布莱切利宣言

人工智能安全峰会 (AI Safety Summit)
2023 年 11 月 1-2 日

中文翻译版 · 仅供参考，以英文原文为准
原文来源：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

参加人工智能安全峰会的国家发表的布莱切利宣言

2023年11月1-2日

人工智能（AI，Artificial Intelligence）为全球带来了巨大的机遇：它有潜力改变和提升人类福祉、和平与繁荣。为实现这一目标，我们确认，为了所有人的利益，人工智能的设计、开发、部署和使用应当以安全的方式进行，应当以人为本、值得信赖且负责任。我们欢迎国际社会迄今为止在人工智能合作方面所做的努力，以促进包容性经济增长、可持续发展和创新，保护人权和基本自由，并培养公众对人工智能系统的信任和信心，以充分实现其潜力。

人工智能系统已在日常生活的许多领域部署，包括住房、就业、交通、教育、健康、无障碍和司法，其使用可能还会增加。我们认识到这是一个采取行动的独特时刻，并确认在我们的国家和全球范围内，以包容的方式，需要安全开发人工智能，并将人工智能的变革性机遇用于造福所有人。这包括公共服务领域，如健康和教育、粮食安全、科学、清洁能源、生物多样性和气候，以实现人权的享有，并加强实现联合国可持续发展目标（United Nations Sustainable Development Goals）的努力。

在这些机遇之外，人工智能也带来了重大风险，包括在日常生活的那些领域。为此，我们欢迎在现有论坛和其他相关倡议中审查和应对人工智能系统潜在影响的相关国际努力，以及对以下方面需要解决的认识：人权保护、透明度和可解释性（transparency and explainability）、公平性（fairness）、问责制（accountability）、监管（regulation）、安全（safety）、适当的人类监督（human oversight）、伦理（ethics）、偏见缓解（bias mitigation）、隐私和数据保护（privacy and data protection）。我们还注意到操纵内容或生成欺骗性内容的潜在风险。所有这些问题至关重要，我们确认解决它们的必要性和紧迫性。

特别的安全风险出现在人工智能的"前沿"（frontier），即那些高度能干的通用人工智能模型（general-purpose AI models），包括基础模型（foundation models），它们可以执行各种各样的任务——以及可能展现造成危害的能力的相关特定窄人工智能——其能力匹配或超过当今最先进模型中存在的能力。重大风险可能源于潜在的蓄意滥用或与人类意图对齐（alignment）相关的意外控制问题。这些问题部分是因为这些能力没有被完全理解，因此难以预测。我们特别关注网络安全（cybersecurity）和生物技术（biotechnology）等领域的此类风险，以及前沿人工智能系统可能放大虚假信息（disinformation）等风险的情况。存在严重甚至灾难性危害的潜在可能——无论是蓄意的还是无意的——源自这些人工智能模型最重要的能力。鉴于人工智能快速而不确定的变化速度，以及技术投资加速的背景，我们确认深化对这些潜在风险及应对行动的理解尤为紧迫。

人工智能引发的许多风险本质上是国际性的，因此最好通过国际合作来解决。我们决心以包容的方式共同努力，通过现有的国际论坛和其他相关倡议，确保以人为本、值得信赖和负责任的人工智能是安全的，并支持所有人的利益，以促进合作应对人工智能带来的广泛风险。在此过程中，我们认识到各国应考虑采用有利于创新和适度的治理和监管方法的重要性，以最大化利益并考虑与人工智能相关的风险。这可以包括根据国家情况和适用法律框架，在适当情况下进行风险分类和归类。我们还注意到在适当情况下就共同原则和行为准则等方法进行合作的相关性。关于与前沿人工智能相关的最可能出现的具体风险，我们决心加强和维持我们的合作，并将其扩大到更多国家，通过现有的国际论坛和其他相关倡议（包括未来的国际人工智能安全峰会）来识别、理解和在适当情况下采取行动。

所有行为者都有责任确保人工智能的安全：国家、国际论坛和其他倡议、企业、公民社会和学术界需要共同努力。注意到包容性人工智能和弥合数字鸿沟的重要性，我们重申国际合作应努力让广泛的合作伙伴适当参与，并欢迎面向发展的方法和政策，以帮助发展中国家加强人工智能能力建设，并利用人工智能的赋能作用来支持可持续增长和缩小发展差距。

我们确认，虽然安全必须在人工智能的整个生命周期中加以考虑，但开发前沿人工智能能力的行为者，特别是那些异常强大且具有潜在危害的人工智能系统，有特别重大的责任确保这些人工智能系统的安全，包括通过安全测试系统、评估和其他适当措施。我们鼓励所有相关行为者就其衡

量、监测和缓解可能出现的潜在有害能力及其相关影响的计划提供适当背景的透明度和问责制，特别是防止滥用和控制问题，以及其他风险的放大。

在我们合作的背景下，为了在国家和国际层面提供信息和指导行动，我们应对前沿人工智能风险的议程将聚焦于：

- **识别共同关注的人工智能安全风险**，建立对这些风险的共同科学和循证理解，并随着能力的持续增长维持这种理解——这是在更广泛的全球方法中理解人工智能对社会影响的一部分。
- **在各国建立相应的基于风险的政策**，以根据这些风险确保安全，在适当情况下进行合作，同时认识到基于国家情况和适用法律框架，我们的方法可能有所不同。这包括：私营部门开发前沿人工智能能力的行为者增加透明度、适当的评估指标、安全测试工具，以及发展相关的公共部门能力和科学研究。

为推进这一议程，我们决心支持一个国际包容性的前沿人工智能安全科学研究网络，该网络涵盖和补充现有的和新的多边、诸边和双边合作，包括通过现有的国际论坛和其他相关倡议，以促进为政策制定和公共利益提供最好的科学支撑。

认识到人工智能具有变革性的积极潜力，作为确保更广泛的人工智能国际合作的一部分，我们决心维持一个包容性的全球对话，参与现有的国际论坛和其他相关倡议，以开放的方式为更广泛的国际讨论做出贡献，并继续关于前沿人工智能安全的研究，以确保该技术的利益能够负责任地为全人类所用。我们期待在2024年再次相聚。

参与国名单

- 澳大利亚（Australia）
- 巴西（Brazil）
- 加拿大（Canada）
- 智利（Chile）
- 中国（China）
- 欧盟（European Union）
- 法国（France）
- 德国（Germany）
- 印度（India）
- 印度尼西亚（Indonesia）

- 爱尔兰（Ireland）
- 以色列（Israel）
- 意大利（Italy）
- 日本（Japan）
- 肯尼亚（Kenya）
- 沙特阿拉伯王国（Kingdom of Saudi Arabia）
- 荷兰（Netherlands）
- 尼日利亚（Nigeria）
- 菲律宾（The Philippines）
- 大韩民国（Republic of Korea）
- 卢旺达（Rwanda）
- 新加坡（Singapore）
- 西班牙（Spain）
- 瑞士（Switzerland）
- 土耳其（Türkiye）
- 乌克兰（Ukraine）
- 阿拉伯联合酋长国（United Arab Emirates）
- 大不列颠及北爱尔兰联合王国（United Kingdom of Great Britain and Northern Ireland）
- 美利坚合众国（United States of America）

2024年10月23日，新西兰（New Zealand）加入了对布莱切利宣言的承诺。

对"政府"和"国家"的提及包括根据其立法或行政权限行事的国际组织。