来提升美国人工智能领导地位的几项建议。首先，我们建议增加对公共部门在人工智能应用方面的社会和生命科学研究资金，并广泛分享发现以促进新的发现。其次，我们建议降低新兴人工智能公司的进入门槛，以刺激创新和健康的市场竞争。推动竞争性市场和开放的AI模型将能够为前沿研究提供新的、多样化的方法，为实现强大模型的好处提供替代途径。最后，我们建议美国进一步发展人工智能人才，以加强我们的研究生态系统并创造新的商业机会。增加劳动力培训计划资金并创建人工智能奖学金服务项目将为私营部门工人和公务员配备最有效地开发和利用人工智能工具所需的技能。保持人工智能领导地位需要持续承诺于那些已被证明对美国具有根本优势的研究和专业知识。

我们接下来提出应对美国和中国之间人工智能竞争的建议。首先，我们建议政府采取双管齐下的策略来加强美国的科技竞争力。防止非法技术转移，并增强美国对关键人工智能组件出口控制的效力，最终将削弱中国的人工智能生态系统。这需要与盟国和合作伙伴紧密合作，制定联合战略并获得广泛支持以实现出口控制目标：合作将转化为更高的效率。我们还建议政府优先考虑缓解美国政府和前沿人工智能公司之间的信息不对称，以避免技术惊喜并与盟国共享情报。对开源情报收集的投资，无论是与国内创造的技能相关还是与中国的AI生态系统发展相关，对于形成积极主动的美国AI政策至关重要。最后，我们建议政府

实施人工智能事件报告制度，以汇总人工智能故障模式和新兴风险的数据，从而更好地指导安全科学研究优先事项和风险缓解策略。

最后，我们建议美国政府采取行动，在最大限度地降低风险的同时实现人工智能的益处。提供对人工智能系统造成的伤害进行补救的公共保证，并保护举报人免受公司报复，可以帮助公众更自信地与人工智能互动。我们建议政府建立标准途径来挑战人工智能驱动的负面决策，并为前沿人工智能公司的员工建立举报人保护，这最终可能通过减少危险行为来提高系统性能。政府还应通过创建用于造成损害的威胁档案以及针对这些风险定制的模型安全措施，积极主动地防范人工智能风险。人工智能评估和标准在阐明系统能力和通过与企业合作推动人工智能进步方面发挥着关键作用。政府应授权联邦机构推进人工智能评估科学和标准，并展示评估采用如何促进人工智能在任务成功中的应用。

RFI响应：人工智能（AI）行动计划的发展

联邦登记文献引用：90 FR 9088

机构名称 国家科学基金会，网络与信息技术研究与发展国家协调办公室
组织：中心安全与新兴技术（CSET），乔治城大学
主要目标原位化合物：米亚·霍夫曼（mh2171@georgetown.edu），杰克·卡尔斯滕（jk2497@georgetown.edu），米娜·纳拉扬南（mjn82@georgetown.edu）

乔治敦大学安全与新兴技术中心（CSET）针对国家科学基金会网络与信息技术研究与开发国家协调办公室就以下评论征求意见的要求，提供如下意见：
人工智能行动计划的发展 乔治城大学内的一个政策研究机构 华尔什外交学院 CSET 为决策者提供关于新兴技术安全影响的数据驱动分析，重点关注人工智能、高级计算和生物技术。我们感谢有机会提供这些评论。

本文件经批准对外公开分发。文件中不包含商业机密或保密信息。政府可在无需注明来源的情况下，将文件内容用于制定人工智能行动计划及相关文件。

概述

乔治城大学安全与新兴技术中心（CSET）根据2025年1月23日总统行政命令，提出以下关于人工智能（AI）行动计划发展的建议。这些建议源自CSET广泛的研究成果，主要分为三类：1）美国可以采取的步骤来推进并保障其在开发尖端AI能力方面的领导地位；2）与中国在AI领域的竞争措施；3）美国政府可以采取的措施，以实现AI的益处同时减轻其风险。

推动人工智能研究与开发

推动人工智能研究与发展（R&D）应在新的AI行动计划中成为优先事项。联邦政府在填补关键性R&D缺口上发挥着关键作用，并且处于独特地位来推动特定领域的研究优先级和生产。鉴于私营部门在推动AI研发中的主导地位，政府可以在没有即时利润动机的重要研究领域发挥关键作用，如理解AI在社交、政治和经济方面的含义。为了促进美国的AI研发领导力，政府应鼓励和奖励采用多学科方法的科研项目，鼓励研究成果公开、广泛地传播，并支持与私营部门创新协调一致的公共部门研究。由于AI是一种通用技术，基础R&D支持下游模型的发展，用于商业、应用，最终实现盈利。我们必须建立一个蓬勃发展、多学科跨行业的AI研究生态系统，以增强美国在AI领域的主导地位。

- **将商业人工智能的开发和创新与由公共部门推动的稳健、持续的研发资金相结合。** 这包括在大学、国家实验室、联邦资助的研究与发展中心和非营利机构进行的研发，有时与私营部门合作，涉及众多技术与非技术领域。这类研究包括：探究人工智能如何与人类行为、过程和结构相互作用和产生影响；旨在提升我们对研究成果转化为商业创新和实际应用的理解和映射的指标；以及创建详实且有广泛可用性的开源数据集。

- **促进人工智能增强的科学进步：** 交点 人工智能与生物技术（AIxBio）具有巨大的发展潜力。 ——不仅为了生物医学和工业创新，而且还作为全球经济优势和科技影响力竞争的战场，这将授予在AIxBio技术发展中领先的国家的。

一个未来的AI行动计划必须包括强大的倡议，以推进以AI驱动的生物技术，以便美国能享有这些利益。这样一个行动计划应包括：

○ **支持AIxBio的资金、能力、基础设施和人力。** 虽然一些最大的、计算密集型的生物人工智能模型由行业开发，但许多更专门化的、领域特定的AIxBio模型来自学术研究小组。为AIxBio研究人员提供额外资源将提高可实现的开发的速度和范围，并促进它们向实际应用的转化。

○ **生物数据基础设施** 目前，未来AIxBio应用能够使用的生物数据类型存在重大限制，包括缺乏大型、标准化、标记和注释的训练集，适用于AI系统。美国应优先发展更强大的数据库，并探索激励措施，以鼓励研究人员以AI可用的格式收集生物数据。

刺激市场、竞争与创新

培养动态和竞争性市场

维持美国在人工智能领域的长期领导地位需要建立一个动态且多元化的国内人工智能系统市场。这样一个生态系统——其中许多不同的开发者被授权追求各种人工智能技术和工具，并将他们的成功转化为可行的企业——将促进创新并确保美国在人工智能领域始终保持领先地位。然而，目前美国的AI行业由少数几家现有企业主导，这些企业拥有权力和动机来阻挠竞争对手的美国AI开发者，并可能抑制颠覆性的人工智能创新。为了促进美国AI行业的长期健康发展，政策制定者应降低新AI开发者的进入壁垒，并确保现有企业不利用其权力来扼杀AI市场的竞争。为此，我们提出以下三条建议：

● **推动一个更加开放和竞争的云计算市场。** 计算资源是构建和部署人工智能系统的关键输入。许多美国人工智能开发者通过以下方式访问这些资源： 云服务提供商 （CSPs）如亚马逊网络服务、微软Azure和谷歌云平台。政策制定者应打击出口费用、限制性合同条款和 其他战术 那些~~电信服务提供商用来~~“锁定”客户的方法，并实施非歧视和开放接入规则，以防止电信服务提供商在大型人工智能开发商和较小的人工智能公司之间偏袒前者。

● **维持人工智能产品的开放分销渠道。** 为了构建成功的商业，人工智能开发者向客户推广和分销他们的产品。许多人工智能产品最突出的“分销渠道”——移动设备、软件套件、在线市场及应用程序商店和云计算平台——都归以下所有： 现任技术公司 内部开发人工智能系统或与知名第三方开发者保持财务联系。政策制定者应禁止公司从事

自我偏好、捆绑和其他允许他们排除竞争对手人工智能开发者从有价值的产品分销渠道的策略。这种行为可能使新的AI开发者更难在市场上立足，从长远来看阻碍竞争和创新。

- **密切监测人工智能行业中的并购（M&A）和企业"伙伴关系"。** 并购交易 可能产生积极的创新效应，使公司能够实现规模经济，获取新技术和人才，同时也可能产生负面效应，可能降低现有企业的创新动力，并使他们能够阻止颠覆性竞争者进入市场。近年来，现有科技企业也开始参与" 合作关系 与外部AI开发者合作。这些安排不受与传统并购交易相同的监管审查，并且他们的细节通常是 不透明 他们可能对创新有类似的正负面影响。联邦贸易委员会和司法部反垄断司应继续密切审查人工智能领域的并购交易和企业合作，并在必要时阻止企业合并，以维护竞争和可竞争的市场。

推广开放人工智能模型

最近人工智能基础模型的进步引发了关于自由发布模型权重的好处和风险的讨论。我们建议特朗普政府在没有超过明确的和可衡量的风险阈值的情况下，避免抑制美国公司发布开源模型。开源模型为美国的优势提供支持：它们促进创新和竞争，促进人工智能安全的进步，并鼓励企业家进入人工智能行业。特别是，我们建议以下事项：

- **支持 开源人工智能模型、数据集和工具的发布** 这可以用于推动美国人工智能发展、创新和经济增长。 开源模型 并且工具启用 更高的参与度 在人工智能领域，允许 低资源组织 他们无法自行开发基础模型以访问、实验和在此基础上构建。刺激经济增长 通过增加竞争和吸引更多企业家。开源数据集允许进行 持续的基准测试 人工智能进步，使对竞争优势或差距的理解更加深入，并激励开发者追求更大的成功。

- **提供 资源用于评估和分析开放模型的影响** 在推进人工智能研究和推动私营部门人工智能增长方面。此类分析可以界定开放模型的益处，可从开放模型发布的潜在风险视角进行比较。 人工智能与中国的竞争 .

- **开发 最佳实践：前沿开源模型发布** 协助 预先识别并最小化 风险，但避免不必要的阻碍或削弱模型开放意愿的法规。与行业合作，建立清晰和可衡量的不容忍风险阈值，以证明关闭模型是合理的，并避免过度强调假设风险。确保这些阈值考虑到边际风险，并在开放带来的好处与风险之间取得平衡。

- **优先考虑 ，伴随人工智能能力的发展，美国人工智能模型在美国和全球人工智能生态系统的扩散** 美国开放模式在国外得到采用，从而建立了对美国技术的依赖，因此赋予了美国政府 软实力，作为与合作伙伴建立更牢固关系和联盟的基础，并鼓励进一步付费使用相关美国AI技术，例如企业订阅服务和企业云计算平台。在国外推广美国AI技术也有助于对抗 影响力不断扩大 中国模型在发展中国家和新兴经济体的应用，以及防止中国为全球数字基础设施的大部分提供基础，对全球产生的影响。 扩散 的 中国意识形态 在世界。

激励多种前沿研究方法的实施。

- **激励美国及其盟友在先进人工智能研究方面采取替代方法。** 美国和西方国家认为，尽管这些模型成本高昂、基础设施需求巨大且存在已知限制，但大型生成式AI模型是通往通用人工智能的主要途径。很可能还有其他途径可以推进AI发展，而不必将大量资源投入一个即将接近规模极限的范式。相比之下，中国也在投资替代方案，如人脑建模、非治疗性脑-机接口和通过与环境价值驱动互动学习的具身AI。这种最后一种方法现在正在中国城市大规模实施。如果美国担心通用人工智能的出现，它应该对可行的途径进行多次押注，承认单一焦点的不稳定，并支持替代方案，这是中国政策制定者认识到的一个真理。

开发并保障人才获取。

加强日益增长的AI劳动力。

学徒制为美国所有背景的工人提供了一条培训、再培训和提升技能的途径，这包括与人工智能相关职业的学徒制。 在过去十年中，数量迅速增加。 此增长与联邦政府资金和连续政府的支持增加相吻合。未来的成功将取决于对这些项目的持续支持。我们建议特朗普政府：

- **增加联邦国家学徒制系统的资金投入，** 在重视技术职业和行业中介的同时。政府还应提供资金用于收集和追踪注册学徒制项目的就业结果，以确定这些项目是否能导致学徒获得高薪工作。

社区学院在培养下一波人工智能工作者方面具有巨大的潜力，但 需要资金和支持以成功 社区学院遍布全国各地，并有着长期培训各年龄层新兴行业工人的历史。然而，它们面临着诸多挑战，包括不确定的、复杂的以及资金流不足等问题。我们建议特朗普政府：

- **全额资助并重新授权职业教育和技术教育项目** 类似地《加强21世纪职业教育法案》（Strengthening Career and Technical Education for the 21st Century Act） 珀金斯诉案), 国家自然科学基金委员会 高等工程技术教育 程序，以及 加强社区学院培训资助 项目。许多学院依赖这些项目的联邦资金来开发和继续提供在人工智能等新兴技术领域的培训。

推动更广泛的AI教育范围

保持美国在人工智能领域的竞争力需要培养和维持必要的工作队伍。对于美国政府来说，能够吸引、招募和保留技术人才对于联邦工作队伍至关重要。在众多将顶级人才引入政府的途径中，奖学金服务计划仍然是政府服务的直接人才渠道。例如，美国国家科学基金会（NSF） 网络军团奖学金-服务项目在很大程度上被认为是一项成功。 由于其长寿和持续的国会资金支持。

- **支持创建一个AI奖学金-服务项目。** 2024年，美国国家科学基金会（NSF）发布了一份报告 详细阐述了在《芯片和科学法案》之后实施人工智能奖学金服务计划的可行性和必要性。国家科学基金会（NSF）的人工智能研究学院提供了一个有前景的场所来培养潜在的人工智能奖学金服务计划，因为这些学院专注于各种领域的人工智能应用，并且与联邦政府已有合作关系。全国共有23个获得NSF认定的学院，其中9个拥有活跃的网络安全部队（CyberCorps）项目，12个被指定为网络学术卓越中心（NCAE）。

人工智能素养通常主导着教育政策讨论。教育工作者、学校系统和教育部已经动员起来，以适应和应对教育系统中人工智能的出现。然而，仅仅关注课堂上的人工智能素养努力，却排除了美国公民的许多部分。人工智能素养可以通过使公民了解人工智能及其局限性，消除恐惧或担忧，并帮助个人对其创造性、原创性工作和思想拥有所有权来支持公民。

- **与美国国会合作，以支持美国人民的人工智能素养努力。** 在2024年，参议员Kelly和Rounds 引入了一项法案。 旨在提升消费者对人工智能产品和服务的认知度和信心。 配套法案 后来，由布莱恩特·罗切斯特众议员提出。这些法案不仅限于课堂教学，旨在为美国公民提供必要的教育和信息，以便在人工智能的使用和消费方面做出明智决定。

与中华人民共和国竞争

禁止向中华人民共和国转移非法技术。

中国的比较优势，现在和历史上，一直在于将中国以外的科学进步商业化。尽管华人社群和全球数据管道模糊了中国真实的能力，但中国科学家仍然承认他们对外国基础科学的依赖。这就是中国法律和非法的技术转移项目越来越关注的领域。尽管认识到中国本土研究的新能力，但该国仍然从其通过盗窃、挪用和其他单方面实践中获得的收益中获益巨大。这在人工智能领域尤其如此，因为产品是数字化的，更容易秘密获取。

- **在ODNI内部创建一个办公室或工作组以追踪技术向中国转移** 美国政府在通过研究安全倡议（即）解决此问题的努力。 NSF-支持的SECURE项目 ）将参与中国的才能计划行为合法化推进了事态的发展。遗憾的是，被揭露的场馆和做法被新颖的获取策略所取代。CSET分析师曾 一套实用的建议 为了减轻中国在相关领域的过度行为，但缺乏一个美国政府在其中的监控和讨论的重点。

评估和监测出口控制的有效性

出口管制是一种重要的经济国家政策工具，旨在对中国和其他竞争者的技术雄心施加延迟和成本，尤其是在半导体技术和人工智能领域。鉴于人工智能硬件和开放式以及封闭式重量级人工智能模型快速发展的能力，确保出口管制政策相应调整至关重要。

- **商务部工业和安全局（BIS）应建立** 场景规划评估 在实施新的出口管制并严格监控其实施效果之前 当前出口管制政策 .

- **情景规划评估应明确阐述。** 他们应包括出口管制政策目标、对基本假设的分析和测试、对美国及其盟国企业的经济影响评估、对中国潜在反制措施和适应性的评估，以及考虑近、长期后果。

- **国际清算银行（BIS）还应进行定期的实施后评估。** 跟踪实现既定控制目标的进展，二级效应，对中国半导体制造设备行业的影响，中国半导体制造能力的发展，以及中国人工智能领域的进步。

与盟友和合作伙伴合作，确保控制措施保持有效。

国际清算银行应继续与盟国紧密合作，制定联合出口管制策略，并改善关于为何需要这些管制以保护共同利益的沟通和信息共享。

- **明确表达 并证明出口管制的目标合理性。** 为了使更广泛的美国出口管制战略得以实施，向盟友提供支持。

- **避免过度使用外国直接产品规则（FDPR）。** 为了扩大美国出口管制的范围。增加使用风险可能激励外国公司在不使用美国技术和组件的情况下进行设计，从而削弱多边努力，并长期损害美国战略。

改善人工智能信息环境

利用开源情报避免技术惊喜

与近几十年来大多数重大技术进步形成鲜明对比，人工智能的开发、部署和使用几乎完全在联邦政府之外进行，并且在很大程度上是在美国以外的国家进行的。这种状况带来了诸多影响，其中包括使美国政府处于固有的信息劣势。美国政策制定者只有在了解当前技术状况及其未来几年发展方向的情况下，才能充分利用人工智能的经济、战略和创新潜力，规避其带来的风险，并确保美国在人工智能领域的领导地位。遗憾的是，目前在政府内部或外部，能够提供对人工智能景观的全面视角的组织。因此，我们建议以下内容：

- **显著扩展人工智能领域开源情报（OSINT）的收集和分析工作。**
这项工作在情报界特别被忽视，该领域仍然专注于机密来源。在其他联邦政府机构中，它的发展严重不足，资源也匮乏。在收集、解读和传播人工智能开放源情报（AI OSINT）方面，需要重大投资，包括研究出版物、供应链数据、市场研究、专利、资本市场数据和劳动力数据等来源。

- **将中国的人工智能生态系统作为本人工智能开源情报计划的特别关注点。** 缺乏一个严肃的项目来跟踪中国的人工智能进步，削弱了在出口管制、贸易政策、研究安全和产业政策等政策领域的联邦努力，以及

提高技术突发的风险。中国本身运行着一个针对美国军事和民用研发的10万人规模的基于开源的监控系统，从而推动其在人工智能和其他关键技术的发展。联邦政府应显著加大监控中国人工智能生态系统的力度，包括中国政府本身（所有相关层面和组织），相关行为者如国有企业、国家研究实验室和国家资助的技术投资基金，以及其他行为者，如大学和科技公司。

共享政府与私营部门间的情报。

关键信息关于前沿人工智能能力是 孤立的 在人工智能公司中，企业的发展实践透明度对于美国政府应对快速的人工智能发展以及预测对国家安全的新兴威胁是必要的。美国政府还应通过利用盟友关于重要人工智能发展的信息来增强其远景扫描能力。相反，美国政府收集的情报有助于公司加强其抵御单一或国家行为者攻击的防御。我们建议特朗普政府与盟国和人工智能公司交换有关人工智能能力的关键信息，并消除参与此类信息共享的障碍。

- **建立 报告程序以收集来自人工智能公司的AI开发过程信息**．报告程序可以要求公司提供详细的文件记录。 培训程序和环境 ， 意外 或 关于 在新模型中找到的能力， 模型规格 (也称为 宪法) 定义公司希望AI模型具备的行为和评估。 详细的关于人工智能开发实践文档将缩小人工智能开发者和政府之间的信息差距，使政府能够迅速应对人工智能系统能力的突然飞跃。

- **合作伙伴 企业共享威胁情报。** 美国政府应与AI公司合作，共享可疑的用户行为模式和其他类型的威胁情报。特别是，情报界和D 美国国土安全部应与人工智能公司合作共享网络威胁情报，并且美国国土安全部应与人工智能公司合作，为由于人工智能系统被恶意使用或失去控制而可能引发的紧急情况做好准备。此外，美国商务部应该 接收、分类和分配报告 在基于前沿人工智能模型的CBRN和网络安全能力支持对新型人工智能威胁的机密评估方面，建立在此基础上 2024 记忆体交换协议 在能源部和商务部之间。

- **贡献并汲取美国盟友在人工智能能力方面的集体智慧。** 人工智能系统的影响跨越国界，关于它们的观察也超越了国界。

人工智能在盟国的进展也可能在本国相关。美国政府应利用来自可信赖盟友，如英国及其人工智能安全研究所的前沿人工智能能力信息，与他们共享信息以保持信任。

鼓励报告AI事件以促进技术采用

公共和私营部门对人工智能系统的日益部署不可避免地导致了对人工智能系统的使用数量不断增加。 失败和有害事件 涉及人工智能。如果美国政府继续不追踪此类人工智能事件，它将错失一个关键的机会来促进人工智能创新和采用。人工智能事件报告和分析加速了对人工智能失败的认知，这些失败出现在人工智能研究最需要的地方，并帮助开发人员创新和改进他们的模型。通过防止重复失败并提高人工智能系统的可靠性，事件报告不仅降低了危害美国公众的风险，还有助于建立消费者和用户对技术的信任。这促进了人工智能的广泛应用，进而实现了人工智能的经济效益。我们建议美国政府：

- **实施一个 强制性的AI事故报告 针对联邦机构敏感应用的制度。** 联邦机构部署人工智能系统，用于广泛的涉及安全和权利影响的使用案例，例如使用人工智能提供政府服务或预测犯罪再犯。在这些情况下，应跟踪和调查人工智能的故障、故障和其他事件，以确定其根本原因，告知风险管理实践，并降低再次发生的风险。当从第三方获取人工智能系统时，供应商应在发现事件后的24小时内向机构报告人工智能事件。

- **直接机构监管高风险领域以实施 混合事件报告 各自行业中的方案。** 高风险领域包括但不限于医疗保健、交通、教育、就业、金融、住房、保险、公用事业和关键基础设施。关于构成人工智能事件或故障的标准应由各机构自行确定。联邦机构应获授权调查此类事件，以确定原因、共性以及新出现的趋势，并传播所学知识和更新的人工智能风险管理建议。

降低人工智能的风险。

保护公众免受AI造成的伤害。

人工智能系统的风险威胁到损害美国人工智能政策目标。如果公民无法确保他们可以从人工智能的危害中寻求救济或人工智能系统正常工作，那么他们可能会不愿意采用人工智能系统，这对实现人工智能益处的相关政策目标将产生不利影响。

- **创建标准途径以质疑人工智能结果** 美国公民如果在受到人工智能辅助决策的实质性影响，而没有高效且可获取的方式来质疑错误的决策，他们将不会信任人工智能系统。此外，受影响个人的报告是识别人工智能错误的有效途径，这是纠正这些错误所必需的前期工作。

- **为举报人工智能公司危险行为的员工建立举报者保护措施。** 告密者保护可以保护员工免受公司报复。 并帮助确保AI公司遵守其承诺和法律。特朗普政府应建立一条安全热线，供员工报告有问题的公司做法，例如未能报告威胁国家安全的系统功能。

防范由人工智能赋能的生物风险

存在担忧，认为人工智能可能会加剧生物学风险，例如，使得非专业人士更易制造生物武器，或者使更严重或针对性的病原体和毒素得以创造。我们建议以下步骤以防范这些结果：

- **构建一个集成的生物安全生态系统。** 尽管恶意行为者可能会在他们的计划中使用人工智能来造成生物伤害，但实现这一目标所必需的基本信息和资源 AI不可用 . 生物安全策略若仅专注于控制人工智能的使用，则无法在未防御人工智能增强型及人工智能不可知生物剂的情况下成功。相反，针对人工智能使用的机制应整合到更广泛的生物安全策略中，并被视为更广泛工具箱中的一件工具。 全面治理工具包 .

- **部署适当的模型保障措施。** 模型安全保障可以被部署以应对安全担忧并针对人工智能生命周期中的各个节点。在 CSET报告 我们确定了潜在的模型治理机制：为开发者提供生物安全培训、数据过滤、限制访问某些数据集和计算基础设施、发布前的评估、模型访问控制、使用监控和伤害报告机制。

- **需要 未来的政策将明确规定它们是否适用于模型 仅在生物或化学数据上训练** 尤其是在使用“基础模型”或“大型语言模型”等术语时。一些术语与通用型和化学生物AI模型都相关（ ） 模型，能够协助分析、预测或生成新的化学或生物序列、结构或功能 ），但在不同的语境中有不同的定义，当它们被引用时容易造成混淆。这对于现有的监管和指导文件尤其具有挑战性，其中大部分对于是否包含化生生物人工智能模型都没有给出明确说明。

- **定义 关注的能力和支持创建不同类型人工智能模型的威胁档案。** A 评估一个AI模型是否能够输出潜在风险的生物信息

信息与量化这一风险谱系具有挑战性，因为许多病原体随时间演变，在某些条件下危险，而在其他条件下则无害。同样，不同用户和人工智能工具的组合会影响潜在的伤害以及最可能有效的政策解决方案。 评估策略和相关缓解措施 。政府部门联盟应制定框架，明确界定风险能力，包括关注的化学生物能力，以便评估者了解需要测试哪些风险。 这些框架可以参考美国国家标准与技术研究院（NIST）草案的第D附录。 对双用途基础模型管理滥用风险 此外，政府机构应构建考虑不同用户组合、AI工具和预期结果的威胁档案，并为这些高度可变场景设计有针对性的政策解决方案。

推动人工智能评估科学与标准的发展

将人工智能评估科学推进至理解模型能力。

评估应提供关于人工智能系统在特定应用中的安全性和适宜性的决策信息，以及是否应使用人工智能系统。评估结果可以提供关于人工智能系统能力以及实施技术护栏或提升人工智能用户技能等干预措施的有效性的见解，这些措施将塑造人工智能的未来发展或采用。鉴于人工智能系统在许多政策领域的日益重要性，确保人工智能评估的严格性和可靠性，以及决策者理解如何解读其结果，至关重要。我们建议人工智能行动计划将评估作为推动人工智能进步的基础性工具。

- **联邦资助机构，如国家科学基金会，应支持与整体提高人工智能评估科学相关的基础研究，特别是针对"代理"系统的研究。** 代理人工智能系统 能够独立在复杂环境中追求复杂目标，并预期在不久的将来将变得更加有能力。 对人工智能代理能力评估和技术机制的基本研究
控制和管理AI代理的行为对于改善其长期性能至关重要。

- **美国人工智能安全研究所（AISI）应与其他利益相关者合作，推进人工智能评估科学，并消除评估前沿人工智能模型的重复工作。** AISI至今已有效与产业、学术界以及其他合作伙伴合作，以改善人工智能模型的评估。特朗普政府应赋予AISI为人工智能开发量化基准的权力，包括测试模型的能力的基准。 对越狱的抵抗 ，
对于制造CBRN武器的实用性 ，并且 欺骗的能力 最近宣布的 美国钢铁公司（AISI）与Scale AI的合作 展示了美国政府如何与第三方合作开发前沿人工智能评估，并有效地利用测试。

基础设施。此类合作伙伴关系还使美国政府能够访问它否则不会拥有的高能效模型，使其能够构建对前沿人工智能模型的全新评估，同时限制重复性工作。

- **政府采购机构应实施对人工智能系统供应商的测试要求。** 要求示例可能包括列出用于衡量模型性能的基准，报告旨在发现特定环境下漏洞或失败的红色团队练习结果，或在政府雇员中进行预采购用户测试。 政府部门也可以根据报告的评估结果，为供应商的AI系统设立最低性能要求。拥有对部署特定模型相关风险的清晰度量，将使美国政府能够有效地利用AI系统并避免AI故障。

- **促进联邦政府内关于人工智能评估和风险的知识与专家经验的交流。** 除美国钢铁协会（AISI）之外，美国政府的其他部分也处于有利位置，以评估由人工智能系统带来的国家安全风险。国家安全局在进攻性网络威胁风险方面拥有专业知识，而能源部则配备了测试核和辐射风险的设备。这些机构可以构建人工智能评估基础设施，补充通过如AISI与Scale AI之间的合作提供的其他测试基础设施。人工智能行动计划应鼓励这些机构利用他们已经拥有的工具和知识——这些工具和知识可能不在政府之外轻易找到——以确保人工智能系统按预期工作。

开发、采用和同步标准

标准可以促进市场功能的顺畅、互操作性和消费者安全。 然而，由于人工智能技术的快速发展和跨部门、应用领域的潜在用例爆炸增长，建立人工智能标准具有挑战性。尽管许多前沿的人工智能公司 发布框架 在管理模型所面临的风险方面，这些框架通常缺乏关于风险缓解措施足够的细节，并且是 易于在竞争压力面前被削弱或完全放弃。 为了有效减轻与人工智能系统相关的风险，美国政府应与其他利益相关方协调制定人工智能标准，并展示如何采用标准可以促进人工智能在任务成功中的运用。

- **开发 并采取标准以减轻AI带来的风险。** 与学术界、民间社会和产业界的利益相关者协调，AISI应制定涵盖以下主题的标准： 模型训练 ，预发布 内部 并且 外部 安全测试 ， 网络安全实践， 如果-那么承诺 ， 人工智能风险评估 ，并且 测试和重测试随时间变化的系统的流程 。标准关于何时进行不同类型的评估，每项评估的最佳实践，以及如何报告模型评估。
应予以开发，以实现模型之间公平的比较。 美国政府可以

通过采用这些标准，将它们纳入采购要求，并分享从采用中汲取的教训，来有效利用人工智能系统的模型。

- **在联邦机构间同步基准人工智能标准。** .提供人工智能工具或将在软件解决方案中包含人工智能工具的公司，目前必须应对多重重叠的要求或标准 在向不同政府部门销售产品时。例如，美国国家标准与技术研究院（NIST）。 人工智能风险管理框架 确定可信赖人工智能的特征，这些特征与不可信赖人工智能不同。 国防部指南 在国防部（DOD）内，不同的军事服务部门有 不同的生成式人工智能政策 如果所有联邦机构都同意遵循一套统一的最小AI标准，以便在采购和部署方面操作，这将大大减轻提供AI解决方案公司的负担，加快标准工具和度量标准的采用，并减少因需要反复起草和回应政府合同中类似但不完全相同的要求而导致的不效率。通过在行业合同方面让美国政府统一遵循标准，美国政府还能够推动有利于美国企业的国际AI标准的采用。

- **通过授权美国国防部（OSD）制定人工智能安全标准和扩大操作授权（ATO）互惠互利，减少军事服务业之间的重复劳动。** OSD尚未授权一个涵盖整个DOD的实体来制定AI政策。这导致军事服务之间工作重复，存在多个备忘录以不同方式指导DOD的工作。例如，在各个服务中，不同的指挥官有不同的网络ATO标准，这要求政府和AI供应商在部署前进行大量修改。需要在OSD和实体之间强制执行持续ATO和ATO互惠，并应授权一个实体来同步政策、快速认证可靠的AI解决方案，并采取措施阻止新兴的安全问题。当DOD建立标准和政策时，这些应与其他政府机构和州机构共享，以进一步同步标准并加速负责任的采用。

致谢

此回应是CSET全体员工的共同贡献，并在CSET先前出版物和数据洞察的基础上进行了大幅扩展。我们特别感谢Mia Hoffmann、Jack Karsten和Mina Narayanan在项目领导方面的贡献，以及Owen Daniels、Igor Mikolic-Torreira和Dewey Murdick在全面审阅方面的努力。此外，还要特别感谢Catherine Aiken、Zachary Arnold、Kendrea Beers、Jack Corrigan、Kathleen Curlee、Jacob Feldgoise、Rebecca Gelles、William Hannas、Jessica Ji、Kyle Miller、Adrian Thinnyun和Vikram Venkatram对团队输入的有价值整合。