2020 年第 10 期（总第 18 期）

# 国际战略与安全研究报告

# INTERNATIONAL SECURITY AND STRATEGY STUDIES REPORT

## 人工智能与国际安全治理路径探讨

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

# 人工智能与国际安全治理路径探讨

傅　莹[①]

**【摘要】**目前很难就全面禁止人工智能武器达成全球共识，更可行的做法可能是要求人工智能赋能武器的发展符合现有国际法规范。为此，各国需要就如何进行风险防范寻求共识，共同努力构建治理机制。现阶段是构建人工智能国际安全规范的关键窗口期。目前中美两国是在人工智能技术研究和应用发展最快的国家，两国需要在这个领域加强协调与合作。

**【关键词】**人工智能　国际安全治理

清华大学战略与安全研究中心在“人工智能国际治理”框架下，于2019年6月启动了一项与美国布鲁金斯学会进行联合研究的项目，研究中心主任傅莹和美国布鲁金斯学会会长约翰·艾伦（John Allen）为项目总牵头人。围绕课题人工智能技术在安全领域的挑战和治理问题。两国专家进行了一系列二轨（非官方）对话和研讨，在讨论的基础上分别就各项议题撰写了论文。据此，傅莹和约翰·艾伦在美国Noema杂志分别发表文章，公布联合研究成果，介绍双方专家的讨论和主要观点，旨在提高国际社会和两国各界对人工智能武器化风险的认知和警惕，并提出治理建议。

约翰·艾伦在文章中提到，美中在人工智能等新兴技术领域的战略竞争日益明显，围绕着“人工智能军备竞赛”的说法不绝于耳。为了避免不可控的军备竞赛，他希望美中两国专家组成的“人工智能与国际安全”研究团队所讨论的安全关切，能够在政府层面对话和非官方论坛中得到进一步探讨，特别是针对人工智能所构成的国家安全风险，以降低这些风险发生的可能性。他认为现在正是规范智能技术发展和应用的关键期。因为一旦这些技术被纳入到军事体系中，呼吁限制其在国家安全领域的应用将极为困难。

① 清华大学战略与安全研究中心主任、外交部前副部长。

以下为傅莹文章正文：

近年来，人工智能技术的快速发展带来了巨大机遇，但是技术革命也往往伴随不可预知的安全挑战，尤其需要关注人工智能技术武器化的道德和技术风险问题。许多国家的专家和学者呼吁禁止发展可以自主识别并击杀人类目标的智能武器，更不应该容许它们执掌人类生死。然而，全面禁止人工智能武器很难达成全球共识，即便能开启相关讨论和谈判也将旷日持久。

从目前的趋势看，人工智能武器化是不可避免的。更可行的做法可能是要求人工智能赋能武器的发展符合现有国际法规范。为此，各国需要就如何进行风险防范寻求共识，共同努力构建治理机制。在与美方进行二轨讨论时，我们的焦点在于如何设定人工智能赋能武器的攻击“禁区”，如何依据国际法律和规范开展对人工智能武器的监管，以及如何鼓励采取克制态度以限制对人工智能数据的军事化滥用。

## 人工智能的军事安全挑战

人工智能赋能武器系统存在诸多潜在挑战。一是人工智能内在的技术缺陷使得攻击者难以限制打击的损害范围，容易使得被打击方承受过大连带伤害，从而导致冲突升级。人工智能赋能武器不仅应该在实施打击时区分军事目标和民用目标，还需要防止和避免对民用目标造成过分的附带或间接损害。然而，现有人工智能技术条件在能否保证武力使用过程中完全满足上述条件方面，是存在不确定性的。

二是当前以机器学习带动的人工智能技术发展需要大量数据，不能完全避免基于大数据训练的算法和训练数据集将偏见带入真实应用系统，因此，不能排除人工智能给决策者提供错误建议的可能性。进而，当训练数据集受到其他国家的污染，致使系统提供错误侦查信息

时，也有可能让军事决策者做出错误判断和相应的军事部署。

三是人机协同的挑战是人工智能军事化的终极难题。机器学习和大数据处理机制存在局限。无论是行为主义的强化学习、联结主义的深度学习，还是符号主义的专家系统都不能如实准确地反映人类的认知能力，比如直觉、情感、责任、价值等。人工智能的军事运用是人—机—环境的综合协同过程，而机器在可解释性、学习性、常识性等方面的不足，将放大发生战场冲突的风险，甚至刺激国际危机的螺旋上升。

## 人工智能安全治理路径探讨

对话双方一致认为，各国需要采取军事克制态度，避免人工智能武器化给人类来带来重大损害。各国应该禁止没有责任和风险意识的辅助决策系统。在使用人工智能赋能的武器时，需要限制其打击的损害范围，防止造成连带伤害，避免冲突升级。此外，军事克制的内容还应该反映在公共教育当中。由于人工智能技术具有易于扩散的特点，它有可能流入某些黑客手里，进而将人工智能技术用于有害公共安全的行为中。

人工智能赋能武器的使用如何与国际法的基本原则保持一致，是安全治理研究中的重点所在。《联合国宪章》规定，除非得到联合国安理会的授权，否则成员国不得使用武力，或者是出于自卫的目的才能使用武力。因此，国家出于自卫目的而使用武力时，所使用武力的强度和规模须与受到的攻击或者受到的威胁的严重性相称。在讨论中，中方专家特别提出，各国须承担法律责任，主动推动和实现在涉及人工智能的军事行动中国际规范的建构。同时需要确定人类参与的阈值，以保证智能武力的使用不会造成过度伤害。因为人工智能赋能的武器平台很难评判什么是必要的、合适的、平衡的攻击，所以人类指挥官

的主观能动性应当得到尊重。

此外，人工智能数据的安全必须得到保证。应该对数据挖掘和采集的过程、数据标注和分类、数据使用和监管进行规范和限制。智能武器训练数据的收集过程和手段应当遵守国际法律，收集的数据数量应达到一定规模。需要确保数据标注和分类的质量和准确性，避免形成错误模型和导致决策者做出错误判断。在数据使用过程中，需要关注使用目标和数据的污染问题。有中方学者建议给智能武器的自主化程度分级。例如，分为半自主化、部分自主化、有条件自主化、高度自主化和完全自主化五级。对自主化程度进行分级，有利于更好地确认和保障人类的作用，从而切实有效地实现对人工智能及自主武器系统的管理和控制。

## 中美人工智能全球治理合作

现阶段是构建人工智能国际安全规范的关键窗口期。目前中美两国是在人工智能技术研究和应用发展最快的国家，两国需要在这个领域加强协调与合作。其他国家也表示出对人工智能应用的安全担忧，说明人工智能治理是人类共同的难题，不是一国两国能够解决的。中美开展对话与合作至关重要，将能够为全球人工智能治理合作贡献智慧。因此，中美两国应就推动构建国际层面的规范和制度进行正式讨论，在各自利益关切的基础上探索合作领域，互换和翻译相关文件，以政策沟通和学术交流的方式降低两国在这一领域影响双边关系和国际安全的潜在风险。

近年来，中国积极释放合作信号，2020 年 11 月 21 日，习近平主席在二十国集团领导人第十五次峰会上强调，中方支持围绕人工智能加强对话，倡议适时召开专题会议，推动落实二十国集团人工智能原则，引领全球人工智能健康发展。2020 年 9 月 8 日，国务委员兼外

长王毅提出《全球数据安全倡议》，包括有效应对数据安全风险挑战应遵循的三项原则，表示希望国际社会在普遍参与的基础上就人工智能安全问题达成国际协议，支持并通过双边或多边协议形式确认倡议中的有关承诺。中国在发展人工智能技术的同时，也高度重视和积极推进相关国内治理建设。2018 年中国发布《人工智能标准化白皮书》列出了四条伦理原则，包括人类利益原则、责任原则、透明度原则、权责一致原则。

中国已经准备好与美国和其他国家地区在人工智能治理方面开展合作。我们相信，人工智能不应成为一场“零和游戏”，技术突破最终应使得全人类受益。

本文 2020 年 12 月 24 日发表于《人民论坛》


扫码关注我们

**清华大学战略与安全研究中心**

办公地点：清华大学明斋 217

联系电话：010-62771388

电子邮箱：ciss@tsinghua.edu.cn