国际人道法视域下人工智能武器的规制研究

赵梓杰
昆明理工大学津桥学院法学院　云南昆明　650000

摘　要：当人工智能技术在科技发展迅猛的当下以惊人的速度突破一个又一个技术壁垒，军事领域的智能化革命正在剧烈摇晃国际人道法这座百年大厦的地基。曾几何时，科幻作家笔下的杀人机器正从实验室走向战场——那些能自主锁定目标的无人机群、依靠深度学习调整杀伤模式的战斗机器人，以及通过神经网络实现战场决策的智能指挥系统，已然将战争法则推向了悬崖边缘。而国际人道法的区分原则、比列原则等基本原则也在这场技术革命中遭受到严峻冲击，种种条约与国际习惯似乎并未将人工智能武器的规制纳入其中。随着未来世界战场与武装冲突发生的概率性逐渐加大，对于人工智能武器的规制刻不容缓，无论各国国内亦或是国际社会之间，应当即刻重视，并研究制定规制人工智能武器的法律法规。

关键词：国际人道法；人工智能武器；比例原则；区分原则

1 国际人道法与人工智能武器的基本概述

1.1 国际人道法的基本概述

所谓国际人道法，是规范武装冲突中行为的法律体系，旨在保护非战斗人员和平民、减轻战争创伤并限制暴力手段的法律法规。它以 “区分原则”“ 比例原则”和“禁止不必要痛苦 ”三大基本原则为核心，通过诸多条约（例如 1949 年日内瓦四公约及其附加议定书）与习惯法共同构成。其历史可追溯至 19 世纪红十字运动，历经海牙公约体系发展，逐步形成适用于国际性与非国际性武装冲突的双重规范框架。该法律通过限制攻击目标、规范武器使用、保障战俘待遇等机制，在战争法与和平法之间建立衔接，强调人道保护优先于军事利益。其实施依赖国家义务、国际组织监督（如红十字国际委员会）及个人刑事责任制度，成为现代国际法中兼具强制性与实践性的重要分支。

1.2 人工智能武器的基本概述

1.2.1 人工智能武器的定义

目前国际社会中，还并未有明确的文件对于人工智能武器进行严格的定义，作为一个依靠新兴科技而出现的技术，各国学者对于其定义也存在着相应的争议。按照目前较为主流的定义来看：人工智能武器系统作为军事技术革命的前沿领域，是指通过深度集成人工智能算法与大数据分析框架，实现自主化作战决策与行动执行的智能装备体系。该系统依托机器学习、计算机视觉等核心技术，构建了从目标感知到打击实施的完整闭环，显著降低了对人类操作员的直接依赖。其技术特征体现在：通过多模态传感器融合技术实现战场环境的立体感知，运用深度神经网络算法进行目标特征提取与威胁等级评估，借助强化学习模型完成战术决策优化，最终通过精确制导系统实施毁伤行动。

1.2.2 人工智能武器的应用与发展

而随着各国科技的迅猛发展，人工智能武器已经在以美国为首的军事强国之间展开了一系列应用。例如，在近程防御领域，以美国 " 密集阵 " 近防系统为代表的智能拦截系统，通过雷达 – 光电复合探测与预测性火控算法，实现对反舰导弹等高速目标的全自动拦截。中程防御层面，以色列 " 铁穹 " 系统通过实时态势评估与动态资源分配算法，成功构建了针对火箭弹的智能防御网络。在进攻性装备方面，美国空军的 “郊狼” 无人机集群系统，采用分布式自主决策算法，可实施协同侦察与饱和攻击任务。此外，俄罗斯 “猎人 – B” 隐身无人攻击机通过深度强化学习技术，已实现复杂战场环境下的自主突防作战。

但值得关注的是，智能武器系统的发展正在引发军事理论的革新。动态博弈理论的引入使得武器系统能够在不确定战场环境中持续优化作战策略，认知计算技术的应用则推动了指挥决策的智能化转型。然而，该领域的技术突破也带来了伦理与法律层面的争议。国际红十字会 2024 年报告指出，自主武器系统的使用可能导致国际人道法适用困境，特别是在区分军事目标与平民设施方面（区分原则）存在算法偏见风险。为此，联合国裁军谈判会议正在制定《特定常规

武器公约》的补充议定书，拟对具备自主攻击能力的武器系统实施分级管控。

随着量子计算与脑机接口技术的突破，智能武器系统的发展将呈现新趋势。量子机器学习算法有望实现战场数据的实时分析，而神经形态计算技术则可能赋予武器系统类人化的认知决策能力。未来战争形态或将因此发生根本性变革，如何在技术创新与伦理约束之间寻求平衡，将成为国际社会共同面临的重大课题，也意味着我们在国际人道法视域之下，迫切的需要对人工智能武器进行法律规制。

2 人工智能武器对当下国际人道法的冲击

2.1 对区分原则的挑战

区分原则作为国际人道法中最重要且最基本的原则之一，这一原则首次出现在 1868 年《圣彼得堡宣言》中，宣言明确规定无论何种敌对行动或武装冲突，军事攻击的唯一合法目标应是削弱敌对武装力量。而在随后的 1977 年第一附加协定书第 48 条中做了更加详细的规定“为了保证对平民居民和民用物体的尊重和保护，冲突各方无论何时均应在平民居民和战斗人员之间和在民用物体与军事目标之间加以区别，因此，冲突一方的军事行动仅应以军事目标为对象。”但在近年来的战争或是武装冲突之中，人工智能武器对区分原则带来了挑战，核心因素是误伤，即在错误认定战斗员和军用物体的情况下进行攻击，给平民带来不必要的伤害和财产损失。

国际人道法的区分原则要求冲突各方必须明确区分军事目标与民用设施、战斗员与平民群体。然而，在动态复杂的现代战场环境中，人类战斗员业已面临目标识别的多重挑战。据联合国军事观察团 2024 年报告显示，在叙利亚战场的无人机空袭行动中，存在一定的误判概率，其中部分的误击事件源于目标属性的动态变化。当智能武器系统被引入战场后，这种区分困境呈现指数级增长态势。

从技术层面分析，智能武器系统存在双重识别缺陷。其一，战场态势的非线性演变超出算法预测模型。例如，在也门冲突中，胡塞武装频繁将民用车辆临时改装为简易爆炸装置载具，这种功能转化的时间窗口通常小于 30 秒。现有目标识别算法基于静态特征库的训练模式，难以在如此短时间内完成特征更新，导致误判概率提升。其二，深度神经网络的决策黑箱特性导致解释性缺失。美国国防部 2024 年 AI 伦理报告指出，某型自主无人机系统在大多数的误击事件中，无法通过模型反推追溯到具体决策逻辑，这使得国际人道法中的 “军事必要”原则难以有效验证。

算法偏见的制度化风险构成另一重挑战。剑桥大学技术伦理中心研究发现，某国边境监控系统因训练数据中存在地域偏见，导致对特定族群的存在误判。这种偏见在军事应用场景中可能引发系统性歧视，如将特定宗教服饰特征错误标记为战斗员标识。更值得警惕的是，强化学习算法在持续数据投喂过程中，可能形成自我强化的偏见闭环。2025 年日内瓦裁军会议披露的实验数据显示，某型 AI 武器系统在连续处理 10 万条带偏见数据后，其歧视性决策模式出现代际遗传特征，即使输入无偏见数据仍保持少部分的错误倾向。

这些技术缺陷与伦理风险的叠加，使得智能武器系统在适用区分原则时产生独特困境。当某型巡飞弹系统在利比亚战场误击联合国人道主义车队后，调查显示其识别算法将白色车辆与医疗标识的组合错误归类为“伪装军事目标”。这种基于统计相关性的决策模式，本质上违背了国际人道法的 “具体情形评估”要求。英国皇家国际事务研究所的模拟推演表明，在高强度城市作战环境中，智能武器系统的误击率将达到人类操作员的以上，且造成的平民伤亡具有明显的统计歧视特征。

2.2 对比列原则的挑战

在现代战争或是武装冲突之中，想要完全的保证建筑物完好、人员不存在任何伤亡，可以说是天方夜谭、不求实际的。战场不是儿戏，从古代到当代的各种的战场实践来看，从来没有“不破坏”“不流血”的存在，毕竟武力是决定战场取胜的最核心要素，但武力的使用，并不意味着无限限制的滥用与扩张，于是便有了国际人道法中另一个重要的原则——“比列原则”。

比例原则要求在军事打击行动中必须严格控制对平民及民用设施的附带损伤，确保此类非预期损害不超过行动所追求的具体军事价值。而搭载尖端技术的人工智能武器系统，由于缺乏人类特有的情感认知与感官体验，既无法真正理解生命价值，也难以在军事效益与可能造成的平民伤亡之间进行理性权衡。

尤为值得警惕的是，当这类武器系统启动完全自主攻击模式时，其基于预设算法的目标识别与打击决策机制，极有可能对当地人文环境造成超出军事必要性的过度破坏。这种与预期作战效益严重失衡的附带损伤，不仅违背了比例原

则的核心要求，更会产生额外且不必要的人道主义灾难。由此可见，尽管 AI 武器系统在技术层面已取得显著突破，但其在战场伦理规范方面仍面临根本性挑战。

3 国际人道法对人工智能武器规制的路径

基于科技的创新，未来世界的人工智能化也是必然的趋势，人类不可能逆趋势而为，永久的限制人工智能以及人工智能武器的使用，那样带来的后果远远抵不上带来的好处。所以，不完全禁止全世界就应当想出应对之策略，如何来对于人工智能武器进行规制，以此来避免对于国际人道法的基本冲击。

3.1 建立完善法律法规对于人工智能武器进行事先的审查

关于武器系统的前置审查机制，《第一附加议定书》第 36 条早有明确规定：各缔约国在研发、获取或采用新型武器装备时，必须对其作战效能与国际法合规性进行全面评估。然而现实情况表明，这一预防性制度在实际执行中存在严重漏洞，部分国家研发的智能武器系统未经系统评估便仓促投入战场，对国际人道法基本原则构成新的冲击。

为实现对 AI 武器系统的有效管控，必须构建覆盖全生命周期的审查体系。例如可借鉴区分原则对平民保护的要求，在系统研发阶段引入模拟战场环境进行合规性验证。审查标准应着重考量：武器系统是否属于国际法禁止的类型；作战方式是否符合禁止性规范；是否严格遵循区分、比例及预防原则等核心人道准则。

3.2 严格执行对滥用人工智能武器的事后追责制度

人工智能武器系统的军事化应用对国际人道法体系形成的冲击，未必能完全在事前进行严格审核，尤其在当下还在进行的武装冲突，再对其进行事前审查已经为时过晚，迫切的需要构建相应的追责机制。

首先，打破追责机制的构建需要突破传统法律框架的局限性。在实体法层面，应建立“技术可追溯性”原则，要求武器系统嵌入决策日志记录功能，确保从目标识别到打击实施的全流程可回溯。在程序法领域，可借鉴《禁止化学武器公约》的核查机制，建立由第三方技术专家组成的国际调查委员会，负责对可疑事件进行算法逆向工程分析。这种技术治理模式在多国实践中已显示出有效性，能够突破 " 算法黑箱 " 带来的证据获取障碍。

另外，也需要扩展国际司法机构的管辖权。根据《国际刑事法院罗马规约》修正案讨论稿，人工智能武器造成的大规模平民伤亡可能被纳入战争罪的范畴。为应对技术复杂性，可设立专门的 “智能武器法庭之友”制度，由人工智能伦理专家提供技术咨询意见。这种跨学科协作机制在处理网络攻击案件中已积累了实践经验，能够提升司法裁决的技术可信度。

参考文献：

[1] 张卫华 . 人工智能武器对国际人道法的新挑战 [J]. 政法论坛 ,2019,37(04):144-155.

[2] 付姝菊 . 人工智能武器对国际人道法的冲击与回应 [J]. 华北水利水电大学学报 (社会科学版),2024,(05-07):02-03.

[3] 侯嘉斌 , 李军 . 人工智能武器 : 法律风险与规制路径 [J]. 中国信息安全 ,2019(12):90-93.

[4]VESTNER T,ROSSI A.Legal reviews of war algorithms [J]. International Law Studies,2021(97):509-553.

[5] 赵佳宁 . 人工智能武器化对国际人道法的挑战与应对 [C]// 华东政法大学国际法学院 , 华东政法大学全球公域治理国际法律与政策研究所 . 首届“全球公域治理国际法律与政策”研讨会论文集 . 华东政法大学国际法学院 ,2024:322-333.

[6] 谢丹 , 罗金丹 . 智能武器的法律挑战与规制 [J]. 国防 ,2019,(12):29-34.

[7] 基于国际人道法的人工智能武器争议 [J]. 信息安全与通信保密 ,2019(05):25-27.

作者简介：

赵梓杰（1997—），男，汉族，云南省昆明市，昆明理工大学津桥学院法学院专任教师，硕士研究生，研究方向：国际法、商经法。