联合国 A/80/78

大 会

Distr.: General
5 June 2025
Chinese
Original: Chinese/English/French/
Russian/Spanish

第八十届会议

暂定项目表* 项目 101

全面彻底裁军

军事领域的人工智能及其对国际和平与安全的影响

秘书长的报告

摘要

本报告综述会员国和观察员国依照第 79/239 号决议提交的答复的要点，但不影响其各自立场。报告载有：军事领域人工智能的相关机遇和挑战；现有和新出现的规范性提议的目录；关于军事人工智能领域倡议和举措的调查；关于今后步骤的思考；秘书长的意见和结论。

* A/80/50。

25-06526 (C) 140725 050825

目录

一. 导言

1. 大会在关于军事领域的人工智能及其对国际和平与安全的影响的大会第79/239号决议第7段，请秘书长就人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战征求会员国和观察员国的意见，特别侧重于致命性自主武器系统以外的其他领域，并提交一份实质性报告，概述这些意见，编目现有和新出现的规范性提议，并在附件中载列这些意见，提交大会第八十届会议，供各国进一步讨论。在同一决议第8段，大会又请秘书长征求国际和区域组织、红十字国际委员会、民间社会、科学界和工业界的意见，并将这些意见以收到的原文列入上述报告附件。本报告根据这些要求提交。

2. 2025年2月12日，裁军事务厅向所有会员国和观察员国发出普通照会，提请各国注意大会第79/239号决议第7段，并征求各国对该事项的意见。裁军厅还向同一决议第8段所述实体发出普通照会和信函，提请它们注意该段，并征求它们对该事项的意见。截至2025年4月11日收到的意见抄录于本报告附件。此日期后收到的任何意见将以呈件所用语文在裁军厅网站上发布。

3. 本报告第二至六节综述会员国和观察员国提交的答复的要点，但不影响其各自立场。秘书长的意见和结论载于第七节。

二. 背景

4. 各国提及科学和技术总体上的快速进步，特别是人工智能领域的快速进步，指出这些进步对社会产生广泛影响。更具体而言，各国指出，人工智能有可能改变军事事务的方方面面，并对国际和平与安全产生重大影响。

5. 若干国家提及目前人工智能在军事领域的应用，以及各自在国防行动中使用人工智能的努力。各国认识到关于致命自主武器系统的讨论的重要性，同时指出军事领域的人工智能问题范围更泛，所涵盖能力涉及更多方面。

三. 机遇与挑战

6. 有国家指出，人工智能既带来机遇，也带来挑战，应当以切合实际的方式予以应对。有国家承认，人工智能的发展速度意味着目前无法预测所有机遇和挑战。有国家建议，这项技术本身不应被污名化。

A. 机遇

7. 速度被认为是人工智能的主要优势，包括在信息分析和决策方面。有国家指出，规模也是一种优势，因此人工智能可用作“战力倍增手段”。若干国家提及人工智能有可能提高效率、准确度和精确度，从而使出错概率低于人类。各国还指出可靠性、安全性和稳健性等其他特征。

应用

8. 若干国家提及人工智能在情报、监测和侦察领域的应用；在这些领域，人工智能可用于有效分析大数据集、推动识别威胁、提升态势感知能力和开展更精准的行动。有国家指出，正是这些特点使人工智能能够支持决策及指挥与控制，从而可能更精准地开展行动，减少平民所面临风险，并为民用物体提供更多保护。但也有国家强调指出，人工智能工具不应取代人类的决策。

9. 若干国家表示，人工智能可融入无人系统。有国家指出，人工智能可以改善军事人员之间以及军事人员与人道主义援助提供者等其他人之间的协调与沟通。总体而言，有国家指出，人工智能可减轻日常任务或重复性任务的负担，增强人在复杂任务中的能力。

10. 一些国家指出，可利用人工智能来检测侵入或其他恶意活动，以加强信息和通信技术安全，包括保护关键基础设施。有国家指出，人工智能可用于检测由人工智能生成的、传播错误信息和虚假信息的内容，识别仇恨言论、宣传或公众情绪的变化。

11. 有国家提及不直接涉及战斗的其他人工智能应用，包括优化后勤、预测性维护、采购、资源配置、行政管理、模拟和训练。

国际和平与安全

12. 若干国家认为，人工智能有助于维护国际和平与安全，例如，人工智能辅助的态势感知可有助于减轻风险及缓和冲突。有国家指出，使用人工智能可降低军事人员面临的风险，例如人工智能可在未爆弹药处置等某些危险任务中取代人类，或者可支持偏远地点的搜救行动。

13. 有国家建议，人工智能可以加强国际人道法，特别是攻击中的区分、相称和预防等基本原则的执行，使平民和民用物体得到更好保护。在这方面，若干国家指出，人工智能能够提升总体的态势感知能力，特别是增强对平民环境的了解，而且人工智能有能力提高精确性，减少人为错误的风险。还有国家指出，人工智能可协助关于平民伤亡情况的调查，从而确保追究责任人的责任。

14. 若干国家建议，人工智能可帮助监测和核查裁军、不扩散和军备控制协定的执行情况。有国家提及人工智能支持维持和平特派团的潜力，包括促进规划、后勤和停火监测。所确定的其他人工智能相关应用包括边境安全、打击恐怖主义、侦查非法武器方案以及优化人道主义援助和应灾。

B. 挑战

15. 若干国家指出，新兴技术、尤其是人工智能领域新兴技术的快速发展对国际和平与安全构成挑战。了解这些挑战很重要，但目前还不可能完全预见所有挑战。

16. 各国强调了与人工智能有关的以下关切：

- “观察-判断-决策-行动”环加速，因此可用于决策的时间缩短
- 自主能力不断提升，失去人类控制，尤其是在使用武力的情况下
- 可能被误用或恶意使用
- 人类对人工智能应用的过度信任
- 各国之间的技术不对称不断加深

国际和平与安全

17. 若干国家指出，将人工智能纳入军事领域可能对国际和平与安全构成挑战。人工智能的使用可能会导致误解、误判和意外升级的风险增加，原因包括人工智能辅助行动的速度增加和规模扩大或出现技术故障。上述原因还可能导致使用武力的门槛降低。若干国家对这一领域出现军备竞赛表示关切。有国家指出，人工智能的使用可能会将平衡点从防御性行动转向进攻性行动，并指出国家之间不断加剧的不平衡状态可能加剧不稳定状况，从而破坏国际和平与安全。

18. 若干国家对人工智能能力扩散、包括扩散至非国家行为体可能产生的破坏稳定影响表示关切。有国家指出，在控制已纳入人工智能能力的武器的扩散方面，目前还没有多边框架。

技术考虑

19. 各国考虑了技术因素带来的风险，其中包括：

- 技术故障和失灵
- 设计缺陷
- 意外行为，偏离设计参数
- 易受网络攻击和数据投毒
- 算法和数据偏见，包括性别偏见
- 自动化偏见，原因是人工操作员的训练不足
- 为训练人工智能模型而收集和处理大量个人数据所引发的隐私关切
- 缺乏训练的人工智能模型所导致的问题
- 测试、评估、核对和验证程序不佳所产生的问题
- 目标选择错误
- 能耗过高
- 过度依赖外部供应商

20. 若干国家对复杂的人工智能能力的透明度和可解释性表示关切，这些复杂能力通常被称为“黑箱”。还有国家对生成式人工智能等民用人工智能应用程序的使用表示关切，原因是这可能会加剧冲突局势的复杂性和不确定性。此外，若干国家对人工智能与其他技术的融合表示关切。

法律和人道主义考虑

21. 若干国家指出，人工智能对遵守国际法，特别是国际人道法和国际人权法构成挑战。使用人工智能可能导致不加区分地使用武力，并引发非法或不法行为情况下的责任和问责问题。各国提出的相关问题包括保护平民和民用基础设施，以及战斗人员可能会面临强度更高、更致命的冲突。

22. 若干国家提出了伦理道德关切，指出人工智能的使用可能会让同情心、道德考量和人类判断发挥作用的机会减少。

可能出现误用的领域

23. 若干国家指出，国家和非国家行为体都可能将人工智能用于网络攻击，包括攻击关键基础设施。人工智能还可能被用于错误信息和虚假信息宣传，包括用于生成虚假信息和深度伪造的信息，并由人工智能驱动的机器人进行传播。使用此种错误信息和虚假信息，例如用于影响选举，可能会破坏稳定。

大规模毁灭性武器

24. 若干国家强调指出，必须保持人类对核武器及其运载系统的控制，并对人工智能可能被纳入核指挥、控制与通信系统表示关切。有国家提及，某些核武器国家承诺保持在对通报和执行有关核武器使用的主权决定至关重要的所有行动中的人为控制和参与。有国家指出了对战略稳定性和升级的潜在影响。

25. 若干国家对人工智能可能助长大规模毁灭性武器扩散，包括向非国家行为体扩散表示关切。在这方面，一个尤为令人关切的问题是，人工智能可能被用于开发和生产生物武器。有国家强调指出，根据现有条约的规定，不得为此目的使用人工智能。还有国家表示，人工智能可用于遏制大规模毁灭性武器的扩散。

四. 现有和新出现的规范性提议

26. 若干国家表示，人工智能应当用于和平目的，包括和平解决争端。各国还强调指出，必须应对并减轻军事领域人工智能引发的风险；其中一些国家指出，应当集体应对军事人工智能引发的挑战。

27. 关于处理军事领域的人工智能问题，各国呼吁采取具备以下特点的办法：

- 灵活、均衡、务实、渐进，从而能够适应技术进步
- 预防性
- 侧重于人工智能的全生命周期，包括预设计、设计、开发、评价、测试、部署、使用、销售、采购、运行和退役

- 基于人工智能的应用和使用，而非技术本身
- 体现现有义务

有国家建议，在该领域的努力应当明确区分致命与非致命用途。

法律考虑

28. 各国回顾了第 79/239 号决议，大会在其中申明，国际法，包括《联合国宪章》、国际人道法和国际人权法，适用于在军事领域人工智能生命周期的所有阶段发生的受其管辖的事项，包括由人工智能启用的系统。有国家指出，国际法、特别是国际人道法，并未明确禁止使用人工智能能力。

29. 各国申明，它们在军事领域使用人工智能时遵守了国际法。有国家建议，遵守法律义务，特别是国际法规定的义务，必须是军事领域人工智能的治理、设计和部署的一个关键考虑因素。此外，有国家认为，人工智能的设计应有助于加强对国际人道法的遵守。若干国家强调指出，必须对这方面的新武器、战争手段或方法实施法律审查。

30. 若干国家强调指出，除法律框架外，还必须考虑到伦理道德因素。

国际和平与安全考虑

31. 若干国家指出，军事领域的人工智能应当加强国际和平与安全，并以不会导致不稳定或局势升级的方式加以使用。有国家表示，各国应当避免通过人工智能来谋求绝对军事优势，并应确保此种技术不会成为发动侵略和谋求霸权的手段。

32. 若干国家指出，人工智能不应破坏现有的裁军、不扩散和军备控制协定。有国家呼吁努力防止人工智能技术向非国家行为体扩散。有国家强调指出，必须避免任意的国际监督机制或歧视性的出口管制。

在军事领域负责任地使用人工智能

33. 若干国家认为，应当在人工智能整个生命周期中以负责任的方式应用人工智能。有国家表示，责任的概念应当与合法性和问责制挂钩。

34. 若干国家强调指出，人工智能的发展必须采用以人为本的办法。许多国家强调指出，必须始终保持人类控制和责任。有国家提及“适合情境的人类控制和判断”和“切实的人类控制”等概念的重要性。相比之下，其他国家指出这些概念并未得到充分界定。有国家认为，使用“切实的人类控制”这一概念可能会妨碍合法研究或不当地限制人工智能在军事领域的使用。

35. 各国强调指出，必须根据国际法确保人类的责任和问责，包括在负责任的、人类指挥和控制系统内的责任和问责。

技术考虑

36. 各国从技术视角审议了各项治理原则，其中包括：

- 保障性，以确保人工智能系统防范外部威胁的稳健性
- 安全性，包括设置护栏机制以便将伤害降至最低
- 可靠性，防止意外后果和故障
- 明确的运行边界和约束，以防止意外行为
- 明确界定的用例
- 治理能力，为此确保适当的人机交互和减少偏见
- 公平与公正
- 隐私保护
- 可解释性、可理解性和可追溯性
- 透明度

37. 有国家强调，所使用的训练数据要能满足充分遵守国际法的要求。若干国家强调，必须在人工智能的整个生命周期进行测试，以发现错误并确保可靠性。还有国家强调指出，需要对人工智能操作人员进行适当培训，以减轻风险并确保遵守国际人道法。有国家着重指出，必须在系统的整个生命周期监测系统性能，并在系统退役时采取安全禁用的措施。

五. 军事领域人工智能方面的倡议和举措

国际论坛

38. 若干国家注意到正在联合国进行的讨论、《未来契约》(大会第 79/1 号决议)及其附件《全球数字契约》，以及大会关于军事领域的人工智能及其对国际和平与安全的影响的决议(第 79/239 号决议)。还有国家注意到安全理事会于 2025 年 4 月 4 日举行的关于利用安全、包容、可信赖的人工智能维护国际和平与安全的阿里亚办法会议。

39. 各国提及了以下方面，即涉及军事领域人工智能的主题的多边讨论，例如裁军审议委员会在题为“关于对国际安全领域新兴技术的共同理解的建议”的议程项目下的讨论、根据《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》设立的致命自主武器系统领域新兴技术问题政府专家组的工作以及大会关于致命自主武器系统的各项决议(第 78/241 和 79/62 号决议)。

40. 各国还提及，它们参加了裁军事务厅和联合国裁军研究所组织的、与军事领域的人工智能有关的各项活动。

国家主导的倡议和举措

41. 若干国家指出，它们发起或参与了与军事领域人工智能有关的倡议和举措，其中包括：

- 《关于在军事上负责任地使用人工智能和自主能力的政治宣言》及其后续落实进程
- 军事领域负责任人工智能进程，其中包括 2023 年在荷兰王国举行的会议(会上核可了一项《行动呼吁》)以及 2024 年在大韩民国举行的会议(会上核可了一份《行动蓝图》)。预计军事领域负责任人工智能全球委员会将在 2025 年在西班牙举行的下一次会议之前发布一份报告。
- 2025 年在法国举行的人工智能行动峰会，会上通过了《关于在人工智能赋能的武器系统中保持人类控制的巴黎宣言》
- 2023 年在大不列颠及北爱尔兰联合王国举行的人工智能安全峰会，会上通过了《布莱奇利宣言》
- 在七国集团范围内开展的人工智能工作
- 人工智能国防伙伴关系
- 2023 年提出的《全球人工智能治理倡议》

42. 有国家表示，这些倡议和举措是有用的，但可能导致各自为政的状况。还有国家表示关切的是，这些倡议和举措的成果并未将所有相关国家的意见考虑在内，可能会破坏这一领域的包容性工作。

区域倡议和举措

43. 各国指出区域倡议和举措发挥重要作用，可促进关于军事领域人工智能的包容各方和针对具体情况的讨论。此方面的事例包括：

- 2025 年东盟国防部长会议非正式会晤期间通过的关于国防领域人工智能合作的联合声明
- 2024 年举行的第十六次美洲国防部长会议，会上通过了《门多萨宣言》
- 北大西洋公约组织范围内的各项活动，包括该组织的人工智能战略(2024 年最后一次修订)及其负责任的使用原则(2021 年制定)
- 2024 年，在军事领域负责任人工智能进程背景下在智利、肯尼亚、荷兰王国、新加坡和土耳其举行的区域协商。

国内倡议和举措

44. 各国提及了各项国内努力，包括现有的人工智能法律、规章、战略和机构，以及为发展上述各方面所作努力。

六. 后续步骤

45. 各国呼吁就军事领域的人工智能开展对话。若干国家呼吁进一步研究军事领域人工智能对国际和平与安全的影响。

46. 许多国家指出，进一步的对话应旨在减轻军事领域人工智能带来的风险。有国家建议，对话目标应当是制定监管或治理框架。若干国家呼吁制定规范、规则和原则，对军事领域人工智能的生命周期进行管理。一些国家表示倾向于制定一个具有法律约束力的框架，但另一些国家认为目前没有必要采取新的法律措施。还有国家表示，规范、规则和原则可为日后的法律承诺奠定基础。若干国家表示反对负责任的开发、部署或使用的规范、规则和原则的概念，指出并未就这一概念达成共识。有国家表示，应避免规章过早地出台。

47. 有国家强调指出，必须避免治理中的重复问题和各自为政状况。有国家认为，关于治理的讨论应当平衡地兼顾人道主义、安全和发展考虑。各国重点指出，必须避免阻碍合法创新和技术进步的各种限制。若干国家认为，不应阻碍人工智能的和平利用，特别是发展中国家的和平利用。

48. 有国家建议，任何治理办法都应考虑到各国在将人工智能纳入军事能力方面处于不同阶段，而且安全环境也不同。有国家强调，所有国家都参与关于军事领域人工智能治理的讨论很重要。许多国家认为，今后的讨论应当采取多方利益攸关方办法，让国际和区域组织、民间社会、科学界和工业界参与。但有国家强调指出，决策应始终是各国的专属特权。

49. 各国审议了今后关于军事领域人工智能的对话的各种优先事项，其中包括：

- 确保遵守国际法，特别是国际人道法
- 保护人的尊严和人权
- 寻求就定义和术语达成共识
- 考虑透明度和建立信任措施
- 处理武力使用方面的自主能力问题
- 处理直接支持作战行动的人工智能系统问题
- 确保适当的数据治理机制
- 加强国际合作与援助
- 支持能力建设，包括通过知识共享、技术转让和良好做法分享，以弥合数字鸿沟和人工智能鸿沟
- 推动持续的区域对话
- 促进国家监管，包括确保私营部门遵守国际法

50. 若干国家建议，关于致命自主武器系统的审议应当是军事领域人工智能讨论的一部分。还有国家表示，正在进行的关于此类系统的讨论是对关于军事领域人工智能的讨论的补充。若干国家回顾了它们关于致命自主武器系统的立场。[1] 一些国家表示，根据《特定常规武器公约》设立的政府专家组是开展关于军事领域人工智能的讨论的最佳论坛，但另一些国家指出，该专家组承担具体任务，成员不具普遍性，因此不是开展此种讨论的适当论坛。

51. 若干国家呼吁在联合国各论坛内讨论军事领域的人工智能问题。有国家建议，本报告可作为此类讨论的基础。各国表示，今后的讨论应当补充正在进行的进程，例如信息和通信技术安全及使用安全问题不限成员名额工作组。

52. 若干国家认为，联合国裁军机制是一个包容各方的有效平台，应当在今后关于军事领域人工智能的讨论中发挥核心作用。有国家建议，裁军谈判会议应当讨论人工智能问题，特别是与核武器有关的问题。有国家表示，还可以在大会第一委员会进行讨论，第一委员会可要求秘书长定期提交关于军事领域人工智能技术发展状况的报告。若干国家建议，可以在裁军审议委员会框架内进行讨论。

53. 此外，有国家建议可在安全理事会的范围内进行讨论。

54. 若干国家建议设立一个专门进程，例如一个不限成员名额工作组。还有国家表示，目前不宜在联合国内设立一个新进程。有国家表示，联合国关于这一主题的任何进程均应遵循协商一致办法。

七. 秘书长的意见和结论

55. 人工智能有可能影响我们生活的方方面面。人工智能若用于和平目的，就可发挥重要作用，推动落实发展承诺和目标，包括可持续发展目标。

56. 在军事领域，人工智能有可能给使用人工智能的军队并给平民带来好处，因为它可提高行动的精确性，降低人为差错的可能性。与此同时，军事领域的人工智能也带来严重挑战，其中最主要的是保持人类责任和问责。

57. 大会第 79/239 号决议申明，国际法，包括《联合国宪章》、国际人道法和国际人权法，适用于人工智能的整个生命周期；该申明是一个重要的基准。然而，关于法律如何适用的重要问题仍有待解决。

58. 需要特别关注涉及使用武力情况下的军事人工智能使用。人工智能在保护平民和战斗人员方面有潜在益处，但当今冲突中据报的人工智能用例，引起了对人类控制及人工智能在助长人口稠密地区敌对行动方面的作用的关切。给予机器剥夺人生命的能力和自由裁量权，在政治上不可接受，在道义上也令人厌憎。

[1] 详情见 A/79/88。

59. 唯有彻底消除核武器，才能根除其风险。在彻底消除核武器之前，我敦促所有拥有核武器的国家商定，关于核武器使用的任何决定都必须由人作出，而非机器。

60. 人工智能可以降低国家和非国家行为体开发或获取化学和生物武器的壁垒。因此，我敦促各国充分履行根据相关裁军、不扩散和军备控制框架承担的义务，并系统地开展评估和做好充分准备，以应对人工智能对这些框架的挑战和影响。

61. 民用人工智能应用可能被纳入军事领域，愈发令人关切。人工智能技术本质上具有多用途适应性，给监督、透明度和追责带来挑战。民用人工智能应用的发展与其在军事领域的潜在用途之间的界限模糊不清，我敦促各国认真审查这一问题。

62. 进一步探索关于人工智能的更多合作机制具有重大价值，在区域和次区域两级尤其如此。区域和次区域组织在制定和实施透明度和建立信任措施方面独具优势，可借此来减轻风险。因此，我鼓励各国考虑在区域和次区域两级制定针对人工智能的独特特点和挑战的透明度和建立信任措施。

63. 在联合国主持下，特别是在实施《全球数字契约》的背景下，关于和平利用人工智能及开展人工智能治理以造福人类的包容性讨论正在进行。尽管如此，会员国对军事领域人工智能的审议主要在联合国各论坛之外进行。大会第 79/239 号决议和本报告是在联合国开展这一重要讨论的重要第一步。我鼓励各国以包容各方的建设性方式进行这些审议，以期增进共识并加强国际合作以减轻风险。

64. 鼓励各国探索如何开展能力建设等各项努力，以确保所有国家都切实参与关于该专题的联合国进程，这对增进共识、拟订共同办法和减轻潜在风险至关重要。

65. 事实证明，大会善于授权开展各项进程，推动就新兴技术与国际安全相关问题开展包容各方的讨论，与此同时促进国际和区域组织、民间社会、科学界和工业界等利益攸关方建言献策。在人工智能领域，这种多利益攸关方办法尤为重要，因为该领域的创新主要由私营部门推动，大量专业知识集中在学术界和科学界，并非由政府掌握。

66. 我建议各国研究本报告所载意见，并在大会第八十届会议上采取具体步骤，以期设立一个包容各方的专门进程，全面应对军事领域的人工智能问题及其对国际和平与安全的影响。

附件一

收到的答复

A. 会员国

阿根廷

[原件：西班牙文]
[2025 年 4 月 10 日]

以下报告是根据大会 2024 年 12 月 24 日通过的题为“军事领域的人工智能及其对国际和平与安全的影响”的第 79/239 号决议提交的。

一般方法

阿根廷共和国认识到，在军事环境中推出人工智能具有重大战略影响。人工智能的使用为各种非致命功能带来了实实在在的好处，但也带来了风险，需要从国际法、道德和行动责任的角度加以应对。在此背景下，在开发和使用此类技术时必须坚持尊重国际人道法和人权，并且必须确保始终维持人类对关键决策的责任和控制。

机遇

人工智能，尤其是其非致命应用，是增强国防能力的合法宝贵工具。优先用途包括以下方面：

- 优化后勤和行动
- 支持情报处理
- 加强网络防御
- 开展模拟、培训和战略规划

这些能力有助于更有效、更安全地开展更适应当前情景的各项行动，在不损害人道主义原则或国家国际义务的情况下加强防御效力。

挑战

人工智能在军事环境中的加速发展带来了必须集体应对的各种挑战，包括以下方面：

- 降低使用武力的门槛，缩短人类决策的时间框架
- 未发现算法偏见的可能性
- 自主系统向非国家行为体扩散
- 固化国家间技术不对称的风险

这些风险凸显了制定共同原则、可核查保障措施和合作框架的必要性。

治理、国际合作和技术包容

我们的理解是，该领域的任何政策制定过程都应建立在以下原则的基础上：

- 应避免制定限制独立开发合法防御技术的一般性规定或不成熟规定。
- 应明确区分致命和非致命用途。
- 应将重要的人类控制作为不可或缺的行动和政策条件加以保证。
- 应促进以加强能力和弥合国家间技术差距为重点的包容性国际合作。

阿根廷已在最近的多边论坛上重申了这些原则，强调必须努力制定在军事领域负责任地使用人工智能的共同标准，特别是在网络防御和网络安全方面这样做。

举一个区域层面倡议的例子：作为 2024 年在阿根廷门多萨举行的第十六届美洲国防部长会议的一部分，负责任地开发、应用和治理军事领域人工智能工作组召开会议，共同致力于制定国际标准。

对《未来契约》的提及

最后，需要指出并记录在案的是，阿根廷共和国已正式表示不赞同大会第 79/239 号决议序言部分引述的《未来契约》。因此，提及《未来契约》并不表示阿根廷承诺、遵守或支持该契约。

奥地利

[原件：英文]
[2025 年 4 月 11 日]

根据大会第 79/239 号决议第 7 段的要求，奥地利愿从国家角度分享以下思考和意见。

与网络安全和网络防御有关的人工智能

人工智能赋能的网络安全软件已被广泛用于帮助检测计算机网络中的入侵和其他恶意活动。此类人工智能工具可能会通过搜索漏洞和可疑活动来提高软件和硬件的韧性，从而促成对信息技术系统进行日益自动化的保护。

与此同时，在网络安全人工智能攻防模型之间的竞赛中，人工智能工具越来越多地被用于增强网络攻击的复杂性和制造新型计算机病毒。此外，包括大语言模型在内的人工智能赋能软件降低了恶意行为体的准入门槛，后者日益无需广泛编程技能即可创建恶意软件。

作为一项混合战略要素与虚假信息宣传有关的人工智能

可以创建和传播伪造内容的人工智能赋能软件越来越多地被用于加强虚假信息活动。所用的方法包括利用生成式人工智能来大规模创建量身定制的本地化内容。此外，人工智能驱动的深度伪造音频和视频软件正在迅速改进，且已得到广泛应用。这些伪造的内容可以利用大规模人工智能驱动社交媒体机器人

网络进行传播，制造舆论转变的表象。因此，人工智能降低了开展大规模虚假信息宣传的门槛，因为所制造虚假内容的数量和质量不再受限于人类操作员的数量或技能。

不过，人工智能算法也可以用来揭露人工智能生成的内容和“伪草根运动”，而使用专门的人工智能工具则可以最有效地揭露以假乱真的深度伪造音频和视频。有必要使用此等人工智能驱动工具来抵消用于虚假信息宣传的人工智能所产生的不良影响。

与武器扩散有关的人工智能

人工智能可以降低获取武器(包括大规模毁灭性武器)的门槛。只需按一下按钮，大语言模型即可提供专业知识，因此，大语言模型以及基于大语言模型的应用程序可能会导致恶意行为体能够更轻而易举地制造武器。应用案例从获取小武器和轻武器的蓝图或打印其组件到转变病原体用于生物战，不一而足。如果现成可用知识缩小了武器方案的范围和规模，那么，发现、预防和防备这些威胁就会更加困难。

与此同时，机器学习算法也可用于打击武器扩散。机器学习算法具备异常检测和模式识别能力，因此有助于识别恶意活动，包括通过检测用于武器方案的非法资金流动或分析卫星数据的模式来做到这一点。

与危机局势中的军备控制核查和决策有关的人工智能

人工智能可以帮助核查军备控制协定。这是因为人工智能有能力分析大量数据(例如，来自卫星图像等来源的数据)并对不同物体进行分类，从而可以识别坦克、导弹和兵营等军事装备或部队调动和演习等军事活动。此外，如前所述，人工智能可以更轻松地发现非法武器方案，从而使违反军备控制协定的行为更加难以实施，而缔约国则可确保各方都遵守协定条款。

基于人工智能对传感器数据进行分析和分类的能力所产生的更多更好的信息不仅可以促进军备控制协定的执行，而且有助于在国家间军事关系特别紧张的情况下做出更好的决策。政治和军事领导人可以在人工智能的辅助下更好地了解局势，从而缓解危机。

和平与安全与《联合国宪章》

在军事领域应用人工智能的一个特殊挑战是，由于使用人工智能造成的意外局势升级和误解，可能会给和平与安全带来风险。机器学习的使用则增加了另一层复杂性，因为系统的运行可能无法被所有行为体充分理解。

在人机交互和人类主体的必要性方面，还需要为在决策支持系统中使用人工智能采取相关措施并设置护栏机制，以确保问责和责任并减少算法偏见。

所有这些风险都必须通过考虑到这些技术带来的具体挑战的监督和措施予以减缓。

应指出，《1949 年 8 月 12 日日内瓦四公约关于保护国际性武装冲突受难者的附加议定书》(《第一议定书》)第 36 条规定，在武装冲突中使用所有新武器、作战手段或方法之前，有义务审查其合法性。

作为一项积极义务和平权行动，人工智能还可用于支持有效履行国际人道法义务，特别是在保护平民方面，包括通过专门为此任务设计的项目、研究和应用程序做到这一点。

多边合作和信息交流框架

军事领域的人工智能问题发展迅速且对所有国家都带来了挑战，因此，交流经验和最佳做法的多边讨论和多边形式非常有现实意义。在这方面，奥地利赞同《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。作为该宣言监督工作组的共同主席，奥地利与德国一道，一直在促进共享在该领域应对挑战和制定相关政策的最佳做法。奥地利还赞同军事领域负责任人工智能峰会《行动蓝图》以及《关于在人工智能武器系统中保持人类控制的巴黎宣言》。

国际社会在军事领域人工智能方面的工作及其在自主武器系统方面的工作之间的关系

在人工智能更广泛的应用范围和军事领域的自主能力方面，需要强调自主武器系统这一具体问题。从法律、道德和安全角度看，自主武器系统尤其令人关切。该问题并非大会第 79/239 号决议的侧重点，因为自 2013 年以来联合国框架内的讨论一直在进行，越来越多的国家表示希望在国际层面为自主武器系统制定规则和限制。因此，对于这份报告，奥地利的评论仅限于强调其赞成就自主武器系统制定一项具有法律约束力的文书的立场，并在此提及政府专家组目前正在《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》框架内开展的重要工作，以及在以下两份决议框架内所作的补充工作：大会有史以来首份关于致命自主武器系统的决议(第 78/241 号决议)——已根据该决议发布了秘书长的报告(A/79/88)；后续决议(第 79/62 号决议)——该决议确定于 2025 年 5 月 12 日和 13 日在纽约举行关于致命自主武器系统的非正式协商。

与人工智能法律框架有关的考虑

《欧洲联盟人工智能法》为欧洲联盟各行各业的人工智能系统建立了立法框架，目的是培养对人工智能应用的信任，并在保障人权、基本自由和民主价值观的同时驾驭人工智能的惠益。它强调了在促进法律确定性、创新和竞争力的同时在人工智能系统的开发和部署方面确保透明度、问责制和人类监督的重要性。该法不适用于为军事、防务或国家安全活动开发的人工智能系统。不过，《欧洲联盟人工智能法》确实采用了基于风险的方法，在处理军事领域广泛的潜在人工智能应用时该方法可能会被证明有用。

前进方向

奥地利珍视在其就人工智能在军事领域的应用提交的资料中所提到的当前以各种形式、在各种论坛上开展的工作，并相信这些工作将有助于形成一套新的国际商定规范和标准，从而确保根据国际法律义务和道德原则在军事领域负责任地使用人工智能。

智利

[原件：西班牙文]
[2025 年 4 月 11 日]

智利以前就曾指出，新技术和新兴技术的迅速发展是国际安全领域的一个重要问题，对所有国家都构成挑战。这些新技术，特别是人工智能，可能会为社会的发展和福祉带来巨大惠益，但同时也提出了其在安全和防务领域的使用会产生哪些影响这一重大问题。新技术可以产生重要惠益，但也会带来风险和困难。

在这方面，智利认为，最好就在军事和安全领域负责任地使用人工智能以及发展和使用所谓的致命自主武器系统达成共识。智利支持为建立和加强促进各国间对话和讨论的论坛所作的多边努力，目的是就使用这些新技术找到可以达成相互谅解和共识的领域。

智利已在人工智能领域占据领先地位，这得益于我国在为部署此类技术创造有利条件方面取得的重大进展以及在人工智能政策和监管讨论方面取得的突破性进展。2021 年 10 月，智利推出了首个国家人工智能政策，该政策是与各种公共和私人利益攸关方合作制定的。该政策重点关注三大基本支柱：促成因素、技术使用和开发、为确保负责任和安全地使用人工智能建立监管和道德框架。

2024 年，智利发布了国家人工智能政策更新版，其中包括国际协调、环境和气候危机、包容性和不歧视、儿童和青少年以及文化和遗产保护等新的分专题。为补充该政策，还制定了一项行动计划，其中包含 100 多项将于 2026 年完成的措施，涉及教育、卫生、环境和文化等领域。联合国教育、科学及文化组织(教科文组织)发布的《人工智能伦理问题建议书》规定的原则也被纳入新的人工智能国家政策，以便该政策与最新国际框架保持一致。

智利是世界上第一个采用准备状态评估方法的国家，该方法是教科文组织开发的一项工具，用于确定一个国家以合乎道德和负责任的方式实施人工智能的准备状况。为此，智利重申其致力于在国家法规中实施教科文组织《人工智能伦理问题建议书》。智利一直在促进合乎道德和负责任地发展此类技术，我国参加了由联合王国(2023 年)、大韩民国(2024 年)和法国(2025 年)组织的历次人工智能峰会就体现了这一点。

在立法方面，智利目前正在讨论一项使用基于风险的方法监管人工智能系统的法案，目的是促进此类系统的开发和实施，同时维护民主原则和个人的基本权利。

在防务和安全领域，智利支持并积极参加了在海牙(2023 年)和首尔(2024 年)举行的军事领域负责任人工智能峰会。智利赞同两次峰会通过的文件(行动呼吁(2023 年)和行动蓝图(2024 年))。智利还支持军事领域负责任人工智能全球委员会的工作。

2024 年 6 月 13 日和 14 日，在智利举行了关于在军事和更广泛安全领域负责任地使用人工智能的区域讲习班。讲习班由智利外交部和哥斯达黎加外交部组织，得到荷兰王国和大韩民国的赞助。智利大学法律、技术和社会研究中心与设在日内瓦的人道主义对话中心也支持并协助组织了此次活动。阿根廷、巴西、智利、哥伦比亚、哥斯达黎加、多米尼加共和国、萨尔瓦多、牙买加、墨西哥、巴拉圭、特立尼达和多巴哥和乌拉圭以及荷兰王国和大韩民国的代表参加了讲习班。智利国防部和武装部队的代表也代表智利参加了会议。

智利认为，人工智能在军事和安全领域的应用可以产生机遇和惠益，如增强决策工作和战略分析、提高后勤业务效率、增强网络防御和网络安全能力——从而加强关键基础设施的安全，同时为复杂的维持和平和人道主义援助任务的规划工作提供便利。人工智能应用还可以提高军备控制和军备控制制度执行情况的核查和监测能力。

智利认为，人工智能技术的开发、部署和使用必须符合国际法，包括在适用情况下遵守《联合国宪章》、国际人道法、国际人权法和其他相关法律框架。

对智利而言，必须建立控制和安全措施，以防止不负责任的行为体在军事领域获得和滥用可能有害的人工智能能力(包括人工智能赋能的系统)，同时铭记任何此类措施都不应损害在其他非军事领域公平获得人工智能能力的惠益。

同样，智利认为，必须共同努力，防止人工智能技术被用于为大规模毁灭性武器在国家和非国家行为体(包括恐怖团体)之间的扩散提供便利，并强调人工智能技术应被用于支持而不是阻碍裁军、军备控制和不扩散努力。尤其必须在不损害实现无核武器世界这一最终目标的情况下维持人类控制和参与对通报和执行关于核武器使用的主权决定至关重要的所有行动。

智利主张制定建立信任措施，如各国就良好做法和经验教训进行信息交流、开展磋商。在这方面，智利认为各国必须制定国家战略、原则、标准、政策、框架和法律并使之制度化，以确保在军事领域负责任地使用人工智能。建立信任措施可以成为在国家和国际层面建立遏制、控制和信誉机制的有效工具，从而提高透明度。

同样，智利认为，必须缩小发达国家和发展中国家之间在数字化和人工智能方面的差距，并认为有必要增强对人工智能在军事领域影响的理解和认识，包括在各国之间开展知识交流、分享良好做法和经验教训。

在这方面，智利认为，必须制定旨在促进能力建设、特别是在发展中国家促进能力建设的倡议和方案，以促进它们充分参与关于军事领域人工智能治理的辩论。智利认识到，能力建设还可以帮助各国更深入地了解军事领域的人工

智能，并促进负责任地合法开发、部署和使用军事人工智能能力。能力建设还将使各国能够更有效地参与国际对话和讨论。

智利认为，必须为能力建设加强国际合作，促进国家、区域、次区域和区域间等层面的对话和辩论，包括为外交、政治和技术官员举办培训课程、会议、讲习班和研讨会，以期弥合在军事领域负责任地开发、部署和使用人工智能方面的知识差距。

智利赞赏就军事领域的人工智能开展区域和次区域讨论和对话，并认为必须促进这样的讨论和对话。为此做出的显著努力包括 2024 年 10 月 13 日至 16 日在阿根廷举行的第十六届美洲国防部长会议，特别是会议成果文件《门多萨宣言》，其中载有以下建议：促进合乎道德地在防务领域使用人工智能；考虑到会议成员国的经济和技术多样性；促进加强相互信任以及半球和区域合作的机制，通过这些机制，会议成员国可以分享知识和良好做法、制定基于共识的标准、建立在防务领域应用人工智能的技术能力。

最后，智利认为，在关于军事领域人工智能的讨论和对话中，包括民间社会、学术界、产业界、私营部门、技术界以及区域和国际组织在内的有关各方的参与至关重要。

中国

[原件：中文]
[2025 年 4 月 11 日]

人工智能技术快速发展并广泛应用于军事领域，改变未来作战方式，给国际和平与安全带来潜在挑战。在世界和平与安全面临多元挑战的背景下，各方应通过对话合作就如何规范人工智能军事应用寻求共识，推动构建开放、公正、有效的安全治理机制，最大限度降低风险，确保人工智能安全、可靠、可控，始终沿着人类文明进步的方向发展。

中国始终以负责任、建设性态度参与军事人工智能全球治理，主张遵循“以人为本的军事人工智能”理念，秉持共同、综合、合作、可持续的安全观，推动构建人类命运共同体。2021 年，中国向《特定常规武器公约》提交了关于规范人工智能军事应用的立场文件，从战略安全、军事政策、法律伦理、技术安全、研发操作、风险管控、规则制定及国际合作等角度，就如何在军事领域负责任地开发和利用人工智能提出系统性看法建议。2023 年，中国提出《全球人工智能治理倡议》，主张各国尤其是大国对军事领域研发和使用人工智能技术应该采取慎重负责的态度。中方具体主张包括：

一是坚持慎重负责。各国尤其是大国在发展正当国防能力的同时，不应借助人工智能谋求绝对军事优势、损害他国合理安全利益，应避免误解误判，防止军备竞赛。

二是坚持以人为本。始终坚持人类是最终责任主体，确保由人控制相关武器系统，尊重和保障人的尊严和人权，遵循全人类共同价值。

三是坚持智能向善。人工智能在军事领域的应用应有利于维护和平，遵守国际人道法和其他适用的国际法，努力减少附带伤亡。

四是坚持敏捷治理。加强前瞻性风险研判和人员培训，实施必要的风险减缓措施，降低扩散风险，同时不妨碍技术创新与和平利用。

五是坚持多边主义。支持联合国发挥应有作用，欢迎各方搭建包容性讨论平台，推动建立普遍参与、具有广泛共识的治理框架。

中方认为，应该客观评价人工智能军事应用的意义，既要引导军事人工智能的正确发展方向，又要防范其毫无约束的发展。下阶段，国际社会应共同努力，趋利避害。中方有以下几点思路和建议：

一要确立指导思想。坚持发展与安全并重，遵守《联合国宪章》宗旨和原则，恪守国际关系基本准则，确保人工智能技术不被用于侵略他国和追求霸权。中方愿就“以人为本的军事人工智能”理念与各方进一步交流，不断凝聚共识。

二要完善治理举措。结合人工智能技术发展及应用现状，推动建立风险评估测试体系，实施敏捷治理，分类分级管理，快速有效响应。各国结合国情，建立健全国内法律和规章制度，完善相关伦理准则，加强教育培训，提升人工智能技术的安全性、可靠性和可控性。

三要加强国际合作。各国应坚持开放包容原则，开展对话交流，增进相互理解，在治理问题上各国应加强政策协调和能力建设合作，不断提升治理水平。

埃及

[原件：英文]
[2025 年 4 月 11 日]

根据大会第 79/239 号决议，阿拉伯埃及共和国政府愿就人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战分享看法。

大会第 79/239 号决议是就军事领域的人工智能这一专题促进多边主义并将其置于政治议程更高位置迈出的重要一步。此前，秘书长呼吁在所有相关利益攸关方的参与下围绕人工智能军事应用的设计、开发和使用制定规范、规则和原则。

我们的理解是，提出这些观点所依据的上述决议旨在特别侧重致命自主武器系统以外的领域。有鉴于此，我们必须重申埃及的坚定立场，即任何关于该主题的切实讨论都不能忽视解决致命自主武器系统所有伦理、法律和安全层面问题的优先级别。就人工智能的军事应用而言，致命自主武器系统是对维护国际和平与安全的最紧迫威胁。

正如联合国秘书长先前所建议的那样，最有效、最现实的行动方针是商定一项具有法律约束力的文书，禁止在不受人类控制或监督的情况下运行且其使用不符合国际人道法的致命自主武器系统。寻求采取禁止和限制和(或)监管的双重方法——包括禁止在没有人类控制的情况下运行的武器系统并管制其他系

统——对于建立必要的普遍法律架构至关重要，此等法律架构会为最大程度地实现人工智能军事应用提供的新机会所产生的惠益提供有利环境，同时以现实、有效、及时的方式应对相关挑战。

围绕军事领域人工智能的国际政策格局远未统一。埃及密切关注多项相关国际倡议，这些倡议表明对相关风险的认识在不断提高。然而，这些倡议过程的审议表明，在观点、威胁认知和优先事项方面存在分歧。因此，我们必须警惕创建碎片化政策框架或相互竞争进程的风险，新技术和新兴技术的其他领域就是这样。

显然有必要精简这些倡议并将其纳入联合国框架，以确保其包容性和有效性。联合国及其裁军机制是制定必要国际规则和规范框架的唯一有效且包容的平台，在技术发展在速度上继续远超国际层面的必要管制时，这点尤其突出。

因此，亟需在联合国的主持下建立一个通用、独立、单轨且值得信赖的平台，探讨军事领域人工智能的未来治理。所设想的联合国主导进程应有所调整，以避免出现某些适得其反的二元对立。其中之一便是确保法律合规和道德规范的合法努力与不顾人道主义影响追求军事利益的倾向之间的二元对立。

还应强调的是，虽然埃及赞赏致命自主武器系统问题政府专家组在《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》框架内进行的讨论，但鉴于政府专家组既不具有普遍性，也没有处理如此广泛多样专题的任务授权，该平台无法取代所设想的关于军事领域人工智能应用的联合国进程。同样令人遗憾的是，政府专家组内部的进展仍然微乎其微，尚未取得任何切实成果。

人工智能技术在带来机遇的同时，其自身特性也蕴含着诸多风险，其运作方式可能无法预测且无法解释。这些风险包括虚假信息、意外升级和网络风险以及非国家行为体的滥用和扩散。这些风险可能是全新的，也可能使现有风险更加复杂。

人工智能的军事应用前景广阔，这一点已得到广泛认可。然而，为详细阐述其未来治理而作出的切实努力应根据其固有风险及其对和平与安全的影响确定正确的优先次序，从而确保讨论有重点、有条理，同时避免不必要的干扰。话虽如此，埃及坚定地认为，除了致命自主武器系统问题外，还应强调其他自主或半自主系统能力，这些能力能够启用武力和(或)降低使用武力的门槛，从而可能引发更多常规和非常规武器军备竞赛动态。核武器和先进常规武器(如高超音速导弹)自主性增强的可能性会带来未知风险，并以不可预测的方式改变未来的冲突。

重点还应放在指挥与控制以及目标选择活动上，而不是放在后勤规划以及情报、监视和侦察上，因为它们的破坏性影响相对较小。同样，应更加注重进攻能力，而不是防御能力。

设想在理想的联合国主导进程中进行的审议应首先致力于就军事领域人工智能的开发、部署和使用的主要要素达成共识。这些要素包括：

- 在军事领域人工智能应用的全周期和各阶段充分遵守适用的国际法，包括必要性、相称性和区别性等国际人道法的基本原则以及其他道德考虑因素。

- 在人工智能军事应用的全周期中保持人这一要素的中心地位，包括作为保持问责之关键推动因素的人类判断、干预、监督和控制。有必要确保所有涉及在军事领域使用人工智能应用程序的软件、算法和设计仍需接受严格的人工审校并遵循可解释性原则。尽管各国政府声称从理论角度看人类对人工智能赋能系统的控制得以维系，但有些政府可能更倾向于不断增强其武器系统的自主性，以进一步实现军事利益。

- 在减轻非国家行为体的扩散风险和遏制恶意使用与维护各国获得人工智能和两用技术的权利之间取得平衡。必须避免采用任何武断的国际监督机制或实行任何类型的歧视性出口管制。

- 设立一个能力建设部分，目的是确保对人力资本、技术转让以及共享知识和最佳实践进行适当投资，以维护发展中国家从各种人工智能军事应用的潜在惠益中受益的权利，另一目的是弥合数字鸿沟。

- 人工智能在军事领域的边界及其与其他新技术和新兴技术的相互作用。鉴于人工智能与网络行动之间的交叉等情况，有必要讨论如何确保与联合国主导的其他进程(包括信息和通信技术安全和使用安全不限成员名额工作组)之间的互补性。此外，除了更广泛的安全领域外，讨论应主要集中在军事领域。

最后，在联合国多方利益攸关方视角内为促进负责任、可问责和以人为本的人工智能制定治理路径时必须确保包容性和公平性，以提供可馈入政策讨论的重要意见投入。但多方利益攸关方的参与不应损害各国在政策制定过程中的主权特权。

萨尔瓦多

[原件：西班牙文]
[2025 年 4 月 10 日]

背景

近年来，人工智能在军事领域的应用发挥了非常重要的作用。诸多报告已确定，这些新技术日益先进、不断得到普及，使得在军事规划和决策过程(包括关于攻击谁或攻击什么的决策过程)中使用这些计算工具成为可能。这引发了关于使用这些技术带来的总体影响、法律含义和对平民的风险等诸多问题。其中一个例子是在多边谈判期间围绕自主武器系统就此类技术的政治、法律和人道主义影响进行的辩论。然而，人工智能的军事应用范围要广泛得多。

因此，我们需要扩大对人工智能在军事环境中的用途和应用的理解，特别是在军事目标选择和使用武力等具体任务方面。

自 2023 年 2 月在荷兰王国举行的首届军事领域负责任人工智能峰会上就其进行讨论以来，在军事领域负责任地使用人工智能这一问题已变得尤为重要。在日内瓦举行的致命自主武器系统问题政府专家组会议上，这一问题也开始具有更加重要的意义。

必须注意到，迄今为止，人工智能的应用和用途主要在讨论自主武器系统的背景下进行辩论，但人工智能在军事应用程序中的使用是一个更广泛的问题，它已具有新的维度，不仅适用于注重武器系统自主性的应用程序，而且特别适用于旨在实现某些军事功能自动化的应用程序。

从广义上讲，人工智能的讨论是一个新兴的议题，仍处于探索阶段且发展非常迅速，以至于国家、区域和多边层面都在制定举措以应对其影响。显然，就促进识别和了解使用人工智能所带来的机遇和风险的技术和能力建设而言，拉丁美洲和加勒比国家与发达国家不在一个水平上。因此，这些国家必须树立国家立场，使其有能力积极参与各种国际论坛和场合上出现的讨论，并在这么做时确保在能力建设和具体方面开展合作，从而可以站在该问题的最前沿，了解国家、区域和全球层面的机遇和潜在的安全相关风险。

萨尔瓦多参与的倡议

- 萨尔瓦多参加了 2023 年在荷兰王国举行的军事领域负责任人工智能峰会并赞同峰会发表的宣言(2023 年 2 月)。
- 萨尔瓦多参加了拉丁美洲和加勒比自主武器社会和人道主义影响会议，会上通过了《贝伦公报》(2023 年 2 月)。
- 在致命自主武器系统问题政府专家组开展讨论时，萨尔瓦多是“16 国集团”的成员。虽然这一议题的侧重点不同，但与人工智能在军事领域的运用有一定的关系。

国家立场

- 某些人工智能应用程序，特别是那些与识别和辨认军事目标的功能无关或与使用武力(给平民带来风险)没有明确关联的应用程序，在军事领域可能有一定的好处。此类应用程序涉及其他行政管理任务，如与军事行动中的人类互动无关的数据分析和自动学习。
- 然而，滥用这些应用程序可能会产生不利影响，尤其会对保护平民和民用基础设施产生不利影响，而平民和民用基础设施是受到国际法(包括国际人道法和国际人权法)规则特殊保护的类别。
- 必须采用基于风险的方法，通过这种方法可以监管或禁止某些人工智能功能，特别是那些限制对使用武力进行重要人类控制的功能以及那些复制因使用不具代表性或包含历史数据的数据库而产生的算法偏见

的功能。此类功能构成人权相关风险，从长远看，还构成国际安全相关风险，而当决定人类生死的权力被交给机器或者当此类工具包含自学等高度复杂的技术元素从而可能带来严重的人道主义、社会、经济、政治甚至环境后果时，这点尤其突出。

- 目前，亟需在人工智能领域引入适足的监管，因为监管对于确保以合乎道德和安全的方式开发此类技术至关重要。这将有助于保护用户和社会免遭潜在的滥用和风险，还将通过为开发人员和研究人员提供清晰安全的环境来鼓励创新。

- 虽然最终目标是制定具有约束力的法律文书，但这些技术的进步速度已经超过了该领域国际法的演变和发展速度。因此，我们认为宜采取一种注重负责任行为的方法，然后可用这种方法为更好地处理该问题建立全面法律承诺的基础。

- 必须考虑到新兴技术在安全问题上带来的挑战。例如，三维打印等材料技术可用于制造小武器和轻武器；在军事领域可使用机器人技术开发具有自主能力的机器人；人工智能某些用途和应用的双重性质意味着它们可能会在武装冲突的指挥和控制功能中复制偏见，从而对平民构成更大的风险。

- 军事领域控制权的丧失或代行可能会引发意想不到的风险。人工智能可用于提高人类能力，但在军事背景下缺乏控制可能会带来其他风险，必须对此进行充分探讨。军事领域的人工智能相关支持应该用于加强特定背景下的决策工作或为其提供信息，但永远不应取代人类的决策和推理。

- 在军事领域使用人工智能必须遵守国际法、国际人权法和国际人道法，必须服务于公共利益。

- 各国必须能够加强其能力，识别因滥用人工智能而产生的风险以及与国际法的相关联系。

- 应将私营企业和学术界等参与创造和开发这类技术的其他行为体纳入多边讨论，并应鼓励利益攸关方之间开展国际合作，以释放和平利用人工智能所带来的各种惠益，支持各国的发展。

芬兰

[原件：英文]
[2025 年 4 月 11 日]

芬兰很高兴提交其对大会 2024 年 12 月 24 日通过的、关于军事领域的人工智能及其对国际安全的影响的第 79/239 号决议的看法；大会在该决议中，请秘书长就“人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战征求会员国的意见，特别侧重于致命自主武器系统以外的其他领域”。

通过关于人工智能在军事领域的应用的国际原则或条例，对于确保遵守国际法、加强安全并减少潜在冲突风险至关重要。与此同时，有必要使符合国际法的国家防卫能力得以发展。芬兰致力于以负责任的方式，根据国际法、特别是国际人道法，以不破坏国际和平，安全与稳定的方式，在军事领域开发、部署和使用人工智能能力，同时在人工智能技术领域致力进行研究、开发、试验和创新。

查明颠覆性技术对外交、安全和国防政策的影响，并制定解决问题的办法，已经变得越来越重要。芬兰积极参与关于技术监管的全球辩论，倡导基本权利和人权，并在人工智能和相关政策的开发和应用中应对相关风险。

除了识别颠覆性技术的风险外，还必须认识到此类技术为安全、防卫能力发展、经济增长、生产力、可持续发展、技术能力和部门投资提供的机会。

机遇

颠覆性技术为推动各部门发展、推动清洁转型、促进可持续经济增长以及提高效率和生产力提供了重要机遇。此类技术还具有在全球一级加强安全、教育、福祉和健康的潜力。

人工智能和其他新兴技术为提高防御能力带来了机会，同时从根本上影响了未来战场、作战手段和方法的形成。技术进步使信息收集和数据处理更有效率，提高了态势感知能力，加快了决策速度，并使接触更精确、范围更广。在现代战争中，远程操作和自主无人系统的重要性日益增加，此类系统将改变战争、作战和战场的未来。随着未来技术发展速度的加快，预测技术进步、将新兴技术纳入防御系统以及利用意外情形，将变得越来越重要。技术优势也可以弥补数量上的劣势。

挑战

与此同时，必须广泛了解以下各方面：安全威胁、滥用的可能性、人权问题以及与人工智能等颠覆性技术发展相关的相互依存关系。随着上述各方面的发展，它们特别是对国防和安全部门，将构成新的挑战。人工智能的发展使网络攻击、信息影响活动及其工具之一——虚假信息，更具针对性和有效性。此外，人工智能已经被用来影响选举。在这样的环境中，还必须更加重视保护机密信息的安全。

国际法，特别是《联合国宪章》、国际人权法和国际人道法，完全适用于网络空间。尊重和遵守网络空间负责任国家行为框架，对于维护国际和平、安全与稳定而言，仍然至关重要。技术发展提出了新问题。例如，这些问题涉及网络环境、人工智能的使用、新武器技术以及关键原料的开发。混合影响活动可能包括旨在阻碍实现国际法规定的问责制的做法。芬兰主张在开发和应用人工智能以及制定相关法规时，充分考虑基本权利和人权以及与之相关的风险。制定本国的原则、标准和规范、政策和框架，对于按照国际法、确保人工智能在军事领域得到负责任的应用，至关重要。

技术发展为敌对行为体提供了新机会，使之能够进行低于公开冲突门槛的混合影响活动。敌意的网络行动已成为强权政治的既定内容，也是国家行为体进行影响活动可用的广泛工具的一部分。平常条件下也进行网络行动、混合行动和信息行动，这可能会模糊战争与和平之间的界限。尽管战争的技术性越来越强，但常规战争能力仍然很重要，特别是在大规模和长期冲突中。

许多国家都面临着激烈的信息影响活动；这些活动中，也运用了人工智能。信息的有害使用已成为广泛影响活动的日常组成部分，信息环境中的竞争更加激烈。

基础设施和技术的发展以及用户数量的增加为网络领域的敌对行动提供了更大机会。许多国家不断面临敌对行为体收集信息网络情报、网络间谍和网络攻击的问题，这些行为体还试图对关键基础设施产生实体影响。除国家行为体发挥作用外，出于政治动机或由国家主导的非国家行为体作为敌对活动的策划者，其作用也在不断增强。

法国

[原件：法文]
[2025 年 4 月 11 日]

一. 军事领域的人工智能及其对国际和平与安全的影响

可利用的机会

帮助规划和决策。法国军队正在利用其弹药和炸药相关事件数据库开发用于预测特定地区潜在威胁的工具。

支持人类。推出了用于机组人员培训的人工智能系统，旨在通过分析从飞行或模拟中收集的数据来改善法国飞行员的培训。人工智能还可以在面对海量数据时帮助人类，例如“金耳”声学分析系系统处理海量声学数据，以便法国操作人员可专注于有附加值的信号。

克服我们在信息和通信技术领域的弱点。人工智能技术可用于支持网络安全，并处理虚假信息扩散问题。法国军队依赖于深伪检测系统。

促进国际人道法的实施，保护人员和财产。人工智能可以为实施国际人道法的基本原则，例如区分、相称性和预防原则作出贡献。人工智能还可用于保护人类，利用配备人工智能传感器的无人机帮助清除地雷。

加强军备控制。人工智能可用于更好地监控和侦测秘密发射、武器生产地点的变化或生化武器试验。人工智能还可以改进武器出口的可追溯性，以加强对此类出口的管控。

加强预防、维持和平及建设和平工作。人工智能将使维和行动更好地适应环境，从而更加有效。法国军队推出即时翻译系统“Resistance”，旨在方便人们与当地居民在无网络连接的情况下进行线下交流，从而打击虚假信息。

需缓解的风险

技术特有的风险。机器学习技术存在各种与偏见相关的风险，源自于无意识偏见、有意识偏见、重组特敏感数据方面的偏见、不透明或难以解释的结果。还存在着能源消耗呈指数级增长的问题。

国际安全和稳定面临的风险日益加大。在不当掌控者手中，人工智能可能会加剧某些国际安全和稳定风险(局势升级、军备竞赛、向非国家行为体扩散、信息宣传和敌对行动延伸至网络领域)，因此需要调整风险缓解措施加以应对。对技术的依赖引起责任缺失的风险，因此，必须确保由人来负责。

二. 在整个生命周期中实现“负责任的人工智能”的关键原则和措施

开发确保遵守国际人道法的人工智能

调整法律审查。这项审查虽然完全适用于军事人工智能，但审查的具体方式必须适应该技术的特点。

进行适当的后续审查。在武器系统生命周期的不同阶段，都必须在必要时进行这种审查。当一个装置进行创新或添加新组件、可能会显著改变其效能时，就必须进行这种审查。

开发可靠、安全的人工智能

对系统进行评估、验证和认证。应在设计阶段，通过风险分析，(根据相关功能的重要性)在适当级别进行评估和验证。这些系统应与明确用途的案例相关联。应根据所涉问题，以适当的时间间隔重复进行审查。

依靠受控的主权数据。应实施对策和适当的禁令，以应对数据泄露的风险。

修正和重新训练系统。务必要指明和描述(在测试或操作使用过程中)所遇到的错误，让操作人员了解需要得到反馈信息，并持续检查该系统是否符合我们的国际义务。

将人工智能置于适当的人类控制和负责任的指挥链之下

确保决策和行动符合法律要求。操作人员或军事指挥官应能够自主判断，对所获得的结果是否符合所下指令和法律义务进行核证。

确保人类责任。人类在设计、部署和使用人工智能技术方面负责，这是不可动摇的原则，需要正式确定负责指挥、控制和执行职能的人员责任链。

调整人类控制。在不限制基于人工智能技术的系统能力的情况下，分析和界定适当的人类控制——这个问题很复杂，需要考虑不同的人类因素、技术因素和环境因素。

培训军事指挥官和人员掌握这些系统。使用前必须有个培训实践阶段，让人员了解相关益处和风险。

发展可持续的人工智能

保护研究。研究项目的目标和范围应该是开放的，不应自动颁布范围过广的禁令。

审议研究对伦理的影响。法国成立了国防伦理委员会，以审议国防领域新技术带来的伦理挑战。

开发节俭的人工智能。提倡节俭行为意味着反思人工智能的使用，提高系统的韧性和可持续性，同时控制成本。

三. 建立实施全球治理的专门进程，以落实负责任的人工智能原则

普遍和包容的进程。相关讨论必须包括所有利益攸关方，特别是各国——尤其是开发和使用这些系统的国家的积极参与是绝对必要的；因此，决策过程应考虑各种相关立场，并制定相应的规则，以确保达成共识——同时也应考虑工业、科学和学术界以及民间社会，不使讨论脱离现实，并保持创新。大会第一委员会可以是进行讨论的适当场所。

经过精简和协调一致的治理架构。单一的框架应能够精简相关工作，提高效率并增强成果的影响。必须确保与致命自主武器系统领域新兴技术问题政府专家组的讨论相辅相成；2026 年后，专家组须能够根据新的任务规定继续工作。

以军事部门特有问题为中心、注重业务的进程。治理应立足于适用于武装冲突的法律体系，主要是国际人道法。任何国际进程的优先事项都应该是确保遵守现有法律准则，讨论指导原则制定事宜以及各国实施这些原则的方式(如促进最佳做法的交流，并以适合军事的方式促进国际合作和援助)，同时促进适当的建立信任措施和减少风险措施。

德国

[原件：英文]
[2025 年 4 月 11 日]

一. 导言

近年来，人工智能技术出现了前所未有的发展，包括在颠覆性技术(如生成式人工智能)的基础上开发应用程序。各国必须能够利用这些技术发展带来的机会，并确保技术进步不会受阻。与此同时，各国需要确保以负责任的方式在军事领域开发和使用人工智能应用程序，并完全遵守国际法、包括国际人道法。要掌握这种平衡术，就务必要开展国际交流。

在此背景下，德国积极参与有关在军事领域负责任地使用人工智能问题的国际进程。除其他外，德国作为共同提案国核心小组的一部分，推动了大会关于“军事领域的人工智能及其对国际和平与安全的影响”的第 79/239 号决议，并全力支持秘书长努力提交实质性报告，说明会员国对“人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战”的意见。

德国欢迎有机会更深入地研究会员国和其他利益攸关方的观点，并分享德国自己在处理这些重要问题时的考虑因素。

二. 原则和工作假设

德国确保负责任地将人工智能用于军事目的的做法是建立在各种国际论坛和讨论框架中确定的以下基本原则基础之上的。

德国在 2021 年积极参与制定北大西洋公约组织(北约)的负责任使用原则，并始终完全遵守这些重要标准：开发和使用人工智能应用程序须具有合法性；须维持人的责任，以确保在军事系统中人工智能设计和操作方面的问责；人工智能在军事领域的应用须具有可解释性和可追溯性；在配备人工智能和具有自主性的系统的整个生命周期中，保持可靠性、安全性、保障性和稳健性；以及确保进行适当的人机交互和减少偏见，以实行可管理性。

此外，德国还批准了 2023 年在海牙以及 2024 年在首尔举行的两次“军事领域负责任人工智能峰会”的成果文件(分别为《行动呼吁》和《行动蓝图》)，以及美利坚合众国 2023 年发起的《关于在军事上负责任地使用人工智能和自主技术的政治宣言》，并积极参与《宣言》的执行进程。

此外，德国也是“人工智能国防伙伴关系”倡议的一部分，在该倡议中，志同道合的国家促进负责任地使用人工智能，促进人工智能伦理实施的共同利益和最佳实践，建立促进合作的框架，并协调人工智能政策方面的战略信息活动。

2025 年 2 月，德国批准了《关于在配置人工智能的武器系统中保持人类控制的巴黎宣言》，其中强调在军事领域应用人工智能时保障人类控制的重要性。

三. 有关德国联邦国防军使用人工智能的关键问题

德国联邦国防军正在研究有无可能使用人工智能，来完成国防军的核心任务并在信息、决策和效率方面获得优势，并且优化行政和后勤流程以及对复杂系统进行预测性维护的相关流程。在军民早期危机检测方面，人工智能还被用于支持专家人员分析海量数据，并对部署进行预测。人工智能是重大国防项目的组成部分，这些项目也在欧洲范围内实施，帮助保持和促进欧洲的卓越技术。就国际军备领域各国发展情况和技术发展情况而言，人工智能有助于确保未来各国和盟国国防所需的能力。部署人工智能的可能性，特别是为保护国家安全和军事目的部署人工智能的可能性——这方面的发展工作是在相关部委和部门的职权范围和职责内进行的。在不影响上述内容的前提下，与安全相关的人工智能技术和人工智能应用已被纳入德国联邦政府的人工智能战略。

联邦国防军对人工智能在武器系统中的使用提出了最高的道德伦理要求，并制定了最高的法律标准。特别是，在这些系统的生命周期内，联邦国防军遵循国际人道法涉及武装冲突的规定，以及联邦政府数据伦理委员会和北约的指导方针，特别是上述关于将人工智能用于军事用途的六项负责任使用原则。

四. 基本考虑

为了保持必要的防御和威慑能力，德国仍然决心在军事领域抓住与人工智能有关的机遇，并且坚信不得阻碍技术进步，特别是考虑到相关技术固有的双重用途属性。

与此同时，德国将继续扩大知识库，评估和解决与在军事领域使用人工智能有关的风险，包括那些与非故意偏见(如基于性别的偏见)有关的风险。在这方面，德国高度重视学术界的重要作用以及在此领域工作的研究机构和智囊团所作的宝贵贡献。为促进相关研究，德国支持相关研究组织，包括联合国裁军研究所(裁研所)，为有目标导向的研究项目提供财政捐助。

确保讨论具有包容性，不仅要考虑地理范围，还要考虑会员国以及工业界、民间社会和学术界的意见——这一点对德国而言至关重要。

在处理与基于人工智能的武器系统相关的机会和风险时，德国特别重视人类控制的概念，并认为存在有效的人类控制框架是确保所有武器系统均符合国际人道法的必要条件。这不仅意味着技术控制，也暗含判断的成分。德国的“人类控制框架”概念包括一套技术上可行的步骤和行动，这些步骤和行动设定了允许该系统算法运行的明确界限。国际法，特别是国际人道法，是上述范围内的核心要素。当涉及到人工智能在战场上的实际使用时，背景至关重要。德国认为，人类控制框架概念是充分考虑到这一点的适当方式。

当人工智能的使用涉及核武器时，需要特别注意，因为相关科学和政治辩论仍处于早期阶段。人工智能在核武器指挥和控制系统中的可能使用可能对战略稳定或核升级产生严重影响。与此同时，人工智能可能为如何遏制大规模毁灭性武器的扩散和使用开辟新的途径。德国旨在推进这些辩论，于 2024 年 6 月 28 日在柏林主办人工智能与大规模毁灭性武器会议，作为德国颇为知名的“捕捉技术——反思军备控制”系列会议的一部分。

《禁止细菌(生物)和毒素武器的发展、生产及储存以及销毁这类武器的公约》和《关于禁止发展、生产、储存和使用化学武器及销毁此种武器的公约》禁止整类大规模毁灭性武器。诸如(生成式)大语言模型等应用程序可促进两用知识的扩散，此种知识可能被滥用于开发、生产或使用生物及化学武器。AlphaFold 等人工智能应用与合成生物学的融合可使恶意行为体设计出新型蛋白质，由于 DNA 序列的变化，这些蛋白质可以逃脱检测。人工智能可以分析大数据云(如人类基因组数据)，并对个体医学疗法的发展大有益处，但也可能被滥用于开发针对特定种族群体的生物武器。

因此，德国将与我们的国际合作伙伴密切合作，继续确定可能的行动方针，以评估人工智能应用对违禁武器开发和生产的影响，并出台可能的法规。与此同时，德国将利用人工智能的优势进行验证、生物取证和降低风险。

五. 德国对国际进程的承诺

自军事领域负责任的人工智能进程启动以来，德国一直积极推动此进程，并将继续这样做。德国是联合国大会关于军事领域的人工智能及其对国际和平与安全的影响的第 79/239 号决议的核心共同提案国之一。德国高度赞扬此项重要倡议的区域间和多利益攸关方方法，并期待 2025 年 9 月这一倡议在西班牙得以继续。

德国为美国的倡议“关于负责任地将人工智能和自主技术用于军事用途的政治宣言”作出了贡献，为之作出全面补充，包括(协同奥地利)共同主持监督工作组。

此外，德国积极参与人工智能国防伙伴关系，并参加了裁研所军事领域人工智能治理专家网络。

德国支持设在日内瓦的致命自主武器系统领域新兴技术问题政府专家组主席罗伯特·因登博斯大使，并继续积极参与此进程，包括在所谓的两级小组框架内协调若干会员国的立场。德国将与其国际伙伴密切合作，继续努力按时，最好是在 2025 年底前，完成政府专家组的任务。

在北约背景下，德国认识到人工智能对进一步发展武装部队和联盟防御能力的潜力，以及人工智能的使用将对联盟国家武装部队的互操作性构成的挑战。必须充分考虑北约、欧盟和德国伙伴国家的多国人工智能发展和人工智能标准化问题，以确保联邦国防军作为军事力量在国际行动中具备互操作性。因此，德国对北约国家在北约人工智能战略的背景下商定负责任使用原则表示欢迎。

六. 前进道路

新兴的颠覆性技术将继续发展并塑造我们的世界，德国认为，务必要就在军事领域负责任地发展和使用人工智能进行包容性的国际协调。现有的国际进程提供了很好的框架，以处理所涉及的有意义的方面，并考虑到各种利益攸关方的意见。德国将继续为这些努力作出积极贡献，以落实和扩大对将人工智能负责任地用于军事领域的政治承诺(如美国牵头的《政治宣言》或军事领域负责任的人工智能进程)的支持。德国期待着审查联合国秘书长就此提交的关于军事领域人工智能的报告的结果。德国将继续在日内瓦政府专家组的框架内为致命自主武器系统进程作出积极贡献。

希腊

[原件：英文]
[2025 年 4 月 10 日]

人工智能融入国防部门，从根本上影响了开展军事行动的手段和方法。人工智能应用于军事领域，带来了显著的作战效益，包括：提高决策速度、增强威胁检测和预测、实时态势感知和评估、优化资源分配和规划、后勤支持、增强人类完成复杂任务中的能力，以及有效处理大规模情报数据。

然而，尽管取得了这些进步，但必须认识到，技术进步也带来了复杂的多层面挑战，需要仔细审查，以确保它们不会破坏区域和全球的和平、安全与稳定。

在这方面，就希腊而言，一个关键的关注领域就是使用具有机器学习功能的军事系统，这带来了一些挑战——包括透明度和可解释性——因为复杂的模型可能会像“黑匣子”一样运行，决策过程不确定，尤其是考虑到战场环境千变万化。

此外，生成式人工智能在军事装备中的潜在应用还带来了一层重要的复杂性和不确定性，因为这些系统可能会自动生成新的解决方案，并通过不断分析和学习新数据能力来适应不断变化的战场条件——这些能力是希腊最为关注的。为应对这些挑战，务必要对其使用规定明确的操作界限和制约，以防止发生意外。

鉴于上述背景，在军事领域使用人工智能最令人担忧的挑战之一在于将其整合到使用核武器的指挥控制和决策支持系统中。将与核威慑有关的决策，甚至是启动使用核武器的相关规程，移交给人工智能系统 —— 这样的前景需要仔细思量，以确保人类监督和参与这些决策，并建立适当的网络安全保障措施，以防止意外升级。

同样令人担忧的是，在当前充满挑战的地缘政治环境中，各国努力保持军事优势——这种努力可能助长缺乏透明度和相互猜疑的军备竞赛。随着力量平衡被打破，先进国家与发展中国家之间的技术差距日益明显，此种竞争会加剧地缘政治的不稳定性，并对全球安全构成重大挑战。

此外，武装部队越来越多地开发和部署人工智能能力，有可能降低武装突的门槛。随着战场上人的因素越来越多地被无人系统所取代，决策速度的加快和战区对无人系统的日益依赖增加了意外升级的风险。

在这种情况下，另一个需要适当考虑的参数是人工智能能力向无视基于规则的国际秩序的国家以及向非国家行为体(包括恐怖组织)扩散和转移。随着人工智能技术越来越容易获取，这些行为体可能会获得并部署这些技术，以追求破坏稳定的目标，从而进一步挑战国际安全。

军事人工智能应用也带来了与心理战和错误信息相关的风险和挑战，因为它们能够大规模生产虚假信息，深度伪造和捏造数据，旨在欺骗公众和破坏机构稳定。自动账户(机器人)和有针对性的宣传算法加强了心理战，影响了公众舆论、选举进程，并造成了社会紧张局势，包括通过虚假宣传活动破坏民众对维和行动的信任。社会偏见，如与性别、年龄、种族和残疾有关的偏见，也会引起人们的担忧，因此必须实施风险评估和缓解措施，以防止算法中的意外偏见和歧视。

此外，人工智能应用于网络安全，既可以用来保护关键基础设施，也可用于恶意目的(如网络攻击和数据拦截)。混合威胁将传统军事行动与进攻性情报战术相结合，需要国家和国际行为体提高警惕和协调，以避免升级，维护区域和国际和平与安全。

鉴于上述情况，希腊强烈支持国际社会努力确保在军事领域负责任地使用人工智能，因为尽管存在上述挑战，人工智能仍可以加强国际人道法的实施，并通过提高目标定位准确性，加强监视和优化人道主义援助，来促进保护平民。

本着这一精神，2025年4月4日，希腊与法国和大韩民国一道，在亚美尼亚、意大利和荷兰王国的宝贵支持下，组织了题为“利用安全、包容、可信赖的人工智能，维护国际和平与安全”的安全理事会阿里亚模式会议。这次会议就联合国如何为维护国际和平与安全作出贡献提出了宝贵的见解，特别是通过以下方面：监管，不扩散、防止人工智能能力在军事领域的转移、加强法治，民主价值观，社会凝聚力和经济发展。

此外，作为其国际参与的一部分，希腊支持在海牙(2023年2月15日至16日)和首尔(2024年9月9日至10日)举行的两次“军事领域负责任人工智能峰会”上达成的关于采取行动促进在军事领域负责任地开发和使用人工智能的联合声明。希腊还批准了美利坚合众国《关于负责任地使用人工智能和自主技术的政治宣言》以及《关于在人工智能武器系统中保持人类控制的巴黎宣言》。

此外，希腊还成立了一个人工智能高级咨询委员会，[1] 以制定全面的国家人工智能战略，并在国防部内建立必要的结构，以应对军事领域内人工智能的应用和自主性所带来的技术、法律、道德和政治挑战。

最后(但并非最不重要的)一点是，为了建设性地促进关于在军事领域负责任地使用人工智能的国际对话，希腊正在组织题为“人工智能时代的武装冲突和危机管理”的国际会议，将于2025年5月22日和23日在雅典举行。

印度

[原件：英文]
[2025年4月1日]

人工智能是一种变革性的技术，它对人类生活的各个方面都产生了重大影响。它正在以前所未有的规模和速度发展，并迅速得到广泛应用和部署。人工智能可以对减少贫困和改善人民生活产生变革性影响。这对印度这样的发展中国家尤为重要。

需要在全球一级作出共同努力，建立人工智能的治理和标准，以维护我们共同的价值观、应对风险并建立信任。人工智能治理和标准应：考虑到跨境的深度相互依存；促进创新；为造福全球而部署，并应增加利用机会和公平，以确保所有方面(特别是全球南方国家)都能受益于人工智能。印度致力于公开讨论创新和治理问题。

[1] 委员会题为“希腊人工智能转型蓝图”的里程碑式研究提供了指导原则和旗舰项目，以推动希腊的人工智能进步，其优先事项包括保障和加强民主、减缓和适应气候以及支持安全。

关于军事人工智能的讨论需要立足于军事现实，因为人工智能正在迅速融入军事理论和行动。世界各地正在发生的冲突表明，越来越多地采用这些技术既带来了风险，也带来了机遇。

人工智能在军事领域的开发、部署和使用带来了伦理、法律和安全方面的挑战。在不淡化这些挑战的情况下，印度支持关于人工智能有可能改善国际人道法遵守情况的观点。

印度支持全球作出集体努力，适当规范人工智能在军事领域的开发、部署和使用。这些努力应处理法律和道德问题，并促进指明和减轻与军事领域人工智能相关的风险。

在军事领域适当监管人工智能的一切集体努力均应着眼于应用和使用，而不是技术及其组成部分。应避免对技术进行污名化。不得限制把技术用于发展的机会。

应根据国际法规定的单独或集体自卫的固有权利，在军事领域合法使用人工智能。国际人道法继续完全适用于军事领域的人工智能。国际人道法所载的区分、相称和审慎等基本原则适用于过去、现在和未来的所有作战手段和方法。

在军事领域使用人工智能时，人类的判断和监督对于降低风险和确保遵守国际人道法至关重要。

有关军事领域人工智能的任何集体努力或适当条例都应考虑到现有的法律义务，尊重国家管辖权和权限，以及相关的国家能力。

印度致力于军事领域负责任的人工智能。

印度正在制定评估国防部门可信赖人工智能的框架，以应对现代人工智能技术带来的复杂挑战。该框架以五项关键原则为中心：(a) 可靠性和稳健性；(b) 安全和保障；(c) 透明度；(d) 公平性；以及(e) 隐私性。以这些原则为基础，可以进一步讨论适当规范人工智能在军事领域的开发、部署和使用事宜。

印度尼西亚

[原件：英文]
[2025 年 4 月 11 日]

印度尼西亚欢迎根据大会第 79/239 号决议第 7 和第 8 段，讨论人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战，讨论过程中特别关注致命自主武器以外的领域。

由于军事领域的人工智能涵盖广泛的系统和应用，要在联合国就这一主题进行包容性的多边审议，这种讨论应该超越动能能力(如致命自主武器系统)，扩展到非动能能力——其中既包括对抗性应用(如自主网络战系统、自适应雷达干扰或电子战能力)，也包括支持性军事职能(如后勤、医疗后送或战术监视)。讨论还应涉及可能对战略平衡有直接影响的其他能力，如增强型传感(如卫星或反潜)、情报或战争规划。

印度尼西亚仍然坚定地致力于维护国际和平与安全，此立场已载入《印度尼西亚宪法》序言。基于这一承诺，印度尼西亚认为，军事领域使用人工智能，必须以促进和平、安全和可持续发展目标的方式进行管理。人工智能应成为和平与安全的促进力，而非不安全、冲突或战略竞争的催化剂。

虽然人工智能本身不是一种武器，但印度尼西亚认识到，它兼具力量倍增器和威胁放大器的双重属性，能够为国际和平与安全带来重大利益和严重风险。人工智能在军事领域的使用带来了各种伦理、法律、道德和技术问题，应从遵循国际法(包括国际人道法和国际人权法)的角度加以审慎考量。

一方面，人工智能被认为提供了广泛的潜力：它可以增强数据处理；提高操作效率、精度和准确性；并有可能改善国际人道法的遵守情况，例如支持相称性评估和预防措施，以减少对平民的伤害。人工智能还可以增强情报、监视和侦察能力，支持后勤和规划，并改善人员管理。

另一方面，人工智能带来了一系列风险和后果，包括可能助长军备竞赛、扩散到非国家行为体、促成犯罪和不负责任的滥用、通过技术优势加剧军事力量的不平衡，以及增加不稳定性、误判、升级和法律模糊性。技术风险还包括网络漏洞、系统故障、数据偏差、目标错误识别和其他操作不确定性。

印度尼西亚特别关切人工智能可能融入核指挥、控制和通信系统而带来关乎存亡的风险。印度尼西亚重申其原则立场，即使用和威胁使用核武器违反国际法，我们需要采取紧急和果断行动，维护和加强禁止核武器的准则。将人工智能引入核武器系统加剧了核武器使用(无论是有意、无意还是意外的使用)所构成的关乎存亡的风险，并增加了核危险。这是对所有国家安全的威胁。印度尼西亚敦促所有核武器国家重新评估其对核武器的依赖，并重申我们对无核武器世界的集体承诺。在彻底消除核武器之前，核武器国家必须在人工智能技术发展的背景下，对核武器及其运载系统保持切实的人力控制、责任心和问责制。

考虑到这些因素，印度尼西亚敦促采取预防措施，应对在军事领域使用人工智能的挑战。印度尼西亚强调，必须对人工智能在军事领域的开发、应用和使用进行管理，以利用其益处并降低其风险。这种治理必须服务于所有国家的集体和平、安全和繁荣。因此，印度尼西亚提出以下要点。

首先，印度尼西亚申明，在人工智能技术的整个生命周期中，都必须维护国际法。这包括《联合国宪章》、国际人道法、国际人权法以及裁军和不扩散条约。各国应当在从采购到评估的所有阶段进行法律审查。各国必须保证对人工智能在军事领域的发展和应用实行问责制，包括人工智能应用于战争或敌对行动的合法性。现在尚未制订规范人工智能在军事领域的使用的此类法律，因此，必须强调指出，人工智能的使用应遵循人道法和遵从公众良知。

在对军事领域使用人工智能的管理进行指导时，除国际法外，伦理方面的考虑也应作为法律框架的补充。在人工智能的开发和应用中，必须促进可追溯性、问责制、责任、可解释性、人道、透明、公平和公正等原则。

其次，印度尼西亚强调，在军事领域设计、开发、部署和使用人工智能方面，人的因素在确保各级(无论是国家、企业还是个人)实行问责制和责任制方面发挥着至关重要的作用。

人工智能在军事领域的开发、应用和使用必须始终以人为本，并以服务于人类利益为宗旨。必须通过培训，保持和加强有效和有意义的人类控制，特别是在涉及使用武力的决定方面。关键决策必须有人的判断、干预、监督和控制。此外，虽然“有意义的人类控制”在管理军事领域使用人工智能方面日益为人接受，但印度尼西亚认为，这一概念尚未满足与此类使用相关的法律、道德、技术和监管问题。需要就“有意义的”人类控制在实际操作中的含义达成一致。

人工智能治理虽然将主要规范国家行为，但也必须涉及民间利益攸关方，特别是参与军事领域使用人工智能的技术公司。各国必须确保私营部门遵守国际法律和道德标准，同时仍然支持人工智能创新生态系统的发展。研究人员和公司有责任确保其人工智能技术是可靠、安全、有保障、可问责的，并在实行问责制的人类控制之下。他们还应负责监测、沟通和解决其产品中的风险。

第三，印度尼西亚强调迫切需要多边、包容和全面的法律和监管治理框架。这必须反映所有国家的利益，无论其人工智能发展水平如何。所有国家都必须在制定军事领域使用人工智能的规则和规范方面拥有平等的发言权，以确保公平的代表性并促进全球信任。

鉴于人工智能涉及多方面的道德、法律和技术问题，广泛的利益攸关方参与至关重要。来自不同学科和文化的参与也是必要的，以确保人工智能系统须符合国际法、人道法、人权和裁军承诺，才可应用于军事领域应用。

第四，务必要保持知情状态，并促进对人工智能在军事中的开发、部署和使用所产生的风险、挑战和影响(无论其为技术性还是非技术性的)进行有意义的讨论。印度尼西亚强调，必须不断评估军事人工智能对国际和平与安全的更广泛影响，特别是在不扩散和裁军的背景下。需要进行更全面的研究，以了解这些影响；这些影响仍未得到充分的探索。

识别与人工智能在军事领域的开发、部署和使用相关的风险，将有助于循证预测、风险评估以及最终制定风险缓解措施。

加强对军事领域使用人工智能相关风险的理解和认识也至关重要。在这方面，除其他外，应分享国家政策和战略，特别是在查明、评估和减轻风险方面这样做，从而提高透明度；酌情在军事领域共享人工智能能力，以加强问责制和建立信任措施；以及跨国界、跨行业和跨部门分享经验教训和最佳做法。

第五，人工智能治理不应阻碍技术发展或限制发展中国家利用人工智能的机会。各种框架应避免强加限制公平利用的条件或障碍。需要采取一种平衡的方法，来解决扩散等风险，同时确保资源有限的国家能够利用人工智能。

最后，人工智能治理必须高度重视弥合数字和人工智能鸿沟。发展中国家不仅在人工智能能力方面，而且在有效管理这些技术的能力方面，都面临着重

大制约。如果这一差距得不到解决，全球治理工作将受到破坏，因为许多国家仍然不具备应对人工智能所带来的复杂、跨境挑战的能力。

印度尼西亚强调，迫切需要处理国家之间和国家内部存在严重的数字和人工智能鸿沟问题，特别是在获得财政、人力和技术资源方面。这些鸿沟有可能加深全球不平等，增加发生冲突的可能性。

作为全球公共产品，和平与安全需要所有国家(包括发达和发展中国家)之间开展国际合作，以应对共同挑战并获得集体利益，包括与人工智能在军事领域的开发、应用和使用有关的挑战。在此背景下，印度尼西亚呼吁加强和平衡国际合作与援助，以促进全球人工智能能力和治理框架。这种合作必须在平等和相互商定的基础上进行，同时考虑到发展中国家的具体需要和情况。这包括，但不限于，能力建设、教育、技术转让、终身学习、技术培训、联合研究和知识共享方面的举措。

这种合作必须是多层次的，不仅在国家和国际组织之间，而且在国家内部的各部门之间。应鼓励建立公私伙伴关系，以促进负责任的创新，并提高业界对其技术可能对国际和平与安全产生影响的认识。

国际合作不仅对于解决数字和人工智能鸿沟至关重要，而且对于为国家之间建立信任创造有利环境也至关重要。它有助于减少人工智能领域的地缘政治分歧和竞争。国际合作必须植根于平等、信任、互利、尊重主权和团结的原则，以便为有意义的合作(包括技术转让和知识共享)铺平道路。

印度尼西亚还认识到，需要加强那些考虑到地方和区域具体情况的区域合作机制。这些机制可以作为达成更广泛全球共识的基石，同时也为更细致和更敏感的审议留出空间。

伊朗伊斯兰共和国

[原件：英文]
[2025 年 3 月 12 日]

大会第 79/239 号决议第 7 段请联合国秘书长就人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战征求会员的意见，特别侧重于致命性自主武器以外的其他领域。伊朗伊斯兰共和国故此提交其意见如下：

人工智能正在成为当今世界变革的主要驱动力之一，为军事工业在不久的将来的发展方式留下不可磨灭的印记，从而从根本上影响国际和平与安全。国家行为体和非国家行为体积极推进其相互竞争的人工智能议程，不对这些议程加以监管是不行的。考虑到非国家行为体发挥主导作用，而且在监管和创新程序与趋势之间必须保持平衡，监管之权务必继续属于会员国主权特权。

从实质性的角度来看，就像对待网络空间和外层空间所用的其他技术一样，伊朗伊斯兰共和国支持人工智能完全用于和平目的，同时铭记，在适当情况下，军事实体亦可和平地受益于人工智能的红利。

鉴于各国发展水平不同，千万要确保数字鸿沟不致演变成为人工智能鸿沟。只有在联合国协商一致框架内，才能保证与人工智能相关的所有监管程序具有包容性。这种方法既维护了会员国的主权，为所有国家营造了公平的人工智能发展环境，也为人工智能行业的蓬勃发展提供了创新灵活性。联合国在人工智能相关监管事务中处于中心地位，可阻止对该问题采取本国排他性做法。在此至关重要的问题上，包容性和基于共识的办法必须占据主导地位。

尽管各种国际论坛都在讨论人工智能问题，但我们对此问题及其对国际和平与安全的影响的认识仍然不全面。断言国际法、人道法和国际人权法完全适用于人工智能，还为时过早。面对这一新的且在迅速演变的现象的巨大影响，国际法律框架可能需要调整和进行演变。

关于国际监管努力，伊朗伊斯兰共和国支持会员国之间作出具有法律约束力的安排(而非制定规范或政治文书)，并以此作为其首选行动方针。

伊朗伊斯兰共和国在其关于裁军的原则立场框架内，反对任何出于政治动机的歧视性、有条件做法或双重标准。因此，大会使用的术语必须反映一种团结和协商一致的意识。在这种情况下，“负责任的应用”等概念过于抽象，无法规范具体性和准确性很强的这一领域。这种抽象的概念会导致误解，并为政治化的做法打开大门。伊朗伊斯兰共和国表示强烈反对使用这种主观的术语，支持并提议在今后的任何文书中将“负责任的应用”改为“和平应用”。

以色列

[原件：英文]
[2025 年 4 月 10 日]

以色列注意到大会通过了第 79/239 号决议，谨根据该决议第 7 段，向秘书长将提交大会第八十届会议供会员国进一步讨论的报告提交本国意见。

以色列认为，目前人工智能的概念有一系列可能的解释，这些解释可能会随着时间的推移而得到完善。

显然，人工智能在军事领域的使用正变得比以往任何时候都更加普遍和频繁。以色列对上述大会决议投了赞成票，并鼓励各国和所有利益攸关方开展讨论，同时保持专业的非政治化性质，将所有国家的合理考虑，包括安全、人道主义、经济和发展因素考虑在内。

我们认为，为了就军事领域的人工智能进行最终可能取得切实成效的认真、负责任的讨论，必须采取务实且平衡的渐进办法。

技术为几乎所有领域，包括军事领域带来各种各样的机遇，因此我们欢迎探讨这些发展可带来的惠益及如何实现这些惠益，并探讨潜在风险及减轻风险的方法。以色列认为，人工智能技术等新兴技术还可有助于推动遵守现行国际人道法。这些潜在的机遇要求我们不应给此类技术贴上负面标签。

在关于人工智能军事用途的全球对话中，以色列始终发出建设性声音。最近，以色列核可了由美国带头提出的《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。我们期待着参加这一倡议今后举行的会议，并继续促进在军事上负责任地使用人工智能和自主能力。

作为该宣言的一部分以及在其他背景下，近年来各国均考虑在国内或国际层面为人工智能在军事领域的开发和使用提供指导。这些指导中一些更基本的共同原则，可能也与在大会第 79/239 号决议背景下的讨论有关；这些原则似乎是：

- 在军事上使用人工智能，必须符合适用的国际法。
- 此种使用应当是负责任的，并加强国际安全。
- 各国应确保根据适用的国际法对人工智能能力的使用承担责任，包括为此在负责任的、人类指挥和控制系统内使用此种能力。

各国为落实这些原则应当采取的实际措施包括：

- 各国应当采取法律审查等适当步骤，以确保将按照国际法、特别是国际人道法为其规定的各自义务使用本国的军事人工智能能力。
- 各国应当采取适当措施，以确保负责任地开发、部署和使用军事人工智能能力。这些措施应当在军事人工智能能力整个生命周期的相关阶段实施。
- 相关人员在开发、部署和使用军事人工智能能力、包括有此种能力的武器系统时，应当以适当谨慎的方式行事。
- 高级官员应当适当有效地监督以下军事人工智能能力的开发和部署，即其应用会带来严重后果的军事人工智能能力，其中包括但不限于有此种能力的武器系统。
- 各国应当支持做出适当努力，确保以负责任的合法方式使用军事人工智能能力，并继续与其他国家讨论如何以此种方式部署和使用军事人工智能能力。

以色列认识到以下方面的价值，即就军事领域的人工智能及其对国际安全的影响开展包容各方的多边讨论，从而在军事必要性与人道考量之间达成适当平衡。

意大利

[原件：英文]
[2025 年 4 月 11 日]

意大利担任七国集团主席国

在意大利于 2024 年担任七国集团主席国期间，人工智能被置于政治和技术讨论的核心。普利亚领导人峰会确认了人工智能对军事领域的影响，并确认有必要建立一个负责任的开发和使用框架。

2024 年 10 月 18 日至 20 日，首次七国集团防卫问题部长级会议在那不勒斯举行。七国集团国防部长在此次会议上重申，他们决心在这一极不稳定的历史时期，以协调一致的具体方式应对安全挑战。此外，各位国防部长强调指出，在防卫相关的研究和开发方面，包括在分享和利用专长及知识方面，需要采取更具协作性的办法，同时为防止恶意获取构建安全环境，以保持竞争优势，包括在新兴的颠覆性技术领域的优势。

最后，七国集团不扩散问题主管小组在声明中确认，人工智能等新兴颠覆性技术对军备控制、不扩散和裁军以及军事行动未来具有深远影响。

一. 军事领域负责任人工智能进程

意大利重视荷兰和大韩民国于 2023 年启动的军事领域负责任人工智能进程；该进程旨在为讨论人工智能军事应用所涉关键机遇、挑战和风险提供一个平台。2024 年，在首尔举行的第二届军事领域负责任人工智能峰会上，意大利核可了《行动蓝图》；这份文件概述了负责任的人工智能治理的各项关键原则，包括遵守国际法的重要性、人类责任与问责、人工智能系统的可靠性和可信度，以及人在军事领域人工智能的开发、部署和使用中的适当参与。

核可了该蓝图的国家强调指出，必须防止人工智能技术被用于助长大规模杀伤性武器扩散，并强调指出不破坏军备控制、裁军和不扩散努力的重要性。此外，为了就人工智能技术及其在军事领域的应用达成共识，该蓝图促请各国承诺开展进一步讨论，并制定有效的法律审查程序、建立信任和信心措施以及适当的降低风险措施。在该框架内，交流信息和良好做法，以及让其他利益攸关方积极参与，对于这一辩论取得进展至关重要。

二. 《关于在军事上负责任地使用人工智能和自主能力的政治宣言》

此外，意大利重视《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。核可该宣言的国家申明，人工智能的军事应用可以而且应当合乎伦理道德、负责任并加强国际安全，并确认在开发、部署和使用军事人工智能能力方面应当实施一系列措施。特别是，各国承诺开展以下工作，即最大限度地减少军事人工智能能力中的无意偏见；确保军事人工智能能力的安全性、保障性和有效性经过适当的严格测试；实施适当的保障措施，以识别和避免意外后果并在出现此种后果时予以有效应对。此外，必须界定负责任的、人类指挥和控制系统，而且军事人工智能能力的使用方式必须与国际义务相符。

三. 《未来契约》

2024 年 9 月，世界各国领导人通过了《未来契约》，重申他们的全球承诺，从而使各国能够应对新的和正在出现的挑战和机遇。在行动 27 中，鼓励各国把握人工智能等新兴技术带来的机遇，但与此同时应对滥用新兴技术所构成的潜在风险。特别是，会员国将与相关利益攸关方协商，继续评估人工智能军事应用中的此类风险，以及在整个生命周期中可能出现的机遇。

四. 《关于在人工智能赋能的武器系统中保持人类控制的巴黎宣言》

意大利最近还核可了《关于在人工智能赋能的武器系统中保持人类控制的巴黎宣言》，该宣言在 2025 年 2 月 6 日至 11 日在巴黎举行的人工智能行动峰会期间获得通过。核可国特别指出，决不能将责任与问责转移给机器，并承诺采用以人为本的方法来开发、部署和使用军事领域的人工智能应用。此外，这些国家承诺确保人工智能在军事部门的部署完全符合国际法和国际人道法，同时促进人工智能技术的研究、开发和创新。

五. 致命自主武器系统领域新兴技术问题政府专家组

人工智能和机器学习的快速发展，也对自主能力在武器系统中的作用产生了重大影响。意大利认为，《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》汇集了各国政府、国际组织和专门机构代表的外交、法律和军事专门知识，迄今为止是处理与武器系统的开发和使用有关的现有问题和新问题的最适当论坛。意大利积极推动在《特定常规武器公约》主持下发起的、致命自主武器系统领域新兴技术问题政府专家组讨论，并致力于根据 2023 年该公约缔约国会议上商定的任务，推动关于拟订未来文书要素的讨论。

意大利认为，这项未来文书应当载有明确禁令和规定，以便最终作为《特定常规武器公约》的附加议定书获得通过。根据这一方针，任何无法依据国际人道法开发和使用的致命自主武器系统，将基于这一事实本身而被禁止。另一方面，可在全面遵守国际人道法的前提下开发和使用、但在关键功能上有决策自主能力的系统将受到监管。意大利认为，事实上，人的因素对致命自主武器系统整个生命周期，即设计、开发、生产、部署和使用都至关重要。此外，应当保留适当程度的人类判断和控制，以确保根据国际人道法承担责任和追责。

日本

[原件：英文]
[2025 年 4 月 11 日]

在第 79/239 号决议中，大会请秘书长就人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战征求会员国和观察员国的意见，特别侧重于致命性自主武器系统以外的其他领域，并提交一份实质性报告，概述这些意见，编目现有和新出现的规范性提议，并在附件中载列这些意见，提交大会第八十届会议，供各国进一步讨论。日本谨就该主题提出以下意见，以便为编写该报告和推动关于该主题的讨论作出贡献。

一. 一般意见

日本致力于维护和加强以法治为基础的、自由开放的国际秩序，从而使所有人都能享有和平、稳定和繁荣，并致力于促进外交，以创建一个人人享有尊严、安全且稳固的世界。根据这些目标，日本积极参与旨在加强国际和平与安全以及推动军备控制和裁军的各项努力。

日本认为，应当对军事领域的人工智能应用进行全面审查，充分了解相关风险和惠益，兼顾人道考量和安全视角。以下做法是有益的，即更深入地了解人工智能在军事领域的应用，促进为负责任的使用而做出切实可行努力，以便最大限度地利用人工智能的惠益并降低风险。

关于人工智能在军事领域的应用，日本支持以下观点：首先，现行国际法适用于人工智能整个生命周期中发生的、受国际法管辖的事项；其次，应当以负责任的方式应用人工智能能力；第三，人始终对人工智能的使用及其影响负责并接受问责。此外，日本强调，有必要提高透明度并将此作为一项重要的建立信任措施，推动在减少风险的同时实现惠益最大化。

二. 日本关于人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战的意见及相关办法

机遇

意见

科学技术、包括人工智能快速发展，正在从根本上改变安全的范式。各国正在努力开发那些可能极大地改变战争性质的尖端技术，进而证明自己是“游戏规则改变者”，而且在实践中，区分民用技术和安保用途的技术已变得极为困难。人工智能具有改变军事事务各个方面的极大潜力，这些方面包括军事行动、指挥和控制、情报及监视和侦察活动、训练、信息管理和后勤支援。考虑到人工智能在军事领域的各种用途，人工智能应用可能带来的惠益包括提高精度、准确性和效率，增强态势感知和认识，促进快速的信息分析，减少人为错误和节省劳动力。妥善适用人工智能，有助于更好地在冲突中保护平民和开展冲突后建设和平工作。

日本利用“机遇”的办法

在军事领域应用人工智能时，需要考虑此类应用能否有效克服人类发现的问题，与此同时牢记人工智能的功能及其局限性。人工智能应用本身不应成为目标，不应仅考虑应用本身而不考虑其功能和限制。因此，各国应当确保军事人工智能能力有清晰、明确的用途，并确保设计和工程工作是以实现这些预期功能为目标。有鉴于此，必须推动国际社会就以下方面达成共识，即人工智能、人工智能在军事领域的功能及其局限性以及人工智能在军事领域的潜在应用。关于防卫当局对人工智能的应用，日本防卫省于 2024 年 7 月发布了《防卫省人工智能活用推进基本方针》，其中阐述了防卫省目前关于人工智能在军事领域的功能及其局限性以及人工智能应用的优先领域的想法。在该基本方针中，考虑

到人工智能的现有能力及其局限性，防卫省确定以下七个领域为人工智能应用的侧重领域：

- 目标探测和识别
- 情报收集与分析
- 指挥与控制
- 后勤支助行动
- 无人载具
- 网络安全
- 提高行政工作效率

此外，该基本方针指出，必须牢记应用人工智能是为了支持人类决策，并牢记在应用人工智能时，人的参与至关重要。

挑战

意见

军事领域的人工智能应用可能存在滥用或恶意使用的风险，以及冲突升级和冲突门槛降低的风险，其原因可能是偏见、意外后果和其他因素。在这方面，日本强调指出，必须防止国家和非国家行为体利用人工智能助长大规模毁灭性武器的扩散，并强调人工智应当支持而不是阻碍裁军、军备控制和不扩散努力。

日本应对“挑战”的办法

考虑到偏见、滥用和恶意使用等风险，日本防卫省将努力降低人工智能带来的风险，并以 2024 年 4 月发布的《日本人工智能企业指导方针》所载以下构想为参考，即以人为本的人工智能、安全、公平、隐私保护以及确保安全、透明度和问责制，与此同时关注国际社会的讨论以及与其他国家防卫当局的讨论。

此外，日本正密切关注人工智能等新兴技术可能对核裁军和不扩散产生的影响。在这方面，日本欢迎美国、联合王国和法国在不扩散核武器条约缔约国 2022 年审议大会上作出的承诺，即保持对通报和执行有关核武器使用的主权决定至关重要的所有行动的人为控制和参与，并促请其他核武器国家效仿。此外，无核武器世界国际知名人士小组在向 2026 年审议大会提出的建议中强调指出，需要共同努力，以应对与新兴技术相关的挑战并把握相关机遇。

三. 关于今后的讨论和国际合作的意见

灵活且平衡的务实办法对于军事领域的人工智能的治理是必要的，以便跟上技术的快速发展和进步。日本强调指出，为推进军事领域负责任人工智能所作努力，可以与推动人工智能技术的研究、开发、实验和创新的努力同时进行，并不会妨碍这些努力。

应当指出，如果在讨论中把特定人工智能技术视作问题，就可能导致民用部门的技术发展和创新受阻，而且可能产生寒蝉效应。此外，应当以包容方式讨论军事领域的人工智能应用，让利益攸关方参与并促进他们之间的交流。

考虑到上述方面，日本坚决支持军事领域负责任人工智能峰会的成果和《关于在军事上负责任地使用人工智能和自主能力的政治宣言》，并期待更多国家加入这些倡议。

关于致命自主武器系统，应当指出，日本坚决支持继续在《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》下进行讨论，并期待关于军事领域人工智能应用的讨论将补充并加强根据该公约设立的致命自主武器系统领域新兴技术问题政府专家组的讨论。

日本确认，人工智能在军事领域应用的透明度是一项重要的建立信任措施，有助于降低风险并推动各国之间的有效协作与合作。日本还确认，能力建设对于促进以负责任办法进行军事领域人工智能的开发、部署和使用发挥重要作用，并致力于加强能力建设方面的国际合作，目标是缩小有关此种办法的知识差距。在这方面，不妨采用良好做法和经验教训交流等方法，日本也将利用各种机会与其他国家交流意见。

最后，关于人工智能在军事领域的应用，日本将继续以积极的建设性态度参与国际讨论，目标是通过兼顾人道考量和安全视角的平衡讨论，在国际社会达成共识。

立陶宛

[原件：英文]
[2025 年 4 月 9 日]

立陶宛感谢有机会根据大会第 79/239 号决议提交意见，供纳入秘书长报告。立陶宛很高兴支持大会于 2024 年 12 月 24 日通过的这项决议。

立陶宛指出，人工智能在军事领域的发展和使用为国际和平与安全既带来机遇，也带来挑战。立陶宛非常重视制定关于负责任使用的各项规范和原则，这将使各国能够利用军事领域人工智能的惠益并减轻潜在风险。立陶宛坚信，确保在军事领域负责任地应用人工智能，符合所有负责任国家的利益。立陶宛深信，若要应对军事人工智能的影响，就需要采取全球行动和多利益攸关方办法，让公共和私营部门、民间社会和学术界参与。

立陶宛高度支持《关于在军事上负责任地使用人工智能和自主能力的政治宣言》，并于 2023 年 11 月 13 日加入该宣言。该政治宣言载有不具法律约束力的原则和最佳做法，推动确保在军事背景下负责任且合法地使用人工智能。该政治宣言考虑了各项措施，例如法律审查、适当监督、最大限度地减少无意偏见以及确保军事人工智能能力有清晰、明确的用例。立陶宛大力鼓励更多国家签署这项《政治宣言》。

此外，立陶宛赞同2021年通过并于2024年修订的《北大西洋公约组织人工智能战略》。该战略为军事领域人工智能的负责任使用制定了六项原则，即合法性、责任与问责、可解释性和可追溯性、可靠性、治理能力和减少偏见。这些不具法律约束力的原则旨在适用于人工智能应用的全生命周期，立陶宛承诺遵守这些原则。

最后，立陶宛谨阐述关于人工智能在军事领域的应用给国际安全带来的机遇和挑战的看法。立陶宛认为，可以而且应当负责任地使用军事人工智能，首先是用于加强国家安全，以及用于促进执行国际法、包括国际人道法，推动履行国家保护平民的各项义务。除了加强武装冲突中的平民保护，负责任的人工智能还为改善决策、后勤、规划和其他提高效率的行动带来机会。

关于人工智能在军事领域的潜在风险，立陶宛强调指出包括但不限于以下方面的挑战，即网络安全、军事人工智能能力的无意偏见以及人工智能赋能系统的意外行为。立陶宛认为，消除这些潜在风险的最佳方法是实施负责任使用的原则，开展能力建设，就人工智能应用程序以及人工智能赋能系统的使用进行适当的人员培训。立陶宛强调，为了在军事领域利用人工智能的惠益，并将人工智能用作关键的防御能力，各国应当避免施加不必要的过度限制，以免阻碍人工智能创新；如果不负责任的国家拒绝接受对军事人工智能的任何此类限制，就更应当避免这样做。

墨西哥

[原件：西班牙文]
[2025年4月10日]

人工智能、自主武器系统以及世界在其监管方面面临的挑战

墨西哥根据题为“军事领域的人工智能及其对国际和平与安全的影响”的大会第79/239号决议提交本文件。

墨西哥确认人工智能在军事领域的应用可带来惠益。但人工智能也对国际和平与安全构成重大挑战，因此需要国际社会对此给予协调一致的紧急关注。

墨西哥重视在联合国框架内开展的多边交流，例如由裁军事务厅协调的题为“机遇、风险以及国际和平与安全”的首次军事人工智能、和平与安全对话；此次对话有助于就新出现的风险和共同责任达成共识。我方也认为，将人工智能融入军事职能对国际和平与安全构成根本性挑战，其中包括冲突意外升级、战略模糊以及武力使用方面的自主性不断提高。

墨西哥认为，应当优先考虑以下方面：加强国际合作；提高透明度；分享良好做法并建设各项能力，以支持遵守规章和尊重国际法的文化；推动制定监管框架，以确保军事背景下的人工智能开发和部署以伦理道德、法律和人道主义原则为准绳，同时防止该技术导致不对称加剧或削弱国际稳定。

国际和平与安全

墨西哥认为，必须采取行动，防止这些技术的扩散和滥用，包括非国家行为体的滥用以及在明确法律框架之外的滥用。

新技术和新兴技术在军事领域的融入，决不能凌驾于国际和平与安全之上。此种融入必须考虑到人类发展和社会赋权，特别是为发展中国家造福。因此，此类技术的目标，应当以和平利用和争端解决为导向，而非致力于打造更有效的军事机器。

以下各方面凸显了积极主动地降低风险的必要性，即数字威胁越来越复杂，新兴技术可能被用作实施国与国之间攻击的手段，在军事环境下难以确保自主系统的可靠性和准确性，人工智能全生命周期各种漏洞的风险敞口、算法偏见、数据投毒以及生成式模型被用于恶意目的。

科学技术的进步，尤其是人工智能、自主系统以及数字和量子技术领域的进步，超出了监管框架管控上述风险的现有能力。因此，墨西哥重申需要制定全面的治理框架，促进国际合作和多边对话，并在这些技术的全生命周期优先关注透明度、问责制和人类实际控制，包括对技术部署实施严格测试和采取道德保障措施。

由于缺乏明确的国际法律框架和必要的多边共识，“负责任”一词在此背景下的使用不应被解释为对使用或开发人工智能赋能的自主军事能力的默许或接受。责任原则必然与合法性和问责制相联系。

在这方面，墨西哥认为必须建立治理和监管机制，以降低人工智能和其他颠覆性技术被用于敌对目的的可能性，同时认识到不仅在行动部署期间存在风险，而且在设计和开发的初始阶段就存在风险。

行动环境

墨西哥注意到，人工智能技术可被融入不同的军事行动环境，因此可产生不同影响。

在武装冲突背景下，必须确保所有基于人工智能的技术的使用均符合国际人道法，特别是区分、相称、预防和人道原则。

在维和行动和应灾领域，人工智能可积极推动后勤协调、风险预测和受灾民众的照护工作，但前提是必须充分尊重人权框架。

关于边境安全，墨西哥确认人工智能可加强监控能力，但强调指出必须确保尊重所有人的尊严，避免自动化决策延续歧视性做法。

致命自主武器系统

墨西哥认为，此次讨论的一个关键层面是致命自主武器系统，该系统是国际和平与安全领域令人尤为关切的事项。在这方面，墨西哥强调指出，关于把

新技术融入军事领域的多边讨论不应各自为政，并认为致命自主武器系统应当成为这些交流中不可或缺的一部分。

墨西哥认为，致命自主武器系统不符合国际人道法，而且存在伦理道德、法律和安全风险，因此国际社会亟需制定相关的明确禁令和管制措施。

墨西哥在大会推动并共同提出了关于致命自主武器系统的第 78/241 和 79/62 号决议，旨在为应对这些挑战创建一个合法的多边论坛。

墨西哥支持秘书长和红十字国际委员会的呼吁，即按照《新和平纲领》的要求，启动关于到 2026 年缔结一项具有法律约束力的文书的谈判，制定关于致命自主武器系统的必要禁令和管制措施。

通过参加圣何塞会议(2023 年)、遵守《贝伦公报》、积极参与“人类处于十字路口：自主武器系统及其监管的挑战”会议(2024 年，维也纳)并核可会议的成果报告，墨西哥已表明本国在这一问题上的政治承诺。

墨西哥认为致命自主武器系统构成多种风险，其中包括：

- 在关于使用武力的关键决策中将人的判断排除在外。
- 取代了军事行动中必不可少的背景评估。
- 削弱了问责和责任归属机制。

使用武力的责任绝不应转移给机器。关于武装系统的部署、启动或人工超控的决定必须始终由人做出，并由人承担相应的法律责任。

墨西哥重申所有军事技术，包括基于人工智能的军事技术，必须遵守源自以下方面的国际义务：

- 《联合国宪章》
- 国际人道法
- 国际人权法
- 国际刑法
- 国际责任法

在这方面，墨西哥认为禁止那些采用以下技术的武器系统至关重要：

- 无法区分军事目标与民用目标
- 无法将相称原则适用于附带损害
- 在认定攻击不必要时，没有解除机制
- 造成不必要痛苦或过度伤害。

墨西哥坚持认为迫切需要启动关于一项具有法律约束力的文书的谈判；这项文书将制定关于致命自主武器系统的具体禁令和管制措施；确保对关键活动保持有意义的人类控制；载有有效落实、监测和问责的机制。

惠益与风险

关于人工智能的具体使用，墨西哥确认以下领域的惠益与风险：

- 指挥与控制：在某些条件下，人工智能可提高作战决策的效率，但这些决策必须始终处于人类的严密控制之下，在涉及使用武力之时尤其如此。人工智能具有处理和分析海量数据和信息的能力，远超人类的能力，因此可以加快、推动和简化关于未来趋势的预测并为战略决策提供实时信息。

- 网络行动：人工智能可为预测和应对网络事件提供宝贵能力，但也会增加紧张局势不断升级的风险，包括在无适当监督情况下的自动化进攻性使用。

- 信息管理和后勤：使用人工智能处理海量数据可方便实时决策，但必须依据各项规程来开展这项工作，确保人工智能的使用合乎伦理道德、可解释和负责任。

尽管如此，墨西哥特别指出将人工智能融入军事领域所涉及的技术风险，因为有证据表明，可能导致冲突升级的技术故障或不可预见的错误持续存在。

荷兰王国

[原件：英文]
[2025 年 4 月 7 日]

荷兰王国欢迎有机会根据大会 2024 年 12 月 24 日通过的第 79/239 号决议，就军事领域的人工智能给国际和平与安全带来的挑战和机遇提出意见。

荷兰确认人工智能的潜在军事应用，并致力于对军事领域的人工智能进行负责任的开发、部署和使用。荷兰的基本立场是，人工智能在军事领域的应用必须符合国际法，包括《联合国宪章》、国际人道法和国际人权法。

2023 年 2 月 15 日和 16 日，荷兰主办了首届军事领域负责任人工智能峰会。自那时起，军事领域负责任人工智能进程为政府、知识机构、智库、行业和民间社会组织的代表提供了一个多利益攸关方平台，以讨论与人工智能的军事应用有关的重大机遇和挑战。全球一级的讨论每年举行，但在军事领域负责任人工智能进程的区域活动期间，全年都有讨论；迄今为止，新加坡、肯尼亚、土耳其、智利和荷兰主办了上述区域活动。

在 2023 年峰会上，荷兰和其他 57 个国家商定了关于在军事领域负责任地开发、部署和使用人工智能的联合行动呼吁。2024 年，荷兰核可了在由大韩民国主办、荷兰共同主办的 2024 年军事领域负责任人工智能峰会上商定的《行动蓝

图》。此外，荷兰核可了《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。

在 2023 年峰会期间，荷兰启动了军事领域负责任人工智能全球委员会，并责成其为各国政府和更广泛的多利益攸关方社区提出短期及长期建议。荷兰正在等待将于 2025 年 9 月发布的委员会战略指导报告。

下一节进一步概述荷兰的立场，载列需要进一步审议的关键问题。

促进国际和平与安全的机会

从军事角度来看，人工智能的主要惠益是速度和规模。人工智能技术使数据处理和分析速度大大加快。此外，人工智能驱动的场景开发和决策支持系统，有助于指挥官制定行动方针，从而提高战略洞察力和迅速有效地应对威胁的能力。

荷兰认为，人工智能还可通过提升洞察力、改善连通性、加强对平民的保护和减少前线行动的风险，为国际和平与安全作出贡献：

- 人工智能驱动的分析和决策支持系统，可增强指挥官掌控和运用信息的主动性，无论在实地局势和长期战略发展方面均如此。这有助于更好地了解冲突区的平民动态、气候安全挑战、性别暴力状况和恐怖主义组织的行为模式。这些信息进而又可用于改善风险和冲突管理，从而促进国际和平与安全。

- 荷兰认识到人工智能在军事领域的应用的价值，这种应用可以改善国防部队之间以及国防部队与人道主义援助行为体、监测组织和地方政府等其他行为体之间的连通性。数据可以在大量用户之间交换，从而利用在安全的网络环境中运行的“智能”传感器创建“单一真实数据源”。此外，人工智能体可被用于以越来越高的速度共享数据。更完善且速度更快的数据共享使连通性得到改善，有利于国际和平与安全，原因是沟通、信息共享和国际合作得到加强，例如在预警系统和危机管理方面。

- 荷兰高度重视人工智能在保护平民方面的潜力。人工智能可以从海量数据中识别出模式和偏差，从而促进更全面地了解平民所处环境。这种更深入的了解有助于减少错标、附带损害和平民伤亡的风险。更广泛而言，人工智能有可能更好地识别对平民和民用物体的潜在威胁，使武装部队能迅速地作出适当反应。此外，人工智能可帮助优化人道主义援助工作，例如在冲突地区提供食品、庇护所和医疗服务。最后，人工智能可以通过收集和分析数据和证据来改善对平民伤亡状况的调查，以确定造成伤害的原因，并确保责任人能被追责。

- 人工智能降低了前线军事人员的风险，原因是在复杂或危险地形，人工智能驱动的自主系统可在某些活动中取代人。例如，水下监视以及在极端天气条件下支持搜救行动。人工智能还可以通过减少军事人员在高风险环境中的暴露来帮助降低医疗和康复成本。

国际和平与安全的挑战

荷兰确定了军事领域的人工智能应用给国际和平与安全带来的各种风险：

- 荷兰感到关切的是，人工智能可被用于扩大和加强网络攻击及信息操纵并使此种攻击和操纵自动化，这两种行为都破坏国际和平与安全。随着生成式人工智能的兴起，信息操纵和自动化网络攻击都更易实施。如果运用在军事领域，它们会扰乱作战通信线路，增加决策难度。从长远来看，大范围的虚假信息传播和自动化网络攻击，可能会削弱对军事通信线路的信任。上述行为还可能影响国家之间的信任，从而可能破坏脆弱的关系，特别是破坏那些已处于潜在冲突边缘的国家之间的关系。

- 人工智能在军事领域的应用存在相关风险，可能导致开发出有可能违反国际法的系统。存在这些不足之处，可能是由于对环境、数据和军事术语的适应不足，进而导致军事决策过于简单化或特定行动环境被忽视等。此外，如果某个应用程序的表现不可预测，基于不相关特征而产生歧视性结果，或提议采取非法的行动方针，那么国家就可能违反国际法律义务。人工智能日益普及，因此自动化偏见、数据集偏见和基于不适当人工智能系统的人类决策所产生的影响，可能给厘清责任、确保追责和采取适当补救措施带来重大挑战。重要的是，不能指望人工智能应用程序采用与人类相同的方式进行推理或运作。

- 可能会出现人工智能驱动的局势升级，从而对国际和平与安全构成潜在风险。人工智能增强了速度和规模方面的能力，使“观察－判断－决策－行动”环加速，因此可能会因为军事意图与人工智能驱动系统所做分析之间的差异而产生误解。所以，人工智能可能在无意中导致局势升级。与人类相比，人工智能系统有能力在更大范围内更快地识别可能的目标，因此其使用也可能导致冲突强度更高、更致命。

- 因此，建立强大的防御系统是一个越来越重大的挑战。新的人工智能应用正在快速涌现，导致难以在军事环境中实施予以有效对抗和防御的战略和战术。人工智能系统使用不断增加所引发的这一具体后果，可能对进攻行动有利，从而对国际和平与安全产生不利影响。

- 随着恐怖主义组织、有组织犯罪网络和其他非国家行为体获得军事人工智能能力，破坏稳定成为另一令人关切的问题。在此方面，荷兰感到关切的是，人工智能可能使这些行为体制造化学、生物、放射性和核武器的难度降低。

考虑到人工智能技术的快速演变，荷兰承认目前尚无法完全预见国际和平与安全领域的挑战和机遇。有些挑战是全新的，而另一些已经存在，但可能会因人工智能的应用而加剧。为了确保所有国家在军事领域负责任地应用人工智能，必须就这一问题开展持续的国际对话。

在军事领域负责任地应用人工智能

为了确保在军事领域负责任地应用人工智能，必须保持适合具体情况的人类判断和控制。人类必须始终承担责任并接受问责。但是，必须注意以下几点。

更多的人类控制并不能确保人工智能更负责任

荷兰认为，对于在人工智能应用中纳入充分的人类判断和控制，并没有一刀切的方法。人类判断和控制可以是直接的人工控制，也可以是较高水平的自动化和自主能力，具体情况如何取决于许多因素。因此，需要在多大程度上对人工智能驱动的应用和系统实施人类判断和控制，须逐案决定。只有这样，才能将以下多种因素考虑在内，例如操作环境，对技术在不利环境中的自主操作能力的影响、系统参数和人机交互。

研究和开发对于在军事领域负责任地部署人工智能应用至关重要

荷兰相信研究和开发具有重要作用。各国必须充分评估本国的人工智能应用是否按照设计的方式运行，以及能否在特定使用环境中部署。上述方面在战斗和其他高风险环境中尤为必要。通过开展一般意义上的研究和开发，以及对具体的人工智能应用适用公认且可靠的测试、评估、核对和验证程序，就可以在部署之前发现并消除或缓解潜在问题。此外，在部署人工智能应用之前，必须让军事人员接受适当培训，使其熟悉这些应用，以确保他们了解应用程序的功能及其局限性。考虑到人工智能应用的快速技术发展，以及这些应用的使用成本不断降低，这一点尤为重要。

军事人工智能的国际治理应当灵活、包容和切合实际

关于军事领域人工智能的国际治理，荷兰确认需要采取灵活且平衡的务实办法。首先，治理框架需要灵活，以便跟上快速的技术发展和战场态势。其次，各方需要努力就以下方面达成共识，即军事领域的人工智能以及随之而来的机遇、风险和可能的解决方案。若要达成共识，就需要开展包容各方的全球对话，并需要让国家、知识机构、民间社会和行业等所有利益攸关方群体积极参与。第三，各国应重点关注为在军事领域负责任地应用人工智能建立保障措施，例如，重点关注确保可追溯性或可理解性等问题。第四，军事人工智能部署的国际治理必须将各国关于监管问题的不同意见都考虑在内。在现有法律义务的范围内，军事领域人工智能的国际治理不应妨碍各国的创新能力。

关于自主武器系统的讨论

人工智能在操作自主武器系统方面具有极大潜力，因此关于人工智能在军事领域的使用的更广泛讨论与关于自主武器系统监管的讨论之间有明显的相似之处。荷兰认为，关于这两个专题的国际讨论相辅相成、互惠互利。

新西兰

[原件：英文]
[2025 年 4 月 11 日]

新西兰提交的这份国家文件是对 2025 年 2 月 12 日裁军事务厅普通照会的回复，应与新西兰对 2024 年 2 月 1 日裁军厅普通照会的回复一并阅读。[1]

新西兰关于军事领域人工智能的立场

新西兰确认，人工智能在军事领域的潜在和现有应用将产生多方面的深远影响。

迄今为止，一些军事组织已将人工智能应用于情报、规划、后勤、导航和通信等广泛的军事职能方面，虽然我们尚不清楚其中许多影响的性质及程度如何。军事领域的人工智能有一定风险，但也可以给使用者带来巨大优势，例如提高速度、效率和准确性以及增强态势感知能力。同其他军队一样，新西兰国防军打算利用人工智能提供的机会，推动改善行动并保持与合作伙伴之间的有效协作。

新西兰重申大会第 79/239 号决议第一段，即“国际法，包括《联合国宪章》、国际人道法和国际人权法，适用于人工智能能力及其在军事领域启用的系统的整个生命周期中发生的受其管辖的事项”。除了具有约束力的法律义务外，还应当在军事领域人工智能的整个生命周期中考虑相关的伦理道德标准。

新西兰确认，人工智能与某些武器系统的开发和使用有关，例如在提高自主能力的水平方面。新西兰关于自主武器系统问题的立场详见新西兰对 2024 年 2 月 1 日裁军事务厅普通照会的回复。

不难想象，人工智能可用来开发大规模杀伤性武器。国际法明确禁止生物武器和化学武器；新西兰申明，若要将人工智能用于开发此类武器，那么《禁止细菌(生物)和毒素武器的发展、生产及储存以及销毁这类武器的公约》和《关于禁止发展、生产、储存和使用化学武器及销毁此种武器的公约》所载通用标准均将适用；除其他外，这意味着不得将人工智能用于此目的。此外，正如包括新西兰在内的《禁止核武器条约》缔约国所指出的那样，在消除核武器和实现无核武器世界之前，必须对核武器及其运载系统保持有意义的人类控制。

现有和新提出的规范性建议

达成共识和建立规范是促进在军事上负责任地使用人工智能的重要方面。2024 年，新西兰与其他许多国家一道，加入了美国带头提出的《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。该宣言申明，“人工智能的军

[1] 可查阅 www.mfat.govt.nz/assets/Peace-Rights-and-Security/Disarmament/New-Zealand-submission-to-the-UN-Secretary-General-on-autonomous-weapon-systems.pdf。

事应用可以而且应当合乎伦理道德、负责任并加强国际安全”。此外，新西兰参加了军事领域负责任人工智能峰会。

新西兰认识到以下方面的重要性，即为拟订和商定关于军事领域人工智能的规范而专门开展多边讨论，包括通过联合国开展此种讨论。在整个进程中，让非国家利益攸关方，包括民间社会、国际和区域组织以及业界参与这些讨论具有重要意义。

挪威

[原件：英文]
[2025 年 4 月 11 日]

挪威欢迎有机会根据大会题为“军事领域的人工智能及其对国际和平与安全的影响”的第 79/239 号决议，就军事领域的人工智能应用给国际和平与安全带来的机遇和挑战提出意见，特别侧重于致命性自主武器系统以外的其他领域。

如秘书长在 2023 年 7 月政策简报《新和平纲领》中确认，人工智能既是一种赋能技术，也是一种颠覆性技术，正越来越多地用于各种民用、军用和军民两用领域。人工智能越来越普遍，加上其快速可扩展性、缺乏透明度和飞快的创新步伐，给国际和平与安全带来了潜在风险，并构成治理挑战。

挪威在国防领域一贯倡导国际法、多边主义和负责任创新，支持在军事领域促进共识、加强治理和发展对人工智能的适当监管。一个最低起点是，军事领域的人工智能应用在整个生命周期必须以负责任的方式予以发展、部署和应用，并遵守适用的国际法，特别是国际人道法。

重要的是，大会在第 79/239 号决议中申明在军事领域使用人工智能适用国际法，包括《联合国宪章》、国际人道法和人权法，并强调了以人为本、负责任地使用人工智能的重要性。

人工智能是一种赋能技术，具有改变军事事务各个方面的非凡潜力，这些方面包括采购、硬件、软件、行动、指挥与控制、战略沟通、监测、情报、培训、信息管理和后勤支援。军事领域的人工智能应用在战术和战略层面均带来了可预见和不可预见的机遇和风险。作为一种通用技术，人工智能是战力倍增手段，有能力重塑战争的进行方式。人工智能、神经技术、合成生物学和量子计算之间的技术融合进一步增加了这方面的复杂性。

对人工智能的发展、部署、使用和管理必须负责任，并符合基本的伦理原则，严格遵守各国根据国际法、包括国际人道法和人权法承担的义务，而这方面最核心的是识别和减少风险。

《挪威国防部门人工智能战略》(2023 年)概述了人工智能可能对致命自主武器系统以外领域做出建设性贡献的关键领域：

- **增强态势感知和决策支持**。在情报、监测和侦察中使用人工智能既有可能，也有必要，原因是数量庞大且不断增加的数据无法通过人工进

行分析。人工智能可用于滤除相关数据(例如通过预处理数据)，自动转换或检测图像中的特殊对象，检测异常和重复情况，并交叉核对信息以检测是否试图编造虚假信息。这方面的改进可以提高行动的效率和精确度，减少生命损失。

- **网络防御**。数字化和更加依赖信息和通信技术在带来益处的同时也带来了脆弱性。数字空间让威胁行为体有可能造成数据外泄、进行间谍活动和破坏以及影响力宣传活动。人工智能可以支持国防部门检测、监控、报告、管理和应对数字威胁的能力。使用人工智能除其他外可以更快地提供关于目标和复杂关系的更完整信息，从相关来源收集信息，并简化对分析的使用。有关人工智能如何构成数字威胁的知识和专门技能发展，对于今后能够发现和避免数字攻击至关重要。因此，通过现有和未来手段，人工智能必定是进一步发展该领域防御数字威胁能力的核心要素。

- **后勤**。成功、有效的军事行动取决于有效的后勤支助。使用那些采用人工智能的系统来简化后勤工作，可以确保改善作业能力和提升防备工作。民用后勤部门的人工智能应用已经取得了很大进展，其中许多有可能会很容易改用于军事部门。

- **支助活动**。许多军事支助活动都有可能利用人工智能加以改进和简化。这些活动包括支持和加强作业能力的任务，如经营和维继物资，采购、管理和处置物资和建筑物，征聘、培训和管理人员，以及提供会计和存档等共同服务。通过更好地利用数据进行分析和决策支持、实现任务自动化以及提高处理信息和知识的能力，人工智能有可能加强支助活动。这会让人工智能有可能转向预见性维护模式，改善信息流，采用新的和更好的人力资源管理支助系统，并改进材料和建筑物成本趋势建模。因此，在支助活动中成功引入人工智能技术可以减少时间消耗并提高效率。

此外，军事领域的人工智能应用有可能加强实施国际人道法，协助在武装冲突中努力保护平民和民用物体。这种应用可有助于建设和平和维持和平活动，并加强军控、裁军和其他合规制度的核查和监测能力。

军事领域的人工智能也带来了前所未有的挑战。人工智能具有固有的脆弱性，可能会产生意想不到的后果，并导致有意义的人类控制、责任制和问责制受损。使用深度学习有可能使人工智能模型难以理解、解释和预测。例如，缺乏了解会让冲突升级动态更加不透明和不可预测。

为确保人类对人工智能的发展、部署和使用保持有意义的控制和监督，必须制定有效的保障监督。人工智能应用越是接近作战行动和使用武力，例如决策支持系统，则这一点就格外重要。人类必须始终对军事人工智能的使用和影响负责和问责。

人工智能系统可能对训练数据的质量和代表性高度敏感。可能存在的偏差、虚假信息和错误信息或不完整的训练数据可能会导致模型产生不准确或歧视性的结果。自动化偏差会让人类用户过度依赖系统的输出。

网络领域的高度自动化能力或自主响应能力，特别是那些没有适当的人在回路机制情况，可能会导致意外反应和情况快速升级。

更多依赖网络技术来完成以前由人工或基础自动化执行的任务，也带来了恶意利用技术漏洞的风险。日益依赖重商主义体系会让人担忧依赖外部供应商、对更新失控以及与专有系统有关的其他漏洞。

上述只是军事领域的人工智能应用所涉潜在风险的一些例子。此外，还有许多不知晓的未知因素。在军事情况中，这些因素相结合或单独都可能破坏任务成果，并构成根本性的法律、伦理、人道主义和军事风险。

《挪威国防部门人工智能战略》(2023年)还概述了负责任地发展和使用人工智能的关键原则：

- **合法性**。人工智能应用的发展和使用必须符合国际法，包括国际人道法和人权法。在研究、发展、获取或采用新的依赖人工智能的武器、作战手段或方法时，各国都有义务确定在某些或所有情况下使用此种武器是否为国际人权法或适用于该国的任何其他国际法规则所禁止。

- **责任制和问责制**。针对使用人工智能要确保的人的责任和问责。必须明确确定关于使用人工智能系统的决策权及其实际使用的责任。

- **可解释性、可理解性、可追溯性**。人工智能应用必须具有适足的可解释性、可理解性、透明度和可追溯性。

- **培训**。人工智能操作员必须接受必要培训，了解人工智能应用的行为，包括如何识别异常行为。

- **可靠性、安全和安保**。人工智能应用应该有明确的、界定清楚的使用范围。人工智能应用的复原力、可靠性和安保在整个生命周期必须在各个使用范围接受测试和验证。人工智能应用必须具备足够的安保级别，并能抵御数字威胁。

- **控制**。有意义的人为控制必须得到确保。人工智能系统必须包括一个足以实现预期用途的人机交互界面，该界面要提供识别和减轻意外后果的能力，并有手段在系统以意想不到的方式运行时采取必要纠正措施。

国际社会有必要深化关于人工智能军事应用及其对和平与安全影响的对话，包括关于确保军事领域负责任人工智能的措施的对话。应特别注意将人工智能用于态势感知和决策的作战行动支持系统，在这种系统中，不希望人工智能应用中出现的输出和行为以及失去有意义的人类控制，可能会产生特别有害的后果。还要探讨混合战争中的人工智能，包括但不限于网络战中的人工智能、电子战中的人工智能和信息战中的人工智能。

挪威致力于加强信息共享和能力建设方面的国际合作。通过发展一个共享知识库，各国将促进共识，缩小差距，提高透明度和建立信任。为此，挪威将鼓励制定和公布与人工智能军事应用有关的国家战略和政策文件。应注意的是减少风险和建立信任措施。

及时发展适当的国际人工智能治理，灵活应对快速技术进步，可有助于防止技术驱动的军备竞赛，同时确保创新为全球安保提供支持。

巴基斯坦

[原件：英文]
[2025 年 4 月 9 日]

人工智能技术在军事领域的快速发展和整合会从根本上改变战争。人工智能通过在自主武器系统、指挥与控制、决策支持系统、情报、监测和侦察、训练、后勤和网络/信息战中的应用，正日益融入军事行动。这些进展虽然提高了作战效率，也对国际和平与安全构成了重大风险。

军事领域中的人工智能相关挑战

战略风险：与核武器的相互作用

人工智能与核武器系统整合会带来战略风险，特别是在核指挥、控制与通信方面。人工智能能力若纳入核力量态势和使用政策，可能会导致误判、事故和灾难性后果。

核威慑的概念在很大程度上依赖于人类的理性、感知和政治决策。纳入人工智能可能会消除或显著减少这些关键的人类因素，增加事态自动或意外升级的风险。一些国家认识到这些深层关切，公开承诺对使用核武器的决定保留有意义的人类控制——巴基斯坦支持并敦促所有核武器国家赞同这一原则。

在有核武器的地区，在常规领域依赖人工智能驱动的决策支持系统和完全自主武器系统也可能导致事态升级风险。在危机期间完全取消人类控制会难以把控冲突规模和持续时间。在动荡、高风险情况下，特别是在核动态紧张地区，自动作出反应可能会加剧常规核纠缠，并对战略稳定产生不利影响。

使用人工智能进行数据评估及情报、监测和侦察可能会给考虑采取先发制人、破坏稳定的反制打击或锚定二次打击能力的国家带来虚假的自信，对区域和全球稳定构成严重风险。

行动风险：失去人类主体性

在军事行动中受人工智能驱动的自主能力有可能减少人类监督，让危机管理复杂化。随着战争加速到“机器速度”，人类决策会受到严重压缩，从而减少缓解危机和外交干预的机会。

人类可能会过度信任人工智能从决策支持系统生成的建议，即使这些建议存在缺陷或不完整，并造成自动化偏差。关键的军事决策可能会过度依赖机器输出，导致指挥官忽视人类的直觉、环境或审慎，从而可能无意中让冲突升级。

人工智能赋能的能力受提高作战效率的诱惑和争夺决定性优势的驱动，可能会提高使用武力倾向，从而降低发生武装冲突的门槛。在危机时期，武力使用门槛低会严重破坏稳定。

技术风险

人工智能的军事应用可能会带来技术漏洞，包括算法偏见、数据中毒和易受网络攻击。由于预警系统失灵或受到操纵或发生数据中毒攻击，可能会爆发冲突。人工智能能力往往像“黑箱”一样运作，生成的决定缺乏透明度或可解释性，使验证和问责变得复杂。这些漏洞可能会导致不可预测的结果和系统故障，并对行动完整性构成显著风险。在特定数据集环境中验证的人工智能能力，在动态更复杂的、完全不同的环境中可能无法可靠地发挥作用。

规范、法律和伦理风险

在军事领域使用人工智能会构成伦理、规范和法律方面的挑战，特别是就遵守国际人道法而言。国际人道法的本质从根本上依赖人的判断、自由裁量和对环境敏感的决策——人工智能系统在本质上难以复制这些品质。将目标选择和交战(包括使用致命武力决定)等关键职能交给自主系统，可能会违反国际人道法的区分、相称性、攻击中的预防措施和军事必要等核心原则。产生不可预测、不可靠或无法解释结果的人工智能系统会使遵守国际人道法进一步复杂化，有可能导致非法或意外伤害。

此外，缺乏人类直接决策或过度依赖人工智能驱动的决策支持系统，会产生重大问责和责任问题，使非法或不法行为的归责和责任问题变得极具挑战性。如果出现问题，指挥官可能会将责任推卸给人工智能，从而使法律问责和潜在的战争罪调查复杂化。

将生死攸关的决定权交给自主系统，可能会削弱同情心、道德推理能力和人类判断力，从而加剧不正当暴力和平民伤亡的风险，这会进一步引发伦理方面的担忧。

扩散和全球安保风险

军事人工智能技术的扩散对国际安保构成重大风险。先进人工智能能力的扩散，尤其是自主武器的扩散，有可能引发新的军备竞赛，破坏地区和全球安全环境的稳定。非国家行为体容易扩散并有可能获得这些能力，进一步加剧了这些担忧。

国际拟采取反应：联合国机制的核心作用

人工智能技术是通用的，和平利用人工智能是实现可持续发展目标不可或缺的一部分。与此同时，人工智能在军事领域的影响贯穿各领域，可能极大影响国际和平与安全，因此需要作出协调一致的国际反应。

巴基斯坦确认人工智能治理倡议在联合国外的价值，但仍知晓其有限制，特别是在普遍参与和正式多边合法性方面存在限制。虽然这些倡议可以通过促进对话和政治意愿为联合国的工作作一补充，但孤立地推行这些倡议有可能造成分裂。因此，有关人工智能军事应用的讨论应在联合国论坛内进行，以确保包容性、合法性并有一个反映所有国家利益的、协调一致的全球框架。

出于这些原因，联合国必须在任何国际反应中仍然处于中心地位。联合国裁军机制在制定军事人工智能方面的国际治理框架和防止规范性格局分裂方面应发挥核心作用。人工智能对军事影响的规模和独特性要求作出多方面的、整体性的多边反应。联合国的会员普遍，这让联合国享有独特地位，是所有国家——既有发达国家也有发展中国家——都可发声的理想论坛。

任何一个论坛或工具都不够。需要一项利用多个联合国裁军机构的结构化战略，每个论坛要从独特的角度和任务授权出发，以互补方式应对这一问题。我们建议利用所有相关论坛，从联合国大会及其第一委员会到裁军审议委员会、裁军谈判会议和《禁止或限制使用某些可被认为具有过分伤害力或滥杀滥伤作用的常规武器公约》。这种做法将全面涵盖战略、人道主义、法律和技术层面，避免出现缺漏和冗余。每个论坛的工作都应为其他论坛提供信息，形成协同效应，促进实现在继续和平使用人工智能的同时降低军事人工智能风险的共同目标。

裁军谈判会议

裁军谈判会议应优先处理与军事人工智能相关的战略风险，特别是在核领域，并与议程项目 1 和 2(“停止核军备竞赛和核裁军”和“防止核战争，包括一切有关事项”)直接保持一致。2023 年，巴基斯坦提议裁军谈判会议就此议题设立一个新的议程项目(CD/2334)。

根据这一新的议程项目，裁军谈判会议应设立一个附属机构或特设小组，专门负责审查与稳定有关的军事人工智能风险，评估其如何助长核风险，并就具体措施进行谈判。这些措施可包括：

- 承诺在有关使用核武器的决定中保持人类控制，不取代人的判断
- 禁止使用人工智能能力篡改数据或针对核指挥、控制与通信系统
- 对某些人工智能能力的部署和使用制定限制措施，这些能力可能发动先发制人的打击，导致核风险升级

裁军谈判会议特别适合开展这些讨论，可平等地让所有军事大国聚在一起，以协商一致方式开展工作，从而保障所有国家的重大安全利益。探讨这一问题可重振裁军谈判会议的工作，表明对新的和正在出现的威胁作出回应。

裁军审议委员会

裁军审议委员会成员普遍，负有审议任务，是就负责任地军事使用人工智能事宜制定切实可行的准则和建议的理想机构。从历史上看，裁军审议委员会有效地制定了类似准则(例如，1988 年建立信任措施和 1993 年区域裁军办法)。

裁军审议委员会在第二工作组可制定全球和区域层面与人工智能军事应用有关的建立信任和安全措施方面的准则和建议。关键要素可包括重申规范基础，建议缓解操作风险和技术风险的措施，制定军事人工智能风险减少战略，应对扩散关切，同时确保有公平机会和平利用人工智能。

联合国大会第一委员会

大会第一委员会应将联合国秘书长的定期评估报告制度化，并根据会员国自愿共享的信息，维护关于军事人工智能能力的技术发展和相关风险的目录。这些定期评估将对不断发展的能力提供权威洞见，提供及时信息，并促进作出知情的国际政策反应。

第一委员会在审查这些报告时，可以举行关于人工智能的专门辩论，并视需要有可能在大会下设立一个不限成员名额工作组，谈判建立一个更制度性平台，例如关于人工智能军事应用的联合国登记册(不过目前最好还是利用现有论坛)。

这些报告还可以确定正在形成共识或需要进一步开展工作的领域，并指导裁军谈判会议、裁军审议委员会和《特定常规武器公约》等论坛的议程。

《特定常规武器公约》

根据《特定常规武器公约》设立的政府专家组对于应对致命自主武器系统所涉人道主义、伦理和法律影响仍至关重要。该公约的包容性(让民间社会和红十字国际委员会作为观察员参与其中)是一个有利条件。

巴基斯坦重视《特定常规武器公约》政府专家组自 2017 年以来完成的工作，特别是 2019 年制定的 11 项指导原则。然而，在《公约》框架下取得的进展一直缓慢，而且主要基于原则，而非侧重于具体规章制度。巴基斯坦同意评估意见，即根据《公约》进行的讨论对人工智能赋能武器的安全层面“关注不够，关注度下降”，这突出表明要在裁军谈判会议和其他论坛采取补充行动。然而，在人道主义方面，根据《公约》所设政府专家组应继续开展工作并加强工作。

巴基斯坦主张完成关于一项具有法律约束力的《公约》议定书的谈判，禁止致命自主武器系统无人控制或无法遵守国际人道法而自行运行。政府专家组现有任务允许会员国拟订此种文书的要素，以提交《公约》缔约国第七次审查会议，并有可能在其后启动正式谈判。

结论

巴基斯坦强调需要采取协调、包容的国际行动，以减轻重大的军事人工智能风险。巴基斯坦设想采取一种兼顾安保与发展的治理办法，确保稳定性，同时促进有益的人工智能发展。通过联合国内的结构化、多论坛战略，国际社会可以建立强有力的规范框架，维护国际安全，保护公平、非歧视地和平使用人工智能。

秘鲁

[原件：西班牙文]
[2025 年 4 月 11 日]

在 2024 年 12 月 24 日通过的第 79/239 号决议第 7 段中(秘鲁对该决议投了赞成票)，大会请秘书长：

> 就人工智能在军事领域的应用给国际和平与安全带来的机遇和挑战征求会员国和观察员国的意见，特别侧重于致命性自主武器系统以外的其他领域，并提交一份实务报告，概述这些意见，编目现有和新出现的规范性提议，并在附件中载列这些意见，提交大会第八十届会议，供各国进一步讨论。

就此，秘鲁在下文提出自己立场的某些方面，以期促进编写上述秘书长报告。

一. 人工智能在军事领域的意义

秘鲁认识到新兴技术在军事领域的快速和动态发展，特别是人工智能的可能应用情况。秘鲁正在密切关注这一领域的发展，包括人工智能似乎正在如何转变从使用自主无人机到决策支持系统的军事行动，并认为必须推动持续多边对话，以制定原则，确保以合乎道德和负责任的方式使用这些工具。

鉴于人工智能可被纳入武器系统和支持军事行动的系统，秘鲁认为必须从人道主义、法律、安保、技术和伦理角度应对使用人工智能提出的挑战和关切，包括与算法偏见有关的风险。使用这种技术可能对国际稳定与安全产生的影响加剧了这些关切。

考虑到使用人工智能对核武器和其他大规模毁灭性武器的影响，这就更加令人担忧。因此，必须强调有意义的人类控制原则。

二. 意见

遵守国际法

在军事领域发展、实施和使用基于人工智能的技术必须遵守国际法，包括国际人权法和国际人道法，还须遵守《联合国宪章》所载基本原则。

在这方面，监管军事领域人工智能的任何规范性活动都须确保负责任和合乎伦理地使用人工智能，还要保证不扩散基于人工智能的军事技术并可公平获取知识和技术能力。

这是为了确保使用人工智能要尊重人的尊严，保护平民，保障国际稳定与和平。

认识到益处和风险

人工智能为更好地了解行动情况、从而改善实施国际人道法并保护平民和民用物体提供了宝贵机会。

然而，使用人工智能可能会在军事领域带来可预见和不可预见的风险，例如因算法偏见、设计缺陷、滥用或恶意使用人工智能等产生的风险。此外，人工智能还可能影响复杂的区域和全球动态，因为人工智能会影响局势升级、作出误判、降低冲突门槛和出现军备竞赛的风险。

负责任发展

在军事领域使用人工智能应促进和平，促进保护平民，并倡导技术进步应该补充而非取代人类能力的概念。

根据适用于自主武器系统的原则，在军事领域应用人工智能必须确保绝不能将责任和问责转嫁给机器。在这方面，秘鲁强调要对涉及使用武力的所有决定保持有意义的人类控制。

在这项技术的整个生命周期，必须全面应对与此技术相关的所有风险和挑战。

可以在不妨碍其他领域人工智能方面研究、发展、实验和创新的情况下，建立控制和保障措施，防止在军事领域滥用这项技术。

执行和透明度

当务之急是制定战略、原则、标准和规范以及国家政策和法律框架，保证在军事领域负责任地使用人工智能。

为利于各国间透明度与合作而制定建立信任和减少风险的措施和交流良好做法的机制，也有重要意义。

讨论形式

秘鲁认为，必须在全球、区域和国家间各级就制定措施确保军事领域负责任的人工智能保持持续对话。

秘鲁还呼吁包容性参与这一领域，其中考虑到各国、特别是发展中国家的意见以及工业界、学术界、民间社会及区域和国际组织等其他利益攸关方的投入。

重要的是考虑到，不同国家和区域处于将人工智能能力融入军事领域的不同阶段，而且，它们是在不同的安保环境中开展行动。

这突出说明要促进发展中国家建设能力并加强国际合作，以缩小现有差距，并促进这些国家参与关于使用这项技术的讨论。

秘鲁参与国际讨论情况

军事领域负责任人工智能峰会

- 秘鲁参加了 2023 年和 2024 年峰会及相关区域讲习班
- 核可了 2024 年军事领域负责任人工智能峰会最后宣言(《行动蓝图》)

《关于在军事上负责任地使用人工智能和自主能力的政治宣言》

- 秘鲁作为观察员参加了关于这一倡议的首次全体会议，随后正式核可了该倡议

人工智能行动峰会-军事会谈(巴黎，2025 年)

- 秘鲁派出高级别代表参会，并签署了《关于在人工智能赋能武器系统中保持人类控制的巴黎宣言》。

大韩民国

[原件：英文]
[2025 年 4 月 11 日]

人工智能是一种赋能技术，有可能从根本上转变军事事务的多个层面——从决策和情报收集到后勤、监测及指挥与控制系统。随着人工智能的快速发展，各国对在军事领域利用这一技术的兴趣与日俱增。

随着人工智能能力和人工智能赋能系统越来越多地纳入军事行动，它们既带来了机遇，也带来了挑战，特别是对国际和平与安全而言。这些发展从人道主义、法律、安全、技术和伦理角度提出了重要问题。

为本呈件目的，下文所述观点特别侧重于致命自主武器系统以外的领域。

军事领域的人工智能机会

人工智能能力以及集成了人工智能的系统，包括用于情报、监测和侦察和决策支持系统的能力和系统，可以通过处理大规模数据、支持优化和生成预见性洞见，提高态势感知，增强精度和准确性并提高效率。这些能力和系统可有助于维护和促进国际和平与安全。

1. *加强实施国际人道法并协助在武装冲突中保护平民和民用物体*

人工智能赋能的情报、监测和侦察系统及决策支持系统可以通过促使作出更准确的战场评估和提高态势感知，加强实施国际人道法基本原则，即区分、相称性和攻击中的预防措施。人工智能可以帮助区分战斗人员和非战斗人员，并利用及时和充分知情的信息评估潜在的附带损害。通过提高战场意识，包括对平民存在的认识，人工智能可帮助必要、适当地采取预防措施保护平民和民用基础设施。

2. *支持维和行动*

人工智能可以支持监测停火协定与和平协议。人工智能还可促进预警机制发现潜在违规行为，加强任务效力和安全。大韩民国在联合国南苏丹特派团“世界之光”分队启动了智能营地试点项目，通过应用人工智能和其他新兴技术，提高联合国维和营地的安全、效率和作业能力。

3. 加强军控和合规制度方面的核查和监测能力

人工智能可以提高国际核查机制监测遵守军控和不扩散协定的能力。国际原子能机构可以利用人工智能提高保障监督流程的效率，特别是那些涉及数据分类、发现模式和识别数据异常值的流程。人工智能赋能系统还可以帮助识别使用化学或生物武器的早期迹象，揭露日益复杂的规避制裁策略，加强国际不扩散制度。

除上述机遇外，人工智能还能通过改进对行为体行为的分析并提升检测和积极作出应对的能力，帮助降低战略风险，如误判、误解和意外升级情况。此外，人工智能能力可促进发展各项能力，以加强网络防御态势、保护关键国家基础设施和打击恐怖主义等。

军事领域的人工智能挑战

人工智能若予以不负责任地发展、部署和使用，则其军事应用可能会带来新的挑战或加剧现有挑战。

挑战可能源于人工智能的技术和运作特性。例如，人工智能的黑箱性质使人难以理解如何以及为何生成特定输出，从而限制了可解释性和可追溯性。数据、算法或系统架构中的设计缺陷和意外偏差可能导致故障或输出偏离预期目标。过度依赖人工智能系统，如自动化偏差，或培训不足，可能会引发与缺乏适当的人类判断和人类参与相关的问题。这些因素可能增加误判、误解或冲突意外升级的可能性，从而对国际和平与安全构成挑战。

人工智能技术的双用途性质可能会增加不负责任的恶意行为体误用或滥用人工智能的风险。例如，在网络领域，人工智能驱动的虚假信息活动和网络攻击(如数据中毒和电邮地址欺骗)可能会加速。此外，不负责任的行为体可能会利用人工智能技术促进发展新型化学或生物武器，从而引发扩散问题并扩大对国际和平与安全的风险。

在军事领域实施负责任的人工智能

为了利用人工智能带来的惠益和机遇，同时应对相关风险和挑战，人工智能能力及其赋能的军事领域的系统必须在整个生命周期中予以负责任地发展、部署和使用。

大韩民国致力于确保和促进在军事领域负责任地应用人工智能。这包括以下主要原则和措施：

- 人工智能应合乎伦理并以人为本。
- 军事领域的人工智能能力必须根据适用的国际法，包括国际人道法和国际人权法加以应用。
- 人类仍要对军事领域中使用人工智能应用及其影响负责并问责，责任和问责绝不能转嫁给机器。

- 需要确保人工智能应用的可靠性和可信度，为此建立适当的保障措施，以降低故障或意外后果的风险，包括数据、算法和其他偏差造成的风险。

- 在军事领域发展、部署和使用人工智能时，需要保持适当的人类参与，包括采取涉及人类判断和控制使用武力的适当措施。

- 相关人员应该能够充分理解、解释、跟踪和信任人工智能能力、包括人工智能赋能系统在军事领域产生的输出。需要继续努力提高军事领域人工智能的可解释性和可追溯性。

大韩民国支持就进一步制定措施确保军事领域负责任人工智能进行讨论和对话，包括利用国际规范框架；严格的测试鉴定协议；全面核查、验证和认证流程；健全的国家监督机制；持续的监测流程；全面培训方案和演练；提升网络安全；明确的问责框架。

建立强有力的控制和安全措施，对于防止不负责任的行为体获取和滥用军事领域潜在有害的人工智能能力、包括人工智能赋能系统至关重要。

大韩民国鼓励各国制定有效的建立信任和信心措施以及适当的降低风险措施，并鼓励各国就良好做法和经验教训交流信息和进行磋商。

大韩民国强调，必须防止国家和非国家行为体利用人工智能能力助长扩散大规模毁灭性武器，并强调人工智能能力不应阻碍军控、裁军和不扩散努力。至关重要的是，在不妨碍实现无核武器世界这一最终目标的情况下，对于为有关使用核武器的主权决定提供信息和执行这些决定而言至关重要的所有行动，必须保持人类控制和介入。

军事领域人工智能能力和人工智能赋能系统的发展、部署和使用应该以维护而不是阻碍国际和平与安全的方式进行。

军事领域人工智能的未来治理

在设想军事领域人工智能的未来治理时，必须促进对人工智能技术——人工智能能力及其局限性——形成共识，并对人工智能在军事领域的可能应用及其对国际和平与安全的影响有共同理解。

特别是对发展中国家而言，能力建设对促进这些国家充分参与治理讨论并促进以负责任的方式对待军事领域人工智能的发展、部署和使用并在这方面达成共识也有重要意义。交流知识、良好做法和经验教训也可促进共识。

鉴于人工智能发展迅速，治理机制应该足够灵活，以适应人工智能的发展。此外，大韩民国支持采取兼顾机遇和风险的平衡办法。过于以风险为中心的或是限制性的治理讨论可能会扼杀创新，掩盖人工智能支持国际和平与安全的潜力。未来治理不应阻碍创新，而应支持创新，并在军事领域负责任地应用人工智能方面发挥推动作用。

由于国际社会对于军事领域人工智能对国际和平与安全影响的认识尚处于早期阶段，而且考虑到技术和政策发展现况，在没有对什么是军事领域负责任的人工智能达成普遍、共同理解的情况下，狭隘地界定人工智能治理路径或制定具有法律约束力的文书或规范为时尚早。大韩民国认为，有关治理的讨论应在持续对话的指导下切合实际，逐步进行。

大韩民国认识到人工智能创新正由私营部门推动，认为未来治理工作必须采取开放和包容做法，与包括工业界、学术界、民间社会、区域和国际组织在内的多利益攸关方互动。

大韩民国认可国家、区域和全球为应对军事领域人工智能带来的机遇和挑战所作努力，包括制定相关国家战略、立法、原则、规范、政策和措施，并确认促进各级对话有重要意义。

为确保军事领域负责任应用人工智能，大韩民国分别于 2022 年和 2025 年在国防部新建了数据政策司和国防人工智能政策工作队。2024 年，国防部启动了国防数据和人工智能委员会，作为最高级别的审议和决策机构。

为了促进对话，大韩民国连同荷兰、新加坡、肯尼亚和联合王国于 2024 年 9 月在首尔举办了第二届军事领域负责任人工智能峰会。2024 年举行的军事领域负责任人工智能峰会和关于军事领域负责任人工智能的系列区域磋商成为交流专业知识、促进包容各方的对话和增进相互理解的孵化器。展望未来，2025 年 9 月将在西班牙举行第三届军事领域负责任人工智能峰会，而且 2025 年即将举行军事领域负责任人工智能区域磋商，它们将继续指导国际社会努力实现军事领域负责任人工智能应用。

大韩民国认为，在联合国框架内、包括在大会第一委员会和裁军审议委员会讨论军事领域负责任人工智能应用，应与联合国外的其他相关倡议互为补充，包括军事领域负责任人工智能峰会进程、《关于在军事上负责任地使用人工智能和自主能力的政治宣言》和致命自主武器系统领域新兴技术问题政府专家组。大韩民国认为，这些倡议相互促进、互为补充。

数据治理也至关重要。由于数据在培训、部署和评估人工智能系统方面发挥着核心作用，相关利益攸关方必须进一步讨论适当的数据治理机制，包括数据收集、存储、处理、交换和删除以及数据保护方面的明确政策和程序。

俄罗斯联邦

[原件：俄文]
[2025 年 4 月 10 日]

俄罗斯联邦欢迎大会通过 2024 年 12 月 24 日第 79/239 号决议，并谨根据该决议第 7 段，对秘书长提交大会第八十届会议供会员国进一步讨论的报告提交本国投入。

导言

俄罗斯联邦高度重视军事领域的人工智能应用相关事项。我们有意在专门国际论坛上对此问题进行进一步实质性讨论。

我们认为《特定常规武器公约》缔约国设立的致命自主武器系统问题政府专家组是进行此种讨论的最佳平台。正是该政府专家组被促请在有关此类武器的人道主义关切和各国合法自卫利益之间达成合理平衡，并在协商一致基础上作出决定。该专家组对人工智能军事应用的审议范围很广，不限于致命自主武器系统问题，还涉及为军事目的使用该技术的若干重要方面(包括法律、技术和军事方面)。

我们注意到在军备控制、裁军和不扩散领域现有制度框架内对这一专题的讨论情况。本文件重点分析在缔约国履行相关国际法律文书规定的义务方面人工智能带来的风险和机遇。

我们欢迎会员国愿意在裁军审议委员会着手讨论人工智能的军事应用议题，这是在国际安全背景下讨论新兴技术的一部分。这种意见交流旨在就其他论坛没有涉及的“军事”人工智能某些方面的建议达成一致。

在上述国际论坛开展工作过程中，要特别关注共同术语的制定、现行国际法的适用、人类控制、问责制以及人工智能技术带来的风险和机遇。

定义

现行国际法对基于人工智能的武器系统和军事装备没有协商一致定义，因而难以探讨该问题。就这类工具以及为军事目的适用此类技术的相关术语形成共同的工作认知，会使该主题以及该议题的讨论前景更为明确。

工作定义应当：

(a) 描述各种基于人工智能的武器系统和军事装备及其应用特有的重要特点；

(b) 不局限于对此种工具的现有理解，而是考虑到此种系统今后会如何发展；

(c) 为专家群体(包括科学家、工程师、技术人员、军事人员、律师和伦理学家)所普遍理解；

(d) 不会被解释为限制技术进步或不利于和平机器人技术和人工智能领域的研究；

(e) 不应仅通过描述功能来界定基于人工智能的武器系统和军事装备。

应避免将这些工具归为“坏”或“好”，即不应根据特定国家集团的政治偏好对其进行分类。

现有高度自动化军事系统不应归入需要进行紧急限制和禁止的“特殊”类别。正是这种自动化水平使此类系统能够在动态作战情况下和各种环境中有效运作，并保证有足够的针对性和准确性，从而确保它们符合国际法、包括国际人道法的原则和规范。

国际法中基于人工智能的武器系统和军事装备

人们普遍认为，国际人道法等现有国际法完全适用于基于人工智能的武器系统。

俄罗斯联邦认为，目前没有令人信服的理由对基于人工智能的武器系统施加任何新的限制或禁止，或是针对此类工具更新或调整国际法，包括国际人道法。开展讨论以期商定关于“负责任”使用基于人工智能的武器系统和军事装备的某些“行为规则”或规范和原则也为时尚早。西方国家倡导的“负责任”使用人工智能的概念基于不为国际法(包括国际人道法)认可的极具争议的标准，并提出了许多问题，没有得到国际社会的一致支持。

人道原则、公众良心要求和人权内容不能被用作对某些类型武器和军事装备施加限制和禁止的绝对和唯一充分条件。针对基于人工智能的武器系统和军事装备的关切，应通过真诚履行现有国际法律准则予以应对。

在武装冲突局势中严格遵守包括国际人道法在内的国际法准则和原则，仍然是俄罗斯联邦的优先事项之一。俄罗斯联邦武装部队严格遵守联邦和部门法律文书中的国际人道法准则。遵守国际人道法的相关问题，包括使用新型武器相关问题，在各类军事人员的条例和培训方案中得到了体现。《俄罗斯联邦武装部队发展和使用基于人工智能的武器系统的概念文件》于 2022 年获得通过。

俄罗斯法律充分考虑了《特定常规武器公约》缔约国 2019 年以协商一致方式核准的关于基于人工智能的武器系统的准则。我们认为，就在国家层面实施这些准则的具体实际措施进一步交流信息，是建立信任和提高透明度的一种方式。

对基于人工智能的武器系统和军事装备的控制

我们认为一个重要限制是，人类应该对基于人工智能的武器系统和军事装备的运作加以控制。因此，此种工具的控制回路应允许操作人员或上级指挥系统进行干预以改变此种系统的运作模式，包括部分或完全停用此种系统。

俄罗斯联邦认为，人类始终要对使用武力的决定负责。所实施的控制以作出决定时所掌握的所有信息为依据。不过，人类控制的具体形式和方法应由各国自行决定，直接控制不一定是唯一选项。

可以通过以下方式对这些系统和装备进行控制：

(a) 提高系统和装备的可靠性和容错性；

(b) 限制目标类型；

(c) 限制系统和装备运作的时间范围、地理覆盖范围和使用规模；

(d) 立即采取干预措施并停用系统和装备；

(e) 在实际运作环境中对系统和装备进行试验；

(f) 允许已成功掌握使用人工智能工具程序的人员操作(控制)这些工具；

(g) 监控单个元件和整个装置的制造；

(h) 监控单个元件和整个装置的拆除和处置。

我们认为，在讨论中引入某些国家倡导的“有意义的人类控制”、“人类介入的形式和程度”、“根据具体情况进行人类控制和评估”、“可预测性、可靠性、可追溯性、可解释性”等概念并不适宜，因为此种概念通常没有法律效力，只会让讨论政治化。

责任

俄罗斯联邦认为，根据国际法，国家和个人(包括开发商和制造商)在任何时候都要对决定开发和使用基于人工智能的武器系统和军事装备承担责任。使用此种工具的责任在于给系统和装备指派任务和下令使用系统和装备的官员。要使用基于人工智能的武器系统，该官员应具备与系统和装备的运作和操作有关的必要知识和技能，并应负责就使用系统和装备的适当性作出决定，而且规划使用系统和装备的形式和手段。

基于人工智能的武器系统和军事装备的机会和限制

众所周知，基于人工智能的武器系统和军事装备在执行指定任务时可能比操作人员更有成效，并且可能降低失误可能性。特别是，此种工具能够大幅减少因操作人员失误及其身心状态以及道德、宗教或伦理信仰造成的国际法(包括国际人道法)范畴的负面影响。使用此种工具可确保更准确地将武器瞄准军事设施，并有助于减少对平民和民用物体造成意外打击的风险。

评估与使用基于人工智能的武器系统和军事装备有关的潜在风险和减轻这些风险的措施，应成为设计、开发、测试和部署任何类型武器系统新技术过程的一部分。

可通过以下方式尽量减少与此种工具有关的风险：

(a) 确保有效的生命周期管理；

(b) 在生命周期的所有阶段进行全面测试，包括在接近现实生活的环境中进行测试；

(c) 提高系统的可靠性和容错性；

(d) 制定准备工作标准；

(e) 确保最大限度地防止未经授权的访问；

(f) 培训操作人员；

(g) 在收集和处理信息时优先使用人工智能技术，以支持军事决策；

(h) 为操作人员连续监视此类系统的运行情况提供便利，并使系统能够在操作人员的指挥下紧急终止战斗任务；

(i) 防止此种工具落入非国家行为体手中，他们可能会将此种工具用于非法目的。

在武器装备、军事装备和特种装备生命周期的各个阶段(开发、生产、运行和处置)都可采取这些措施。

后续步骤

我们认为，有益做法是，各国在裁军审议委员会并在致命自主武器系统问题政府专家组中继续审议关于将人工智能用于军事目的的问题，该专家组是在军备控制、裁军和不扩散制度框架中进行此种讨论的最佳国际平台。与此同时，在一个论坛上的相关讨论不应重复已在类似论坛中进行的意见交流。

我们反对分散这方面的努力。将人工智能用于军事目的问题转到任何其他国际平台进行讨论、建立其它论坛审议这个问题或在没有联合国绝大多数会员国(包括基于人工智能的武器系统的主要开发国，包括俄罗斯联邦)参与下在某个狭隘的论坛讨论这个问题，似乎都会适得其反。

特别是，在一些西方国家集团组织的非包容性的“负责任地将人工智能用于军事目的峰会”以及一般性的人工智能峰会上讨论这一议题没有建设性。这些活动及其成果文件没有考虑到所有利益攸关方的观点，不能被视为表明各方对该主题达成共识的进一步工作的基础。它们会引起分歧，无助于在这一领域共同努力。

企图绕过专门多边论坛，在包括此种“峰会”在内的其他论坛上“巩固”对这些问题的单方面做法，将产生极为不利的后果。它们有可能严重破坏正在进行的关于“军事”人工智能的建设性和包容性工作，并可能分裂在这一领域形成共识和提出建议的努力。

在上述国际论坛的讨论过程中，我们认为必须主要侧重于就现有国际法、包括国际人道法适用于基于人工智能的武器系统和军事装备事宜商定共同的专门术语和做法，侧重于确保人类控制此种工具，并侧重于这种技术带来的风险和机会。

俄罗斯联邦请秘书长在根据大会第 79/239 号决议第 7 段提交的实务报告中考虑到上述提议，并将本文件列入该报告附件。

塞尔维亚

[原件：英文]
[2025 年 4 月 4 日]

人工智能的发展与应用是当今世界军事行动方式的重要变革因素。它提供了新的可能性，同时也给国际稳定和军事领域的和平与安全带来新的挑战。因此，有必要着手建立一个适当的国际框架以规范其应用。

1. 军事领域人工智能的可能性和优势

在非致命性军事背景下应用人工智能可以改善军事行动的诸多方面：

(a) 提高行动认识水平；

(b) 提高决策过程的质量和速度；

(c) 通过快速的数据处理提升情报数据和侦察的质量，从而迅速识别威胁；

(d) 支持在军事冲突中保护平民和非战斗人员；

(e) 通过监督停火和预测冲突动态，支持和平行动和任务；

(f) 通过降低成本和节约资源，改善预测性维护和物流优化的流程和程序。

2. 军事领域人工智能的主要挑战和威胁

在作战和非作战系统中开发和纳入人工智能对国际和平与稳定及国际人道法构成严峻挑战，主要体现在以下领域：

(a) 因在动态环境中应用错误而导致的技术风险和功能故障，可能威胁人的生命、造成物质损失，并影响国际人道法的执行；

(b) 在遵守国际法方面，特别是在执行其原则(如区分、比例以及确定目标时采取预防措施)方面存在的法律和道德风险；

(c) 缺乏为人工智能操作的行为和活动确定责任的明确规则；

(d) 算法缺陷可能导致决策和区分过程中的偏见和错误，因为使用非代表性数据组可能导致对平民的错误识别或者对族裔或民族群体的威胁；

(e) 人工智能算法的应用可能会造成参与行动人员责任减轻的错觉；

(f) 基于错误的前提通过人工智能做出决策的战略风险；

(g) 不加筛选地融合和纳入新技术，特别是在信息和网络行动或使用核、化学和生物手段的领域这样做；

(h) 缺乏开发、组织和在冲突中负责任地应用人工智能系统的专业人员；

(i) 通过制造和散布虚假信息的方式在信息行动中滥用人工智能，这可能会挑起冲突、加剧紧张局势。

3. 建立法律和道德操守框架

考虑到所评估的风险和挑战，国际社会有必要建立强制性的法律和道德操守框架，以便：

(a) 促进并着手在联合国内部开展对话，加强对国际人道法规范的遵守，包括制定国际法律准则、规则和原则，以确保人工智能系统的开发和应用符合国际人道法原则(区分、比例和采取预防措施保护未参战人员)；

(b) 对人工智能应用于系统和武器的核准适用性启动合法性评估程序；

(c) 确保在武装冲突期间保护个人生命和自由，并在和平时期，特别是在监控背景下，保护个人隐私；

(d) 加强联合国机制，对出于军事目的使用人工智能的风险引入强制性审议，升级裁军谈判会议，协调裁军审议委员会的工作，设立新的联合国专门机构，并扩展负责任使用人工智能方面的现有联合国举措；

(e) 发起联合国对话，为在军事领域负责任地使用人工智能界定定义，并为其应用制定安全协议(测试、评估、验证与核查)；

(f) 制定措施，在开发、建立和应用军事领域人工智能系统和服务的过程中，使私营部门招募符合国际人道法原则；

(g) 扩展现有联合国机构和文件关于人工智能开发和应用道德操守问题的建议，纳入处理冲突的具体准则。

在国际武装冲突背景下应用人工智能系统，需要国际社会采取广泛多边行动以促进使用责任。联合国应在发起对话、制定规范和建设国际社会能力方面发挥主导作用，以防止各自为政、实现适当管理。

新加坡

[原件：英文]
[2025 年 4 月 11 日]

作为小国，新加坡一贯支持基于规则的多边体系和联合国的作用。联合国提供了国际法和国际规范的基础。多边机构、制度和法律关乎所有国家，特别是小国的生存。

新加坡认为，包括人工智能启用的系统在内的军事领域人工智能能力，应在其整个生命周期中以负责任的方式加以应用，并遵守适用的国际法，特别是国际人道法。

人工智能可能给军事领域带来益处，提高精确性和态势感知，从而减少对平民和/或民用物体的附带伤害。然而，倘若缺乏适当治理框架，人工智能或将带来冲突升级和误判的风险。有鉴于此，新加坡认为国际社会需就此议题开展讨论。

新加坡对军事领域人工智能的治理办法

新加坡《国家人工智能战略 2.0》的主要目标之一是建立值得信赖的环境，保护用户并促进创新。为此，包括国防在内的各政府部门正在制定人工智能治理框架，以利用人工智能的益处，同时确保减轻其潜在危害。

通过与国防技术专家、军事规划人员、国际法专家和政策专业人士的磋商，新加坡制定了军事领域人工智能国家原则，该原则于 2021 年发布，应对四个主要关切领域。

(a) **负责**：首先，必须应对人工智能突发行为的风险。人工智能系统必须有明确定义的预期用途，开发者和使用者都对人工智能系统的结果负有责任；

(b) **可靠**：其次，必须应对人工智能系统输出结果错误或不准确的风险。人工智能系统应经过测试，确保达到适合其预定用途的水平。其设计应尽可能减少意外偏见并生成一致的输出结果；

(c) **稳健**：第三，必须应对恶意行为者利用人工智能的风险。人工智能系统的设计应考虑网络和对抗性人工智能的威胁。为解决“黑箱效应”，应妥善记录人工智能系统的开发过程以支持可解释性；

(d) **安全**：第四，我们必须重点关注关键安全背景下人工智能失效的风险。人工智能系统，无论在部署平台方面还是在周边资产和人员方面，其使用都应当安全。

新加坡对出于军事目的开发、测试、训练和部署由人工智能启用的系统采取的治理办法参考了上述指导原则。

有关军事领域人工智能的国际和区域倡议

新加坡积极参与军事领域人工智能治理的国际倡议。2023 年，新加坡核可了《军事领域负责任人工智能行动呼吁》和《关于在军事上负责任地使用人工智能和自主能力的政治宣言》。2024 年，在大韩民国首尔，新加坡共同主办了军事领域负责任人工智能峰会，期间新加坡核可了《军事领域负责任人工智能行动蓝图》。

新加坡还认识到区域倡议的重要性，以确保就人工智能军事应用开展包容各方和因地制宜的讨论。新加坡共同主办了 2024 年军事领域负责任人工智能亚洲区域磋商，为区域各国提供了交换意见的平台，话题包括人工智能为军事领域带来的机遇和风险。

2025 年 2 月，在马来西亚槟城举办的东盟国防部长会议非正式会晤上，新加坡和其他东南亚国家联盟(东盟)成员国通过了《关于国防领域人工智能合作的联合声明》。在声明中，东盟各国国防部长承诺促进负责和问责地使用人工智能，通过信息交换加深区域对人工智能在国防部门影响的了解和认识，并在东盟成员国之间分享最佳做法和经验教训。

在联合国讨论人工智能与国际和平与安全的未来方向

新加坡认为，在国际社会支持本决议的基础上开展的任何进一步讨论均应秉持开放和包容的性质。有鉴于此，我们赞成在联合国范围内围绕军事领域的人工智能建立一个不限成员名额工作组。如果设立此不限成员名额工作组，则该工作组应采取多利益攸关方的办法，除其他外，包括技术专家、军事规划人员、国际法专家和政策专业人士。新加坡重申致力于与所有会员国合作，推动人工智能在军事领域的负责任应用。

西班牙

[原件：西班牙文]
[2025 年 4 月 11 日]

导言

人工智能引发包括安全和国防在内的所有领域的革命。其开发与应用既带来巨大进步和机遇，也伴随诸多挑战。

武装部队对这项技术的采用不仅重新定义了军事行动的开展方式，也在改变全球战略平衡。

西班牙国防部开发和纳入人工智能基于负责任、合乎道德和合法的军事用途，符合国际人道法并确保尊重人权。

人工智能正在改变关于军事力量和安全的传统理念，为多域环境下的数据收集和分析、决策制定和行动执行提供先进能力。这涉及到各国处理国防和安全问题的方式发生范式转变，促进更快速、更精准地应对新出现的威胁。

在军事领域，人工智能正在对不可预测的战场产生颠覆性影响，军事行动的规划和开展因而发生范式转变。人工智能还影响着军事领域的其他方面：后勤、训练、信息管理和解读、情报、监视、目标捕获和侦察。

值得注意的是，西班牙秉承对负责任地使用人工智能的承诺，主办 2025 年军事领域负责任人工智能峰会，并核可了《行动呼吁》(2023 年，海牙)和在 2024 年上一次峰会上提出的《行动蓝图》。

西班牙国防部的概念和监管框架

西班牙国防部对人工智能的开发、部署和应用遵循一系列基本原则，以确保在符合国家和国际法规的前提下安全、合乎道德地使用人工智能。这些原则载于国防部《人工智能开发、纳入与使用战略》(依据国防国务秘书处第 11197/2023 号决议制定)，并符合《北大西洋公约组织(北约)2021 年人工智能战略》(2024 年修订)，旨在最大限度地利用人工智能为国防领域提供的机遇，同时减轻其军事应用的相关风险：

- 合法性：人工智能应用程序的开发和使用应遵守包括《世界人权宣言》和国际人道法在内的适用国内和国际法。

- 人类责任和问责：任何人工智能的开发或使用都应允许明确的人工监督，以确保适当问责和明确责任归属。
- 可解释性和可追溯性：人工智能应用(包括使用可审计的方法、来源和程序)应当对相关人员易于理解并保持透明。
- 可靠性和透明度：人工智能应用应针对精确、定义明确且有限的用例进行定制，并应提供信息以促进所有利益相关方对这些应用的普遍理解。应在其整个生命周期内根据这些用例测试并保证这些能力的安全性、可靠性和稳健性。
- 可治理性：人工智能应用程序的开发和使用应符合预期设计功能，这些功能应包括发现和避免意外后果的能力。当发现计划外或非预期行为时，应启用断开或停用机制。
- 减少偏见：应采取一切必要措施，尽量减少人工智能开发和使用中的错误和主观倾向。
- 隐私：开发、实施和使用基于人工智能的应用程序，从设计到整个生命周期均须尊重个人隐私。

在监管框架方面，国防部正在就军事领域人工智能的开发、实施和使用制定一套标准和最佳做法，以确保在国家和国际法律框架(特别是严格遵守国际人道法和人权)下负责且高效地使用人工智能。

机遇

国防部侧重于多个不同领域的人工智能能力开发，以提高武装部队的效能。根据该战略，人工智能的使用集中在行动、情报、后勤、网络安全和决策支持领域。

人工智能将有助于提高军事行动中决策的准确性、速度和效能(始终以遵守国际人道法为前提)，以期更高效地执行任务、降低部队风险，并帮助加强武装冲突中对平民和民用物体的保护。

人工智能实时分析海量数据的能力可改善态势感知和威胁应对能力，提升行动安全性。所有这些能力提升都保障人类始终保持控制，而不是将责任下放给机器。

在军事培训和教育方面，在欧洲总参谋学院(“C5 指挥官小组”)(英国、法国、德国、意大利和西班牙)的框架内，正在建立一个与军事教育中的人工智能有关的协作空间。

西班牙还与北约数据和人工智能审查委员会就在军事领域负责任地使用数据和人工智能开展协作。

此外，国防部还宣布对特定地区进行战略投资，以促进与人工智能和其他先进技术相关的项目。这些投资不仅旨在加强该产业，同时也意在促进新区域的工业振兴。

挑战

人工智能在军事领域的发展和应用必须遵守国家和国际监管框架，包括尊重国际人道法的执行，从而整合努力以确保人类有效控制与军事行动应用人工智能有关的关键决策。

在隐私和数据保护方面，训练人工智能模型所需的海量数据收集和处理会带来个人数据保护和信息安全方面的风险。

安全性和可靠性

最大的挑战在于如何安全可靠地使用人工智能。主要相关风险如下：

- 人工智能算法的训练数据可能包含偏见，这可能导致错误决策或意外后果。
- 人工智能模型训练不足可能导致解读错误，从而给军事行动造成潜在灾难性后果。
- 人工智能系统可能成为网络攻击目标，这些攻击可能操纵其行为或使其失效。
- 存在数据投毒风险，即恶意行为者篡改训练数据集以引发算法故障。

在西班牙，军事领域人工智能的发展遵循问责制和持续监控的原则，并在系统生命周期的每个阶段实施风险评估、审计和追溯机制。人工智能的任何开发或使用都应使人类得以进行明确监督，以确保适当问责和责任归属，与人工智能活动相关和并行的人类行动需留下明确的可追溯痕迹，而不是交由机器做出最终决策。

人工智能还必须是可靠和可预测的，其自主程度须受到训练有素的操作人员控制和监督。

任何人工智能解决方案都应在不同于其训练环境的环境中进行评估，并接受非功能性测试——在规定的、不断变化的场景下进行负载、压力和性能测试——以研究其行为和允许的偏见。

此外，这些人工智能能力应在其整个生命周期中接受严格测试和不断审计，以便及早发现潜在错误并提高操作可靠性。在部署的所有阶段均应实施人类监督和控制协议，确保关键决策不完全委托给人工智能。为此，正在努力确保人工智能的发展获得公认实体的认证。

为提高基于人工智能系统的稳健性并防范外部行为的影响，在设计阶段就纳入安全性至关重要，以保证其能够抵抗网络攻击和对抗性操纵，并确保所使用的数据和模型的完整性。

人工智能可能成为数据投毒或模型操纵等攻击的目标；因此，需持续监控系统性能并定期验证测试和进行审计。应促进备份和灾后恢复计划，保证不利情况下的系统可操作性。

还应加强与网络安全机构和人工智能专家的合作，确保武装部队拥有最佳工具和策略，以保护系统免受外部威胁，保障系统的运行可靠性。

作为国防部四大重点领域之一，这些技术领域的人才和人员培训至关重要，以确保操作人员了解这些系统的范围和局限性，并能在其行为出现偏见时进行干预。对人员开展合法和合乎道德地使用人工智能方面的培训和宣传，对于减少与偏见有关的风险、确保人工智能在武装部队中的使用客观、可靠且符合国家和国际法规(特别是国际人道法)至关重要。

目前正在编制一份良好做法指南，可为国防部各部门共同编写的文件提供基础。北约提出的在军事领域负责任地使用人工智能的良好做法已得到传播。例如北约负责任人工智能评估和工具包，其目的是落实北约通过的负责任使用人工智能的原则，这些原则包括合法性、责任、可追溯性、可靠性、可治理性和减少偏见。

瑞士

[原件：英文]
[2025 年 4 月 11 日]

1. 机遇和风险

人工智能或将改变军事事务的诸多方面。它有望通过提高可靠性、效率、精确性，安全性和稳健性等方式支持军事任务和行动。关键领域包括态势感知、决策、情报、监视和侦察、后勤和供应链、培训和模拟以及指挥与控制，通过分析大型数据集实现更迅速、更知情的决策。例如，在监视和侦察中，人工智能可以分析无人机和卫星图像，比人类分析师更快探测到动态。人工智能还可以通过处理传感器数据来支持目标识别，以区分友军和敌军。在后勤领域，人工智能可以优化供应链、预测设备故障，并确保资源在正确时间抵达正确地点。在决策支持方面，人工智能模拟可以为指挥官提供预测性见解和潜在结果，以指导战略规划。人工智能驱动的训练和模拟系统提供逼真的自适应环境，让士兵做出更好战备。最后，人工智能可以通过简化信息流、改进决策和加强跨单位协调来支持指挥与控制。人工智能还可以通过预警系统、预测分析和监测机制，助力威胁检测、网络安全、维持和平、军控核查和冲突降级，从而有助于促进稳定和安全。然而，尽管这些发展可能会给武装部队带来益处，在军事领域纳入人工智能也会引发若干重要关切和潜在风险。

若在武装冲突中负责任地使用人工智能，则其有可能通过改善风险评估或提高瞄准精度从而减少附带损害等方式，促进对国际人道法的遵守并加强对平民和民用物体的保护。然而，武装冲突中若干形式的人工智能军事应用，特别

是涉及高风险应用时，也引发法律、人道主义、道德操守，安全和战略稳定性方面的严重关切，必须加以应对。例如：

- **目标选择错误**：尽管人工智能在技术上可以根据其训练数据识别物体或个人，但遵守国际法所需要的背景了解和价值判断构成一项特别挑战，这可能导致将物体或人员误判为军事目标，从而导致非法或意外打击。

- **升级风险**：在瞬息万变的危机中，黑箱决策支持工具可能建议采取攻击行动，而未提供明确理由。如果缺乏可解释性，指挥官可能盲从错误指令，或浪费关键时间质疑指令。

- **误判意图**：在使用人工智能系统评估人员或物体行为的风险时，特别是当根据基于过往行为和情境推导的模式进行评估，而缺乏适合情境的人类控制和判断时，可能引发(法律和安全)关切。例如，监控对手行为的人工智能系统可能因错误数据而将常规部队调动错误归类为敌对行为，这可能引发抢先行动和意外升级。

这些风险凸显了确保遵守现行国际法(特别是国际人道法)义务的重要性，以及亟需就这一问题开展进一步对话和研究，以更好地了解风险与挑战、可能采取的必要措施，并审议建立额外规范性治理结构的必要性、附加值和可行性。这可包括建立国家立法、拟订最佳做法、国际规范、标准或文书，或制定业务准则。

2. 法律框架

人工智能与任何其他技术一样，其开发和使用并非发生在法律真空中。军事领域的人工智能在开发、部署和使用时必须完全遵守现行国际法，特别是《联合国宪章》、国际人道法和人权法以及其他相关法律框架。任何技术都不得挑战国际法的有效性。国际法——特别是《联合国宪章》全文、国际人权法和国际人道法——适用且必须得到遵守和履行。

各国和冲突各方必须在所有情况下，包括在军事行动中使用人工智能时，尊重国际人道法并确保其得到尊重。因此，军事领域人工智能的设计应旨在加强对国际人道法的遵守以及对平民和民用物体的保护。例如，可以通过严格的目标选择、验证和核查程序等方式确保人工智能系统优先考虑精确性、尽量减少伤害以及问责制，以实现上述目标。此外，人工智能的使用应有助于加强履行在军事行动中采取一切可行预防措施的法律义务，包括通过改进风险评估等方式支持指挥官在整个敌对行为中保护平民和民用物体，以避免或尽量减少附带伤害。

一个关键的行动领域是，确保用于设计和训练军事领域人工智能的数据集能够支持其使用完全符合国际法。在敌对行为以外，如果军事领域的人工智能被用于执行国际人道法管辖的其他任务，例如与人员羁押和拘禁有关或与被占

领土的人群控制和公共安全有关的任务，必须遵守国际人道法的所有相关规则和原则。

在开发和使用军事领域的人工智能时，系统设计或训练数据中可能内嵌过度宽松的法律解释，例如扩大合法目标定义或提高可接受附带伤害的阈值。此类解释如被大规模应用，可能逐渐削弱国际人道法的保护宗旨，并显著增加对平民的伤害。这一风险凸显了维护法律规范完整性的重要性，在未来军事领域人工智能的治理、设计和部署中，必须始终将其作为核心考虑因素。

3. 理解和原则

基于上述法律框架，并考虑到人道主义、道德、安全和战略稳定的关切，应进一步制定以下理解和原则：

1. 人类责任、问责和参与

- **责任与问责**：各国必须确保，人类始终根据适用国际法对涉及军事领域人工智能的决策负责并承担问责。
- **适合情境的人类控制和判断**：关键军事决策(从会议室到战场)，特别是那些涉及使用武力的决策，必须始终在适合情境的人类控制和判断下制定。军事领域的人工智能可以辅助决策，但不应取代法律和道德考量及判断，例如决策的认知自主性。各国必须将这些系统仅纳入一个人类得以保持判断并行使适当程度控制的指挥控制链。应尽可能应对意外偏见的问题。

2. 可靠性、可预测性/可解释性、稳健性

- **可靠性**：军事领域的人工智能必须可靠，以避免意外后果或故障，当其可能产生负面影响或伤害平民和民用物体时尤其如此。仅在可以合理预见影响和后果时，方可使用军事领域的人工智能。
- **可预测性/可解释性**：人工智能的决策过程应对于负责部署的人员可预测并可解释，使其得以理解和预测系统行为。
- **稳健性**：军事领域的人工智能还必须在技术和操作层面具备稳健性，以便在部署和使用时始终有保障且安全。

3. 风险缓解

- **增强态势感知**：应利用人工智能提高战场感知，特别是通过检测平民的存在来降低伤害概率。
- **预测分析**：应利用人工智能驱动的预测模型协助评估风险、制定冲突降级战略等，以防止平民伤亡。
- **内置防护机制**：军事领域的人工智能应纳入保护措施以尽量减少伤害，并应支持在系统故障时进行充分的人工干预。

4. 避免新的升级路径

- **稳定性**：军事领域人工智能的设计、部署和使用不得加剧国际紧张局势或创造新的升级路径。
- **军备控制**：人工智能可支持军备控制，不得破坏现有的不扩散、军备控制与裁军的规范和文书，或阻碍对这些规范的遵守，特别是有关生物武器和核武器的规范。
- **危机管理**：军事领域的人工智能可支持降级和危机管理。

5. 军事人工智能系统的生命周期管理

负责任地军事使用人工智能需要采取全面和风险敏感的办法处理军事领域人工智能的整个生命周期。这包括此类系统的设计、开发、测试、部署、运行、更新和退役。在每个阶段，必须系统地综合考虑相关的法律、人道、业务和技术因素。这种基于生命周期的办法对于高风险的军事领域人工智能尤为重要，例如那些涉及自主武器、目标选择或决策支持、可能造成人员伤亡或物体损毁，以及广而言之，决策受国际人道法管辖的军事领域人工智能。对于风险较低的系统，如行政支持工具或后勤规划系统，则需根据适合情境的风险评估应用生命周期管理。

- 在设计和开发阶段，各国必须确保使用高质量、代表性的数据集(基于最小偏见)对系统进行训练以使其使用完全符合国际法、规范和标准，以尽量减少非预期偏见。
- 在测试鉴定阶段，必须执行严格的验证和核查程序，以确认现实条件下的可靠性、法律合规性和操作稳健性。
- 在部署和运行使用阶段，必须采取保障措施以监控系统性能，确保适合情境的人类控制和判断，并支持充分的人工干预。
- 在整个更新和学习阶段，各国必须制定严格的系统修改规程，包括版本控制、重新验证和正式批准程序。
- 针对退役阶段，必须采取措施以安全禁用或存档系统，以防滥用、意外激活或重新部署。

4. 国际治理

瑞士强调，必须建立一个包容持续的联合国进程，就军事领域人工智能的益处、风险和挑战凝聚共同理解，并制定负责任地使用人工智能的原则。因此，需要纳入所有会员国、相关利益攸关方以及科学家和技术行业、民间社会及学术界代表的参与，以确保其合法性、专业性和广泛支持。相关联合国进程应当透明、定期召开会议，并与其他相关倡议保持一致。

在军事领域负责任地使用人工智能方面的所有国际治理努力，其总体目标必须是确保遵守国际法，特别是国际人道法。此外，这些努力的核心必须是人

道主义和道德关切、保障稳定性以及减少安全风险。有效的治理框架、共同的规范和持续的多边对话应有助于防止意外升级、促进透明度和相互信任，并加强国际法在技术颠覆时期的作用。各国通过基于这些原则治理军事领域的人工智能，为构建更可预测、更具韧性、更加和平的安全环境做出贡献。

具体努力可包括：

- 促进与军事领域人工智能相关的共同理解、定义和术语以及共同范围；
- 确定人道主义、法律、安全和道德操守的机遇和关切，并增进对其理解；
- 探讨透明度和建立信任措施；
- 制定原则、规范、最佳做法和其他建议；
- 为实施上述努力提供指导。

乌克兰

[原件：英文]
[2025 年 4 月 11 日]

乌克兰一直在包括军事领域在内的各项活动领域积极开发和应用人工智能。乌克兰清楚地认识到这一技术在提高人类福祉和军事能力方面的潜力，以及在民用、特别是在军事领域滥用该技术的显著危险。在俄罗斯联邦无端且无理地全面入侵乌克兰的背景下，这些风险尤其严重。在此期间，俄罗斯联邦系统地违反了战争法规和习惯以及国际人道法。

乌克兰支持并参与就负责任地开发、部署和使用民用和军用人工智能建立全球共识的国际努力。

迄今为止，除其他外，乌克兰已经签署 2023 年《布莱奇利宣言》；成为《关于在军事上负责任地使用人工智能和自主能力的政治宣言》的签署国之一，该宣言发布于在海牙举办的 2023 年军事领域负责任人工智能峰会；支持在 2023 年军事领域负责任人工智能峰会上达成的《军事领域负责任人工智能行动呼吁》，以及作为 2024 年军事领域负责任人工智能峰会成果文件通过的《军事领域负责任人工智能行动蓝图》；在巴黎举行的 2025 年人工智能行动峰会上加入了《关于发展包容、可持续的人工智能造福人类与地球的声明》；并共同提案联合国大会迄今为止关于人工智能通过的所有三项决议，包括第 79/239 号决议《军事领域的人工智能及其对国际和平与安全的影响》。

乌克兰随时准备积极参与新的全球倡议，以鼓励安全、合乎道德和负责任的人工智能发展。乌克兰还支持在包括联合国安全理事会在内的整个联合国系统内就人工智能的不同方面开展讨论。

作为一个爱好和平的国家，乌克兰对其他国家没有领土主张，同时作为俄罗斯军事侵略的受害者，乌克兰不承认对本国的任何此类主张。乌克兰发展和使用军用人工智能完全旨在通过行使《联合国宪章》规定的自卫权来加强自身防御能力。

乌克兰就军事背景下使用人工智能识别了以下对国际和平与安全的关键风险：

- 在作战和武器系统中纳入人工智能的竞争，特别是无需人工干预即可运行的完全自主武器系统的出现，可能引发新一轮更危险的全球军备竞赛，阻碍可持续发展目标的实现。
- 与其他数字技术一样，随着网络攻击威胁的不断增加，以及应用领域的日益复杂和扩大，军事系统中的人工智能越来越容易受到网络干扰和当事方的操纵，其目的是剥夺人工智能的预期应用特性和选择性使用功能。
- 过度依赖人工智能进行决策可能导致人类失去对关键军事流程的控制。
- 将未开发成熟的人工智能仓促集成到武器系统中，特别是在目标识别能力存在缺陷时，可能造成滥杀滥伤和增加平民伤亡。
- 目前缺乏控制人工智能集成武器扩散的多边框架。
- 使用人工智能集成武器而不遵守战争法规和习惯及国际人道法的行为引发严重的法律和道德关切。

大不列颠及北爱尔兰联合王国

[原件：英文]
[2025 年 4 月 11 日]

人工智能是一系列通用技术，其中任何一种技术都可以使机器执行传统上需要人类或生物智能执行的任务，尤其是当机器从数据中学习如何完成这些任务时更是如此。人工智能技术正在以非凡的速度不断成熟和得到采用。作为一组包含不同系统、方法和应用的技术，它们有着不同的发展轨迹和影响。可以肯定的是，它们具有推动社会、经济和政策(包括国防和安全)各方面转型变革的潜力。

联合王国欢迎大会第 79/239 号决议提供的机会，以审议军事领域人工智能的影响，而不仅限于已得到广泛和宝贵讨论的与致命自主武器系统有关的影响，包括根据《禁止或限制使用某些常规武器公约》设立的政府专家组正在进行的讨论。对军事人工智能更广泛的战略影响进行严格评估，将非正式和正式国际论坛上就这一议程所讨论的思想、理念和良好做法汇集起来，将有助于全面讨论如何充分利用人工智能在军事领域带来的机遇，同时有效应对相关风险。

军事领域人工智能的机遇

在军事领域纳入人工智能可能会改变国防、全球安全动态和战争特性。由人工智能启用的先进技术可以更快、更全面地对不同来源的海量数据进行分类和提炼，有助于提高效率和改善决策，加快业务规划的节奏和严谨性。情报、监视和侦察系统中的人工智能可以提供更准确的作战环境信息，并使规划者得以减少对平民的影响，从而为平民和民用基础设施提供更多保护。自主后勤和未爆弹药功能将减少对实地部署军事人员的需求。因此，在军事背景下使用人工智能可以加强国家和国际安全，降低人类生命风险并减少伤亡。

联合王国国防部关于人工智能和维持和平的研究确定了和平行动可以从人工智能增强的能力和系统中受益的方式，包括：

- 分析能力将提高态势感知、业务决策、情景规划和情感分析能力。
- 无人驾驶航空器(无人机)等自主系统可以扩大对广阔地理区域或高风险区域(维持和平人员长期驻扎可能有风险)的覆盖。
- 后勤工作能改善向当地民众提供医疗保健和援助的工作，支持任务目标并建立社区信任。

这种能力可用于加强对军备控制与和平协定的监测与核查，从而更容易及时、可信地发现违规事件或确认遵守情况。人工智能工具可以更好地探测、识别、归因和核查各种敌对的亚阈值行动，从而降低此类活动的效力，并有可能从一开始就阻止它们。它们还可以帮助实时监测和识别可能加剧紧张局势或破坏任何和平谈判或停火的网上仇恨言论、宣传或公众情绪变化。

挑战和风险

在军事背景下使用人工智能可能加剧现有风险，并在武装冲突门槛之上和之下构成额外威胁。急于采用人工智能能力来获得战略优势，可能导致各国以基于法律、道德或安全原因无法接受的方式使用人工智能。人工智能引发升级或因人工智能系统的故障或脆弱性、薄弱性、不成熟或不安全引起事故，这些新的风险将需要新的协议和降级机制。敌对行为体可能会试图攻击国家人工智能系统，破坏对其性能、安全性和可靠性的信心(例如，通过“毒化”数据源、破坏供应链中的硬件组件，以及干扰通信和指令)，这可能会在危机时刻和其他行动环境中扰乱系统并歪曲军事决策。

在冲突时期，这些技术及其带来的行动节奏可能会极大地压缩决策时间，使之几乎超出人类理解力的极限，可能需要以机器速度做出反应。许多人工智能能力的黑箱性质意味着人类通常无法辨别特定输出结果是如何或为何产生的。人工智能驱动的行动可能会导致不可预测和不透明的行为，使准确推断和判断对手意图变得困难，或者可能产生误解或引发意外后果。操作人员可能过度信任算法输出，而不完全了解人工智能系统的基本假设、制约和缺陷。如果缺乏适当的保障措施、规范和协议，人工智能驱动的系统可能会加剧误解、误判和意外升级的风险。

先进人工智能能力/工具和其他两用技术的广泛可用性可能会增加扩散危险以及国家和非国家行为体的新型武器开发。人工智能还可能被用于增强或推进虚假信息的企图以产生对国家的敌意，从而可能导致冲突和紧张局势升级。

联合王国对军事领域安全和负责任人工智能的承诺

联合王国认识到，人工智能引起了人们对公平性、偏见、可靠性以及人类责任和问责性质的深切关注，在军事背景下尤其如此。各国在采用新技术方面历史悠久，并将继续依赖于长期确立的法律、安全和监管制度，但我们必须认识到人工智能的性质带来的特殊挑战，以及积极证明我们负责任和值得信赖的重要性。

联合王国通过《国防人工智能战略》和相关的《人工智能道德原则》阐述了其对安全和负责任人工智能的承诺。这些《人工智能道德原则》载于联合王国“雄心、安全、负责”政策中，建立了道德框架的考虑因素，即以人为本、责任、理解、偏见，减少伤害以及可靠性。2024 年 11 月发布的《可靠国防人工智能》联合服务出版物，为国防部内外的团队指明了方向，指导他们如何实施这些《人工智能道德原则》，以交付稳健、可靠和有效的由人工智能启用的服务和能力。

联合王国通过《人工智能道德原则》力求培养对人工智能技术及其应用的信任，实现人机协作的全部潜力，同时减轻与其使用、滥用或停用相关的风险并防止意外后果。这种办法使联合王国能够利用整个国防和产业的创新和创造力，得以雄心勃勃地采用人工智能启用的解决方案。

联合王国政府清楚，联合王国使用人工智能增强国防流程、系统或军事能力的任何行为都受到国家和国际法的管辖。联合王国武装部队在所有各项活动中始终致力于遵守其法律义务，从就业法到隐私、采购，以及武装冲突法，也称国际人道法。联合王国武装部队有健全的做法和程序以确保其活动和人员遵守法律。这些做法和程序正在并将继续应用于由人工智能启用的能力。在武装冲突中部署由人工智能启用的能力需要完全遵守国际人道法，满足区分、必要、人道和比例四项核心原则。我们清楚，使用任何不符合这些基本原则的系统或武器都将构成违反国际法。

通过适合情境的人类参与来履行人类的责任制和问责制也至关重要。这种适合情境的人类参与对于满足我们的政策、道德原则和国际人道法义务是必要的。人类参与的性质将视能力性质、业务环境和使用背景而不同。联合王国将确保始终维持人类对核武器的政治控制。

联合王国对国际倡议的贡献

全球稳定需要有雄心但负责任的军事人工智能发展。国际社会对于在军事背景下使用人工智能的风险、保障措施和标准的认识不断发展。鉴于这些风险本质上是国际性的，需要采取全球应对措施。

联合王国一直站在国际努力的最前沿，支持安全和负责任地发展和使用人工智能。联合王国自豪地主办了首届人工智能安全峰会，就人工智能安全达成了《布莱奇利宣言》。联合王国还在委托编写《国际人工智能安全报告》方面发挥了作用。该报告发布于 2025 年 2 月，是世界首份关于先进人工智能系统的风险和能力的现有文献全面综述。该报告所建立的认识对于为利用人工智能促进和平与安全等国际讨论提供信息至关重要。我们支持在《全球数字契约》框架下，为缩小数字鸿沟、加强人工智能国际治理以造福人类所作的努力。

联合王国积极支持推动军事领域行动的国际倡议。我们支持兰德研究与发展公司欧洲部、加州大学伯克利分校和在军事领域负责任地使用人工智能全球委员会等组织的工作，汇集各种不同的、受到广泛认可的专家来探讨这些问题、理解最新思想，为决策者指明前进方向并提出可行建议。

联合王国继续积极参与关于人工智能相关国防和安全问题的国际对话，并继续分享其在制定和落实在军事领域安全和负责任地采用人工智能的办法方面的经验。联合王国欢迎通过军事领域负责任人工智能峰会(联合王国于 2024 年共同主办了该峰会)以及美国领导的《关于在军事上负责任地使用人工智能和自主能力的政治宣言》等举措取得的进展，以增进对机遇和战略风险的理解，以及对如何通过适当措施支持安全和负责任地使用人工智能来应对这些机遇和风险的理解。人工智能的道德和保证是动态的领域，需要持续的参与、协作和迭代。

展望未来

联合王国期待在现有进程迄今取得的进展基础上再接再厉，包括基于秘书长的报告在联合国开展侧重于切实行动的讨论。鉴于军事背景下人工智能的性质，必须采取包容各方的多利益攸关方办法，参考来自各国、行业，学术界和民间社会的技术、军事和法律专门知识。

尽管我们拥有大量信息，我们对军事应用和影响的集体认识仍然不足，对人工智能的性质和能力仍然存在巨大的知识差距和误解。需进一步开展工作以建设各国的能力，加强我们对军事人工智能在战略层面的影响和潜在风险及挑战的集体认识，并确立普遍商定的术语以利于建设性的讨论。讨论应侧重于有助于应对风险的切实、有效和适当措施和做法，包括保障措施和行为规范、旨在减少误解风险的新沟通渠道和透明度机制、最新的理论、建立信任的措施以及反映军事人工智能影响的军备控制协定。

B. 欧洲联盟

[原件：英文]
[2025 年 4 月 11 日]

欧洲联盟欢迎有此机会，根据大会于 2024 年 12 月 24 日通过的第 79/239 号决议，就军事领域的人工智能对国际和平与安全构成的挑战和机遇提交意见。

首先，欧洲联盟愿重申其长期立场，即军事领域的人工智能必须遵循国际法，特别是《联合国宪章》、国际人道法和国际人权法。

同样，欧洲联盟希望重申另一长期立场，即必须始终保持人类对使用武力的判断和控制。在军事领域使用人工智能方面，人类亦须保持负责和问责，确保以负责任的方式应用该技术。

欧洲联盟认识到，在军事系统中应用人工智能既产生机遇，也带来挑战。人工智能发展如此迅猛，以至于当前尚无法预判其全部优势或风险。

就此而言，欧洲联盟欢迎联合国对此事项的持续关注和在相关国际论坛内开展的讨论。在此方面，欧洲联盟特别赞赏继续开展“军事领域负责任人工智能”进程，该进程始于 2023 年在荷兰举行的首届军事领域负责任人工智能峰会，继而由大韩民国主办 2024 年军事领域负责任人工智能峰会。欧洲联盟欢迎在西班牙举行的 2025 年军事领域负责任人工智能峰会继续这一进程，并对西班牙组织下一届军事领域负责任人工智能峰会表示感谢。

欧洲联盟注意到，所有欧洲联盟成员国均已核可 2023 年《军事领域负责任人工智能行动呼吁》和2024年《军事领域负责任人工智能行动蓝图》。欧洲联盟认为，军事领域负责任人工智能峰会采取的多利益攸关方、包容性进程的理念是针对在军事上负责任使用人工智能问题的一种颇具前景的办法。在此方面，欧洲联盟认识到近期其他贡献的价值，例如国际人工智能峰会和 2025 年 2 月 10 日和 11 日由法国主办的人工智能行动峰会。欧洲联盟还肯定在《关于在军事上负责任地使用人工智能和自主能力的政治宣言》框架内开展的工作，这些工作为有关人工智能对国际和平与安全影响的广泛国际辩论做出了宝贵贡献。

欧洲联盟认为，所有欧洲联盟成员国签署的军事领域负责任人工智能的成果文件和《政治宣言》互为补充，对于进一步发展负责任的军事使用人工智能的全球思考、治理和切实可行的解决方案非常重要。

欧洲联盟认识到，在军事领域应用人工智能具有军事优势。在军事行动的速度、规模和精确度方面尤其如此。通过管理和预处理来自监测和武器系统、无人机和卫星图像的大量数据集，人工智能可以提供战术优势，使人类操作员得以更快更好地做出决策。通过改进后勤或使用预测性维护管理改善设备维护，人工智能应用可以降低成本。同样，在不确定的环境中，人工智能可以提供更远距离和更精确的军事行动。

与此同时，军事领域应用人工智能的速度和规模优势本身也带来了挑战。人工智能加速了观察-定向-决策-行动环。由于军事意图与人工智能驱动的系统生成的分析之间存在不一致，速度和规模的能力提升可能会造成误解。因此，人工智能可能在无意中导致升级。速度也对保持人类对使用武力的判断和控制这一目标构成挑战。

在此背景下，欧洲联盟强调必须开展国际合作，研究军事领域人工智能的影响和潜在治理框架。

Annex II

Replies received from international and regional organizations, the International Committee of the Red Cross, civil society, the scientific community and industry[1]

A. International and regional organizations

African Commission on Human and Peoples' Rights

[11 April 2025]

I. Introduction

The African Commission on Human and Peoples' Rights (the African Commission), as the premier treaty-based human and peoples' rights body of the African Union (AU), is entrusted with the mandate of promoting and protecting human and peoples' rights in Africa under the African Charter on Human and Peoples' Rights (African Charter). In the African Commission's study on Addressing Human Rights Issues in Conflict Situations, the African Commission's Focal Point who led the study observed that 'it is … in conflict and crisis situations that the most egregious violations and abuses of rights are perpetrated…With the changes in the nature of conflicts and the attendant heightened threat to human and peoples' rights, there is a greater need for the human rights system to pay increasing attention to and provide effective responses to the challenges that these new dynamics present to the protection and observance of rights.' In the current context, one of the major new dynamics that carries serious implications for peace and security and therefore human and peoples' rights relate to Artificial Intelligence (AI) and in particular its rapid development and use in the military domain.

During its 1214th meeting, the AU Peace and Security Council (PSC), in requesting the AU Commission to conduct a study to assess the adverse impact of AI on peace and security, underscored the necessity of ensuring African perspectives in shaping global AI governance frameworks. Against this background and having regard to its work on AI and other technologies and human and peoples' rights[2] and human rights in peace and security, the African Commission is pleased to share its views in response to the invitation of the Secretary-General for submission of inputs on AI in the military domain and its implications for international peace and security.[3]

II. AI in the military domain and peace and security

The development and use of AI technologies in the military domain particularly to automate military functions such as surveillance, targeting, and the deployment of

[1] In accordance with operative paragraph 8 of General Assembly resolution 79/239, the replies received from international and regional organizations, the International Committee of the Red Cross, civil society, the scientific community and industry are included in the original language received. The Secretary-General remains committed to multilingualism as a core value of the United Nations.

[2] Resolution ACHPR/Res. 473 (EXT.OS/ XXXI) 2021 on human and peoples' rights and artificial intelligence (AI), robotics and other new and emerging technologies in Africa, available at https://achpr.au.int/en/adopted-resolutions/473-resolution-need-undertake-study-humanand-peoples-rights-and-art.

[3] The Focal Point of the African Commission on its study on human and peoples' rights and AI, robotics and other technologies acknowledges with appreciation the contribution of Professor Thompson Chengeta, who is the consultant providing technical assistance in the development of the study, through the Centre for Human Rights, University of Pretoria.

lethal force have far reaching consequences for peace and security and hence for human and peoples' rights. The AU Continental AI Strategy, endorsed during the 44th Extraordinary Session of the Executive Council of the African Union, highlights AI governance and regulatory challenges, particularly in military applications, warning that AI could exacerbate conflicts through inaccurate predictions or deployment of autonomous weapon systems. Additionally, the framework raises concern about disinformation, misinformation, cybersecurity threats, and military risks.

From the perspective of the development and use of AI in the military domain, peace and security should not be seen just from the perspective only of what it means for stability of states and societies. Beyond its conception under the UN Charter and public international law associated with friendly relations of states, peace and security is also a fundamental right of all peoples. The African Charter thus stipulates that 'All peoples shall have the right to national and international peace and security. The principles of solidarity and friendly relations implicitly affirmed by the Charter of the United Nations and reaffirmed by that of the Organization of African Unity shall govern relations between States.'[4]

The framing of peace and security as a right of peoples compels states to assess and govern the development and deployment of AI technologies in the military domain through a human rights lens that prioritises the prevention of harm, suffering, and injustice. Together international law conception of peace and security, it places an affirmative duty on states to ensure that AI systems do not contribute to conflict, perpetuate structural inequalities, or violate the rights and dignity of individuals and communities. By embedding peace and security within the framework of human rights, states are not only accountable for avoiding direct acts of aggression, but also for proactively creating and maintaining environments in which human flourishing, security, and justice are protected from the potentially disruptive or harmful impacts of emerging military technologies.

The implication of AI in the military domain to peace and security, farmed comprehensively, thus goes beyond how it shapes the obligation of states for non-aggression. It also covers how algorithm-driven systems may dehumanise individuals, introduce bias, and lead to unaccountable or disproportionate harm. It raises critical questions about the erosion of human oversight, the potential for unlawful killings or violations of international humanitarian law, and the targeting of vulnerable or marginalised populations.

By transforming military capabilities, the application of AI in the military domain can also have implications for peace and security by heightening tendencies for engaging in hostilities. The resultant escalation of tension and violence will be inimical not only to stability and peace between and within states but also most importantly carries more adverse consequences for the development needs of the less developed parts of the world such as Africa. While AI may contribute to advancing the development needs of Africa, its development and use in the military domain can have devastating consequences for development detrimental in particular to the right to development enshrined in Article 22 of the African Charter.[5]

This link between peace and development is also central to the Sustainable Development Goals (SDGs), especially SDG 16, which promotes peace, justice, and strong institutions. Without peace and security, sustainable development cannot be achieved. Recognising this link is critical in the governance of military AI, as the

[4] Article 23(1) of the African Charter.

[5] All peoples shall have the right to their economic, social and cultural development with due regard to their freedom and identity and in the equal enjoyment of the common heritage of mankind.

militarisation of AI can aggravate instability, particularly in fragile regions, and undermine Africa's developmental aspirations. By reaffirming the interconnectedness of peace and development, the African Commission calls for a governance approach that upholds peace as both a human right and a developmental imperative.

III. The need for a human and peoples' rights-based regulation of the development and use of AI in the military domain

Given the ways in which the use of AI in the military domain transforms the conduct of hostilities and how the development of AI relies on the extraction of natural resources particularly critical minerals such as rare earth minerals, it is the submission of the African Commission that both the process of extraction of resources in the development of AI in the military domain and the use of AI in the military domain need to be in full compliance with human and peoples' rights standards and international law principles, including international humanitarian law.

First and foremost, it is of paramount significance that the development and use of AI in the military domain complies with the right to peace and security enshrined in Article 23 of the African Charter on Human and Peoples' Rights. As a right that is born out of the recognition of the inseparability of the enjoyment of other human rights states from peace and security, this right entails that the use of AI in the military domain should be consistent with the international law prohibition of the use of force enshrined in the UN Charter and the Constitutive Act of the African Union.

Second, the use of AI technologies in conflict settings need to ensure respect for applicable human and peoples' rights and international humanitarian law principles, including most notably needs to adhere to the principles of precaution, necessity, distinction, proportionality and legitimacy. These requirements apply irrespective of whether the context in which the use of AI in the military domain relates to international armed conflicts or non-international armed conflicts. As established in the African Commission's study,[6] parties to conflict are obliged to observe human rights standards where such conflicts do not meet the IHL threshold of armed conflict. As such, those who use AI technologies in conflict situations that do not meet the IHL threshold of armed conflict are legally obliged to respect and ensure respect for the human and peoples' rights standards established under treaty and customary international human rights law.

Third, the development of AI in the military domain and the use AI technologies in hostilities need to comply with the principle of transparency. This is fundamental because it is the basis for ensuring effective regulation of the development and use of AI in the military domain and for compliance with applicable human rights and international law standards. Additionally, transparency is critical for ensuring compliance with the obligation for respecting the dignity, privacy and data protection of individuals. The principle of transparency is also a pre-requisite for addressing some of the concerns that arise from use of AI in the military domain including bias (owing to the source and type of data used) and explainability. Transparency is also critical not only with the development of AI in the military domain but also with respect to the transfer of AI technologies in the military domain.

Fourth, from the perspective of human and peoples' rights and IHL, the other standard key to human rights and international law-based regulation of the development and use of AI concerns accountability. In the event of the occurrence of violations of human and peoples' rights standards or IHL principles from the development and use of AI in the military domain, there has to be both institutional and individual accountability. Accountability in this instance encompasses not only

[6] ACHPR, Addressing human rights issues in conflict situations, https://achpr.au.int/en/node/895.

the measures that are taken against perpetrators but also the remedial steps that need to be put in place for redressing victims.

Firth, building and sharing of technical knowhow critical to ensuring regulation by states is the other principle. Recent developments including the jamming of GPS systems affecting flights reported in Eastern DRC and the deployment by the Islamic State of West Africa of armed drones, highlight not only the need for effective regulation but also the need for developing the requisite infrastructure and technical capacity for ensuring effective regulation.

IV. The link between the development of AI in the military domain and Africa's natural resources and its implications for peace and security

The African Commission is also of the view that when discussing peace and security, stakeholders must be aware of the link between development of military AI, Africa's natural resources – particularly critical minerals – and the notion of peace and security. Article 21(1) of the African Charter on Human and Peoples' Rights affirms: "All peoples shall freely dispose of their wealth and natural resources. This right shall be exercised in the exclusive interest of the people. In no case shall a people be deprived of it."[7] Article 21(5) further provides that "States parties to the present Charter shall undertake to eliminate all forms of foreign economic exploitation particularly that practised by international monopolies so as to enable their peoples to fully benefit from the advantages derived from their national resources."[8]

This provision is particularly important in the context of military AI, which depends heavily on critical minerals such as cobalt, lithium, and rare earth elements – resources abundantly found in Africa. The 2924 Report of the Chairperson of the African Commission's Working Group on Extractive Industries, Environment and Human Rights Violations, stressed the "significance of critical minerals for new and emerging technologies" and highlighted that Africa has been burdened by a "resource curse phenomenon."[9] The report of the Chairperson noted that "extraction of minerals and other resources not only fuels but also at times becomes the site where contestation over whose control and use triggers conflicts. In some instances, this has created a vicious cycle of insecurity and violence, a condition that not only leads to major human and peoples' rights violations but also the perpetuation of a vacuum of effective governance and the concomitant exploitative, socially and environmentally costly extraction of the resources of the continent."[10]

Therefore, governance of military AI must not only ensure the legal use of force but also address the exploitative chains of extraction that power such technologies. This requires strict oversight, equitable benefit sharing, and regional solidarity to prevent Africa's resources from being used to fuel further conflict and inequality.

V. Conclusion

The African Commission is of the view that the development and use of AI in the military domain carries far reaching consequences for international peace and security in general and for less developed parts of the world such as in Africa that historically suffered violations and remain vulnerable to the adverse impacts of the development and use of AI in the military domain without robust and effective legal regime for such development and use in the military domain. The African

---

[7] Article 21(1) of the African Charter.

[8] Article 21(5) of the African Charter.

[9] African Commission's Working Group on Extractive Industries, Environment and Human Rights Violations (2024), https://achpr.au.int/en/intersession-activity-reports/extractive-industries-environment-and-human-rights-violations (accessed 08 April 2025).

[10] As above.

Commission affirms that the development and use of AI in the military domain needs to be regulated on the basis of international law, human and peoples' rights and international humanitarian law standards with particular regard to the development and peace and security interests and human and peoples' rights needs of less developed parts of the world.

More specifically, beyond and above the right to peace and security, the governance of AI in the military domain needs to ensure respect for applicable human and peoples' rights and international humanitarian law principles, including most notably needs to adhere to the principles of precaution, necessity, distinction, proportionality and legitimacy, the principles of transparency, accountability and redress for victims and the obligation to build and share technical knowhow necessary for enabling societies to avert the risks that the development and use of AI in the military domain carries for peace and security. Only by ensuring that the development and use of military AI are aligned with international legal standards including those relating to the right to peace and security, the right to development, the right to privacy and protection of personal data, the right to remedy and the responsibility for exercising human control, the right to and control over natural resources and by addressing the structural inequities underpinning global technological advancement, can states uphold their duties to their peoples and advance genuine peace, justice, and security in relation to the development and use of AI in the military domain.

B. International Committee of the Red Cross

[19 March 2024]

Summary

The full submission is available at: https://www.icrc.org/en/article/artificial-intelligence-military-domain-icrc-submits-recommendations-un-secretary-general.

The International Committee of the Red Cross (ICRC) welcomes the opportunity to submit its views for consideration by the United Nations Secretary-General, in accordance with resolution 79/239.

The recommendations that the ICRC makes in this submission are in line with its long-standing mandate and practice of promoting respect for and the development of IHL, including its application to new technologies of warfare. This submission is intended to support States in ensuring that military applications of AI comply with existing legal frameworks and, where necessary, identifying areas where additional legal, policy, or operational measures may be required.

1. Normative proposals: Reaffirming existing IHL as the starting point

The ICRC has consistently emphasized that, while IHL does not explicitly prohibit or regulate the use of AI in military applications, it does restrict its development and use, and places strict constraints on AI when it is integrated into weapon systems or used in some way to conduct warfare.[1]

Existing and emerging normative proposals on the military application of AI should build upon established international legal frameworks and mechanisms, including IHL. Where necessary, these frameworks can be reinforced through the development of additional legal instruments, operational guidance or policy measures to address specific risks or challenges posed by emerging technologies. The form and content of such measures may vary depending on the specific use case. The ICRC

[1] This has also been affirmed by States, including in the UN General Assembly with Resolution 79/239.

encourages the international community to engage in concrete discussions on particular applications of AI in the military domain and to prioritize consideration of those that pose the greatest risks to people affected by armed conflicts.

2. A Human-centred Approach to military AI

In line with the resolution, the ICRC advocates for a human-centred approach to the development and use of AI in armed conflict.[2] This approach has at least two key dimensions: first, ensuring a focus on the humans who may be affected by the use of AI; and second, emphasizing the obligations and responsibilities of the humans using or ordering the use of AI in military operations.

Despite the growing development of AI-related technologies in the military domain, IHL requires individuals to make legal determinations. Humans must, for instance, determine the lawfulness of attacks that they plan, decide upon or execute, and they remain accountable for those determinations. The ICRC considers that human judgement is crucial for reducing humanitarian risks, addressing ethical concerns and ensuring compliance with IHL. Accordingly, while certain technical tasks may be carried out by machine processes, it is not the system itself that must comply with the law, but the humans using it.[3]

This does not mean that commanders and combatants cannot or should not use tools, including AI-decision-support systems. However, these tools must only be designed and used to support, rather than hinder or replace, human decision-making.[4] Further, States and parties to armed conflicts must ensure that human control and judgement are preserved in decisions that pose risks to the life and dignity of people affected by armed conflict. This is essential for ensuring respect for applicable laws, including IHL, and upholding ethical standards.[5]

3. Specific Applications of ai in the military domain

The ICRC has identified three specific applications of AI in the military domain that pose particularly significant risks to those affected by armed conflict:

1. AI in Autonomous Weapon Systems

Resolution 79/239 acknowledges the increasing integration of AI into weapons and weapon systems, a development that raises significant legal and humanitarian concerns. The integration of AI, particularly machine learning (ML) techniques, into autonomous weapon systems (AWS) exacerbates existing challenges posed by AWS in ensuring compliance with IHL. In particular, it increases difficulties for human users to understand, predict, and control the system's functioning and effects.

Users of AWS must be able to, with a reasonable degree of certainty, predict the effects of that weapon in order to determine whether it can be directed at a specific military objective, and take steps to limit those predicted effects, as required by IHL. This entails the ability to understand the functioning of the AWS: the nature and functioning of its sensors, the definition of its target profile and the potential effects in the circumstances of use, including any risk of error or malfunction. This is

[2] ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021).

[3] ICRC, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflicts (IHL Challenges Report), 2024, p. 61.

[4] *Ibid*.; ICRC, IHL Challenges Report – Chapter 2: Contemporary and future challenges in the conduct of hostilities, 2019, p. 32.

[5] ICRC, Decisions, Decisions, Decisions: computation and Artificial Intelligence in military decision-making, ICRC, 2024, p. 8.

particularly relevant for AWS that function in opaque ways (the "black box" challenge), such as AWS relying on AI techniques, which prevent the human user from being able to understand, predict or explain the system's output. This impossibility effectively results in a lack of control over the weapon's effects, rendering it indiscriminate by nature.

In this regard, we reiterate the joint call made by the ICRC President, with the UN secretary-general,[6] for new, legally binding rules prohibiting certain AWS and constraining the use of others.[7] In particular, we recommend a prohibition on

- unpredictable autonomous weapons – those that, due to their design or the circumstances and manner of use, do not allow a human user to understand, explain or predict the system's functioning and effects;
- autonomous weapons designed or used to target humans directly. This is required because of the significant risk of IHL violations and the unacceptability of anti-personnel autonomous weapons from an ethical perspective.[8]

The ICRC supports all efforts by States to urgently adopt a legally binding instrument to regulate AWS, in whichever forum they choose.[9] The integration of AI into AWS should also be considered when discussing normative proposals on military applications of AI. Doing so is essential to ensure a consistent and comprehensive approach to the regulation of military AI, to avoid normative gaps, and to effectively address the serious legal, ethical, and humanitarian risks that are exacerbated by the integration of AI into AWS. In this regard, the ICRC considers it important that binding prohibitions and restrictions on AWS, including AWS that incorporate AI, are integrated into broader discussions on the governance of military AI.

2. AI in Military Decision-Making

AI decision-support systems (AI-DSS) are computerised tools that bring together data sources – such as satellite imagery, sensor data, social media feeds or mobile phone signals – and draw on them to present analyses, recommendations and predictions to decision makers.

The use of AI-DSS raise concerns related to system functioning, data quality, and human-machine interaction. These systems risk increasing the rate of unforeseen errors, perpetuating problematic biases – particularly those based on age, gender, ethnicity, or disability, and making it difficult for the users to understand how and why the system generates its output from a given input.

Generally, AI-based systems will perform better when given well-defined goals and access to representative and high-quality data. However, armed conflict environments are marked by uncertainty, volatility, and deliberate deception techniques by adversaries, which makes it extremely difficult to obtain reliable or transferable data. Even where good data exists, it may not reflect the specific operational or humanitarian dynamics of a particular context.[10] Moreover, for AI systems that rely on training data, the utility of those data can rapidly diminish once

---

[6] ICRC, Joint call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to establish new prohibitions and restrictions on Autonomous Weapon Systems, 2023.

[7] ICRC, ICRC Submission on AWS to the UN Secretary-General, 2024, p. 6.

[8] *Ibid*.

[9] *Ibid*.

[10] ICRC, IHL Challenges Report, 2024, pp. 64–65; ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021); ICRC, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, ICRC, 2024, pp. 31 and 54.

a conflict begins. Parties to armed conflicts will continuously seek to maintain the initiative and operate in a manner that is not anticipated by their adversary, adapting their strategies and tactics accordingly. This can fundamentally alter the environment in which the system was expected to operate, making the original data no longer representative of the new operational conditions. In such cases, the system's outputs may become unreliable, and the AI model may require re-evaluation or retraining in order to remain fit for purpose.

Human interaction with these systems raises further concerns, such as "automation bias" – a propensity to rely on machine outputs even when other available information may call those outputs into question – which is particularly pronounced in high-pressure or stressful environments like in armed conflicts.[11] Taken together, these factors can hamper a user's ability to scrutinize the information available. The practical consequence might be, for instance, that someone plans, decides upon or launches an attack based solely on an AI-DSS's output, thereby effectively serving as a human rubber stamp rather than assessing the lawfulness of the attack by considering all the information reasonably available including the AI-DSS output.[12]

On the positive side, the careful use of AI-based systems may facilitate quicker and more comprehensive information analysis, which can support decisions in a way that enhances IHL compliance and minimizes risks for civilians. In the context of urban warfare in particular, the ICRC has recommended that online open-source repositories should be used to gather information about the presence of civilians and civilian objects.[13] Importantly, IHL imposes obligations to take constant care to spare the civilian population and to take all feasible precautions in attack. Therefore, in developing and using AI-DSS, armed forces should be considering not only how such tools can assist them to achieve military objectives with less civilian harm, but also how they might be designed and used specifically to protect civilians. However, the important point is that these computer outputs can inform but must not displace the need for legal determinations.

Beyond targeting decisions, militaries are also exploring the use of AI to support other operations traditionally carried out by humans, including detention operations. While technology deployed responsibly and with robust human oversight can contribute to IHL compliance, it also carries risks including bias, lack of transparency, and faulty programming and analysis, all of which can undermine compliance with IHL.[14]

To support efforts by States and other actors to ensure that military uses of AI-DSS remain consistent with IHL and humanitarian principles, the ICRC has formulated a non-exhaustive set of preliminary recommendations relating to the development and use of AI-DSS in armed conflict. They focus on 1) ensuring human control and judgement; 2) system design requirements; 3) testing, evaluation, verification and validation; 4) legal reviews; 5) operational constraints on use; 6) user training; 7) after-action reviews; and 8) accountability, among others. The recommendations are annexed to the full version of this submission.

---

[11] ICRC and the Geneva Academy, Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts: Current Developments and Potential Implications, ICRC, 2024, p. 17.

[12] ICRC, IHL Challenges Report, 2024, p. 65.

[13] *Ibid*., p. 66; ICRC, Reducing Civilian Harm in Urban Warfare: A Handbook for Armed Groups, 2023, p. 15.

[14] ICRC, IHL Challenges Report, 2024, p. 22.

3. AI in Information and Communications Technologies

AI is expected to change how actors defend against and conduct information and communications technology (ICT) activities, including in armed conflict. In particular, States have noted with concern that the use of AI and other emerging technologies in malicious ICT activities may further increase their scale and speed, as well as the harm they may cause.[15] For example, AI enables tools to identify and develop exploits for new vulnerabilities in software or networks, or to conduct harmful ICT activities autonomously, whether in offence or in defence. The ICRC is concerned that this could increase the risks of indiscriminate attacks, incidental civilian harm, including damage to critical civilian infrastructure, as well as the uncontrolled escalation of conflict, particularly in complex and interconnected digital environments.[16]

Similarly, information or psychological operations are not a new feature of armed conflicts; however, AI is changing how information is created and spread. AI-enabled systems, particularly generative AI, have been widely used to produce harmful content – text, audio, photos and video – which is increasingly difficult to distinguish from authentic, original content.[17] The ICRC is concerned about the consequences for civilians that might result from the creation and spread of such information through ICT, including information that contributes to or encourages violence, causes lasting psychological harm, undermines access to essential services or disrupts the operations of humanitarian organizations.

In light of these concerns, the ICRC underlines the importance of applying existing international law, including IHL, to the use of AI in ICT activities. The ICRC urges States to ensure that the development and use of AI-supported ICT activities respect the protections afforded to civilians and civilian infrastructure in armed conflict. Moreover, in light of the emergence of increasingly autonomous ICT capabilities, the ICRC further encourages States to address the serious challenges posed by these tools, particularly by considering whether existing international law, including IHL, provides sufficient safeguards against the harm such tools can cause, or whether additional limits are needed.

4. Conclusion

The ICRC is grateful for the opportunity to share its above views and recommendations on ways to address the challenges and concerns raised by AI for the secretary-general's consideration, and stands ready to contribute further to assist States in taking effective action to address the risks posed by AI applications in the military domain.

C. Civil society

Autonorms

[10 April 2025]

The following is the AutoNorms project's submission pursuant to Resolution 79/239 on "Artificial intelligence in the military domain and its implications for international peace and security" adopted by the United Nations General Assembly on 24 December 2024. The resolution requests the UN Secretary-General to seek

[15] 34th International Conference of the Red Cross and Red Crescent, Resolution 2 "Protecting civilians and other protected persons and objects against the potential human cost of ICT activities during armed conflict", 2024.

[16] ICRC, IHL Challenges Report, 2024, pp. 66–67.

[17] *Ibid*., pp. 58–59.

views, including those of Member States, civil society, the scientific community and industry, on "opportunities and challenges posed by the application of artificial intelligence in the military domain, **with specific focus on areas other than lethal autonomous weapons systems**". The AutoNorms team welcomes the opportunity for representatives of academia to submit their views on this important and timely topic.

The AutoNorms project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 852123). Led by Professor Ingvild Bode and hosted by the Center for War Studies at the University of Southern Denmark, the project examines how the integration of artificial intelligence (AI) technologies into weapon systems and military targeting shapes international norms governing the use of force.[1]

Introduction

Over the past 2-3 years, the international debate about applications of AI in the military domain has been characterized by two significant, near-simultaneous changes. First, there has been a **move away from its predominant focus on autonomous or AI technologies in weapon systems** towards considering AI technologies across a wider range of military decision-making tasks, especially in relation to targeting. To reflect his move, this submission focuses on the employment of **AI-based decision support systems (AI DSS),** or systems that are meant to be used as tools to directly or indirectly inform the complex process of use-of-force decision-making, for example, by analyzing large volumes of data, recognizing patterns within the data, predicting scenarios, or recommending potential courses of action to human decision makers.

Second, there has been a **growing emphasis on human-machine interaction** in the context of using AI in the military domain.[2] This emphasis results from the broad recognition that, even when humans are 'in' or 'on' the loop of targeting decision-making, they need to exercise a sufficient level of oversight, control, and agency over the targeting process. Human oversight is a governance principle featuring prominently across various international initiatives, including A/RES/79/239. However, **dynamics of human-machine interaction as part of the use of AI DSS both introduce new issues and solidify existing sets of challenges that require governance attention.** Our submission highlights these challenges and the need to ensure the exercise of human oversight and agency throughout the full targeting decision-making spectrum. It is structured in three parts, starting with explicating challenges of human-machine interaction, then commenting on the relative under-development of the international debate about AI DSS, and finally, sketching a way forward.

Challenges of human-machine interaction in the use of AI DSS

The use of AI DSS involves various dynamics of human-machine interaction because military personnel such as operators and intelligence analysts routinely and increasingly interact with a network of AI systems throughout the targeting process. These interactions involve multiple challenges **which have the potential to affect**

[1] The members of the AutoNorms team are Professor Ingvild Bode, Dr Hendrik Huelss, Dr Anna Nadibaidze, Dr Guangyu Qiao-Franco, and Dr Qiaochu Zhang. The AutoNorms project is based at the Center for War Studies, University of Southern Denmark, Odense, Denmark. For more information, please visit our website: www.autonorms.eu.

[2] Ingvild Bode and Anna Nadibaidze, "Symposium on Military AI and the Law of Armed Conflict: Human-Machine Interaction in the Military Domain and the Responsible AI Framework," *Opinio Juris*, April 4, 2024, https://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-human-machine-interaction-in-the-military-domain-and-the-responsible-ai-framework/.

the exercise of human agency, or humans' capacity to understand a system's functions and its effects in a relevant context; deliberate and decide upon suitable actions in a timely manner; and act in a way where responsibility is guaranteed.[3]

Dynamics of human-machine interaction result in **distributed agency between humans and AI systems, where they are not separated into two distinct entities but rather form part of a socio-technical system**.[4] As part of this system, both sides may influence each other in different ways, which then translate into various forms of distributed agency located along a spectrum. In some instances, dynamics of human-machine interaction will offer more opportunities for exercising human agency in targeting decisions. In other instances, however, the humans involved in use-of-force decision-making will be more constrained in their ability to exercise agency.

For example, **humans' ability to exercise agency might be limited by cognitive biases such as automation bias or anchoring bias**. Humans could over-trust AI DSS even when knowing that there might be malfunctions or unintended errors involved, risking an overreliance on algorithmic outputs without engaging in the critical deliberations and assessments that are needed to exercise human agency, especially in critical targeting decisions that might inflict death, destruction, and severe harm. Such biases are typically exacerbated by the increased speed of AI-assisted military decision-making, especially in contexts where there are high levels of pressure to act rapidly. They can also be exacerbated by AI DSS that are used for prescription or recommendations, because such systems restrict the options or courses of action available to human decision makers.

Moreover, given that AI DSS are likely to be employed not individually but rather as part of a network of systems, the increased complexity of interactions can result in situations where humans act upon some outputs suggested by AI DSS, but do not overall exercise a high quality of agency. Due to these and many other concerns related to interactions between humans and AI DSS, **there is a need to further investigate challenges of human-machine interaction that result in AI DSS not positively 'supporting' humans but rather undermining humans' ability to exercise agency**.[5]

The risks of not addressing challenges of distributed agency are substantial. First, situations where humans are restricted in their exercise of agency **raise questions about compliance with international humanitarian law**, which requires that humans be held accountable and legally responsible for violations of legal principles. Although humans remain officially in control of the selection and engagement of targets, there are concerns about the exact role played by humans in context of using AI DSS in practice.

Second, these concerns also extend to the risk of **negatively affecting moral agency and responsibility in warfare**. Challenges of human-machine interaction that result in distributed agency would allow humans to feel less morally responsible

---

[3] Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang, *AI in Military Decision Support Systems: A Review of Developments and Debates* (Odense: Center for War Studies, 2024), https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates/.

[4] Ingvild Bode, *Human-Machine Interaction and Human Agency in the Military Domain*, Policy Brief No. 193 (Waterloo, ON: Centre for International Governance Innovation, 2025), https://www.cigionline.org/publications/human-machine-interaction-and-human-agency-in-the-military-domain/.

[5] Anna Nadibaidze, "Do AI Decision Support Systems 'Support' Humans in Military Decision-Making on the Use of Force?" *Opinio Juris*, November 29, 2024, https://opiniojuris.org/2024/11/29/do-ai-decision-support-systems-support-humans-in-military-decision-making-on-the-use-of-force/.

for decisions that could affect other people's lives. They also risk making the human role a nominal, 'box-checking' exercise which can *de facto* be compared with AI DSS playing an 'autonomous' role because the human role is substantially reduced.

Third, there are **security and operational risks related to distributed agency dynamics**, especially when they give too prominent roles to AI DSS and algorithmic outputs. AI systems often malfunction, are trained on biased sets of data which do not apply beyond the training context or specific contexts of use, as well as integrate assumptions that might not be strategically or operationally beneficial.

Various types of biases, issues of trust, uncertainties, targeting and military doctrines, political and societal contexts in which AI DSS are used – all these aspects **can lead to dynamics of distributed agency which limit the exercise of human agency and prioritize algorithmic outputs**. It is important to investigate these dynamics and ensure that distributed agency provides more opportunities than limitations to human decision makers in warfare.

Relative under-development of the international debate on AI DSS

Despite increasing reports about the use of AI DSS in recent and ongoing armed conflicts, and the significant challenges and risks they pose to the effective exercise of human agency, **the international debate on human-machine interaction in the use of AI DSS remains insufficiently developed**, particularly within intergovernmental UN settings. Current discussions on AI in the military domain, including those within the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (GGE on LAWS), have focused on the use of AI at the tail-end of the targeting process, specifically autonomy and AI in weapon systems. This narrow focus risks overlooking or failing to address critical normative, legal, ethical, security, and operational risks that can proliferate and compound throughout the entire targeting decision-making process.

An increasing, albeit still limited, number of stakeholders are raising this issue at international multistakeholder forums, such as the Summits on Responsible Artificial Intelligence in the Military Domain (REAIM). Some international non-governmental organisations and research institutes – such as the International Committee of the Red Cross (ICRC), the Stockholm International Peace Research Institute (SIPRI), the UN Institute for Disarmament Affairs (UNIDIR), and the Asser Institute – have initiated discussions on challenges posed by AI in the military domain, beyond the issue of autonomy in weapon systems. Despite this progress, there remains **a clear need to develop a more comprehensive and inclusive international multistakeholder debate to guide the responsible development and deployment of AI DSS in military contexts.**

Way forward

In closing, we sketch three ways intended to move the international debate about applications of AI in the military domain forward:

1. **Increase awareness for the implications of practices of designing, developing, and using AI DSS**. States and other stakeholders across industry, civil society, and academia engaged in the governance, development, and use of AI DSS for military targeting must consider the implications of their practices. These practices influence what counts as 'appropriate' ways of considering and employing AI DSS and thereby shape what becomes the accepted, requisite quality of human oversight and agency exercised over the whole process of use-of-force decision-making. To increase such awareness, the debate pursuant to A/RES/79/239 at the UNGA First Committee should centrally focus on the issue of AI DSS in the military domain.

2. **Consistently map both ‘best’ practices and ‘problematic’ practices associated with the design, development, and use of AI DSS**. To get a better sense of the direction that the design, development, and use of AI DSS take, states and other stakeholders need to closely map their own (and others’) practices. While there have been some limited efforts to exchange potential best practices, we also need to be attentive to practices with potentially problematic effects. This should encompass practices exercised across the full life cycle of AI systems from development to use and post-use review. Mapping such practices would offer stakeholders a better overview of which practices may be beneficial, i.e., provide opportunities for the exercise of human agency, and which practices may be problematic, i.e., limit the exercise of human agency, and therefore assess the desirability of particular practices.

3. **Pursue the debate on AI DSS within a multistakeholder format**. States should work with diverse stakeholders – including academics across social sciences and technical disciplines, civil society representatives, and international organizations – to develop normative guidance and regulation, especially regarding the human role in military decision-making. Moreover, top-down processes towards governing AI DSS should be accompanied by a bottom-up, standard-setting process focused on establishing operational standards. Such an inclusive approach could strike a balance between national security and humanitarian concerns, while reinforcing the need to ensure that humans can exercise agency in use-of-force decisions.

Global Commission on Responsible Artificial Intelligence (AI) in the Military Domain

[11 April 2025]

1. Introduction

The Global Commission on Responsible Artificial Intelligence (AI) in the Military Domain (GC REAIM) welcomes the opportunity to contribute to the United Nations Secretary-General’s report pursuant to resolution A/RES/79/239.

GC REAIM recognises that military applications of AI present both opportunities and challenges for global peace and security. Accordingly, the establishment of responsible and ethical governance – consistent with States’ obligations under applicable international law – is essential. The global community must take proactive steps to ensure that military AI is developed and deployed in a manner that de-escalates rather than escalates conflicts; respects and enhances, rather than compromises, the sovereignty and territorial integrity of states; promotes rather than threatens the security and safety of civilians; constrains and supports rather than erodes the existing rules-based international order.

In line with GC REAIM’s resolute commitment to advancing international governance efforts, this note outlines some of the – non-exhaustive – views expressed by GC REAIM Commissioners and Experts on the implications of AI in the military domain to peace and security. The views presented are general in nature and will be further elaborated in the forthcoming GC REAIM report. While the Commission plans to present substantive and actionable recommendations for stakeholders in September 2025, this note does not yet include concrete proposals. As discussions among Commissioners and Experts are still ongoing, it instead highlights some of the key opportunities, challenges, benefits and risks posed by AI in the military domain to peace and security.

2. Technological Foundations

GC REAIM holds that meaningful policy deliberations on AI in the military domain must be grounded in a shared, foundational understanding of the underlying technologies and their potential trajectories. The complexity of AI technologies often gives rise to misunderstandings, inflated expectations, or misguided applications. Consequently, it is imperative to demystify AI through formal and well-defined frameworks that distinguish between current capabilities and speculative future developments. To support this objective, GC REAIM is developing a taxonomy which seeks to map the full spectrum of AI applications across military and broader peace and security contexts. The taxonomy differentiates between the implications of AI in operational activities – such as warfighting and intelligence – and administrative activities – such as logistics and personnel training and helps identify the specific applications of AI that should be prioritised in governance deliberations.

In its approach to the creation of a taxonomy, GC REAIM highlights the need for and contributes to a concerted effort to clarify, standardise, and encourage the accurate use of technical language with different layers of abstraction for policymakers, experts, and the public, thereby enhancing transparency, mutual understanding, and public trust. GC REAIM also cautions against the uncritical multiplication or adoption of new terminologies in AI governance discourse, unless these are clearly defined; and to ensure such terms are not used to circumvent or obscure existing legal obligations. Precision and consistency in language are the basis of responsible AI governance.

3. Implications for Peace, Security, and Stability

GC REAIM recognises that the integration of AI into the military domain presents benefits as well as both foreseeable and unforeseeable risks to international peace and security. A balanced approach to the range of opportunities and challenges emerging throughout the AI life cycle lies at the core of GC REAIM's method and is essential for responsible AI governance.

AI in the military domain may contribute to international peace and security in several important ways. At the developmental stage, the advancement of military AI capabilities may act as a deterrent to violence, as the mere development and presence of advanced technologies by responsible actors can encourage restraint by aggressors. Military AI may enhance early warning systems, strengthening conflict prevention strategies, and supporting arms control verification through AI-driven tools that foster transparency, trust, and cooperation among states – fundamental elements in conflict prevention. AI can also bolster national security and defence by improving the precision, accuracy, and efficiency of intelligence analysis and situational awareness, enabling real-time threat detection, and facilitating more efficient counterterrorism operations through predictive analytics and autonomous systems. AI-powered systems can rapidly process vast amounts of complex data, enabling military forces to make timely, informed decisions that may prevent escalation and support conflict de-escalation efforts. These traits can also help improve targeting accuracy and precision, potentially reducing the risk of collateral damage or fratricide – attacks on one's own forces – and aiding compliance with International Humanitarian Law (IHL) to protect the security of protected persons, such as civilians and non-combatants, during armed conflict. Military AI may also reduce certain forms of human bias and enhance accountability by providing precise data, surveillance, and real-time monitoring, enabling clear attribution of actions to specific actors. In these ways, AI offers meaningful opportunities to reinforce adherence to international law and ethical standards, strengthening the normative foundation of the rules-based international order underpinning global peace and security.

AI in the military domain also presents a range of risks. As with the development of other general-purpose technologies, the development of AI in the military domain may accelerate arms races. AI technologies driven by the commercial market may be repurposed by militaries or soldiers in need or increase the access of violent non-state actors to AI-enabled military capabilities, which may intensify ongoing conflicts and contribute to broader instability. There are also concerns that states could employ AI technologies to suppress human rights, entrench internal repression, and destabilise both regional and global peace.

Concurrently, as with AI more broadly, the environmental consequences of military AI – such as the energy-intensive demands of AI systems, resource extraction, and ecological damage from AI-enabled military systems – could aggravate resource scarcity and environmental degradation, fuelling tensions and undermining long-term peace. However, given the impact militaries have on civilian technology development, efforts to reduce the environmental impact of AI in defence settings could have far-reaching beneficial consequences for all uses of AI. As such, considerations of environmental impacts should be a component of responsible AI governance in the military domain.

The large-scale data extraction required for AI development could intensify geopolitical rivalries, facilitate intrusive surveillance, and create distrust through opaque and exploitative data practices. Such deployment of military AI may perpetuate discrimination and exacerbate social divisions, undermining stability and ultimately international peace and security.

There are simultaneously significant concerns regarding the potential of integration of AI within the command, control, and communication (C3) structures of nuclear weapons. A number of Commissioners and Experts have emphasised that this is a red line that must not be crossed. The commitment of several nuclear-armed states to human decision-making surrounding the employment of nuclear weapons is therefore applauded. Further, the development of large-scale lethal autonomous weapon systems – such as swarms of anti-personnel devices – risks creating a new category of weapons of mass destruction, posing serious threats to global peace and security. Relatedly, AI may lower the barriers to creation and use of nuclear, chemical, or biological weapons by state or non-state actors, thus generating new challenges for arms control and non-proliferation regimes.

Beyond these strategic risks, AI may affect the character of war and lower the thresholds for armed conflict. By increasing the speed of armed escalation and driving changes in the capabilities of weapons systems, AI in the military domain may reduce states' confidence in their deterrent capabilities – particularly in the face of cyber infiltration risk – thus influencing how decision makers receive, process, and act on information. AI in the military domain could also exacerbate asymmetric warfare and violence by widening technological disparities that could increase the likelihood of force being used prematurely or disproportionately.

Operationally, inaccurate AI systems used for targeting can undermine the security of protected persons under IHL by increasing the risk of indiscriminate attacks, violations of proportionality, and failure to distinguish between combatants and civilians. Closely related to this is the risk of fratricide due to potential errors in target identification or decision-making, which can undermine operational effectiveness, escalate conflict, and erode trust within militaries and alliances. Finally, there are views that the use of certain AI systems in the military domain can create accountability gaps absent clear rules. By complicating the attribution of responsibility for unlawful actions, the deployment of AI in the military domain could undermine key principles of international law and state responsibility for internationally wrongful acts. This may complicate efforts to hold individuals or

states responsible for violations, leading to a reduced deterrent effect against unlawful conduct. Without avenues to hold actors legally responsible, the enforcement of international law weakens, potentially destabilising peace, encouraging impunity, and exacerbating global insecurity.

4. Decision-Making and Responsibility

GC REAIM acknowledges the ethical and legal challenges that arise from integrating AI into military decision-making which may have a direct impact on preservation of peace and security. The relationship between human judgment and machine outputs is complex and without measures to ensure lawful, responsible and effective development and deployment, there can be an erosion of accountability and increased risks of unintended harm. As AI systems become more sophisticated and integrated within military capabilities, it is plausible that algorithmic decisions may become more commonplace across global battlefields, introducing moral and legal challenges regarding human control, oversight and judgment in diverse contexts.

To address these risks, GC REAIM promotes the need for context-appropriate human judgement over specific uses, capabilities and decisions of AI in military applications. The GC REAIM report will list considerations and conditions that underpin and support human responsibility, judgment and means of adequately evaluating relevant actions and decisions. This could include the introduction of technical standards for explainability, as well as maintaining appropriate human oversight in targeting decisions, assessments of precautions, proportionality and distinction, and other critical operational choices. However, given that the very definition of autonomy in machines suggests the minimisation or removal of the human, ensuring human responsibility and accountability may require focusing on human decision-making at earlier stages of a system's life cycle, as the systems structure the behaviour of all who work with it. Human oversight is essential to uphold state obligations under applicable international law, in particular, IHL.

Military AI systems must be designed not only to support all individual and collective agents in the military domain to be effective in safely carrying out their lawful tasks, but also to do so responsibly and without compromising or undermining their status as moral human agents. GC REAIM suggests that military AI based socio-technical systems need to be explicitly and demonstrably designed to adequately attribute and apportion responsibilities and is determined to contribute to this process. For the security of protected persons, parties to armed conflicts should at all times be able to demonstrate that everything possible has been undertaken to create the conditions under which military personnel can effectively apply extant and widely shared principles and laws of armed conflict to their own situation, when using or relying upon AI components in the execution of their tasks.

5. Governance and Regulation

In light of both the opportunities and risks associated with military AI, GC REAIM supports a comprehensive governance framework that implements authentic international law. GC REAIM reiterates that existing legal regimes provide a solid foundation for regulating AI technologies. Governance must incorporate and account for procedural safeguards (due diligence and legal reviews, transparency of testing, evaluation, and validation, accreditation, and verification), substantive obligations drawn from various branches of international law, and soft law tools (military doctrines, national policies and strategies, norms and standards). In principle, all relevant international legal frameworks must be considered and applied. These include, but are not limited to, the following: (1) international law (*jus ad bellum*) which regulates when and how states use force, codifying a general prohibition on the use of force and exceptions such as in the case of self-defence, (2) international

humanitarian law (*jus in bello*) which governs conduct during armed conflict and ensures the security of protected persons, (3) international human rights law.

GC REAIM further emphasises the critical role of international, regional, and domestic institutions in implementing and enforcing these legal norms. Effective governance requires collaboration across these levels and the inclusion of both binding (hard law) and non-binding (soft law) instruments. Soft law mechanisms, such as codes of conduct and ethical principles, can complement existing treaties and facilitate rapid, flexible responses to technological developments.

To address the diverse range of challenges surrounding the integration of AI into the military domain, GC REAIM supports proactive risk-mitigation and confidence-building measures. While binding regimes are challenging for general purpose technologies, there may be opportunities for rigorous monitoring, verification, and enforcement mechanisms inspired by successful global arms control regimes. For example, Commissioners and Experts have discussed ideas such as an Autonomous Incidents Agreement to reduce the risks of miscalculation among AI-enabled autonomous systems, or a committee or consortium that could set guidelines and recommendations surrounding the testing and evaluation of AI systems, including generative AI. GC REAIM also suggests that states and industries should consider adopting human-centred safety-by-design principles, implement red-teaming practices throughout AI system life cycles, and maintain clear chains of accountability for all actors. Only through robust multilateral dialogue and inclusive multi-stakeholder cooperation can AI be effectively governed to enhance peace and security rather than exacerbate global instability.

GC REAIM acknowledges that the development of a comprehensive governance framework for military AI faces several key challenges. First, there is the challenge of diverse interests and perspectives, with states, private companies, and civil society holding varying and sometimes conflicting views on the regulation of military AI. Second, the sensitivity surrounding national security and defence poses a significant barrier, as many states are reluctant to subject their military technologies to international scrutiny or regulation due to legitimate security interests. Third, achieving meaningful and substantive inclusivity in discussions is often difficult, as key stakeholders may be excluded or marginalised in decision-making processes. Fourth, a trust deficit between states, international organisations, and the private sector complicates efforts to establish cooperative governance. Fifth, the presence of crosstalk, incommensurability, and discursive dissonance arises due to the diverse backgrounds and expertise of stakeholders, making consensus-building challenging. Finally, these obstacles are compounded by the lack of clear frameworks that address the complex ethical, legal, and technical issues at the nexus of AI and the military domain. In light of these challenges, the final GC REAIM report will offer strategies to navigate and overcome these barriers in developing a robust governance framework.

6. Conclusion

GC REAIM observes that the rapid advancement and deployment of AI technologies in military contexts poses opportunities, challenges, benefits and risks for global peace and security. Balancing these considerations must be met with a technologically sound, inclusive, principled, and legally grounded approach to governance.

A clear understanding of AI's technological foundations is necessary to properly address its role in modern warfare. Ethical and legal responsibility should remain human-centred, and governance frameworks must rely on the robust application of international law, supplemented by cooperative multilateral efforts and soft law

instruments when appropriate. In its formation and deliberations, GC REAIM has had the opportunity to reflect upon the conversations happening in broader governance processes, finding ways to effectively bridge gaps between disciplines and regional perspectives.

GC REAIM urges the United Nations and all State Parties to place these principles at the heart of global discussions on the implications of AI in the military and broader peace and security, for the present and future generations. Only through concerted international cooperation, guided by a shared commitment to human dignity, peace, and justice, can we ensure that the future of AI in the military domain is one that strengthens our common security.

InterAgency Institute

[11 April 2025]

The InterAgency Institute was established in December 2020 as a digital think tank, founded by expatriate and Global South women as a collective of researchers. It is in this condition that we address this submission on "opportunities and challenges posed to international peace and security by the application of artificial intelligence in the military domain, with specific focus on areas other than lethal autonomous weapons systems," following A/RES/79/239. With this, we seek to craft a complementary set of suggestions to develop the policy discussion in points where understanding that AI encompasses a wide array of data-processing techniques, and may be integrated into different types of warfare, in multiple parts of the organization, and at different levels.

The InterAgency Institute would like to point to overarching trends that fall within our areas of expertise, namely: (1) a focus on the global south, specially in how to prevent furthering the security gap; and (2) in how interagency cooperation in a time of greater mistrust may be leveraged to ensure the integration of AI in the military does not. Additionally, we make the point that Decision Support Systems (DSS) create analogous problems when compared to Autonomous Weapon Systems (AWS).

1. Addressing the security gap between the Global South and the Global North

The increasing technological intensity and digitalization of the battlefield are likely to increase the capacity gap between countries in the Global North & South. The "optimization of war" entails furthering this discrepancy, augmenting threats, and deteriorating the global security landscape. The wide range of AI-enabled solutions represents discrepant utility levels across tools.

While some tools require a low threshold (thus providing usually an equally low ceiling), the systems that pose the biggest military advantage require a high knowledge threshold to be implemented, therefore, will likely not be open source, and will only be available to entities with sufficient means to develop or acquire them. Given the experience in past decades on multilateral forums, it is important to recognize that interest in access to these technologies will play a role in the negotiations.

In the long term, the current trend of "technological sovereignty" (or more specifically of restricted technological access due to global inequalities) may be transformed to undermine such technological control, creating far-reaching implications of this new revolution in warfare, involving stakeholders that may be reluctant to shape modern discussions due to a lack of current development of these technologies in their ecosystem.

2. InterAgency cooperation in times of distrust

These issues call for interagency cooperation at both the strategic and operational levels. The lack of interagency cooperation might lead to threat escalation and the eroding foundations for peace and security. Interagency cooperation should focus on formalizing specific channels for communication between different States, developing strategies for AI implementation that will not damage diplomatic relations, and generating more transparency in the interactions between agencies and contractors. The participation of different branches of government at the UN-level discussions is pivotal for a whole-of-government perspective in the deliberations. Beyond interagency cooperation at the governmental level, the wide array of applications of military AI calls for different sets of Confidence Building Measures (CBMs).

Since AI may be integrated in different warfare types and at different levels, its applications for different contexts have different ethical implications and consequences. Therefore, a monolithic understanding of risks posed by AI in the military context and consequently a unique set of CBMs would be inadvisable. CBMs for AI use in the strategic level of cyberspace will not be the same as CBMs for AI use in the tactical level of aerial warfare. Therefore, thinking about CBMs for military AI as a monolith will lead to inaccurate and in some cases inapplicable measures, undermining its effectiveness.

There is a necessity for sharing best practices in the introduction of AI into military procedures. In this sense, a trade-off should be made, prioritizing best practices that contribute to strengthening the aforementioned points of interagency cooperation and CBMs, and other practices that fall within the larger umbrella of strengthening international peace and security. Sharing of best practices relating to cybersecurity and reliability of the technology could also take place, but they should give priority to CBMs that focus on integration of AI at the strategic level and in manners that avoid the escalation of threats.

3. Decision Support Systems

Target identification or recognition via AI-enabled Decision Support Systems (DSS) entail analogue problems to Autonomous Weapon Systems (AWS). Digital dehumanization, lowering the threshold of violence, and automation bias are byproducts of that process that may only be avoided by the creation of red lines prohibiting such systems that replicate those concerns.

This problem stems not only from AI, but from a wider trend. Other data processing techniques that involve deterministic sorting of data that is not adequately processed by human operators also generate these problems. This caveat should be made to understand that not only systems with AI-enabled technology in DSS pose these kinds of threats, but a wider array of data gathering/processing techniques.

Conclusions and recommendations

- Formal interagency bodies to interface with multilateral AI/military tech negotiations
- Funding and support for academic research in the Global South focused on military AI implications;
- Regular technical-diplomatic summits focused on transparency, shared definitions, and threat perception;
- Prioritize capacity-building initiatives for Global South actors;

- Red lines and confidence building measures could be tailored to the specific technology and operational context;
- The discussions on Autonomous Weapon Systems encapsulate worries around AI-enabled Decision Support Systems. The creation of red-lines for these systems could benefit from building upon recommendations of the GGE on LAWS;

International Committee for Robot Arms Control

[11 April 2025]

The International Committee for Robot Arms Control (ICRAC) values the opportunity to submit our views to the United Nations Secretary-General in response to Resolution A/RES/79/239 "Artificial intelligence in the military domain and its implications for international peace and security."

Founded in 2009, ICRAC is a civil society organization of experts in artificial intelligence, robotics, philosophy, international relations, human security, arms control, and international law. We are deeply concerned about the pressing dangers posed by AI in the military domain. As members of the Stop Killer Robots Campaign, ICRAC fully endorses their submission to this report, and wishes to provide further detail regarding the concerns raised by AI-enabled targeting.

Increasing investments in AI-based systems for military applications, specifically AI-enabled targeting, present new threats to peace and security and underscore the urgent need for effective governance. ICRAC identifies the following concerns in the case of AI-enabled targeting:

1. AI-enabled targeting systems are only as valid as the data and models that inform them. 'Training' data for targeting requires the classification of persons and associated objects (buildings, vehicles) or 'patterns of life' (activities) based on digital traces coded according to vaguely specified categories of threat, e.g. 'operatives' or 'affiliates' of groups designated as combatants. Often the boundary of the target group is itself poorly defined. Although this casts into question the validity of input data and associated models, there is little accountability and no transparency regarding the bases for target nominations or for target identification. AI-enabled systems thus threaten to undermine the Principle of Distinction, even as they claim to provide greater accuracy.

2. Human Rights Watch research indicates that in the case of IDF operations in Gaza, AI-enabled targeting tools rely on ongoing and systematic Israeli surveillance of all Palestinian residents of Gaza, including with data collected prior to the current hostilities in a manner that is incompatible with international human rights law.

3. The increasing reliance on profiling required by AI-enabled targeting furthers a shift from the recognition of persons and objects identified as legitimate targets by their observable disposition as an imminent military threat, to the 'discovery' of threats through mass surveillance, based on statistical speculation, suspicion and guilt by association.

4. The questionable reliability of prediction based on historical data when applied to dynamically unfolding situations in conflict raises further questions regarding the validity and legality of AI-enabled targeting.

5. The use of AI-enabled targeting to accelerate the scale and speed of target generation further undermines processes for validation of the output of targeting systems by humans, while greatly amplifying the potential for direct and collateral

civil harm, as well as diminishing the possibilities for de-escalation of conflict through means other than military action.

Justification for the adoption of AI-enabled targeting is based on the premise that acceleration of target generation is necessary for 'decision-advantage', but the relation between speed of targeting and effectiveness in overall military success, or longer-term political outcomes, is questionable at best. The 'need' for speed that justifies AI-enabled targeting is based on a circular logic, which perpetuates what has become an arms race to accelerate the automation of warfighting. *Accelerating the speed and scale of target generation* effectively *renders human judgment impossible or, de facto, meaningless.* The risks to peace and security – especially to human life and dignity – are greatest for operations outside of conventional or clearly defined battlespaces. Insofar as the use of AI-enabled targeting is shown to be contrary to international law, the mandate must be to *not* use AI in targeting.

In this regard, ICRAC notes that the above systems present challenges to compliance with various branches of international law such as international humanitarian law (IHL), jus ad bellum (UN law on prohibition of use of force), international human rights law (IHRL) and international environmental law. In the context of military AI's implications for peace and security, jus ad bellum, a framework that prohibits aggressive military actions and regulates the conditions under which states may lawfully resort to the use of force, is the most relevant. In the same manner IHRL is important in this context because it is designed to uphold human dignity, equality, and justice – values that form the foundation of peaceful and secure societies.

International Humanitarian Law and Youth Initiative

[11 April 2025]

Artificial intelligence (AI) has gained a universal recognition during the 1950s'. Technological emergence has assisted humans in almost all facets of their lives thereby making work easier and faster. Moreso, the rapid growth of Artificial intelligence in technological field enthralling commercial investors, law makers, defense intellectuals and international competitors can be evidential in theoretical premises of international security. The use of Artificial intelligence (AI) in modern warfare particularly in the In the Middle East and North Africa, Ukraine/Russian armed conflict which has resulted in the killings of thousands of innocent civilians with women and children being the most vulnerable. The emergence of AI is expected to be utilized in improving all sectors in our daily lives However, its Negative application in the military domain continues to create Humanitarian crisis between warring parties making it of regional and international concern. The war in Gaza is one of the deadliest and most destructive war in history with technology playing a central role in enabling mass slaughter and destruction ranging from supplying the dystopian systems used to automate the killings and bombing. [1] Following the October 7 2023, there have been extensive reports evidencing the Israeli occupation forces use of surveillance technology, artificial intelligence, and other digital tool to determine who, what and when to attack in Gaza trip. Thus, this violates the principles of international humanitarian law which emphasize the necessity of distinguishing those in active combat and not[2] and to take necessary precautions when conducting an attack to minimize civilian harm.

[1] Accessnow. (October 2024) Big Tech and the risk of genocide in Gaza: what are companies doing? Available at https://www.accessnow.org/gaza-genocide-big-tech/.

[2] Article 48 of Additional protocol I of the Geneva convention.

IHLYI in this paper, responding to the request of the UN Secretary-General pursuant to a resolution A/RES/79/239, adopted by the General assembly on 24 December 2024 on Artificial intelligence in the military domain and its implication to international peace and security therefore, it analyzes AI In modern warfare, its implication to international peace and security and the role of technological companies in armed conflict.

Artificial Intelligence in Modern Warfare: A Legal and Humanitarian Perspective

The rules of international humanitarian law do not explicitly address the use of modern technological tools and artificial intelligence (AI) during armed conflicts. However, its core principles – such as distinction, proportionality, and precaution – remain applicable and binding on all parties. These principles require the differentiation between military objectives and civilians, and oblige parties to take all feasible measures to avoid or minimize harm to civilian populations. In recent years, militaries have contracted private companies to develop autonomous weapons systems. However, the armed conflict in Gaza stands out as one of the most prominent cases where commercially developed AI models – originally created in countries like the United States – have been employed in actual combat operations, despite the fact that these systems were not initially designed to make life-or-death decisions.

This shift highlights a troubling rise in the militarization of technology without clear legal or ethical oversight. While some of these tools may enhance operational efficiency, their unregulated use poses serious risks of human rights violations, especially amid a lack of transparency about how these tools function, the origin of the data they rely on, and the accuracy of their outcomes[3].

One of the most pressing concerns recently raised is the deployment of digital military tools based on unreliable data or flawed algorithms. Some of these systems depend on mass surveillance of Gaza's[4] population, including the collection of personal data prior to the outbreak of hostilities. Such practices raise legal and ethical questions regarding their compatibility with international obligations to safeguard privacy and prevent the misuse of personal information for the purpose of direct targeting.

Among the tools reportedly in use is a system that tracks population movement through mobile phone data to monitor evacuations from certain areas. Another generates lists of structural targets to be hit militarily. A third tool classifies individuals based on levels of suspicion regarding their affiliation with armed groups, while a fourth seeks to determine the precise location of a target in order to carry out a strike at the opportune moment. These tools largely rely on data extracted from mobile devices – whether through cell tower location information or GPS[5]. However, from a technical perspective, such data is insufficiently precise to confirm an individual's presence at a specific location at a given time, particularly in conflict zones where individuals frequently change phones or numbers. Over-reliance on this technology may lead to fatal mistakes, especially when a mobile phone is used as a substitute for verifying a person's actual presence in a targeted area. Legally, the use of such systems without taking all feasible precautions to protect civilians constitutes

[3] Human Rights Watch, "Israel: AI-Powered Targeting Systems May Be Committing War Crimes in Gaza", 2024.

[4] Associated Press, "Documents Reveal Israel's Use of AI Tools in Targeting Gaza", Investigative Report, 2024.

[5] Human Rights Watch (2024). Questions and Answers: Israeli Military's Use of Digital Tools in Gaza Available at Questions and Answers: Israeli Military's Use of Digital Tools in Gaza | Human Rights Watch.

a clear violation of international humanitarian law – particularly Article 57[6] of Additional Protocol I to the Geneva Conventions, which obliges parties to take constant care to spare civilian lives during military operations.

Given this reality, urgent questions must be raised about the future of AI in warfare and the legislative and legal mechanisms needed to regulate it. Without proper oversight, these tools risk becoming instruments of systematic human rights abuses rather than technologies aimed at ensuring greater protection for those affected by war.

Implications of Artificial Intelligence on International Peace and Security

Armed conflicts in various regions around the world, such as Gaza, Lebanon, Syria, Ukraine, and Libya, have had catastrophic humanitarian and security consequences. These conflicts have led to the mass displacement of civilian populations, depriving thousands of people of their basic rights such as food, water, shelter, and healthcare. These individuals live in dire humanitarian conditions, with a significant increase in deaths due to famine, thirst, and diseases caused by contaminated water, in addition to exposure to harsh weather conditions without protection.

In this context, the increasing use of artificial intelligence and drones as weapons in conflicts, particularly by Israel in the Gaza Strip[7], stands out. Since October 2023, there has been a notable escalation in the use of "quadcopters" to carry out precise and targeted strikes against civilians. These drones are equipped with data analysis algorithms and offensive capabilities, enabling them to target individuals based on tracking their movements or mobile phone signals.

According to documented reports, this technology has led to the death of more than 1,000 Palestinians by May 2024, including a significant number of women and children. This constitutes a grave violation of international humanitarian law, particularly Articles 51 and 57 of Additional Protocol I to the Geneva Conventions, which prohibit attacks on civilians and obligate parties to the conflict to take all necessary precautions to avoid harming them.

The concerns are not limited to the use of artificial intelligence against individuals but extend to the misuse of data. Relying on mobile phone tracking technologies (either through GPS data or cell tower signals) to pinpoint individuals' locations presents serious risks. Recent studies have shown that these systems do not provide enough accuracy to reliably determine someone's location, especially in conflict zones where phones may be swapped or disconnected frequently. This means that relying on these methods without field verification can lead to erroneous decisions, resulting in unlawful killings.

In a well-known case, a Palestinian woman named "Silah" was killed while carrying a white flag and leading her family to safety. After stepping onto a main street, she was targeted by a small drone that shot her in the head. This incident, witnessed by those around her, serves as a stark example of the disastrous outcomes of unregulated use of technology on the battlefield[8].

In Libya, drones played a decisive role in the battles between conflicting parties, particularly as many of these drones, including Turkish and Chinese models, were operated using data analysis systems to target objectives. Some of these systems are

---

[6] Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I), Article 57.

[7] TRTWORLD (2024) Quadcopter strikes: 1000 Palestinians killed by Israeli drones in one year. Available at Quadcopter strikes: 1000 Palestinians killed with drones in a year.

[8] Gaza grandmother gunned down by Israeli sniper as child waved white flag," *Times Kuwait*, November 2024, https://timeskuwait.com/news/gaza-grandmother-gunned-down-by-israeli-sniper-as-child-waved-white-flag.

believed to rely on artificial intelligence techniques for targeting, without legal oversight. The use of these tools in urban areas like Tripoli and Sirte has led to the deaths of civilians and extensive damage to infrastructure[9].

All of these events indicate that integrating artificial intelligence into managing and directing armed conflicts without an internationally binding legal framework to regulate its use could open the door to widespread violations, especially if these systems are not subject to independent and transparent oversight to ensure compliance with international humanitarian law and human rights.

Roles of Companies Developing AI in Armed Conflicts

Through a rapid increase in artificial intelligence and computer services, U.S. tech corporations have discreetly given Israel the ability to monitor and kill many more militants in Gaza and Lebanon more quickly. However, the death toll among civilians has also skyrocketed, raising concerns that these instruments may be causing the deaths of innocent people. Israel's recent wars are a leading example of commercial AI models developed in the United States being used in active warfare, despite concerns that they were not originally designed to help decide who lives and who dies.

For years, militaries have hired private companies to create customized autonomous weapons. Numerous American software companies have backed Israel's battles in recent years, including Microsoft and the San Francisco-based startup OpenAI. Under "Project Nimbus," a $1.2 billion contract signed in 2021[10] when Israel first tried out its in-house AI-powered targeting systems, Google and Amazon offer cloud computing and artificial intelligence services to the Israeli military. The military has made use of Dell and Cisco data centers and server farms. Palantir Technologies, a Microsoft partner in U.S. defense contracts, has a "strategic partnership" that provides AI systems to support Israel's war efforts, while Red Hat, an independent IBM company, has also supplied cloud computing technologies to the Israeli military.

Furthermore, through a number of programs, Microsoft also supplies Israel's government with services that have allegedly been used to help the Israeli military, police, Israeli Prison Service (IPS), and illegal settlement operations. Over 10,000 Palestinians are being held by the IPS as of October 2024; half of them have been detained without being charged or having a trial date scheduled. At least 310 medical professionals, UN employees, women, and children are among the Palestinian prisoners from Gaza who are presently detained in prolonged, secret, and incommunicado detention, where they are subjected to torture, mistreatment, and sexual violence and abuse, according to the UN Human Rights Office.

Companies are under obligation to respect human rights within their scope of operations. Companies that directly aid the offender – for example, by offering financial, logistical, military, or intelligence support – may be held criminally responsible for a crime committed during an armed conflict. Companies and their managers or executives may be held accountable in certain situations even if they had no direct involvement in the crime or no intention of supporting it. As the Office of the High Commissioner on Human Rights (OHCHR) noted, companies "should treat

[9] France 24. (2021). "Have Killer Drones Been Deployed in Libya?". France 24. Retrieved from https://rb.gy/1m6k43.

[10] APNEWS (2025). As Israel uses US-made AI models in war, concerns arise about tech's role in who lives and who dies. Available at How US tech giants' AI is changing the face of warfare in Gaza and Lebanon | AP News.

this risk in the same manner as the risk of involvement in a serious crime, whether or not it is clear that they would be held legally liable[11]."

In light of the concerns raised in this submission and their implications for international peace and security, IHLYI urges states to:

1. **Refrain from the use of AI in military applications**: States should immediately halt the use of artificial intelligence in military activities and establish national regulations and laws to prevent its deployment in warfare.

2. **Work towards a global ban on the military use of AI**: States should actively pursue international agreements and frameworks to ban the use of AI in military contexts, ensuring that no country utilizes AI for warfare.

3. **Avoid the development of autonomous and AI-enabled weapon systems**: States should refrain from developing autonomous weapon systems or AI-powered weaponry that could be used to target humans, ensuring human oversight and decision-making in military actions.

4. **Ensure the protection of personal data**: States must guarantee that personal data is protected from misuse by military forces, law enforcement agencies, border control, and private contractors collaborating with these entities.

5. **Promote accountability in AI development**: Technology companies, researchers, engineers, and financial institutions should commit to not supporting the development or funding of AI technologies designed for military applications, advocating for responsible innovation in line with humanitarian principles.

Peace Movement Aotearoa and Stop Killer Robots Aotearoa New Zealand

[21 May 2024]

Peace Movement Aotearoa and Stop Killer Robots Aotearoa New Zealand welcome the opportunity to contribute our views to the UN Secretary-General's report on artificial intelligence (AI) in the military domain and its implications for international peace and security. Our submission briefly outlines our involvement in this issue, and has three sections summarising our position on: a) A new international instrument on military use of AI and autonomy in weapon systems is urgently needed; b) Key focuses of a new international instrument; and c) Scope of a new international instrument. The points below are based on discussions with our member and supporting groups about the content of this submission.

Introduction

Peace Movement Aotearoa is the national networking peace organisation in Aotearoa New Zealand, established in 1981 and registered as an Incorporated Society in 1982. Our purpose is networking and providing information and resources on peace, humanitarian disarmament, human rights and social issues; and we have extensive national networks of member and supporting groups and individuals. We are a founding member of the Stop Killer Robots campaign and coordinate the national Stop Killer Robots Aotearoa New Zealand (SKRANZ) campaign.

SKRANZ was launched in April 2013 to support the global campaign, with a specific national focus on urging New Zealand to take national action to prohibit the development, production and use of autonomous weapon systems; and to take

[11] Accessnow (2024) Big Tech and the risk of genocide in Gaza: what are companies doing? https://www.accessnow.org/gaza-genocide-big-tech/.

international action to support negotiations on a new treaty to prohibit autonomy in weapon systems. Since 2023 we have widened our focus to include military use of AI as its perils became increasingly obvious.

(a) A new international instrument on military use of AI and autonomy in weapon systems is urgently needed

As outlined in our submission for the UN Secretary-General's report on autonomous weapon systems (A/RES/78/241) last year, it has been clear for some years now that rapidly developing technological advances in the use of force and increasing autonomy in weapon systems pose an unprecedented threat both to humanity and to the foundations of international human rights and humanitarian law, which are based on respect for human life and dignity, protection of humanity in times of oppression and armed conflict, and human responsibility and accountability for harm.

The serious ethical, humanitarian, legal, and security concerns posed by these developments have been discussed for more than a decade within United Nations bodies – including the Human Rights Council, meetings related to the Convention on Certain Conventional Weapons and in the UN General Assembly – as well as in regional and national governmental and non-governmental forums.

Even as these discussions have taken place, some states have increasingly incorporated autonomy into military use of force in ways that have already resulted in gross violations of international law with disastrous consequences for civilian populations. It is apparent that the absence of specific international law on autonomy in weapon systems, and with differing interpretation by some states as to how existing law applies to new technological developments, the risk of proliferation of ever more dangerous and uncontrollable weapon systems is increasing rapidly.

The need for urgency for international action on this has been highlighted over the past eighteen months by, for example, Israel's use of AI-powered target suggestion systems in Gaza to make high explosive strikes on numerous targets possible in a short time frame, resulting in indiscriminate slaughter of civilians and systematic destruction of life-sustaining infrastructure. The reality of digital dehumanisation with catastrophic consequences is now very evident, as is the increasing tendency towards the development and use of autonomous weapon systems that will remove any remaining vestige of humanity from war.

We have noted with concern that states who brought forward A/RES/79/239 include states that have armed and supported Israel's genocidal attacks on Gaza, and where big data tech companies contributing data storage and AI capabilities to Israel's military systems are based.

Similarly, 'responsible AI in the military domain' (surely an oxymoron) is being promoted by states already developing their own AI targeting and autonomous weapon systems, as a way of undermining the push towards a binding instrument to prohibit these critical threats to international peace and security.

The US 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' has highlighted for us the risk of horizontal proliferation of both military use of AI and autonomous weapon systems as states that do not have their own capability in this regard move from interoperability to integration with the states of armed forces that do or that are developing it. In the case of New Zealand, for example – as it seeks to be a 'combat capable force multiplier with enhanced lethality' [12] – this involves closer military integration with the armed forces of

[12] See, for example, the 2025 Defence Capability Plan released this week.

Australia and the US: New Zealand endorsed the US 'Political Declaration' early last year specifically to be compliant with US military doctrine.

These unfortunate developments underscore the urgent need for a new international instrument on military AI and autonomy in weapon systems to clarify and strengthen existing law. The instrument must include both prohibitions and regulations, as outlined below, and must include military use of AI in combat.

As emphasised in the UN Secretary-General's 2024 report on autonomous weapon systems[13], negotiations on a new instrument must begin without any further delay, in a multilateral forum where states can come together to work constructively, where the voices of those whose lives have already been impacted by military use of AI and increasing autonomy in weapon systems can be heard, and where UN agencies, the International Committee of the Red Cross (ICRC), and NGOs are active participants.

(b) Key focuses of a new international instrument

While much of the work around military use of AI and autonomous weapon systems has focused on the issue of meaningful human control over the use of force, it is our view that the key underlying ethical imperative is preventing human beings from being targeted or attacked by any system utilising digital code and/or sensors. A prohibition on military use of AI and autonomy in weapons systems that are designed or used to target human beings must be the starting point.

Meaningful human control over the use of force clearly has an ethical component, but it is also a practical and legal means to ensure accountability for any autonomy in weapon systems that breach the key dictates of humanitarian law.

(c) Scope of a new international instrument

It is our view that a new international instrument should include overarching rules to establish a framework for evaluating current and future technological developments, while promoting increased compliance with international human rights and humanitarian law.

Such overarching rules would prohibit autonomous weapon systems that are designed or used to target humans, and lay out specific obligations to ensure meaningful human control over other systems: for example, that the human operator/s understand the capabilities and limitations of the system, are able to fully evaluate the context in which the system will be used, and are making mindful firing decisions rather than assuming the technology is accurate – this would act to regulate autonomy in weapon systems. It would be useful to specify that decisions made by states on their assessment of new or altered weapon systems that incorporate autonomous features or functions must be transparent.

Furthermore, in the context of the UN Secretary-General's forthcoming report on AI in the military domain and in the light of the awful consequences of military use of AI in Gaza, the scope of a new international instrument must go beyond autonomous weapon systems. It is very clear that there is a spectrum of harmful military use of autonomy, ranging from target decision support systems (as some have described systems such as Lavender), data-based targeting systems, generation of target lists by algorithm or AI, sensor-based targeting systems, through to weapon systems that combine these elements and incorporate varying degrees of machine learning to make target selection decisions and attack autonomously.

We note the 2023 Joint Call by the UN Secretary-General and ICRC President stated *"The autonomous targeting of humans by machines is a moral line that we must*

[13] Lethal autonomous weapons systems: Report of the Secretary-General (A/79/88), 1 July 2024.

not cross"[14], yet that has already happened – a point reiterated in the UN Secretary-General's 2024 report[15].

It is therefore our view that a new instrument must cover military use of AI – including systems that automate significant decision-making in the use of force, such as target generation, force deployment, and engagement – as well as autonomous weapon systems.

Finally, although we have referred in this submission to military use of AI and autonomy in weapon systems, prohibitions and regulations in a new international instrument must also apply to all coercive agencies of the state, including those used for policing and internal security, for border control, in corrections facilities and in places of detention.

Ploughshares

[11 April 2024]

Project Ploughshares, a Canadian peace research institute, has for over a decade focused its advocacy and research on the military applications of emerging technologies, including artificial intelligence (AI) and autonomous weapons. As AI systems are rapidly advancing and being tested in contemporary conflict zones, international governance frameworks have struggled to keep pace. Meanwhile, intensifying geopolitical competition increases the likelihood that AI technologies will be deployed in complex, dynamic environments for which they are not suited – raising significant risks for civilians.

The wide-ranging use of AI in military applications demands urgent and coordinated international attention. We encourage the Secretary-General and member states to focus on three particularly pressing areas: the use of AI in decision-support systems related to the use of force, the dual-use nature of AI technologies, and the widening capacity gap among states engaging in multilateral discussions.

AI decision-support systems

One area that remains insufficiently addressed in current international discussions is the use of AI in military decision-making, especially decisions about the use of force. Of particular concern are AI-enabled targeting tools such as "Lavender" and "Gospel," reportedly used in in Gaza. These systems are classified as "decision support" because a human is technically required to approve target selections. However, there is little transparency regarding how these decisions are made, how frequently AI-generated recommendations are rejected, or whether human operators fully understand how the AI systems reach their conclusions.

In practice, these systems raise the risk of "rubber-stamping," in which human oversight becomes superficial, thereby undermining the principle of meaningful human control and increasing the likelihood of harm to civilians. The potential use of such AI systems in early-warning, surveillance, reconnaissance, and nuclear command-and-control systems further amplifies these concerns.

To mitigate these risks, states must work toward clear norms, regulations, and training requirements that enhance operator understanding, counter automation bias, and ensure genuine human engagement in decision-making processes.

---

[14] Joint call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to establish new prohibitions and restrictions on Autonomous Weapon Systems, 5 October 2023.

[15] As at note 3.

Dual-use challenges

AI's dual-use nature – its applicability to both civilian and military domains – creates further governance complexity. Civilian-developed technologies can be repurposed for military use without appropriate testing or safeguards, increasing the risk of conflict escalation, misuse, and error. Additionally, the accessibility of certain AI tools means that nonstate armed groups may also gain access, potentially using them to target civilians and infrastructure.

We urge states to develop policy mechanisms, including export controls, technology impact assessments, and multistakeholder engagement, to account for dual-use risks and promote responsible innovation.

Capacity- and knowledge-building

Current multilateral discussions reveal stark capacity disparities among states, many of which do not have the resources or technical expertise to participate meaningfully in governance efforts. To ensure inclusive and equitable global engagement, we recommend that states collaborate with the UN Office for Disarmament Affairs to strengthen capacity-building initiatives.

The scientific and academic communities also have a role to play in supporting the development of accessible resources and training materials. International forums, such as the upcoming REAIM Summit in Spain, should include dedicated sessions for knowledge-sharing, especially to support representatives from under-resourced states.

Final thoughts

The international community is at a crossroads. The accelerating militarization of AI demands robust diplomatic responses. We can – and must – move from aspirational principles to concrete, enforceable frameworks, by employing political will, inclusive dialogue, and cross-sector collaboration.

AI-powered warfare is no longer a theoretical risk; it is a present reality. Whether this new era enhances global security or undermines it will depend on the steps states take now to strengthen governance, manage technological competition, and uphold international humanitarian norms.

Without timely, coordinated action, the risks of accidental escalation and unintended conflict will only increase.

Soka Gakkai International

[10 April 2025]

The Soka Gakkai International (SGI) welcomes the opportunity to share our views on the important issue of artificial intelligence (AI) in the military domain. As an NGO whose work is guided by Buddhist principles, we urge that the United Nations, its Member States and other stakeholders take into careful consideration the impact of AI in the military domain from a standpoint of upholding and respecting human dignity.

Introduction

AI in the military domain is rapidly evolving and transforming modern warfare and international peace and security. These systems are being used for various purposes, including surveillance, autonomous weapons, decision-making support, and logistics. With such wide-ranging applications, the integration of AI technologies in military systems poses significant challenges. To better ensure compliance with

international humanitarian law (IHL) and uphold protection for civilians and combatants alike there are several issues that we may consider.

Lack of transparency and accountability

- If an AI system were to make an error – such as identifying a target incorrectly – it could be difficult to pinpoint the cause of the error, "the black box problem". Was it a flaw in the data used to train the AI, an issue with the algorithm, or a problem in the operational context? Without transparency within these systems, assigning responsibility is difficult.
- International laws and treaties, such as the Geneva Conventions, were created before AI systems became commonplace in warfare. Without global norms and legal frameworks, there is no consistent approach to ensuring accountability for AI decisions made in warfare.
- With inadequate accountability mechanisms in place, AI could be used for military strategies that violate human rights, suppress civil liberties, or engage in unethical operations.

Speed of decision-making and risk of escalation

- The ability of a military force to make decisions and execute actions faster than its opponent is increasingly viewed as having a strategic advantage. However, the drive for speed can lead to unintended and costly consequences.
- Decisions made too quickly without proper analysis or consideration can lead to poor outcomes, including tactical blunders, strategic missteps, or ethical violations.
- Instead of diffusing a tense situation or negotiating, if combatants react too quickly it could provoke an even greater confrontation, further escalation and prolonged conflict resulting in more human suffering including amongst civilians.
- The acceleration of decision-making processes closes down the possibility of meaningful human control, the growing trend to automate decision-making threatens the ability to achieve human oversight which is essential to facilitate compliance with IHL.

Bias in AI in the military domain

- AI bias refers to the presence of systematic and unfair discrimination in AI systems, such as historical bias, where systems may reinforce harmful stereotypes, bias in data processing and algorithm development which can lead to making biased decisions and bias in how the systems are used.
- AI bias in the military domain is a significant concern, particularly as AI systems are increasingly being integrated into defense and security operations. The potential for AI bias to emerge in these areas can result in human rights implications, exacerbating existing inequalities and lead to deadly consequences for certain groups.
- AI heavily relies on vast amounts of high-quality and reliable data for decision-making. There are several potential violations when it comes to obtaining this data including issues around privacy and surveillance, challenges of bias also arise when dealing with incomplete and inaccurate data.
- When AI systems are biased, they not only perpetuate inequalities but also contribute to the digital dehumanization[16] of marginalized groups.

[16] Digital dehumanization is a process where humans are reduced to data, which is then used to make decisions and/or take actions that negatively affects their lives.

Proliferation

- Nations may rush to develop AI-based military technologies to outpace their adversaries, which could lead to a destabilizing arms race and increased global tensions.
- Without regulation autonomous weapons systems in particular, could proliferate globally, including amongst non-state actors which could increase crime nationally and regionally, exacerbating social inequalities, overwhelm resources and infrastructures of countries, as well as undermine social and national security.

Conclusion

The issue of AI within the military contexts is complex, and without regulation, it could lead to serious consequences for global peace and security. The desire to speed up decision-making processes within this context has yet to be proven as an effective way of resolving conflicts and achieving peace and security. Furthermore, you cannot divorce AI in the military and AI in civil uses, a failure to address AI in a military context could have widespread repercussions in all spheres of civil life including law enforcement, border control, education, housing and health care. Fundamentally, AI is here to stay, how we utilize it in the military and in our lives will shape the course of humanity. We have the possibility and the responsibility to decide how we want to use technology, knowledge, and the world's resources. To use it in a way that uplifts humanity or degrades it? This is an urgent question that requires moral, ethical and courageous leadership.

Stop Killer Robots

[11 April 2025]

The Stop Killer Robots campaign welcomes the opportunity to submit our views to the United Nations Secretary-General in response to Resolution A/RES/79/239.

Established in 2012, we are a coalition of more than 270 non-governmental organisations working across 70 countries.[1] We seek to counter threats to humanity and human dignity through the adoption of a new international treaty to prohibit and regulate autonomous weapons systems.[2] We support the development of legal and other norms that ensure meaningful human control over the use of force, counter digital dehumanisation, and reduce automated harm.[3]

Building an effective international response to emerging technologies

Autonomous weapons systems, 'AI in the military domain,' and trends and developments in increasingly automated decision-making and action in the use of force – as well as in our lives and societies more broadly – are all part of the same concerning picture:

The growing influence of computer processing and algorithmic thinking increasingly shapes our interactions in the world and the outcomes available to us. There are clear threats to peace, justice, dignity, human rights, equality, responsibility and accountability, and respect for law. We are getting closer to machine processes determining whom to kill.

[1] See www.stopkillerrobots.org/about-us and www.stopkillerorobts.org/a-global-push/member-organisations.

[2] See https://www.stopkillerrobots.org/our-policies/.

[3] See www.stopkillerrobots.org/vision-and-values/.

To address these challenges effectively, a comprehensive and holistic response is needed from the international community.

Adopting a legally binding instrument on autonomous weapons systems will be one critical component: we must draw basic red lines for humanity against the automation of killing, which brings under jeopardy both international humanitarian law and international human rights law, in particular the presumption of innocence, the right to equality and non-discrimination, dignity, and wipes away contextual circumstances of the target(s) in question. The UN Secretary-General's comprehensive report last year reiterated his urgent call on states to negotiate a legally binding instrument to prohibit and regulate these systems by 2026.

But, a new international treaty on autonomous weapons systems alone may not be enough. States must also reach agreement on preventing and addressing grave harm from other uses of emerging technologies. A whole set of strong international rules are needed that stop the erosion of meaningful human control and the slide towards greater digital dehumanisation and automated harm, across international and domestic practice, in armed conflict and in civilian life.

'Military applications of AI' are already contributing to civilian harm

The risks of integrating AI into the use of force in armed conflict reach far beyond those to peace and security between states: a holistic consideration of peace and security that considers dimensions such as ethical, legal, and humanitarian issues must be taken into account in the UN Secretary-General's report under resolution 79/239.

We are already seeing grave threats to civilian protection and human rights and huge harm being caused by AI and automation in the use of force. This is arising from the quest for speed in warfare, the reduction of people to objects, and issues such as automation bias that Stop Killer Robots has raised the alarm about for years.

We have been horrified by reports of the use of AI-powered 'decision support systems' by Israel in Gaza, which suggest human targets to strike.[4] According to reports, human approval of these suggestions in vast volumes at high speed has been minimal – entailing digital dehumanisation, the erosion of meaningful human decision-making and control (including through automation bias), and directly contributing to massive and devastating harm to civilians in Gaza, alongside other tools.[5]

Many other states are developing and using such 'decision support systems', which raise concerns around international humanitarian law, human rights law, and ethics. So far there are few reports on how these are being deployed, with what constraints and with what impacts. Nevertheless, the push by many states to develop and integrate AI and autonomy into decision-making and the use of force is a huge concern. The further use in hostilities of these kinds of tools by any state in the unacceptable ways that we have seen in Gaza must be prevented. Stop Killer Robots struggles to see how such uses could meet the definition of the responsible application of AI in the military domain given in resolution 79/239.

Further risks to peace and security, rights, and human dignity

The quest for greater speed through AI and automation – towards the goal of increasing the tempo of conflict to a point beyond human cognition in the pursuit of a military and strategic edge – is an extremely dangerous one for international peace

[4] 'Lavender': The AI machine directing Israel's bombing spree in Gaza, +972 Magazine https://www.972mag.com/lavender-ai-israeli-army-gaza/.

[5] Questions and Answers: Israeli Military's Use of Digital Tools in Gaza, Human Rights Watch, https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza.

and security. These risks are further to the impact 'AI in the military domain' is already having on civilian protection. Risks include unwanted escalation, lowered political thresholds to the use of force, and arms race dynamics.

Technologies that can contribute to target selection (such as threat detection tools) and remote biometric surveillance (such as facial recognition) have already had documented negative impacts on human rights such as the rights to privacy, equality and non-discrimination, freedom of expression and peaceful assembly, and the freedom of movement. In the case of facial recognition for identification (1:n), the technology is considered by many legal experts as wholly incompatible with international human rights law.

That AI systems inevitably encode and reproduce the biases of our societies – including racism, sexism and ableism – and that such bias cannot be eliminated, is also well established. The use of such systems to process people in the use of force will inevitably lead to disproportionate – and multiplied – impacts on already marginalised and minoritised people. Integrating automation and AI into decisions and actions in the use of force against people contributes to digital dehumanisation – the process where humans are reduced to data, which is then used to make decisions and/or take actions that negatively affects their lives.

The relationship with autonomous weapons systems

Stop Killer Robots notes that the UN Secretary-General's report will be on the "application of artificial intelligence in the military domain, with specific focus on areas other than lethal autonomous weapons systems." It is important nevertheless to highlight that various applications beyond the boundary of autonomous weapons systems are closely linked to them.

Firstly, such tools could be integrated as components of autonomous weapons systems now or in the future. For example, a 'decision support system' could be used as an autonomous targeting system, connected to a platform tasked to strike targets on the list generated, based on processing sensor data. Secondly, these tools are linked not only practically, but raise and are part of the same picture of concern. Strikes undertaken based on the nominal human approval of targets generated by a decision support system do not sit far from strikes undertaken with an autonomous weapons system.

It is therefore important that states consider these issues in dialogue: many of the rules and principles developed for autonomous weapons systems on keeping control and rejecting automated killing will need to be extended (with adaptations) to other tools; and, how the development of AI in the military domain more broadly will impact the direction and challenges posed by autonomous weapons systems will need consideration.

Recommendations

Technologies incorporating AI and automation into the use of force in armed conflict are currently being deployed without specific agreed rules; the principles various states have proposed and committed to so far have been too weak and vague to prevent civilian harm and risks to peace and security.

All developments in autonomy and AI in the use of force which threaten our safety, security, and humanity must be urgently and adequately addressed through strong regulation by the international community, with unacceptable uses prevented.

States must:

- Move with urgency to negotiate and adopt a new international treaty to prohibit and regulate autonomous weapons systems;

- In International discussions, critically and meaningfully engage with the implications and real-world consequences of current practice in the use of tools that fall under ‘AI in the military domain,’ including acknowledging and examining humanitarian harm;
- Fully consider the legal, ethical, humanitarian, and peace and security risks of further development and use of such systems, whatever the perceived ‘benefits’ may be
- Work urgently to prevent unacceptable uses of technology and trends in development, through committing to develop strong norms for meaningful human control and against digital dehumanisation:
  - This should take place domestically, regionally, and internationally.
  - It must involve a comprehensive and holistic international response, including a legally binding instrument prohibiting and regulating autonomous weapons systems alongside other measures.
  - It should include consideration and development of the other legal instruments necessary to preserve meaningful human control and to protect human dignity against AI in the use of force.

Stop Killer Robots Youth Network

[10 April 2025]

The Stop Killer Robots Youth Network welcomes the opportunity to submit recommendations for consideration by the United Nations Secretary-General in response to Resolution 79/239 “Artificial intelligence in the military domain and its implications for international peace and security” adopted by the General Assembly on 24 December 2024. As a global network of young people under age 30 in over 50 countries working to secure a future free of automated killing, we have consistently advocated for the creation of a new treaty on autonomous weapons systems (AWS) – in particular, we insist on a total prohibition of anti-personnel autonomous weapons as we wish to build a world without such dehumanising weapons. While youth will inevitably face the risks of new weapons technologies, we remain underrepresented in the decision-making process and are often sidelined in forums that shape our interests. As youth who have grown up in an increasingly digital world, we wish to create a future where technology is used to promote peace, justice, equality, and human rights, not perpetuate violence.

With escalating conflicts and the rapid deployment of new weapons technologies around the world, there is an urgent need to reinvest in international law as a measure to build trust and achieve sustainable peace and security. The application of artificial intelligence (AI) in the military domain presents numerous challenges that concern us as youth, including digital dehumanisation, the gamification of violence, and the further erosion of human control and involvement over the use of force.

Military AI & AI systems already in use

Artificial intelligence has been progressively implemented in the military domain over the past decade, however, due to the opacity of military activities and development, the wide public has not been aware of this issue until recently when the active uses of AI systems have been mediatized. We have seen and monitored the use of AI systems to support the targeting of both objects and people. Unfortunately, the use of such systems have not been able to alleviate civilian suffering, for example, in Gaza where one third of victims are children and where too many civilian infrastructures, including critical infrastructures such as humanitarian camps,

hospitals[1], and schools[2], have been either directly targeted or indirectly impacted by the hostilities.

There have been other concerning uses[3] of AI systems outside of the military which need to be considered as they might be implemented in the military domain, mainly predictive AI and facial recognition. Predictive AI technologies have been used in the police and judicial systems since the early 2010s and have been shown to be ineffective, incorrect, and subject to reinforcing discriminatory behavior.[4] If predictive AI were to be implemented in the military domain, it could lead to the increasing risk of civilians being targeted as they could be labeled as possible fighters or being indirect victims of military activities due to the multiplications of targets with predicted military advantages. Facial recognition technologies (FRTs) are also of concern as they are also unreliable especially when it comes to the identification of non-white males. Facial recognition-enabled targeting in military operations must be prohibited as those systems cannot comprehensively analyse every factor that makes military personnel or civilians a target or not.

Digital dehumanisation

One of the main concerns we have about the use of AI systems in the military domain is the proliferation and banalisation of **"Digital dehumanisation"**. We define digital dehumanisation as the process whereby humans are reduced to data, which is then used to make decisions and/or take actions that negatively affect their lives. This process deprives people of dignity, demeans individuals' humanity, and removes or replaces human involvement or responsibility through the use of automated decision-making in technology.[5] Additionally, the increased speed and scale of target production through military AI erodes moral restraints in war and lowers the impact and capacity of decisions from human operators[6], thus enabling the AI systems to make decisions without meaningful human control, which further dehumanises the decision-making process.

Relying on (Big) data leads to problems

We also believe that the use of (big) data in the military leads to multiple issues which need to be considered.

One of the primary issues is the challenge of data labeling – the process of categorizing and tagging data to train algorithms. Inaccurate or biased labeling can have far-reaching consequences, particularly in the context of distinguishing between

[1] World Health Organization (2025), 'oPt Emergency Situation Update'. https://www.emro.who.int/images/stories/Sitrep_57.pdf.

[2] Save the Children (2025), 'Education Under Attack In Gaza, With Nearly 90% Of School Buildings Damaged Or Destroyed'. https://www.savethechildren.net/blog/education-under-attack-gaza-nearly-90-school-buildings-damaged-or-destroyed.

[3] Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica (2016), 'Machine Bias'. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[4] Will Douglas Heaven, MIT Technology Review (2020), 'Predictive policing algorithms are racist. They need to be dismantled'. https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

[5] Automated Decision Research (2022), 'Autonomous weapons and digital dehumanisation'. https://automatedresearch.org/news/report/autonomous-weapons-and-digital-dehumanisation-a-short-explainer-paper/.

[6] Marta Bo and Jessica Dorsey, OpinioJuris (2024), 'Symposium on Military AI and the Law of Armed Conflict: The 'Need' for Speed – The Cost of Unregulated AI Decision-Support Systems to Civilians'. http://opiniojuris.org/2024/04/04/symposium-on-military-ai-and-the-law-of-armed-conflict-the-need-for-speed-the-cost-of-unregulated-ai-decision-support-systems-to-civilians/.

combatants and non-combatants in conflict zones. If the data used to train military AI systems is flawed or biased, it can lead to disastrous mistakes, such as the targeting of innocent civilians or misidentification of threats.

A critical issue when relying on big data is that the nature data itself is often broken and is incomplete. This means that the data used to train AI models can be incomplete, outdated, or unrepresentative of real-world situations. Such flaws in data can lead to systems that fail to generalize properly, resulting in inaccurate or incorrect predictions and decisions. For example, in combat situations, a lack of diversity in the data used to identify individuals could lead to inaccurate targeting, with devastating consequences. Important data might be missing or poorly represented, such as the exact location of civilians or combatants, which can lead to AI failing to make informed and balanced decisions. In a war scenario, a system trained with data from a specific past conflict may not be capable of handling a new, unpredictable situation. For instance, an AI system that has been fed data from one particular type of conflict might struggle to apply that data to a war with entirely different characteristics, resulting in errors in target identification or incorrect decision-making.

Another significant problem is that many AI systems operate as black boxes. This means that while these systems make decisions and predictions based on the data they process, the decision-making process is not transparent or easily understood. In military scenarios, where the consequences of decisions are extremely serious, the lack of transparency is particularly concerning. If an AI system makes an error, such as wrongly identifying a civilian as a combatant, the absence of clarity about how the system reached that conclusion makes it nearly impossible to understand the origin of the error. This makes accountability difficult, as we cannot determine why the system acted in a particular way. The lack of explanation regarding the decision-making processes of AI also makes it impossible to correct or adjust the system's behavior, potentially perpetuating errors without the ability to fix them effectively.

Linguistic and cultural bias embedded in data which is used to train AI systems can create security vulnerabilities and catastrophically misinterpret communications, behaviors, and intentions across diverse cultural contexts, potentially triggering lethal automated responses to misunderstood signals.[7] These systems risk automating and amplifying existing prejudices at unprecedented scale and speed with life or death consequences in conflict zones where cultural misunderstandings could rapidly escalate into devastating military actions causing dire consequences.

Accountability

The inclusion of AI systems in the command and decision-making chains will indubitably lead to a lack of accountability and liability for those relying on these systems to make decisions. It will create a sense of distance and lack of liability on the consequences of a decision which mean that decisions may be made without specific, consistent and thorough analysis of the lawfulness and humane characters of the decision. Then, if an action taken using AI systems violates IHL, the people involved in the implementation and those involved in the decision-making should be held accountable and the use of an AI system shall never exempt people from their responsibilities.

We recognize that military operations are bound by multiple bodies of law – national law, International Humanitarian Law (IHL) and International Human Rights

[7] Jimena Sofía Viveros Álvarez, Humanitarian Law & Policy (2024), 'The risks and inefficacies of AI systems in military targeting support'. https://blogs.icrc.org/law-and-policy/2024/09/04/the-risks-and-inefficacies-of-ai-systems-in-military-targeting-support/.

Law (IHRL) – which need to be respected and implemented in order for operations to be lawful. Unfortunately, rules of engagement and of targeting – and all the exceptions – cannot be fully understood and implemented by AI systems. Concepts like doubt, proportionality, and the balance between humanity and necessity are inherently human judgments that cannot be captured by an algorithm. Machines cannot be trusted to uphold these standards on their own. Therefore, it is critical that AI systems never act in a vacuum and that humans retain oversight and decision-making power at all times.

What the future might look like

While AI theoretically has the potential to enhance precision and efficiency in military operations, its integration into warfare raises significant concerns about the future of global security. Autonomous weapons systems, capable of making life-or-death decisions without human control, introduce ethical dilemmas and risks of unintended consequences. The use of AI in military technology is likely to aggravate the existing arms race, as nations compete to develop increasingly sophisticated AI systems, widening the power gap between technologically advanced countries and those less developed, leaving them vulnerable in terms of military readiness. The deployment of autonomous weapon systems and AI-driven tools makes conflict more unpredictable, scalable, and asymmetric, granting certain nations the ability to unleash devastating technologies that smaller states or non-state actors may not be able to counter. The proliferation of AI in the military sphere also raises the threat of terrorism, as organized actors could easily access advanced AI-powered systems. Moreover, the fast-paced, constantly evolving nature of AI development turns military strategies into a “cat and mouse” game, where advancements are met with equally rapid countermeasures. In light of these challenges, the future of military AI must be handled with extreme caution, emphasizing robust ethical frameworks, international regulations, and stringent human oversight to prevent these technologies from destabilizing global peace.

What we need

We call for the establishment of a meaningful legally binding instrument for the use of AI-driven systems in the military requires comprehensive integration of the technical sector alongside state actors, addressing the urgent need for standardized verification protocols and trust-building mechanisms between nations. Such an instrument should define clear autonomy thresholds that specify permissible levels of independence in target selection and engagement, mandate extensive documentation of algorithmic decision processes and testing methodologies and establish explicit red lines that cannot be crossed including prohibited deployment scenarios, target categories, and operational environments. This framework should apply consistently across developing and developed nations, incorporate independent verification bodies with appropriate technical expertise to conduct regular compliance audits, and establish enforcement mechanisms with meaningful consequences for violations, all while facilitating technical data sharing and research that builds confidence between stakeholders in this domain.

These systems present an unprecedented threat to global security and human rights, and the risks they pose to non-combatants are immense. It is crucial that it implements a robust framework of monitoring, accountability and oversight. Firstly, the states need to be bound by positive obligations to ensure the responsible use of AI in the military domain. Accountability is a fundamental aspect of this framework. We call for comprehensive mechanisms that oversee every stage of the AI system life cycle, from development and updates to transfers and research. States must ensure that any uses of AI systems are monitored, with clear reporting structures in place to

address incidents promptly. Furthermore, it is vital that human operators using these systems receive thorough training and guidance to make ethical decisions in the field. The principle of meaningful human control must remain central when it comes to the use of AI in the military domain to ensure that ultimate responsibility for any actions remains with human decision makers.

Unione degli Scienziati Per Il Disarmo

[6 April 2025]

Introduction

USPID (***U****nione degli* ***S****cienziati* ***P****er* ***I****l* ***D****isarmo, Union of Scientists for Disarmament*) is an association of concerned scientists – founded in 1983 and based in Italy – which promotes arms control and disarmament initiatives based on scientific understanding of risks posed by military applications of science and technology. USPID submits to the United Nations Secretary-General its views on "Artificial intelligence in the military domain, with specific focus on areas other than lethal autonomous weapons systems, and its implications for international peace and security", in accordance with the invitation formulated in operative paragraphs 7 and 8 of Resolution 79/239 adopted by the UN General Assembly on 24 December 2024.

Hazards for peace and security arising from AI military applications

USPID expresses its deep concern about new hazards for peace, international security, and the respect of International Humanitarian Law (IHL) which arise on account of the ongoing and accelerating military efforts to incorporate Artificial Intelligence (AI) into multiple facets of warfare. Major sources of these hazards have been identified in current limitations of our capability to understand, predict precisely, and control the behavior of AI systems developed by machine learning methods and their interactions with other human or artificial agents. Initially identified in connection with the operation of AI-enabled Autonomous Weapons Systems (AWS), these hazards are now spreading to AI systems supporting intelligence collection, the achievement of situational awareness, and human decision-making in warfare.

Exceptionally grave concerns are raised by proposals to integrate AI in Nuclear Command, Control, and Communication (NC3) and in adjacent systems supporting nuclear decisions, and to let AI perform tasks that might directly or indirectly affect nuclear decision-making. A significant case in point is the proposal to use AI technologies in nuclear early warning and decision-support systems, which is being advanced with the expectation that AI accuracy will reduce potential errors, and its processing speed will buy more time for nuclear decision makers. However, on account of the probabilistic nature of AI information processing, one cannot exclude the risk of AI perception leading to false positives of a nuclear attack or producing perniciously unreliable recommendations given the impossibility of ensuring that the underlying models are aligned with human values and the UN overarching goal of preventing and removing threats to peace. If such mistakes occur, no matter how infrequent, large-scale and even existential implications for humanity might ensue. Accordingly, it would be imperative to proceed with time-consuming verifications of AI responses in nuclear early warning. But these verifications would be hindered by the black-box nature of much AI information processing and by the reliance on mostly simulated data, eventually thwarting the expectation of buying more time for human decision makers.

Additional concerns are raised by proposals to exploit the rapid pace at which AI operates to speed up battlefield decision-making and targeting cycles. These proposals are fueled by the goal of gaining military advantage over potential adversaries.

However, fighting at machine speed jeopardizes both the effectiveness of human oversight on AI-enabled decision support systems and the fulfilment of ethical and legal roles that are attached to human oversight of warfare action. Indeed, overly tight temporal windows for decision-making hinder effective human control over IHL threats raised by machine suggestions. Human interventions which aim at preventing inadvertent conflict escalations prompted by fighting at machine speed are similarly hampered. In addition to this, excessive speed in human-machine interactions has been identified as a factor inducing automation biases on the battlefield, and potentially skewing human decision-making even in the absence of AI failures.

Further hazards arise in connection with inherent vulnerabilities of AI learning methods and systems. Malicious manipulation of input data might be exploited to induce classification mistakes by AI systems. Moreover, poisoning attacks corrupting learning datasets may impair learning processes and the accuracy of resulting AI systems. These risks are compounded by our current inability to fully align AI systems with human goals and values, potentially causing them to deviate from strategic objectives.

Recommended actions

Mindful of these and other emerging hazards posed by the rapid adoption of AI technologies and systems in the military domain, USPID recommends

- to integrate discussion of AI in NC3 into the Non-Proliferation Treaty framework and in dedicated high-level dialogues and forums such as the Summit on Responsible Artificial Intelligence in the Military Domain (REAIM);
- to develop sustained international dialogue, good practices, and confidence-building measures concerning new and emerging risks for peace and IHL respect raised by AI warfare applications;
- to support a comprehensive and detailed inquiry aimed at identifying actual and potential AI applications in the military domain, jointly with situations of use that pose serious threats to peace, international stability, and the respect of IHL;
- to consider and investigate the need to introduce international regulations or prohibitions for those AI military applications that pose serious threats to peace, international stability, and the respect of IHL.

Women’s International League for Peace and Freedom

[24 May 2024]

The Women’s International League for Peace and Freedom (WILPF) has opposed war and the development of technologies of violence since its founding in 1915. WILPF has consistently condemned military spending and militarism as detrimental to human life and wellbeing. Our concerns with artificial intelligence (AI) in the military domain and its implications for international peace and security are grounded within our wider opposition to weapons, war, and violence, as well as in our opposition to patriarchal, racist, and colonial power relations that are embedded within AI technology.

While there are many perils of the military use of AI; WILPF’s submission is focused on the following issues:

1. The need for human emotion, analysis, and judgement in relation to the use of force;

2. The existence of gender, racial, and other bias in AI technology and the implications for digital dehumanisation;

3. The impacts of military use of AI on privacy and personal data;

4. The environmental harms exacerbated by the military use of AI; and

5. The dangers of war profiteering and arms racing.

Due to the concerns raised in this submission and in other spaces, WILPF opposes the military use of AI. This technology, rather than placing limits on violence or harm, expands both. Governance is insufficient in the face of the profits and power the developers of these technologies seek.

In light of the concerns raised in WILPF's submission and the implications for international peace and security, WILPF urges states:

- To refrain from using AI in the military domain and to develop national laws and regulations to this end;
- To pursue a global prohibition on the military use of AI;
- To not develop autonomous weapon systems or AI-enabled weapon systems, including those that can be used to target human beings;
- To ensure protection of personal data from use by militaries, police, border enforcement, and private companies and contractors collaborating with these institutions;
- To uphold human rights and dignity online and offline; and
- To address the environmental harms generated by data centres, cloud computing, and AI by reducing the number of these centres and energy consumption and water use, which will include reducing the overall use of AI.

WILPF also urges:

- Technology companies, tech workers, scientists, engineers, academics and others involved in developing AI or robotics to pledge to never contribute to the development of AI technologies for military use;
- Financial institutions such as banks and pension funds to pledge not to invest money in the development or manufacture of AI for military use; and
- Activists, academics, affected communities, and other concerned about privacy rights, digital dehumanisation, environmental and climate justice, gender-based violence, and other issues to collaborate and strategise to oppose the development and use of AI in the military and other violent domains.

D. Scientific Community

AI, Automated Systems, and Resort-to-Force Decision Making Research Project, The Australian National University

[11 April 2025]

Introduction

This executive summary highlights policy recommendations outlined in ***AI, Automated Systems, and Resort-to-Force Decision Making – Policy Recommendations: Submission to the UN Secretary General Pertaining to A/RES/79/239*** **(11 April 2025)**, available on the UNODA website. For a complete account of the underlying research and associated research papers, please refer to the full submission.

Underlying Research Project

This research has arisen from a **two-and-half-year research project (2022-2025),** entitled ***Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making***, led by Professor Toni Erskine (Australian National University) and funded by the Australian Government through a grant by the Department of Defence.

Its focus is **distinctive and critical**. While the attention of academics and policy makers has been overwhelmingly directed towards the use of AI-enabled systems in the *conduct of war* – including, prominently, on the emerging reality of 'lethal autonomous weapons systems' ('LAWS'), this project has addressed the **relatively neglected prospect of employing AI-enabled tools at various stages and levels of deliberation over the *resort to war.*** In other words, 'it takes us from AI on the battlefield to **AI in the war-room'**.[1]

This research project has brought together **leading scholars and practitioners** working on different aspects of international politics and security, strategic and defence studies, and artificial intelligence (AI) to contribute to a multi-disciplinary study and set of **policy recommendations on the risks and opportunities of introducing AI, machine learning (ML), and automated systems** into state-level decision making on the **initiation of war**. Our interventions are made from the perspectives of political science, international relations, law, computer science, philosophy, sociology, psychology, engineering, and mathematics.

Project participants presented and discussed their research at two workshops (June 2023 and July 2024) at the Australian National University (ANU), convened by Professor Toni Erskine and Professor Steven E. Miller (Harvard). Participants also received feedback on their initial research-based policy recommendations from senior Australian Government delegates from the federal civil service as part of a one-day policy roundtable (July 2024) at the ANU.

'Four Complications'

For all the potential **benefits** of AI-driven systems – which are able to analyse vast quantities of data, make recommendations and predictions by uncovering patterns in data that human decision makers cannot perceive, and respond to potential attacks with a speed and efficiency that we could not hope to match – challenges abound. Through this project, we have sought to address **four thematic 'complications'** that we propose will accompany the gradual infiltration of AI-enabled systems in **decisions to wage war:**[2]

- **Complication 1** relates to the displacement of human judgement in AI-driven resort-to-force decision making and possible implications for deterrence theory and the unintended escalation of conflict.
- **Complication 2** highlights detrimental consequences of automation bias, or the tendency to accept without question computer-generated outputs – a tendency that can make human decision makers less likely to use (and maintain) their own expertise and judgement.

[1] T. Erskine and S. E. Miller, 'AI and the Decision to Go to War: Future Risks and Opportunities', *Australian Journal of International Affairs*, Vol. 78: 2 (2024), pp. 135–147 (p. 138).

[2] For an account of these 'four complications', see T. Erskine and S. E. Miller, 'AI and the Decision to Go to War: Future Risks and Opportunities', *Australian Journal of International Affairs,* Vol. 78: 2 (2024), pp. 135–147 (pp. 139–40).

- **Complication 3** confronts algorithmic opacity and its potential effects on the democratic and international legitimacy of resort-to-force decisions.
- **Complication 4** addresses the likelihood of AI-enabled systems impacting organisational structures and chains of command, whether degrading or enhancing strategic and operational decision-making processes.

Contributors to this project have explored these proposed complications in the context of either **automated self-defence** or the use of **AI-driven decision-support systems (DDS)** that would inform human resort-to-force deliberations. We have identified risks and opportunities of using AI-enabled systems in these contexts and make recommendations on how risks can be mitigated and opportunities promoted.

Complication 1: Displacement of human judgement

AI in Nuclear Crisis Decision Making

One key area of research undertaken in response to this complication is the nuanced interplay between AI and human decision making in the high-stakes context of **nuclear crisis management**. **Risks (including the increased fragility of nuclear deterrence relationships, crisis signalling becoming more complex**, and **unintended escalation)** have been explored in two broad areas: i) automation in military deployments, or taking the human 'out of the loop' in the decision to use nuclear or strategic non-nuclear weapons (SNNW); and, ii) the integration of AI into human decision-making (particularly in early warning threat assessments). Although much of this research has focused on risks, **novel benefits** of introducing AI-driven decision-support systems (DSS) into human-led nuclear crisis management have also been proposed.

Policy Recommendations:

- **Always incorporate human-in-the-loop safeguards**: Ensure AI systems in nuclear command and control are always overseen by human operators and that human decision-making remains central to determining when and how nuclear-weapon states resort to the use of their arsenals.
- **Promote a holistic approach to AI-safety**: AI safety should account for both technical and socio-technical dimensions. Assess safety challenges in AI-enabled DSS comprehensively, including issues of security, trust, and liability.
- **Broaden the scope of risk assessments**: Apply risk assessments relating to the deployment of AI and ML not only to obvious areas such as nuclear launch orders, but also to less obvious areas such as early warning intelligence assessments (including by non-nuclear allies) and SNNW capabilities (including by non-nuclear allies).
- **Restrict the use of AI-assisted warning data**: The key to balancing the benefits of incorporating AI into early warning against the risks is limiting what AI-assisted warning data is used for. In AI research, prioritise tasks such as calculating effective evasive manoeuvres in the event of an attack and using pattern recognition and anomaly detection to improve arms control verification.
- **Pursue** informal arms control and confidence-building: Advance informal measures such as regular dialogue, red-line agreements, and information-sharing mechanisms. Expand unilateral initiatives like moratoriums where feasible.
- **Explore AI's potential to promote empathy and enhance decision making**: Decision makers must exercise 'security dilemma sensibility' (SDS) in times of crisis. Decision makers and diplomats exercise SDS when they are open to the

possibility that the other side is behaving the way they are because they are fearful and insecure, and crucially, recognize the role that their own actions may have played in this. Explore ways that the balanced integration of AI and human judgement could enhance SDS during nuclear crises by promoting empathy and trust.

AI Mistakes in the Resort to Force

Another area addressed in relation to this complication is **state responsibility** when **errors** occur in AI-driven or autonomous systems involved in resort-to-force decisions. Such errors may arise from poor system training, data poisoning by adversaries, or two AI-driven systems interacting in unintended ways. It is essential to develop legal standards and practices that reduce the risk of unintended conflict resulting from such failures.

Policy Recommendations:

- **Adopt robust security and cyber hygiene**: States should adopt robust protections against AI data poisoning and cyber attacks to meet jus ad bellum standards of good faith and reasonable conduct.
- **Clarify legal guidelines on delegating the use of force to autonomous systems**: Senior leadership within states should set clear domestic legal standards regarding when and how autonomous systems may be authorised to use force.
- **Commit to transparency in after-action reviews**: States should commit to being transparent and deliberate about after-action reviews of any AI errors that occur in the field, potentially drawing on civilian casualty review processes as a model.

Complication 2: Automation bias

Our research in response to the second complication focuses on the relationship between human actors and **AI-driven DSS** in resort-to-force decision making. It includes a detailed survey-based study of **military trust in AI** during strategic-level deliberations and a robust account of the importance of ensuring that there are human **'experts-in-the-loop'** when AI-driven systems contribute to decisions on war initiation. This body of work also addresses the **benefits** of employing DSS to **enhance our cognitive capacities** in strategic decision making and, conversely, uncovers the potential **dangers** of such reliance if these systems **dull our sensitivity to the tragic qualities of war** or contribute to the **erosion of restraint** by creating the illusion that they replace us as responsible actors.

Policy Recommendations:

- **Consider the multidimensionality of trust**: Recognize that soldiers' trust in AI is not a forgone conclusion. Rather, it is complex and multidimensional, and further complicated by biases, uncertainty, and lack of education.
- **Interrogate norm compliance**: In terms of governance, explain how policies on increasingly autonomous capabilities coincide or diverge from international norms and laws informing their use.
- **Embed experts in decision structures**: Enshrine an 'expert-in-the-loop' organisational structure – i.e., high-level experts as core decision makers.
- **Prohibit automation**: Prohibit automation of resort-to-force decisions.

- **Increase AI literacy of domain experts**: Provide and require basic technical training for high-level domain experts so they understand the logics of AI and can thus incorporate AI decision inputs from an informed position.
- **Provide on-going, substantive training for domain experts**: Sustain substantive training for, and assessment of, high-level experts to bolster and ensure substantive competencies.
- **Regulate non-autonomous AI**: While autonomous AI agents, e.g., lethal autonomous weapons systems (LAWS), need regulation, so do non-autonomous AI systems, which leave humans vulnerable to new forms of influence, moral and cognitive atrophy, and undermined responsibility.
- **Design AI-driven DSS to promote more accurate perceptions of their capacities**: Ensure AI-driven DSS are not easily mistaken for responsible agents in themselves by avoiding anthropomorphism, building in warnings about system limitations, and incorporating features that emphasise human agency and accountability.

Complication 3: Algorithmic opacity

Our research in response to the third complication addresses **how the lack of transparency of AI-driven decision making can threaten the legitimacy** of AI-informed decisions on the resort to force. This body of work includes original research on **large language models (LLMs)** and their potential to exacerbate existing **pathologies in intelligence analyses**. It also examines the role that the **'architecture of AI'** and its hidden vulnerabilities play in deliberations surrounding the resort to force. Moreover, research within this pillar conceives of military decision-making institutions as **'complex adaptive systems'** – a conceptual framework that yields a range of insights, including that human-machine teams possess a form of **'cognitive diversity' that could be leveraged for more efficient decision-making,** but also **exploited to poison information flows**, and that technical explanations for algorithmic opacity will not solve accountability concerns.

Policy recommendations:

- **Develop policy to limit epistemic pathologies of LLMs**: States should clearly determine defence and intelligence policy towards either a) procurement of LLMs, b) state development of LLMs, or c) a combination of both. They should use this guidance to develop policy which seeks to limit the epistemic pathologies of LLMs in autonomous decision-making.
- **Commit to sector-wide procurement guidelines and oversight of generative AI tools** used in decision-making chains.
- **Commit to regulating data markets** and access to those markets through alliance relationships.
- **Promote understanding of the tech ecosystem and its fragilities**: Increase understanding of the inherent interdependencies and vulnerabilities in the tech ecosystem, including by creating technology literacy training programs designed specifically for politicians and policy, intelligence, and military leaders.
- **Invest in research** to develop a comprehensive picture of the architecture – physical and digital – that underpins AI, including critical dependencies and vulnerabilities and how access and power are distributed.

- **Invest in research on social media** and its impact on functions of government, including its potential to disrupt democracies, facilitate foreign interference, and influence decision making on the use of force.
- **Recognize AI's current influence**: Significantly increase awareness of government reliance on the architecture of AI, especially for critical government functions, including resort-to-force decision making.
- **Invest in research and development** to maximize the benefits of human-machine cognitive diversity.
- **Implement responsible AI governance programs** that carefully balance accountability with operational efficiency.
- **Perform regular red-team exercises** to ensure that the integration of AI in decision-making institutions does not induce systemic blind spots and vulnerabilities in military decision-making.

Complication 4: Impact on organisational structures

Our research regarding the fourth complication explores both the **beneficial and damaging effects that AI-driven systems can have on institutional structures** in the context of resort-to-force decision making. Studies focus on how AI-driven **DSS can improve 'adaptive culture'** within military organisations, thereby improving wartime decisions, and how the urgent need to **upgrade AI literacy and educate human analysts** should lead us to **reform institutional structures and cultures**. The novel notion of **'proxy responsibility'** is proposed as an institutional response to ensure that responsibility can be meaningfully assigned to humans for resort-to-force decisions that are informed by AI systems. Moreover, original research highlights the significance of the **neglected category of AI 'integrators'** – sandwiched between the 'developers' and 'users' of AI within organisational structures – when it comes to strategic military applications of AI.

Policy recommendations

- **Set (and evolve) measures of effectiveness**. If AI-enabled adaptive capacity is to work effectively, measures of military effectiveness must guide which direction adaptation might take. Establish such measures at the tactical (battlefield) and strategic (war-room) levels to guide development and implementation of AI-enabled adaptation.
- **Know where adaptation relevant data is found, stored and shared**. An enhanced adaptive stance in military institutions must have enhanced data awareness as a foundation. Data awareness and management must become one of the basic disciplines taught to military personnel.
- **Scale AI support from individual to institution**. There is unlikely to be a one-size-fits-all algorithm or process that can enhance learning and adaptation at every level of military endeavours. Create a virtual 'arms room' of adaptation support algorithms as part of an institution-wide approach to adaptation.
- **Routinely question AI-enabled outputs**: Build mindsets, protocols, institutional cultures, and inter-agency structures in 'normal' pre-crisis times to routinely question AI-enabled output from human-machine teams.
- **Institute an advisory body**: In order to support the notion of 'proxy responsibility' as an institutional response to 'responsibility gaps' when decisions on war initiation are informed by AI-enabled systems, establish and/or strengthen state-level 'AI departments'. These departments would integrate

technical, political, and ethical competence and expertise and advise on resort-to-force decision-making processes.

- **Support research on AI integration**: Fund research on the integration of AI in strategic decision-making.
- **Provide standards**: Outline minimum standards for the responsibilities of AI developers and integrators.
- **Facilitate inter-group discussions** between developers, integrators and users during development, integration, and longer-term maintenance processes.
- **Create accountability guidelines**: Provide well-defined guidelines and rules indicating who is accountable if something goes 'wrong'.

Queen Mary University of London, T.M.C Asser Institute, University of Southern Denmark, University of Utrecht

[11 April 2025]

Views of members of the scientific community and civil society; specifically, we are a group of academics with expertise in ethical, legal and political dimensions of military Artificial Intelligence and herewith put forward our shared views pursuant to resolution 79/239 "Artificial intelligence in the military domain and its implications for international peace and security" adopted by the General Assembly on 24 December 2024, in accordance with the request of the UN Secretary-General contained in Note Verbale ODA/2025-00029/AIMD.

Introduction:

The rapid advancement and integration of AI technologies into targeting operations have sparked ongoing debates surrounding their ethical, legal, and operational implications. Over the past decade, the discourse on AI in warfare has largely centered on autonomous weapon systems (AWS),[1] driven in part by the initiation of discussions in 2013 and the formalization of a regulatory process under the UN Convention on Certain Conventional Weapons (CCW) and the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE LAWS), which exclusively focuses on lethal AWS.[2] However, the increasing integration of AI-based decision-support systems (AI-DSS) into targeting practices[3] introduces new layers of complexity that demand closer attention from a broad range of stakeholders. This submission responds to that need, structured around three key components: (1) a

[1] The latest definition of AWS from the CCW GGE LAWS Rolling Text (26 November 2024): "A lethal autonomous weapon system can be characterized as an integrated combination of one or more weapons and technological components that enable the system to identify and/or select, and engage a target, without intervention by a human user in the execution of these tasks." On file with authors.

[2] For a brief overview of some of the latest developments of the GGE LAWS see Jeroen van den Boogaard, *Warning! Obstacles Ahead! The Regulation of Autonomous Weapons Systems in the GGE LAWS*, *Opinio* Juris, 4 March 2024 found at: https://opiniojuris.org/2024/03/04/warning-obstacles-ahead-the-regulation-of-autonomous-weapons-systems-in-the-gge-laws/.

[3] There have been several reported uses of AI-DSS by Israel in Gaza and potentially in Lebanon, by both Ukraine and Russia in the ongoing conflict, and by the United States in its actions against Houthi rebels in the Red Sea and in Yemen, to name a few. For a comprehensive overview of literature in this space, see e.g., Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang, *AI in Military Decision Support Systems, A Review of Developments and Debates,* Centre for War Studies, University of Southern Denmark, November 2024. Found here: https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates/.

brief overview of how AI-DSS are currently used in targeting decisions; (2) an analysis of key concerns, including how these systems shape the potential exercise of human judgement and control and underline fundamental gaps in global governance; and (3) a concluding set of recommendations.

1. Overview of AI-DSS and the joint targeting cycle

Defined as "the process of selecting and prioritizing targets and matching the appropriate response to them, considering operational requirements and capabilities,"[4] targeting is a core military function at the very heart of warfare. While the potential range of use cases for AI-DSS in military decision-making is broad, in targeting, AI-DSS can be understood to serve as **tools** that use AI techniques to collect and analyze data, provide information about the operational environment as well as actionable recommendations, with the aim of aiding military decision makers in evaluating factors relevant to legal compliance such as taking precautions and ensuring proportionality in attacks.

More specifically, AI-DSS are increasingly integrated across multiple phases of the joint targeting cycle (JTC), including within target development and prioritization, capabilities analysis, and mission execution. The JTC is a reflective example of a structured process used by military forces to identify, evaluate, and engage targets while ensuring compliance with operational, legal, and ethical standards,[5] generally consisting of six (non-linear) phases:

1. **End-State and Commander's Objectives**: Defining strategic military goals and desired outcomes.

2. **Target Development and Prioritization**: Identifying, verifying/validating, and prioritizing targets based on intelligence and mission goals.

3. **Capabilities Analysis**: Assessing the available strike options and their effectiveness.

4. **Force Assignment**: Allocating specific military assets (e.g., airstrikes, artillery, cyber operations) to engage the target.

5. **Mission Execution**: Carrying out the targeting operation while ensuring compliance with relevant laws and the rules of engagement.

6. **Assessment**: Evaluating the effectiveness of the operation and adjusting for future operations, if necessary.

Within this framework, AI DSS are assumed to serve primarily as informational and analytical tools which support human decision-making rather than supplant it. However, this assumption and framing obscures how AI-DSS influence human cognitive processes within the JTC. This impact on human decision-making is often

[4] United States Department of Defense, *Dictionary of Military and Associated Terms*, March 2017, found at: https://www.tradoc.army.mil/wp-content/uploads/2020/10/AD1029823-DOD-Dictionary-of-Military-and-Associated-Terms-2017.pdf.

[5] Michael Schmitt et al, *Joint and Combined Targeting: Structure and Process*, Chapter 13 in Jens David Ohlin (ed) *Weighing Lives in War* (Oxford, 2017). See also, Jessica Dorsey and Marta Bo, *AI-Enabled Decision-Support Systems in the Joint Targeting Cycle: Legal Challenges, Risks, and the Human(e) Dimension*, forthcoming 2025, *International Law Studies.* "Targeting generally involves four key steps: (1) objectives and guidance, (2) planning, (3) execution, and (4) assessment. Encapsulating these four key steps, the United States and NATO outline their targeting processes through similar six-phase cycles [addressed in this submission]. As the reader can discern, different states employ different doctrines for targeting. What is important … is not necessarily the specific labels for various steps followed by any given state, but rather how and when compliance with the principle[s of IHL are] incorporated into the targeting process."

underestimated and remains insufficiently examined, leaving critical discussions about the role of AI-DSS largely absent from current policy debates.

2. Analysis of Key Concerns

(a) (Meaningful) Human Judgement and Control

AI-DSS are often portrayed as enhancing human decision-making and the quality of decisions therein. The perception of AI-DSS as mere subsidiary tools has led to a narrative that the integration of AI-DSS poses fewer challenges than AWS, given that these systems do not directly "engage" targets (i.e., they do not have an inherent capability to directly carry out the use of force) and are tools that assist human commanders. The outputs are ostensibly ultimately reviewed through (several layers of) human oversight, such as processes of verifying and validating targets using additional intelligence sources.[6] As a result, errors or inaccuracies in AI-DSS outputs are often seen as non-critical, based on the assumption that robust human oversight and appropriate control will compensate for them. However, closer examination reveals that this control is frequently superficial, offering only the appearance of, rather than actual meaningful, or context-appropriate, human judgement and control.

This is because AI-DSS structure and condition the quality of human control and oversight and limit the ways control and oversight can be exercised. The use of AI-DSS creates a shared decision-making space between human military personnel and AI technologies. States appear to have recognized and focused on many of the advantages of this shared decision-making space for military personnel, i.e., how the use of AI-DSS advances human decision-making through offering data-driven insights. But using AI-DSS also delimits the capacity to exercise human oversight and control because of the technologies' complexity and the increased speed (and therefore scale) it can bring to decision-making processes. Rather than supporting human oversight, using AI-DSS may risk humans becoming little more than reactive cogs in socio-technical systems.[7] Moreover, this configuration risks amplifying adverse human biases, such as automation bias, anchoring bias, or cognitive action bias, to the detriment of exercising qualitatively high levels of human control.[8] Considering AI-DSS as a distinct form of technology therefore reveals significant challenges associated with military AI and human oversight, challenges that extend beyond those that arise when simply integrating the technology in weapon systems.

Recent conflicts have shown the risks associated with AI-DSS being employed in critical functions, such as target selection and even nomination, and their conditioning and constraining of human involvement, affecting the fulfilment of core legal obligations embedded within the JTC. The use of AI-DSS raises fundamental concerns about whether human decision makers can retain adequate cognitive autonomy over the JTC process or whether humans will become overly reliant on algorithmic outputs for critical judgements in the context of armed conflict.[9] Consequently, there are significant legal concerns regarding the effects of such systems on decision-making processes and use of force decisions and ability for users to comply with IHL obligations, especially with respect to the obligation to take all

---

[6] Alexander Blanchard and Laura Bruun, *Automating Military Targeting: A Comparison Between Autonomous Weapon Systems and AI-Enabled Decision Support Systems*, Stockholm International Peace Research Institution (SIPRI) forthcoming 2025 (draft on file with authors).

[7] Ingvild Bode, *Human-Machine Interaction and Human Agency in the Military Domain*, Policy Brief No. 193 (Waterloo, ON: Centre for International Governance Innovation, 2025).

[8] Dorsey and Bo, supra n. 5.

[9] *Ibid*; see also Anna Nadibaidze, Ingvild Bode, and Qiaochu Zhang, *AI in Military Decision Support Systems, A Review of Developments and Debates,* Centre for War Studies, University of Southern Denmark, November 2024. Found at: https://www.autonorms.eu/ai-in-military-decision-support-systems-a-review-of-developments-and-debates/.

feasible precautions to minimize civilian harm to the greatest extent possible in attack and comply with the principles of distinction and proportionality.[10]

Importantly, these concerns are not new. There is extensive debate around how to preserve meaningful human judgment and human agency when conducting IHL-evaluative legal assessments, in the context of AWS. These discussions – which include expert analysis on accountability, human-machine interaction, automation bias, and the effect of AI systems on legal and ethical reasoning[11] – provide valuable lessons that must inform discussions around military AI and specifically the use of AI-DSS in military contexts.

(b) AI-DSS: Understudied, Under-Addressed and Unregulated

Framing AI-DSS as mere tools, has led to an underestimation and lack of analysis on the way their use affects the cognitive decision-making process within the JTC. The relative lack of attention paid to AI-DSS so far can partly be attributed to the fact that such systems are seen to be used with a human *in* or *on* the loop framework, with their outputs ostensibly reviewed by one or more individuals during the targeting process. As a result, current understandings of AI-DSS use appear to align with widely supported principles of human control and oversight. However, this gap in the debate is also caused by a lack of transparency around how specific AI-DSS function, and a consistent failure to comprehensively examine how they are being used in practice.

Additionally, the persistent focus on AWS at the expense of AI-DSS obscures the growing reliance on AI in shaping operational and strategic outcomes. Unlike AWS, which have been debated in the framework of the CCW for the past decade, AI-DSS lack a comparable institutional platform. Attention to AI-DSS remains scattered across various initiatives but these efforts have yet to provide the dedicated regulatory focus or coordination needed.

3. Recommendations:

i. **Reassert** the central role of human cognitive and legal reasoning in military operations by implementing safeguards that ensure key legal assessments remain grounded in human(e) judgment. Leverage existing insights from debates on AWS and research on human-machine teaming and human-computer interaction to inform discussions on AI-DSS.

---

[10] Article 57 of the First Additional Protocol to the Geneva Conventions. See also Dorsey, Bo supra n. 5 (on AI-DSS and their effects on the principle of precautions); Jessica Dorsey, *Proportionality under Pressure: The Effects of AI-Enabled Decision Support Systems, the Reasonable Commander Standard and Human(e) Judgment in Targeting*, forthcoming *International Review of the Red Cross* (2025) (on AI-DSS and their effects in the context of IHL proportionality assessments).

[11] Marta Bo, *Autonomous Weapons and the Responsibility Gap in light of the Mens Rea of the War Crime of Attacking Civilians in the ICC Statute*, 19 Journal of International Criminal Justice 2021; Bo, M., Bruun, L. and Boulanin, V., *Retaining Human Responsibility in the Development and Use of Autonomous Weapon Systems: On Accountability for Violations of International Humanitarian Law Involving AWS* (SIPRI: Stockholm, Oct. 2022), p. 41; Boulanin, V., Bruun, L. and Goussac, N., *Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human-Machine Interaction* (SIPRI: Stockholm, June 2021), p. 54; and Bruun, L., Bo, M. and Goussac, N., *Compliance with International Humanitarian Law in the Development and Use of Autonomous Weapon Systems: What Does IHL Permit, Prohibit and Require?* (SIPRI: Stockholm, Mar. 2023), p. 24. Elke Schwarz, "The (im)possibility of meaningful human control for lethal autonomous weapons systems," *Humanitarian Law and Policy*, 29 August 2018, found at: https://blogs.icrc.org/law-and-policy/2018/08/29/im-possibility-meaningful-human-control-lethal-autonomous-weapon-systems/.

ii. **Recognize** and address the incremental effects of AI-DSS design and use on human cognitive reasoning and critical deliberation. Promote awareness and attentiveness as a crucial part of reasserting and strengthening the exercise of human agency in targeting decision-making.

iii. **Reinforce** calls for greater attention to the implications of AI-DSS in armed conflict. Utilize platforms such as the UN General Assembly's First Committee on Disarmament and International Security and the Global Commission on the Responsible Use of AI in the Military Domain to foster inclusive and complementary discussions on the associated risks and systemic changes AI-DSS introduce.

United Nations Institute for Disarmament Research

[11 April 2025]

Artificial intelligence (AI) is rapidly transforming the military domain and profoundly influencing international peace and security. Initiatives such as the summits on Responsible AI in the Military Domain (REAIM) and the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, while not being universal processes, have significantly elevated international attention on the military applications of AI. In particular, they have moved the debate beyond lethal autonomous weapon systems (LAWS) and have successfully highlighted the multifaceted impacts of AI, fostering broader international policy engagement. Building on the political momentum generated by these initiatives, resolution 79/239 adopted by the United Nations General Assembly in December 2024 represented a significant milestone as the first UN resolution on AI in the military context and has offered Member States, international and regional organizations and the multi-stakeholder community the opportunity to share their views on opportunities and risks.

For many years, the United Nations Institute for Disarmament Research (UNIDIR) has played an important role in shaping and informing discussions on the broader impact of AI in the military domain, both within and beyond applications of this technology in weapon systems. It has undertaken research, facilitated multilateral dialogues, and offered policy insights that underline AI's transformative potential for international peace and security. This policy note draws from all the work conducted to summarise opportunities and risks and to offer a potential roadmap for future policy action.

The international community can now shape how AI is used in the military domain, putting principles of responsible AI at the core. A central challenge is the complexity of defining the "military domain". States and regions interpret the scope of this domain differently based on their unique security landscapes, realities and operational practices. For some countries, military roles extend to internal security tasks such as policing, border control, combating organized crime, protection of critical infrastructure or humanitarian relief in response to natural disasters. Others maintain a stricter definition, limiting military functions to battlefield engagements. These variations, rather than serving as barriers, offer important context for multilateral discussions. International governance frameworks should remain flexible and inclusive, acknowledging and adapting to diverse national and regional security perspectives.

In the many operational contexts within the military domain, AI acts as a force multiplier across several military tasks, including command and control (C2), information management and intelligence, advanced autonomy, logistics, training and simulation, and organizational and support functions. In C2, AI enhances the speed and quality of decision-making, thereby helping commanders rapidly analyse battlefield scenarios. It has the potential to improve adherence to international humanitarian law (IHL), for example by integrating detailed proportionality and other legal assessments. AI-driven intelligence tools analyse large volumes of data at speed, and so improve

situational awareness and threat detection. In logistics, AI optimizes supply chains and predictive maintenance, enhancing operational readiness and improving the sustainability of military operations over time. AI further supports advanced autonomy in drones, cybersecurity, and operations in the information domain. Training and simulation benefit from AI by creating personalized, realistic synthetic environments and scenarios. In short, if developed, deployed and used responsibly, AI could increase operational effectiveness while offering new ways to mitigate risks and reduce harm.

However, integrating AI in military contexts also presents significant risks and challenges – technological, security, legal, policy and ethical.

Technologically, military AI systems face issues related to the quality, availability and inherent biases of data. These may lead to unpredictable and potentially harmful outcomes, including violations of international law. The "black box" nature of AI systems, often coupled with their adaptiveness and highly context-dependent nature, complicates trustworthiness assessments and may, at times, challenge the conduct of effective investigations into alleged violations of IHL. Cybersecurity vulnerabilities also expose AI systems to adversarial attacks, requiring stringent security measures.

Security challenges include risks of miscalculation and unintended escalation, particularly through AI-enabled rapid decision-making processes and AI-enabled autonomy, which may result in escalatory responses. The potential for an AI arms race exacerbates international and regional tensions, possibly leading to destabilizing outcomes similar to historical arms competitions. The proliferation of AI technologies to non-state actors further complicates threat landscapes and necessitates robust life cycle management of military AI systems. Additionally, AI-generated disinformation threatens societal stability by undermining trust in information and can have a direct impact on military operations.

Legal challenges revolve around ensuring compliance with international law, particularly IHL and international human rights law. Key debates focus on, among other things, accountability and both state and individual responsibility for AI-driven actions, especially regarding lethal decisions. States diverge on whether existing legal frameworks are sufficient or if new, specialized regulations are required. Beyond international law, ethical considerations emphasize maintaining human judgment in critical decision-making and preventing societal biases from infiltrating AI systems. The latter requirement calls for greater diversity and inclusivity in AI development. Additionally, bridging gaps between government, academia and the private sector remains challenging yet crucial for effective governance.

Addressing these challenges requires a comprehensive road map with actions at the multilateral, regional and national levels.

Multilaterally, establishing a United Nations-led comprehensive platform that enables a regular institutional dialogue to address military AI's broader implications on international peace and security is key as it would provide an institutional framework to advance policy discussions. This platform could build on the existing internationally developed AI principles and frameworks, such as UNESCO's recommendations or the commitments made in the Global Digital Compact (e.g. safe, secure and trustworthy AI) and further refine them for application in the military domain. These principles could be further developed into voluntary norms of responsible behaviour in the development, deployment and use of AI in the military domain and provide a solid foundation for future multilateral instruments. In addition, such platform could be leveraged to develop practical confidence-building measures (CBMs), lead inclusive multi-stakeholder engagement, and deliver global capacity-building programmes that enhance global security via transparency, cooperation and predictability.

Regionally, existing organizational frameworks can be used to tailor CBMs and guidelines that reflect local security contexts. Cross-regional dialogues would

facilitate mutual learning, prevent information silos, and include diverse perspective which would encourage globally coherent responses.

Nationally, states should develop comprehensive AI strategies that detail vision, priorities and governance frameworks, ensuring compliance with international norms and ethical standards. Robust governance structures (e.g., dedicated AI steering committees and ethics boards), alongside iterative legal reviews, would enhance accountability and safety. Transparent communication and clearly defined accountability protocols would further support responsible AI implementation. High standards of data governance, life cycle management approaches, rigorous training programmes and updated military operational guidelines complete these proposed national measures, ensuring the responsible integration of AI in the military domain.

Table A below provides an overview of the proposed roadmap for policy action.

Table A: A roadmap for future policy action

Level	*Action*	*Rationale*
Multilateral	**Establish a multilateral process under United Nations auspices to provide a comprehensive platform for discussion on military applications of AI and their impact on international peace and security.** This process could be leveraged to: a. Develop a set of overarching, core principles of responsible AI in the military domain to help align national efforts and reduce risk. b. In the future, further develop these core principles into international voluntary norms or guidelines for responsible state behaviour in the development, deployment and use of AI in the military domain. These guidelines could take the form of a code of conduct or a political declaration supplemented by more technical instruments as required (e.g., on AI assurances, and robust protocols for testing and evaluation). c. Develop confidence-building measures (CBMs) for military AI. States could agree on and implement practical CBMs to increase transparency and trust regarding AI in the military domain.	Collectively, these multilateral actions aim to foster cooperation, set common rules and share knowledge on military AI at the international level with a view to increasing predictability. They aim to shape the global landscape so that all states move towards safer and more transparent integration of AI in the military domain, thereby reducing the risks. While clustered under a single umbrella recommendation, each of the actions above could be implemented on its own, although their mutually reinforcing nature would amplify the impact achieved if they are implemented in combination.

Level	*Action*	*Rationale*
	d. Promote multi-stakeholder engagement in support of multilateral policy action. e. Develop and implement a coherent capacity-building programme.	
Regional	**Leverage regional and subregional organizations and dialogues to discuss the issue of AI in the military domain.** Regional and sub-regional organizations could: a. Develop region-specific CBMs, norms or guidelines that reflect local contexts. b. Set up networks for information-sharing on AI-related best practices suited to their security landscape. c. Develop joint AI-development projects, aligning operational, legal and technical requirements. **Initiate cross-regional dialogues Initiate cross-regional dialogues on AI**, where two or more regional groups exchange lessons and possibly align their approaches.	Regional and subregional approaches allow tailoring to specific security realities and threat perceptions, which could lead to concrete results that are more aligned with specific needs. In addition, regional and subregional approaches could be leveraged to inform and shape global dialogues and strengthen context-specific capacity-building. Cross-regional dialogue can be a useful tool to enable mutual learning and avoid echo chambers.
National	**Implement a comprehensive approach to AI governance in the military domain** to include the following actions: a. Develop a comprehensive national strategy or policy on AI in security and defence. b. Establish robust governance structures and review processes. c. Implement transparency and accountability measures d. Implement robust data practices and governance frameworks for all military AI applications.	A national strategy clarifies roles and responsibilities, and provides a clear direction for the development, acquisition, integration and use of AI in the military domain. Dedicated structures provide focus and accountability. They create effective checkpoints that AI projects must pass and comply with consistently (e.g., ethical approval, legal clearance, safety certification), reducing chances of unsafe or unlawful deployment. Transparency builds public trust and international confidence that a state is using AI responsibly.

Level	*Action*	*Rationale*
	e. Manage AI capabilities throughout their entire life cycle – from design and development, through testing and deployment, to updates and decommissioning – with continuous risk assessments and mitigation at each stage.	Accountability ensures that the presence of AI does not create a vacuum of responsibility – maintaining the ethical and legal norm that humans are accountable for military actions.
	f. Invest in human capital and training by developing extensive training programmes for military personnel on AI and cultivating a new generation of AI-literate officers and specialists. This includes not only technical training but also training on the ethical and legal aspects of AI use in operations.	By prioritizing robust data governance and the provision of the necessary infrastructure to enable it, militaries can improve the performance and trustworthiness of their AI systems and reduce error rates. A life cycle view ensures that safety and compliance are ongoing commitments reducing chances of failure in the field and ensuring that accountability is maintained throughout the system's use. Human expertise and judgment remain critical. Training reduces misuse and enables more effective human-machine teaming.
	g. Review military operational guidelines to strengthen AI governance in military contexts, including military documentation (e.g. doctrines, standard operating procedures and others), and rules of engagement.	Existing military governance tools and instruments can be used to strengthen the governance of AI in the military domain at a more practical, tactical level, thereby offering an impactful complement to the highest levels of governance and the associated obligations emanating from international, regional and national laws and regulations.

E. Industry

Microsoft

[24 May 2024]

Microsoft welcomes the opportunity provided by the United Nations General Assembly resolution A/RES/79/239 on "Artificial Intelligence in the Military Domain and its Implications for International Peace and Security", and UNODA's invitation to share perspectives on the opportunities and challenges posed to international peace and security by the application of artificial intelligence (AI) in the military domain, with specific focus on areas other than lethal autonomous weapons systems.

Our perspectives reflect Microsoft's deep commitment to our Responsible AI Principles and our Secure Future Initiative, emphasizing cybersecurity, safeguarding international norms, and promoting trust in technology, and our active participation in multi-stakeholder initiatives including the UNIDIR-led Roundtable for AI, Security, and Ethics (RAISE).

I. Opportunities

Microsoft recognizes substantial opportunities in responsibly applied AI within the military domain, particularly:

- *Enhancing cybersecurity and defense capabilities:* AI significantly strengthens cybersecurity defenses by automating threat detection, enabling faster and more accurate responses to cyber threats. Technologies such as Microsoft Security Copilot illustrate the transformative potential of AI in defense, empowering cybersecurity professionals to identify and mitigate risks efficiently. Initiatives like Microsoft's Zero Day Quest and collaboration with MITRE ATT&CK demonstrate proactive industry efforts to enhance global cybersecurity preparedness and resilience.
- *Broad spectrum of military applications:* Beyond cybersecurity, responsibly designed AI can significantly enhance efficiency and effectiveness across logistics, command and control systems, intelligence processing, military training, peacekeeping, humanitarian assistance, and disaster relief operations. Diverse applications underscore AI's transformative potential beyond combat scenarios alone.
- *Improving compliance with international humanitarian law:* AI technologies should improve the accuracy and effectiveness of targeting processes, aiding militaries to better adhere to principles of distinction, proportionality, and necessity. AI should significantly enhance protections for civilians and civilian infrastructure, thereby reducing unintended collateral damage in conflict.
- *Capacity building and international cooperation:* The adoption of AI in the military domain presents opportunities for global knowledge-sharing and capacity-building initiatives. International partnerships should support developing nations by sharing security capabilities, knowledge, and best practices, thus bridging technological divides and fostering global stability.

II. Challenges

Microsoft equally acknowledges significant challenges and risks associated with AI applications in the military domain:

- *AI-enhanced cyber threats:* AI has escalated cyber threat capabilities, empowering state-sponsored and criminal actors to carry out increasingly sophisticated cyber operations. These AI-driven threats include advanced phishing campaigns, automated exploitation of vulnerabilities, and adaptive malware, significantly increasing global cybersecurity risks.
- *Risks of escalation and miscalculation:* Integrating AI into military decision-making risks unintended escalation and/or miscalculation. Rapid, automated decision-making processes may inadvertently lower conflict thresholds, amplifying risks of destabilization or accidental conflict.
- *Proliferation and uncontrolled diffusion:* Uncontrolled diffusion, especially through open-source models and decentralized development, heightens the risk of malicious use by both state and non-state actors, including terrorist groups and cyber mercenaries. Increasingly accessible dual-use and proprietary AI

systems enable actors even with limited resources can gain access to capabilities that previously required significant investment or expertise, posing additional threats to international security and stability.

- *Algorithmic bias and ethical implications:* Algorithmic biases embedded within AI systems pose ethical and humanitarian concerns. Biases related to gender, race, age, or socioeconomic factors in AI datasets can intentionally and unintentionally perpetuate inequality and discrimination, particularly within sensitive military and security applications.
- *Digital divides and inequality:* Without deliberate policy actions, disparities between developed and developing nations in AI capabilities could deepen, increasing geopolitical tensions and socio-economic inequalities, thus undermining long-term global stability.

III. Relevant normative proposals

Microsoft recognizes several existing and emerging normative frameworks relevant to AI governance in the military domain, including:

- UNIDIR's RAISE initiative, facilitating international multi-stakeholder dialogues and governance proposals.
- The Responsible AI in the Military Domain (REAIM) Summits, emphasizing transparency, accountability, and human oversight at the international level.
- The US Department of Defense Responsible AI Strategy, highlighting responsibility, equitability, traceability, reliability, and governability.
- NATO's Principles of Responsible Use for AI in Defence, emphasizing reliability, governability, and traceability among member nations.

IV. Microsoft recommendations

To maximize opportunities and mitigate the challenges, Microsoft proposes several key recommendations:

- *Establish clear international norms and standards:* Develop explicit international norms and industry standards governing responsible use and development of military AI. These norms should delineate acceptable and unacceptable behaviors, providing robust frameworks to deter misuse and foster transparency and accountability, supported where appropriate by monitoring or compliance mechanisms. AI governance frameworks should explicitly differentiate operational contexts, such as peacekeeping, humanitarian assistance, crisis management, and conflict scenarios, to appropriately address varied ethical, legal, and humanitarian considerations. To ensure continued relevance, such norms should be periodically reviewed and updated to reflect evolving technological developments and operational realities.
- *Ensure human-centric oversight and accountability:* Adopt policies ensuring meaningful human judgment, oversight, and accountability remain central to military decisions involving AI, particularly regarding the use of force. Clear oversight mechanisms and enforceable accountability structures, including rigorous human control and review processes, are necessary to maintain ethical standards, avoid automation bias, and mitigate unintended consequences.
- *Advance secure and transparent AI development practices:* Promote rigorous technical standards and comprehensive life cycle management protocols covering pre-design, development, testing, deployment, operation, acquisition, and decommissioning. Robust vulnerability management, security audits, and

transparent development and deployment processes should be integral components, alongside clear capacity-building measures, ensuring AI systems remain secure, responsible, and resilient throughout their operational life cycle.

- *Enhance responsible data governance practices:* Establish clear international guidelines on responsible data governance specifically tailored to military AI applications. Transparent and accountable data management practices addressing collection, sharing, storage, training, and operational usage are crucial for managing dual-use risks, preventing misuse, and maintaining strict compliance with international legal and ethical frameworks.
- *Address and reduce algorithmic bias:* Prioritize addressing algorithmic bias through rigorous testing, transparent data practices, and inclusive AI development processes. Developers and users should establish clear policies to proactively identify, mitigate, and remediate biases, especially when AI systems are deployed in sensitive military or security contexts.
- *Promote responsible innovation and risk-based regulation:* Support regulatory frameworks that are risk-based, outcome-oriented, and balanced, ensuring they encourage innovation while adequately addressing security and ethical risks associated with AI deployment. Industry should advocate for flexible, adaptive regulations that keep pace with technological change, without imposing overly prescriptive or impractical requirements. Industry-led initiatives, such as voluntary codes of conduct, vulnerability disclosure standards, and collaborative red-teaming exercises, should be actively supported and integrated into broader international normative frameworks.
- *Strengthen international governance and alignment:* Support and actively engage in international initiatives, including REAIM Summits and dialogues at the UN General Assembly and UN Security Council. Robust international governance frameworks, characterized by transparency, clear accountability measures, and trust-building mechanisms, are essential for coherent and inclusive approaches to AI governance. Member States and stakeholders should coordinate closely through these forums to reduce fragmentation and ensure global alignment.
- *Support knowledge-sharing and awareness-raising with the UN system:* Encourage and actively contribute to efforts by the UN Secretariat and relevant UN entities to convene meaningful multi-stakeholder expert dialogues, workshops, and knowledge-sharing on AI in the military domain. Exchanges through voluntary contributions, technical expertise, and collaborative initiatives should aim at enhancing global understanding of AI's implications for international peace and security.
- *Strengthen international cooperation and information sharing:* Encourage robust international cooperation, emphasizing real-time threat intelligence sharing and joint attribution mechanisms. Industry actors should actively participate in collective cybersecurity efforts, enhancing global cybersecurity preparedness and response.
- *Foster multi-stakeholder dialogue and collaboration:* Actively participate in and support forums such as RAISE, involving states, international organizations, academia, civil society, and industry. Such inclusive dialogues are essential for mutual understanding, shaping responsible AI practices, and developing collaborative governance structures.

V. Conclusion

Microsoft is deeply committed to proactive collaboration with Member States, the UN system, industry, and civil society to implement these recommendations swiftly and effectively. Through sustained collective efforts and ongoing engagement in multi-stakeholder initiatives, Microsoft will continue supporting responsible AI governance, innovation, and practices that meaningfully contribute to international peace and security.