布莱切利宣言 — AI 安全峰会 2023

外交部 (MFA)

2023 年 11 月

中文翻译版

布莱切利宣言——AI 安全峰会 2023（The Bletchley Declaration by Countries Attending the AI Safety Summit）

概述

2023 年 11 月 1-2 日，首届全球 AI 安全峰会在英国布莱切利庄园（Bletchley Park）举行。新加坡作为 28 个签署国之一，签署了《布莱切利宣言》。该宣言宣布了一项新的全球努力——通过确保 AI 安全来释放 AI 带来的巨大益处。

核心理念

与会各国确认，AI 应当以安全、以人为本、可信和负责任的方式设计、开发、部署和使用，以造福全人类。

机遇认知

宣言肯定 AI 具有变革和增进人类福祉、和平与繁荣的巨大潜力。AI 系统已部署于日常生活的众多领域，包括住房、就业、交通、教育、健康、无障碍和司法等。

与会各国欢迎将 AI 的变革性机遇用于公共利益，包括： - 公共服务（健康和教育） - 粮食安全 - 科学研究 - 清洁能源 - 生物多样性和气候 - 人权实现 - 联合国可持续发展目标

风险共识

一般性风险

AI 在带来机遇的同时也带来重大风险。与会各国认识到需要在以下方面采取行动： - 人权保护 - 透明度和可解释性 - 公平性 - 问责制 - 监管 - 安全性 - 适当的人类监督 - 伦理、偏见缓解 - 隐私和数据保护

前沿 AI 风险

特别关注前沿 AI 带来的安全风险——那些高度通用的 AI 模型（包括基础模型）可能执行各种任务，其能力匹配或超越当今最先进的模型。

重大风险来自： - 蓄意滥用 - 与人类意图对齐相关的控制问题 - 网络安全威胁 - 生物技术风险 - 虚假信息扩散 - 潜在的严重甚至灾难性伤害

合作承诺

国际合作

- 以包容方式共同确保 AI 安全
- 通过现有国际论坛和相关倡议推进合作
- 采用促进创新且适度的治理和监管方法
- 根据各国情况进行风险分类

具体行动议程

1. **识别共同关切的 AI 安全风险**：建立共同的科学和循证理解
2. **制定基于风险的政策**：确保安全，同时尊重各国不同的方法和法律框架
3. **支持国际科研网络**：促进前沿 AI 安全的科学研究

各方责任

- 各国政府、国际组织、企业、公民社会和学术界都有责任确保 AI 安全
- 开发前沿 AI 能力的主体承担特别强的安全责任
- 鼓励提供适当的透明度和问责

签署国

共 29 个签署方：澳大利亚、巴西、加拿大、智利、中国、欧盟、法国、德国、印度、印度尼西亚、爱尔兰、以色列、意大利、日本、肯尼亚、沙特阿拉伯、荷兰、尼日利亚、菲律宾、韩国、卢旺达、**新加坡**、西班牙、瑞士、土耳其、乌克兰、阿联酋、英国、美国。

2024 年 10 月，新西兰加入布莱切利宣言。

意义

布莱切利宣言标志着全球 AI 治理进入新阶段——各国首次在前沿 AI 安全问题上达成共识并做出集体承诺。新加坡的签署体现了其作为负责任 AI 发展参与者的国际立场，也为后续参与首尔 AI 安全峰会奠定了基础。