2026年人工智能与国际安全研究

中欧人工智能与国际安全二轨对话阶段性报告

2026 年 6 月

清华大学战略与安全研究中心
CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

Centre for
Humanitarian
Dialogue

中欧人工智能与国际安全二轨对话阶段性报告

一、概述

2024 至 2025 年期间，清华大学战略与安全研究中心（CISS）与瑞士人道主义对话中心共同四轮举办“中欧人工智能与国际安全二轨对话”，聚焦人工智能技术发展带来的潜在风险，及其对国际安全的影响。该系列对话广泛汇聚了来自中欧双方的国际组织、高校、智库机构以及科技企业等多个机构、不同领域的专家学者，持续开展深度研讨与交流，旨在：

- 深化对人工智能应用潜在安全风险的认知和理解，包括人工智能赋能武器与决策支持系统（AI-DSS）本身存在的风险，以及由此可能引发的冲突无意升级风险；
- 确定可能的建立信任措施（CBMs），降低人工智能在军事领域应用过程中可能导致的冲突升级风险；
- 积极探索中欧双方在人工智能治理方面的合作路径。

该系列对话从初期对人工智能相关安全风险分类、彼此共同关切等议题讨论展开，逐步延伸至基于场景式研讨的案例分析，最终形成了关于具体建立信任措施和安全治理框架的相关思考与明确建议。尽管与会专家学者承认中欧在部分议题的优先事项和观点上存在差异，但在两大核心问题上达成了持续且高度一致的共识：其一，应对人工智能带来的安全风险具有紧迫性；其二，持续开展对话交流既是增进彼此相互认知和理解的重要基础，也是积极探索人工智能全球治理路径的重要依托。

二、核心议题

1. 人工智能安全风险的认知

与会专家学者探讨了人工智能相关安全风险议题，包括：

- **技术漏洞风险：**主要体现在算法存在偏见、技术缺乏透明度，同时易遭受对抗性攻击和数据投毒等恶意操作，进而影响军事人工智能系统的可靠性与安全性。
- **人为因素风险：**核心表现为易产生自动化偏见、过度依赖人工智能给出的建议，且人工智能会对决策人员的速度以及判断产生的影响，可能导致决策偏差或误判。
- **应用场景风险：**在实际应用场景中，可能出现对人工智能行为的误解或人工智能本身的故障导致的误判。虚假信息和深度伪造技术还会加大信息战和认知战的负面影响，进一步加剧地区及国际局势的紧张。
- **战略不稳定风险：**人工智能的应用有可能挑战传统战略稳定框架，降低国家使用武力的门槛，加深各国之间的战略互不信任。

2. 基于场景推演的讨论

与会专家学者结合具体场景开展针对性推演与分析，核心讨论场景如下：

- **场景 A（海上领域）：**围绕军事人工智能决策支持系统（AI-DSS）可能引发的民用船只误认危机展开深入探讨，强调系统训练数据缺陷、自动化决策偏差，以及突发情况下跨机构、跨国家间缺乏统一的沟通协调协议带来的风险问题。此类失误不仅可能引发不必要的海上摩擦，更可能直接升级局势，对全球海洋安全构成威胁。
- **场景 B（陆地 / 冲突战区）：**探讨冲突环境中人工智能武器系统的应用伦理与实操风险，明确算法、操作人员还是指挥决策层的责任归属，同时需厘清降低风险在系统技术设计规范与实战操作政策层面的区别。

3. 建立信任措施（CBMs）

该系列对话的核心重点是探讨具有可操性的建立信任措施，就人工智能全生命周期管理等层面进行探讨交流，不断增进国际社会的互信与协作，降低人工智

能应用的潜在安全风险。与会专家学者强调，构建人工智能相关的信任建立措施，需采取分阶段、循序渐进的推进方法。一是聚焦解决技术层面的核心风险，如算法偏差、系统可靠性等问题，不断增强彼此的认知和理解，减少因技术路径差异或信息不对称带来的误解与风险；二是促进国际合作，强化风险识别和防范能力，增强信息交流与共享，针对人工智能技术漏洞、重大安全风险等突发情况，能够实现及时沟通和处理。

三、主要挑战

尽管中国和欧洲专家学者在应对人工智能带来的风险方面有着诸多共识，认为持续对话是构建人工智能全球治理框架的重要基础，但也强调当前人工智能安全治理仍面临较多挑战，主要体现在如下方面：

- **透明度问题：**人工智能系统往往具有“黑箱”特征，其决策过程缺乏可解释性，不仅削弱了公众信任，也增加了监管难度。不同国家在数据获取、算法披露及监管标准方面存在差异，也进一步加剧了信息不对称问题。因此，有必要增强算法可解释性相关技术研究，促进信息沟通交流，缩小监管政策带来的差异。
- **决策速度：**人工智能技术显著提升了决策效率，但也可能因自动化与高速度决策机制带来误判风险甚至冲突升级隐患。因此，有必要适度放缓人工智能辅助决策节奏，确保关键环节中保留充分的人类判断与干预空间，以最大限度降低误判及冲突升级风险。同时，在紧急或高压决策环境下，往往难以预留充足时间进行审慎的人类评估，必须寻求时效和安全管控的平衡，坚持适当的人类判断和控制，保留人类的最终决策权。
- **私营部门的作用：**随着私营企业在人工智能技术研发与应用中的深度参与，传统以国家为主体的治理模式正面临结构性冲击。企业在技术创新中的主导作用提高了技术演进速度，但也在一定程度上削弱了国家对关键技术的直接控制能力，增加了跨境治理与监管协调的复杂性。因此，有必要进一步明确政府与企业之间的权责边界，推动建立多方参与的协同治理体系。

四、对未来合作的建议

1. 持续开展对话交流增进相互认知和理解

针对当前中欧双方专家学者在相关议题上存在的观点分歧，建议考虑保持常态化专家对话机制。通过定期、高规格的专家级交流，持续深入探讨人工智能带来的国际安全风险与治理路径，进而可以有效弥合信息不对称，减少误解与分歧，逐步积累并凝聚领域共识，为后续深化合作筑牢基础。

2. 进一步提升对话的政策导向和国际影响力

中欧既是维护世界和平的“两大力量”，也是促进全球安全与稳定的重要支柱。建议对话成果需紧密贴合政策实际，适当邀请官方代表以观察员身份参与，共同探讨人工智能潜在国际安全风险挑战对不同地区国家的感知，进一步增强彼此间的共同认知和理解，提升对话的政策转化能力与国际影响力。

CISS 人工智能项目简介

清华大学战略与安全研究中心（CISS）成立于2018年11月7日，是聚焦国际战略与安全领域研究的高校智库。自2019年以来，CISS聚焦人工智能技术发展前沿与国际安全治理问题，专门设立人工智能项目专家组，扎实推进人工智能与国际安全相关研究。同时，CISS持续与美国布鲁金斯学会、瑞士人道主义对话中心开展中美、中欧人工智能与国际安全二轨对话，不断拓展同联合国裁军研究所、红十字国际委员会等国际组织和智库机构在人工智能全球治理方面的项目合作，并通过联合研究和政策交流，形成一系列重要的研究成果和政策报告，为推动人工智能国际安全领域的交流合作积累国际共识。此外，CISS还积极承接来自外交部、科技部、财政部等国家部委委托的研究项目，持续深化人工智能治理的区域与国别研究，为相关政策制定与国际合作贡献力量。

亨利·杜南人道主义对话中心简介

亨利·杜南人道主义对话中心（Henry Dunant Centre for Humanitarian Dialogue，以下简称“HD”）于 1999 年成立，致力于通过对话来预防、缓和并解决冲突。HD 总部位于瑞士日内瓦，是瑞士联邦法律下注册的国际组织，也是联合国官方合作伙伴。2022 年，HD 被授予卡内基沃特勒和平奖（Carnegie Wateler Peace Prize），以表彰 HD 在调解冲突和促进和平方面做出的突出贡献。此前曾获该奖项的有：联合国难民署、联合国儿童基金会、联合国前秘书长达格·哈马舍尔德、国际联盟首任秘书长埃里克·德拉蒙德等。

HD 遵从人道、公正和独立原则，工作风格严谨、务实且低调，注重创新，并始终坚持以成果为导向开展工作。拥有 300 余名员工，足迹遍布亚洲、非洲、拉美、中东和欧亚地区，在全球 80% 以上的暴力冲突中发挥调解作用。HD 自 2008 年开始在中国开展工作，为中国与相关国家和组织提供沟通平台，旨在促进各方坦诚对话，交换意见，增进互信，并致力探索有效预防、和平解决争端分歧的恰当路径。

清华大学战略与安全研究中心
CENTER FOR
INTERNATIONAL SECURITY AND STRATEGY
TSINGHUA UNIVERSITY

Centre for
Humanitarian
Dialogue

ciss@tsinghua.edu.cn
010-62771388
清华大学明理楼 428A 室
http://ciss.tsinghua.edu.cn

微信公众号　官方网站　联系我们