**[[1]](#footnote-2)**A blue and black sign

Description automatically generated

**政府行动建议：防范人工智能与生物技术风险，推动人类健康与福祉**

本行动建议由参与宏理国际与新美国安全中心举行的三边[对话](https://inhr.org/responsible-ai-standards)的代表及[约翰斯·霍普金斯大学健康安全中心](https://centerforhealthsecurity.org/our-work/aixbio)的生物安全专家共同提出：

认识到未来十年人工智能与生物技术的进步具有为人类健康、动植物健康、粮食安全、气候以及经济福祉带来变革性的改善的潜力；  
认识到人工智能与生物技术的融合，尤其是在与AI驱动的设计和自动化结合时，可能带来深远风险，亟需全球关注从而防止此类技术被滥用或产生意外后果，例如开发危险的合成病原体和增强性生物武器；  
认识到必须着重防范人工智能可能引发全球大规模传染病的后果，以及对国家安全、经济安全和公共卫生安全造成的风险，例如其可能显著降低非专业人员设计、合成、获取和使用生物武器的技术门槛；  
确认各国和民间社会有必要更深入地了解这些潜在风险，并提升对最佳实践和有效防控措施的认知，以防止或减轻可能产生全球性影响的重大风险；  
赞赏各国政府在制定国家政策时发挥的独特作用，即既推动人工智能和生物技术的持续创新，又能建立健全防范与缓解重大风险的保障措施，从而确保这些变革性技术以负责任的方式研发和应用，以此造福全人类。

据此我们提出以下建议以供各国政府参考。这些建议旨在为制定涉及人工智能与生物技术融合的国家治理措施提供指引，适用范围涵盖前沿人工智能模型以及大量使用生物学数据训练的高性能人工智能模型，尤其聚焦于预防或减轻可能产生全球性影响的重大人工智能与生物技术风险，而非涵盖所有类型和层级的相关风险。以下建议仅代表本小组的集体观点，并不构成任何国家、机构或个人的官方立场或背书。

各国政府可在以下领域探索并付诸行动：

### **意识提升、培训与人员能力建设**

1. 在政府、执法部门、私营部门和学术界加大教育与培训投入，提升相关人员的技术与风险认知能力，培养多学科专家队伍，以评估并降低前沿人工智能模型被滥用以及相应的生物安全威胁风险。
2. 与模型开发者、生物安全专家及其他专业人士合作，持续完善模型开发、风险评估和安全测试的前沿时间，保障安全创新并防范人工智能与生物技术交叉所带来的重大后果。
3. 鼓励并在必要时激励私营部门和投资方，将生物安全风险培训、红队与蓝队演练以及有效安全防护的建立，纳入支持 AI 初创企业和相关投资的必备条件。

### **安全评估、测试与行业最佳实践**

1. 将人工智能和生物技术交叉企业的发展纳入未来国家治理措施的讨论与分析，制定适用于前沿人工智能模型和基于生物数据训练的高性能人工智能模型的行为准则和必要法规，涵盖安全评估与基于能力阈值的负责任扩展标准。
2. 鉴于识别和缓解人工智能与生物技术相关风险的挑战，应优先推动统一的风险评估与安全测试标准，以识别并控制可能导致全球严重生物威胁风险的人工智能模型。
3. 将高优先级安全评估制度化，对发现的具有危险能力的人工智能模型及时采取补救措施，并制定共享指引，并且帮助模型开发者、部署者及相关方识别和减轻潜在风险。
4. 在政府内部监督或建立对前沿人工智能模型及基于生物数据训练的高性能模型开展红队与蓝队演练的能力。相应团队应由跨学科人才组成，并采取防护措施，避免红队活动在实验室中生成真实危险的生物构建物；如有必要，可与私营部门合作开展安全代理实验，并利用人工智能赋能工具进行漏洞识别与修补。
5. 制定并推广安全评估与红队演练的最佳实践，从整体生态系统风险出发进行评估，而不仅限于单一模型的应用。
6. 研究并建立适当的激励措施（包括资金奖励或其他形式），鼓励产业界和学术界开发安全与防护机制，以降低人工智能和生物相关的重大风险。
7. 深入分析开源前沿模型及基于生物数据训练的高性能模型的潜在益处与风险，审慎考虑对特定类型的开源模型采取何种监管措施，例如如何限制模型及其权重的访问，以防恶意行为者绕过安全防护或“越狱”等行为。
8. 当企业间竞争妨碍行业自愿采纳最佳安全实践时，应考虑引入强制性要求或激励机制，在保障公共安全的同时，兼顾防止过度抑制创新，并确保人工智能持续创造积极的社会效益。

### **蛋白质设计的安全规范**

1. 为防止利用人工智能设计并在实验室合成危险生物分子，应制定覆盖合成基因序列制造商、相关产品使用者及核酸合成桌面设备生产商的国家级安全政策。该政策可包括“了解你的客户”规定和订单筛查要求，以确保核酸合成技术仅用于促进科研和其他有益目的，防止被恶意滥用。

### **国际合作加强安全措施以应对人工智能与化学生物威胁**

1. 重申并加强反对制造生物武器的国际规范，研究建立相应机制，为《禁止生物武器公约》缔约国提供关于人工智能及其他新兴技术相关生物武器风险的专家与科学支持。
2. 开展多边对话，探索防范或应对人工智能和生物技术重大风险的国际协定、专门机构或其他合作机制。
3. 设立或指定国家权威机构，作为政府间就人工智能安全与安保议题开展正式技术沟通的联络点。
4. 建立并常态化旨在应对人工智能和生物技术威胁的国际信息共享机制，以支持反恐合作并应对对国际和平与安全的新兴威胁。

注：本文系对 [2024 年 5 月在泰国举行的宏理国际人工智能与生物技术建议的原始版本](https://docs.google.com/document/d/1WsGnpgYoa9zd3FNLLTWeOIpF0-hP2Sm6fUG0HW4nLao/edit?tab=t.0)所作的修订，并且得益于 Founder’s Pledge 的资助。

1. 就本文件而言，“前沿人工智能模型”是指具备极高能力、且通常以巨大计算资源训练的人工智能系统。此类模型有时也被称为通用型人工智能（GPAI）模型。 [↑](#footnote-ref-2)