ICRC

武装冲突中的人工智能与机器学习：以人为本的方法

日内瓦，2019 年 6 月 6 日

一、引言

红十字国际委员会是一个公正、中立和独立的组织，其特有的人道使命是保护武装冲突和其他暴力局势受难者的生命与尊严，并向他们提供援助。红十字国际委员会还通过推广和加强人道法与普遍人道原则，尽力防止苦难发生。

在冲突不断增加、技术迅速变化的时代，红十字国际委员会既需要了解新技术对受武装冲突影响民众所造成的影响，也需要设计人道解决方案，以满足最弱势群体的需求。

与不同行业和地区的其他组织一样，红十字国际委员会正在努力应对**人工智能**和**机器学习**对其工作所产生的影响。人工智能是指利用计算机系统开展以往需要人类智力、认知或推理的工作[1]；而机器学习则涉及使用大量数据改善自身性能并从经验中“学习”的人工智能系统。[2]这些软件工具或算法能够应用于多种不同工作，其潜在影响可能十分深远，但尚未被完全理解。

在人工智能与机器学习的应用领域当中，红十字国际委员会特别关注两个宽泛而又截然不同的问题：第一，相关技术**在作战行为**或其他暴力局势中的**应用**[3]；第二，相关技术在援助及保护武装冲突受害者的**人道行动中的应用**。[4]本文阐述了红十字国际委员会对下列议题的观点：人工智能与机器学习在武装冲突中的使用，其潜在人道后果，以及规范其发展与使用的相关法律义务和道德考量。不过，本文还提及了包括红十字国际委员会在内的组织在人道行动中使用人工智能工具的情况。

二、红十字国际委员会对待新作战技术的方法

评估武装冲突中当前与此后短期内一些新发展的影响，是红十字国际委员会的一贯做法。评估内容包括考虑新的作战手段和方法；特别是其与国际人道法（又称武装冲突法或战争法）规则是否相一致，以及受保护人员遭受不利人道后果的风险。

红十字国际委员会并不反对新作战技术本身。某些军事技术，例如能够提高攻击精准度的技术，可能有助于冲突各方尽量减少战争的人道后果，尤其是对平民的人道后果，并保证尊重战争规则。然而，与任何新作战技术一样，精准技术本身并无益处，而一线所遭受的人道后果将取决于新武器的实际使用方式。因此，有必要对新技术就其技术特性以及使用或预计使用方式进行实际评估。

---

[1] 《牛津词典》，“人工智能”：https://en.oxforddictionaries.com/definition/artificial_intelligence.
[2] 《牛津词典》，“机器学习”：https://en.oxforddictionaries.com/definition/machine_learning.
[3] ICRC, “Expert views on the frontiers of artificial intelligence and conflict”, *ICRC Humanitarian Law & Policy Blog*, 19 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict.
[4] ICRC, *Submission to the UN High-Level Panel on Digital Cooperation*, January 2019: https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UN-Panel-Digital-Cooperation.pdf.

任何新作战技术必须，也必须要能够按照现有国际人道法规则加以使用，这是最低要求。[5]然而，新作战技术之特性、其计划与预期使用情况及其可预见的人道后果可能会产生一些问题，即考虑到其可预见的影响，现有规则是否充分，或是否需要加以澄清或补充。[6]显而易见的是，新兴技术用于军事领域并非不可避免。这是国家作出的选择，必须受到现有规则的约束，并考虑到对平民和不再参与敌对行动的战斗员的潜在人道后果，以及更广泛的“人道”与“公众良心”的考量。[7]

三、冲突各方对人工智能与机器学习的使用

武装冲突各方（无论是国家还是非国家武装团体）在作战行为中可能采取的人工智能与机器学习的诸多使用方式及其潜在影响尚不明晰。然而，**从人道视角而言**，包括对于遵守国际人道法而言，至少有**三个重叠的领域具有相关性**。

1. 实体机器人系统（包括武器）自主性的增加

一项重要的应用是数字**人工智能与机器学习工具用于控制实体军事装备**，特别是数量日益增加、规格不同、功能各异的海、陆、空无人机器人系统。人工智能与机器学习或可增强此类武装或非武装机器人平台的自主性，并控制整个系统或特定功能，如飞行、导航、监视或目标选择。

对红十字国际委员会而言，考虑到人类对武器和武力的使用失控的风险，**自主武器系统**（在选择和攻击目标的“关键功能”上具有自主性的武器系统）从人道、法律和道德视角来看是我们最为关切的问题。[8]这种失控会给平民带来风险，因为后果无法预测；存在法律问题[9]，因为按照国际人道法的要求，战斗员进行攻击时必须根据实际情况作出判断；还会面临道德问题[10]，因为武力使用决策中的人类主体作用对维护道德责任和人类尊严是必要的。基于上述原因，红十字国际委员会一直敦促各国确定人类控制的实际要素，作为武器系统自主性限制国际标准的基础，其重点如下[11]：

---

[5] 日内瓦四公约《第一附加议定书》（1977 年 6 月 8 日）的缔约国有义务在发展及取得新武器期间并在将其用于武装冲突之前对其进行法律审查。对其他国家而言，法律审查是常识性措施，可帮助确保该国武装部队能够根据该国承担的国际义务从事敌对行动。

[6] 红十字国际委员会，《国际人道法及其在当代武装冲突中面临的挑战》，第 32 届红十字与红新月国际大会报告，日内瓦，2015 年 10 月，第 37-45 页：https://www.icrc.org/zh/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts.

[7] “人道原则”与“公众良心要求”见于《第一附加议定书》第 1 条第 2 款以及《第二附加议定书》的序文，称为“马斯顿条款”，是习惯国际人道法的一部分。

[8] ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019: https://www.unog.ch/__80256ee600585943.nsf/(httpPages)/5c00ff8e35b6466dc125839b003b62a1?OpenDocument&ExpandSection=7#_Section7.

[9] Davison, N., “Autonomous weapon systems under international humanitarian law”, in *Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament* Affairs *(UNODA) Occasional Papers No. 30*, November 2017: https://www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law.

[10] ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, report of an expert meeting, 3 April 2018: https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control.

[11] ICRC, *The Element of Human Control*, Working Paper, Convention on Certain Conventional Weapons (CCW) Meeting of High Contracting Parties, CCW/MSP/2018/WP.3, 20 November, 2018: https://www.unog.ch/80256EDD006B8954/(httpAssets)/810B2543E1B5283BC125834A005EF8E3/$file/CCW_MSP_2018_WP3.pdf.

- 无需人类干预就能选择并攻击目标的武器在运行过程中所需**人为监督、干预及失效能力**的程度如何？

- 在运行和使用后果方面需要在多大程度上确保**可预测性？**在故障或失灵的可能性方面需要在多大程度上确保**可靠性**？

- 特别是在**任务**、**目标**（如装备或人员）、**使用环境**（如无人区或人口稠密地区）、**自主运行时间**（即时间限制）和**打击范围**（即空间限制）方面，还需要哪些武器**运行限制**？

重要的是要认识到**并非所有自主武器都包含人工智能和机器学习技术**；例如具备自主模式的空防系统等在关键功能上具有自主性的现有武器通常使用基于规则的简单控制软件进行目标的选择和攻击。然而，**人工智能与机器学习软件**——特别是针对“自动目标识别”而研发的软件——**可能构成未来自主武器系统的基础，使这些武器的不可预测性增加了新的维度**，并造成对缺乏可解释性和存在偏差的担忧（见第五部分第二点）。[12]同类软件也可用于选择目标的“决策支持”中，而非直接控制武器系统（见第三部分第三点）。

反过来，并非所有使用人工智能和机器学习的军用机器人系统都是自主武器，因为该软件可能用于除目标选择之外的控制功能，如监视、导航和飞行。虽然，从红十字国际委员会的角度来看，武器系统（包括人工智能系统）的自主性产生了最为紧迫的问题，但使用人工智能和机器学习技术增加一般军事装备（如无人机、陆上车辆和海船）的自主性也可能出现人机交互及安全性问题。民用领域有关确保自动驾驶汽车或无人机等自主交通工具安全性的讨论可能会为其在武装冲突中的使用提供经验教训（另见第三部分第三点）。

2. 网络战和信息战的新手段

人工智能与机器学习在网络武器或网络能力发展中的应用是另一个重要领域。并不是所有的网络能力都包含人工智能和机器学习技术。然而，这些技术预计将**改变防御网络攻击能力及攻击能力的性质**。例如，人工智能和机器学习支持的网络功能可以自动搜索漏洞以利用或防御网络攻击，同时自动发起反攻。此类进展可能会扩大攻击规模，改变攻击性质，也许还会改变攻击的严重程度。[13]其中一些武器系统甚或可称为“数字自主武器”，可能会引发自主武器所面临的类似的人类控制问题。[14]

红十字国际委员会在网络战方面的重点仍然是确保现有国际人道法规则在武装冲突的任何网络攻击中均得到遵守，并确保保护民用基础设施和服务所面临的特殊挑战由实施或防御此类攻击的人进行应对[15]，以尽量减少人道代价。[16]

人工智能和机器学习在数字领域的一个相关应用是**其在信息战中的使用**，特别是用于编造并传播意图欺骗的虚假信息，即**刻意发布的虚假信息**，以及传播没有欺骗意图的虚假信息，即**错误信息**。并非所有虚假错误信息均涉及人工智能和机器学习，但这些技术似乎有可能会改变战争中信息操纵的性质和规模及其潜在后果。人工智能系统已广泛用于制作文本、音频、照片或视频等各类虚假信息，越来越难以与真实信息区分开来。冲突各方使用这些系统以扩展古老

[12] ICRC, *Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems under agenda item 6(b)*, Geneva, 27-31 August 2018: https://www.unog.ch/80256EDD006B8954/(httpAssets)/151EF67AD8224E14C125830600531382/$file/2018_GGE+LAWS+2_6b_ICRC.pdf.
[13] Brundage, M. *et. al.*, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, February 2018.
[14] United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, UNIDIR, 2017.
[15] 虽然红十字国际委员会主张国际人道法适用于网络行动，但切不可将此视为容忍网络战，或容忍网络空间的军事化：*ICRC,* 2015, *op. cit.,* pp. 38–44.
[16] ICRC, *The Potential Human Cost of Cyber Operations*, report of an expert meeting, May 2019: https://www.icrc.org/en/document/potential-human-cost-cyber-operations.

的宣传方法，以操纵舆论并影响决策，可能对一线战场产生重大影响。[17]对红十字国际委员会而言，我们担心平民可能会由于数字虚假信息或错误信息而遭到逮捕或虐待、歧视，或无法使用基本服务，或人身或财产遭到攻击。[18]

3. 武装冲突中决策性质的变化

也许，最为广泛而深远的应用是利用**人工智能和机器学习进行决策**，包括对数据源的广泛收集分析，以识别人员或物体、评估生活或行为模式、为军事战略或行动提出建议，或预测未来行动或局势。

这些“**决策支持**”或“**自动决策**”**系统实际上是智能、监视和侦察工具的扩展**，使用人工智能与机器学习自动分析大型数据集，以便为人类作出特定决策提供“建议”，或是既进行自动分析，又自动执行系统随后的决策或行动。相关的人工智能和机器学习应用包括模式识别、自然语言处理、图像识别、面部识别和行为识别。**这些系统的可能用途极为广泛**，涉及范围从决定攻击对象、攻击时间[19]，拘留对象和拘留时间[20]，到决定军事战略，甚至是核武器的使用[21]，以及具体行动，包括试图预测敌方行动或先发制人。[22]这些系统在决策领域的应用，视其使用或滥用情况，以及技术能力和局限性，可能会增加平民居民的风险。

基于人工智能和机器学习的**决策支持系统**可以通过增加现有信息收集分析的速度和范围，可以使人类在依照国际人道法开展敌对行动方面作出更好的决策，并最大限度地降低平民的风险。然而，同样的算法生成的分析或预测也可能对决策产生负面影响、助长违反国际人道法的行为，并加剧平民的风险，尤其是考虑到目前技术的局限性，例如不可预测性、缺乏可解释性及偏差，情况就更是如此（见第五部分第二点）。

从人道角度来看，冲突各方**各种不同的由人工智能介导或受其影响的决定**可能是相关联的，在此类决定可能具有造成人员伤亡或摧毁物体的风险，且受到国际人道法具体规则的规制时尤为如此。例如，人工智能和机器学习用于**武装冲突中的目标选择决策**，在会造成严重人员伤亡的情况下，需要进行具体考量，以确保人类始终能够在遵守敌对行动相关法律规则的基础上作出基于局势的判断（见第五部分）。用于直接发起攻击（而不是为人类决策者提供分析或“建议”）的人工智能系统实际上会被视作自主武器系统，也会产生类似问题（见第三部分第一点）。

使用决策支持和自动决策系统还可能**在其他应用领域产生法律和道德问题，例如武装冲突中实施拘留的决定**。这也会给民众生活造成严重后果，并受到国际人道法具体规则的规制。就这一点而言，民用领域也存在类似的探讨：人类判断的作用、偏差和缺乏精度的问题、警察在逮捕决定中使用的风险评估算法，以及刑事司法系统关于判刑和保释的决定。[23]

---

[17] Hill, S., and Marsan, N., “Artificial Intelligence and Accountability: A Multinational Legal Perspective” in *Big Data and Artificial Intelligence for Military Decision Making*, Meeting proceedings STO-MP-IST-160, NATO, 2018.

[18] ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, March 2019, p. 9: https://www.icrc.org/en/event/digital-risks-symposium.

[19] USA, *Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems*, Working Paper, Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts, March 2019.

[20] Deeks, A., “Predicting Enemies”, Virginia Public Law and Legal Theory Research Paper No. 2018-21, March 2018: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

[21] Boulanin, V., (ed.), The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk. Vol. 1, Euro-Atlantic Perspectives, Stockholm International Peace Research Institute (SIPRI), May 2019.

[22] Hill, S., and Marsan, N., *op. cit.*

[23] McGregor, L., “The need for clear governance frameworks on predictive algorithms in military settings”, *ICRC Humanitarian Law & Policy Blog*, 28 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings; AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.

更广泛地说，此类人工智能和机器学习工具可能会导致**战争个人化**（与民用领域中的服务个性化相似）程度增加。这指的是数字系统将传感器、通讯系统、数据库、社交媒体和生物识别数据等多个来源的个人验证信息汇总，由算法生成关于一个人及其状态和成为攻击目标可能性的判定，或预测其未来的行动。

一般而言，由于滥用人工智能所支持的**数字监控、监测和入侵技术**，平民居民可能面临的人道后果（**数字风险**）可能包括成为攻击目标、遭到逮捕、面临虐待、身份遭到盗用而无法使用服务、资产失窃或因害怕受到监视而承受心理影响。[24]

四、人道行动中人工智能与机器学习的使用

人道行动中（包括红十字国际委员会）对人工智能与机器学习的应用方式可能也非常广泛。人道组织正在探索将此类工具用于对特定行动地点的公共数据来源进行舆情监测、监控与分析，并可以**为人道需求评估提供信息**，如所需援助类型（食物、供水、避难所、经济援助、医疗援助）及需要援助的地点。

类似的人工智能数据聚合与分析工具可能会用来帮助**了解**一线战场的**人道后果**，包括平民的保护需求。例如，图像、视频或其他模式分析工具可评估民用基础设施遭到的破坏，人口迁移模式，粮食作物的活力，或武器污染（未爆炸弹药）程度。这些系统可能还会用于分析图像及视频，以发现并评估敌对行动及其人道后果。

例如，红十字国际委员会已经开发了**舆情分析系统界面**，使用人工智能与机器学习捕捉并分析大量数据，为特定行动背景下的人道工作提供信息与支持，包括使用预测分析帮助确定人道需求。

各类人道服务或可受益于人工智能与机器学习在特定工作中的应用。例如，人工智能面部识别以及借助自然语言处理进行姓名匹配等能够**改善失踪人员身份辨认结果**的技术就受到了关注；红十字国际委员会一直在探索使用此类技术，支持其中央寻人局的工作，帮助因冲突而离散的家庭重聚。该组织还在探索将基于人工智能与机器学习的**图像分析及模式识别用于卫星影像**，从而绘制人口密度图表，支持城市地区的基础设施援建项目，或补充其保护平民工作中尊重国际人道法行为的相关记录。

这些**人道行动中的应用也会带来潜在的风险**，以及法律与道德问题，在涉及数据保护、隐私、人权、问责以及确保人类参与对民众生活及生计有重大后果的决策等方面时尤为如此。任何人道行动中的应用在制定与落实上都必须遵照数字环境下**“零伤害”**的原则，并（包括在涉及个人数据保护时）尊重隐私权。

红十字国际委员会还会确保在考虑到对技术能力及局限进行实际评估的基础上，在其人工智能与机器学习应用的设计与使用方面体现出**中立、独立、公正的人道行动的核心原则与价值观**（见第五部分第二点）。红十字国际委员会现与布鲁塞尔隐私中心共同主导一项人道行动中数据保护的相关活动，目的是制定人道领域人工智能与机器学习等新技术使用的指南，在效益最大化的同时也将上述核心考量纳入其中。红十字国际委员会/布鲁塞尔隐私中心《人道行动中的数据保护手册》第二版随后即将面世。[25]

[24] ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, *op. cit*., p. 8.
[25] ICRC, *Handbook on Data Protection in Humanitarian Action,* 2nd Edition, 30 October 2018: https://www.icrc.org/en/document/handbook-data-protection-humanitarian-action-second-edition.

五、以人为本的方法

作为努力为受武装冲突及其他暴力局势影响民众提供保护和援助，履行国际人道法所赋予的职责，并受基本原则中人道原则指导的人道组织[26]，**红十字国际委员会认为，针对人工智能与机器学习的发展与使用，确保采取真正以人为本的方法至关重要**。要做到这一点，首先就要考虑人类的义务与责任，以及为确保这些技术符合国际法以及社会与道德价值观所应达到的要求。

1. 确保人类控制与判断

红十字国际委员会认为，**在可能**对民众生活**产生严重影响**（尤其是生命威胁），且受到国际人道法规则规制**的工作中，必须保留人类控制；在类似性质的决策中，必须保留人类判断**。**人工智能与机器学习必须用于服务人类，并帮助而非取代人类决策者。**考虑到正在开发此类技术用于本由人类执行的任务，人工智能和机器学习应用的发展与武装冲突中人类所处的中心地位之间就存在固有的矛盾，需要持续予以关注。

人类控制与判断对可能导致人员伤亡或民用基础设施的破坏或损毁的工作及决策尤为重要。这可能会产生最为严重的法律与道德问题，或将需要采取政策性应对措施，如制定新的规则与法规。**最为重要的是关于使用武力的决策，也就是在武装冲突中选择何人何物为目标，并对其发起攻击。**然而，可能使用人工智能的更广范围的工作与决策也会对受武装冲突影响的民众造成严重后果，例如逮捕及拘留相关的决定。在考虑人工智能用于敏感性工作及决策这一方面，或许可以借鉴民用领域“安全关键”人工智能系统治理的相关经验，即那些在出现失灵的情况下会造成人员伤亡或严重财产或环境破坏的系统。[27]

另一矛盾领域是**人类与机器在执行不同工作时速度上的差异**。由于人类是武装冲突中的法律以及道德主体，他们在作战行为中所使用的技术必须在设计与使用上使战斗员能够履行其法律与道德义务及责任。这可能会对用于决策的人工智能及机器学习系统产生重大影响；为保留人类判断，可能需要确保系统的设计与使用按照“人类速度”帮助决策，而非按照“机器速度”加速决策，超越人类干预。

（1）武装冲突中人类控制的法律基础

对冲突各方而言，**对用作作战手段和方法的人工智能与机器学习系统保留人类控制是确保遵守法律的要求**。国际人道法规则规制的对象是人类。遵守并实施法律的是人类，对违法行为负责的也是人类。特别要提出的是，战斗员需要根据国际人道法规制敌对行动的规则履行作出所需判断的特殊义务，这一责任不可能转而由一台机器、一个软件或一个算法来履行。

这些规则要求应由负责计划、决定及执行攻击的人员**作出符合局势的判断**，以确保符合：**区分原则**——可合法攻击的军事目标以及不得攻击的平民或民用物体；比例原则——攻击预期

[26] 红十字国际委员会和红十字会与红新月会国际联合会，《国际红十字与红新月运动基本原则：人道行动道德规范和工具》，2015 年 11 月：https://shop.icrc.org/les-principes-fondamentaux-de-la-croix-rouge-et-du-croissant-rouge-2757.html.

[27] 例如，见人工智能合作组织对人工智能与机器学习技术的安全表示关注，认为这是“一个迫切的短期问题。此类技术在医药、运输、工程、计算机安全及其他领域的应用取决于使人工智能系统能够在无法确定、难以预期且可能存在敌意的环境中安全运作的能力。”The Partnership on AI, *Safety-Critical AI: Charter,* 2018: https://www.partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions.

对平民造成的附带损害与预期的具体和直接的军事利益相比并不过分；以及**攻击中的预防措施原则**——从而尽量进一步降低对平民的风险。

如人工智能系统用于攻击——无论是用于实体武器或网络武器系统，还是用于决策支持系统——**其设计与使用必须确保战斗员能够作出相关判断。**[28]就自主武器系统而言，《某些常规武器公约》的缔约国已认识到，“必须保留”武器系统使用及武力使用方面的“人类责任”[29]，许多国家、国际组织（包括红十字国际委员会）以及民间团体都在强调需要保留人类控制，以确保遵守国际人道法，并符合道德价值观。[30]

除武力使用及目标选择之外，人工智能系统在可能用于其他受国际人道法具体规则规制的决策时，也需要对必要的人类控制与判断进行审慎考量，拘留就是这种情况。[31]

（2）人类控制的道德基础

人工智能与机器学习的新兴应用还使得道德问题成为公众讨论的首要议题。在由政府、科学家、伦理学家、研究机构及技术公司制定并达成一致的**一般性“人工智能原则”中，一个共性层面就是人类因素**在确保遵守法律及道德可接受性上**的重要性**。

例如，2017 年《阿西洛马人工智能原则》强调奉行人类价值观，符合“人类尊严、权利、自由和文化多样性”，进行人类控制；“为实现人为目标，人类应该选择如何以及是否由人工智能代为决策”[32]。欧盟委员会人工智能高级专家组强调“人类主体作用与监督”的重要性，认为人工智能系统应“支持人类自主与决策”，并通过人为干预、监督或指挥的方法确保人类的监管。[33]2019 年 5 月由经合组织全部 36 个成员国，以及阿根廷、巴西、哥伦比亚、哥斯达黎加、秘鲁以及罗马尼亚一致通过的《经济合作与发展组织人工智能原则》突出了“以人为本的价值观及公平”的重要性，具体指出人工智能用户“应实施人类决策能力等适应局势情况，且与技术发展水平相一致的有关机制与保障性措施”[34]。2019 年 5 月，由中国领先科研院所及技术公司一致通过的《人工智能北京共识》指出“人工智能及其产品的研发者应不断提升模型与系统的成熟度、鲁棒性、可靠性、可控性”，并鼓励“探索人机协同……更能发挥人类优势和特点”[35]。大量技术公司也各自发表了“人工智能原则”，突出人类控制的重要性[36]，尤其是针对产生伤害风险的敏感性应用的重要性[37]，并强调“人工智能的目的……是增强，而非取代人类智能”[38]。

---

[28] ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, *op. cit.*

[29] United Nations, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, CCW/GGE.1/2018/3, 23 October 2018, Section III. A. 26(b) & III. C. 28(f): http://undocs.org/en/CCW/GGE.1/2018/3.

[30] See, for example, statements delivered at the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems, 25–29 March 2019: https://www.unog.ch/80256EE600585943/(httpPages)/5C00FF8E35B6466DC125839B003B62A1?OpenDocument.

[31] Bridgeman, T., “The viability of data-reliant predictive systems in armed conflict detention”, *ICRC Humanitarian Law & Policy Blog*, 8 April 2019: https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention.

[32] 生命未来研究所，《阿西洛马人工智能原则》，2017 年：https://futureoflife.org/ai-principles-chinese/.

[33] European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

[34] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, 22 May 2019: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[35] 北京智源人工智能研究院（BAAI），《人工智能北京共识》，2019 年 5 月 28 日：https://baip.baai.ac.cn.

[36] Google, *AI at Google: Our principles*, 7 June 2018: https://www.blog.google/technology/ai/ai-principles. “我们会设计能够提供适当反馈机会与相关解释，并具备吸引力的人工智能系统。我们的人工智能技术会根据适当的人类指导及控制进行运作。”

[37] Microsoft, “Microsoft AI principles”, 2019: https://www.microsoft.com/en-us/ai/our-approach-to-ai; Sauer, R., “Six principles to guide Microsoft’s facial recognition work”, 17 December 2018: https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work.

[38] IBM, “IBM’s Principles for Trust and Transparency”, 30 May 2018: https://www.ibm.com/blogs/policy/trust-principles.

一些政府还在制定面向军队的人工智能原则。美国国防部在其 2018 年《人工智能战略》中呼吁“以人为本”使用人工智能，向国防创新委员会下达了制定《国防人工智能原则》的任务。[39]法国国防部则致力于按照三大指导原则——遵守国际法、维持充分的人类控制以及确保永久性指挥责任——使用人工智能技术，并将成立部长级道德委员会以应对新兴技术问题。[40]

在红十字国际委员会看来，在对民众生活造成严重后果的工作中保留**人类控制**，并在类似决策中保留**人类判断，也对在战争中维护一定程度的人道标准至关重要。红十字国际委员会强调，需要在武装冲突中武力使用决策上保留人类主体作用。**[41]这一观点源自对更为广泛的人道精神、道德责任、人类尊严及公众良心要求的道德考量。

然而，人类主体作用的道德考量或可更广泛地适用于人工智能与机器学习在武装冲突及其他暴力局势中的其他用途。或许可以借鉴社会上关于人工智能与机器学习两用技术敏感应用（尤其是安全关键系统）的探讨，以及私营部门科学家及开发人员提出的相关治理建议。例如，谷歌曾表示可能存在“敏感情况，在这种情况下，无论人工智能系统精度多高，社会都会希望由人类作出最终决定”；同时，如果犯罪行为的法律判决或生命攸关的医疗决策等高危决策完全交由机器负责，“那么人们有充分理由将此类行为视作对人类尊严的侮辱”[42]。微软在考虑人工智能面部识别时，强调“针对可能对人产生重大影响的用途”，要确保“适当程度的人类控制”，要求对涉及“个体的身心伤害风险，对个体的就业前景或使用金融服务的能力造成负面影响、可能对人权产生影响，或可能使个体的个人自由受到侵犯”的敏感用途采取“人为操控”方法，或进行“有意义的人类审查”[43]。由于武装冲突中的应用可能是敏感度最高的，上述宽泛的探讨之中可能就包含人工智能所需限制的相关洞见。

保留**人类控制与判断将是**确保守法，并缓和人工智能与机器学习某些应用所引发道德关切的**重要要素。但**如果没有充分考虑到下列人机交互问题，**其本身并不足以抵御潜在风险：对情况的认识**（对人类干预介入时系统状态的认识）；有效人类干预的**可用时间**；**自动化偏差**（人类过度信任系统的风险）；以及**道德缓冲**（人类将责任转移至系统的风险）。[44]而且，确保有意义的有效人类控制及判断既需要审慎考量人工智能和机器学习技术的能力，又需要考量其局限性。

2. 理解人工智能与机器学习的技术局限性

虽然已就人工智能与机器学习的新功能进行了广泛探讨，**但对此类技术的能力与局限性进行实际评估也是必要的**，在将其用于武装冲突中时尤为如此。首先，应当承认在将人工智能与机器学习用于进行特定任务或决策时，我们并非在进行同类替换。这需要**理解人类与机器在开展工作时的根本区别，以及两者各自具备的优劣势**；人类与机器做事方式不同，分工也不同。我们必须明确的是，作为供人类使用的无生命物体与工具，“机器无论多么善于伪装，永远也无法在其交互中体现真正的人性。”[45]

[39] US Department of Defense, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, 2019.
[40] France Ministry of Defence, “Florence Parly wants high-performance, robust and properly controlled Artificial Intelligence”, 10 April 2019, https://www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee.
[41] ICRC, *ICRC strategy 2019-2022*, 2018, p. 15: https://www.icrc.org/en/publication/4354-icrc-strategy-2019-2022.
[42] Google, *Perspectives on Issues in AI Governance*, January 2019 p. 23–24: http://ai.google/perspectives-on-issues-in-AI-governance.
[43] Sauer, R., *op. cit.* “我们会鼓励并帮助我们的客户使用面部识别技术，针对可能对人产生重大影响的用途确保适当程度的人类控制。”
[44] ICRC, *Ethics and autonomous weapon systems: An Ethical Basis for Human Control?*, *op. cit.* p. 13.
[45] Google, 2019, *op. cit.* p. 22.

在谨记这一点的基础上，考虑武装冲突中（以及事实上在人道行动中）的应用时，有几项技术问题值得关注。**人工智能，尤其是机器学习，因其存在不可预测性及不可靠性**（或者说是安全问题）[46]**、缺乏透明度**（或者说是可解释性）**且存在偏差，而令人担忧。**[47]

机器学习系统并非按照预编程指令顺序工作，而是**根据其所接触到的数据制定自己的规则**——或是训练数据，或采用试错与环境交互的方法。**因此，**就（有特定输入的）特定局势中如何运作（达到输出）这一方面，与预先编程系统相比，**机器学习系统的不可预测性要高得多，**而且其运作高度依赖于一项工作中可用数据的质量和数量。开发人员很难得知训练何时完成，甚至无从知晓系统的学习内容。同一个机器学习系统即使面对同样的局势，都有可能作出不同应对，有些系统可能针对某项特殊工作提供意料之外的解决方案。[48]这些核心问题在系统部署用于执行具体任务之后在持续“学习”并改变其模式的情况下会加剧。机器学习系统的不可预测性在解决问题上可能是优势，对棋类游戏等不具备危险性的工作不会造成问题[49]，但对于自主武器系统、网络战及决策支持系统等武装冲突中的应用而言可能会成为重大关切的问题（见第三部分第一点至第三点）。

使情况更为复杂的是，许多机器学习系统**是不透明的；这些系统会提供无法解释的输出。**这种“黑匣子”一样的特性使得使用者很难理解系统在接收到特定输入之后如何以及为何实现其输出，在当前很多情况下甚至无法理解；换言之，机器学习系统缺乏可解释性。

这些不可预测性及缺乏可解释性的问题，**使得人工智能与机器学习系统建立信任面临着重大挑战**。然而，影响信任的另一问题是**偏差**，其中涉及诸多方面，有可能强化现有的人类偏差，也有可能会给系统的设计和（或）使用带来新的偏差。一种普遍性的偏差源于训练数据。也就是说，为某项具体工作可用于训练算法的数据在数量、质量及性质方面的局限可能会为该项工作相关系统的运作引入偏差。就武装冲突中的应用而言，可用于具体任务的优质代表性数据往往十分稀缺，这种偏差就可能会成为重大问题。不过，其他形式的偏差也可能源自系统为数据的不同要素所赋予的权重，或是系统在执行任务期间与所处环境的交互。[50]

对于不可预测性、缺乏透明度或可解释性，以及偏差问题的担忧已体现在人工智能与机器学习的**各类应用中**，如图像识别[51]、面部识别[52]以及自动决策系统。[53]然而，人工智能与机器学习的应用（如计算机视觉）存在另一重大问题——**语义鸿沟**，显示出人类与机器在执行工作中的巨大差异。[54]计算机视觉算法在使用特定对象的图像进行培训之后或许能够在新的图像当中识别上述对象并进行分类。然而，算法并不理解其所分析对象的含义或概念，也就是说算法可能会犯人类绝不会犯的错误，比如在对某物体进行分类时归入完全不同、毫不相关的类别。显然，这在武装冲突的某些应用中会引发严重关切，自主武器系统或是目标选择的决策支持系统都会受到影响（见第三部分第一点及第三点）。

在武装冲突中，尤其是可以假定敌对各方会试图采取对他方系统进行电子欺骗等对策的局势中，对人工智能与机器学习的使用可能就更加难以建立信任。**在对抗状况下**，无论是为欺骗系统而改变环境，还是使用另一个机器学习系统生成对抗图像或制造对抗状况（生成对抗网络，GAN），**机器学习系统极易受到影响**。广为人知的一例是，研究人员欺骗了图像分类算法，使

[46] Amodei, D., et al., *Concrete Problems in AI Safety*, Cornell University, 2016: https://arxiv.org/abs/1606.06565.
[47] ICRC, *Autonomy, Artificial Intelligence (AI) and Robotics: Technical Aspects of Human Control,* report of an expert meeting, 2019 (forthcoming).
[48] Lehman, J., et al., *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities,* Cornell University, 2018: https://arxiv.org/abs/1803.03453.
[49] Silver, D., et al., Mastering the game of Go without human knowledge, *Nature*, Vol. 550, 19 October 2017, pp. 354–359.
[50] UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer,* UNIDIR, 2018.
[51] Hutson, M., “A turtle – or a rifle? Hackers easily fool AIs into seeing the wrong thing”, *Science*, 19 July 2018: http://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.
[52] AI Now Institute, *op. cit.*, pp. 15–17.
[53] *Ibid.*, pp. 18–22.
[54] Smeulders, A. *et al.*, Content-Based Image Retrieval at the End of the Early Years, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, December 2000, pp. 1349–1380.

其将 3D 打印的乌龟认作“步枪”，将 3D 打印的棒球认作“浓咖啡”[55]。如果人工智能图像识别系统用于武器系统或进行目标选择决策，那么此类问题的风险同样显而易见。

六、结论与建议

人工智能与机器学习系统可能会**对人类在武装冲突中的角色**产生**深远影响**，尤其是在以下方面：武器系统及其他无人系统的自主性日益增强；网络战与信息战的新形式；以及更广泛而言，决策的性质。红十字国际委员会认为，政府、军方及其他武装冲突中的相关参与方**在使用人工智能与机器学习系统时**必须采取真正**以人为本的方法**。

一条普遍原则是**当人工智能与机器学习应用于可能对民众生活产生严重影响**（尤其是生命威胁）**的工作与决策时**，且这些工作或决策受国际人道法具体规则规制时，**保留人类控制与人类判断至关重要**。人工智能与机器学习系统，作为工具，**必须用于服务人类，并帮助而非取代人类决策者。**

确保产生上述风险的人工智能实体及数字系统中的人类控制与判断，**对于遵守国际人道法，以及道德层面上在武装冲突中维护一定程度的人道标准而言，是十分必要的。**为使人类有意义地发挥自身作用，可能需要使这些系统的设计与使用**按照人类速度帮助决策，而非按照“机器速度”加速决策，**超越人类干预。此类考量或将最终促成制定人工智能与机器学习系统设计与使用方面的限制，从而基于法律义务及道德责任实现有意义的有效人类控制与判断。

人类控制与判断总的原则是一个必要的组成部分，但其本身并不足以抵御武装冲突中人工智能与机器学习的潜在风险。**其他需考量方面**包括要确保：系统运行及相应后果的**可预测性**与**可靠性**——或者说是安全性；系统运行方式及系统产生特定输出之原因的**透明度**——或者说是**可解释性**；以及在系统设计与使用中**减少误差**——或者说是保证公平。为对给定系统的使用**建立信任**，需要解决这些问题，可以在投入使用之前**在现实环境中进行严格的测试**。[56]

所需人智交互的本质可能要取决于道德考量，国际人道法的特定规则，以及其他在相关局势中适用的法律。因此，**一般性原则可能需要由人工智能与机器学习在特定应用领域或特殊情况下的具体使用原则、指南或规则加以补充。**

红十字国际委员会认为，最为紧要的关切是在造成人员伤亡及破坏与毁灭的决策上人类与机器之间的关系，以及**确保**武装冲突中**人类对武器系统及武力使用之控制的重大意义**。随着武器系统自主性日益增强，无论这些系统是否使用人工智能技术，事实上由传感器和算法进行最终决策的风险依然存在。这种情况会引发法律与道德关切，迫切需要加以解决。

红十字国际委员会强调需要确定遵守国际人道法与解决道德关切所必需的**人类控制关键要素**，以作为武器系统自主性限制国际标准的基础，其中应包含人类监督的程度（包括干预及使系统失效的能力）、可预测性及可靠性水平，以及行动限制。[57]

这一针对自主武器系统**基于人类控制的方法也会适用于人工智能与机器学习在武装冲突**，尤其是存在重大人身安全风险，且国际人道法具体规则适用的局势**中决策层面上更为广泛的应用**，如在目标选择与拘留工作中决策支持系统的使用。

---

[55] Hutson, M., *op. cit*.
[56] Goussac, N., “Safety net or tangled web: Legal reviews of AI in weapons and war-fighting”, *ICRC Humanitarian Law & Policy Blog*, 18 April 2019: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting.
[57] ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, *op. cit.*