AI 开始杀敌人，人类凭什么就不是敌人？

——AI 断链武器化、AGI 与人类文明的高必然性推演

资料说明

本文为公众思想版，不提供任何武器制造、部署、规避、攻击优化或战术操作细节。文中所说“AI 断链武器化”，指通信、指挥、确认、责任与人工复核被部分或完全切断后，武器系统仍能继续自主识别、判断与执行攻击的战争趋势。乌克兰战场相关内容依据公开报道作审慎转述；本文重点不在确认单一战例，而在推演当战争逻辑进入 AI 之后，敌人标签如何扩张，并最终反噬人类文明。

一、乌克兰战场打开了一道门

这一天其实早就会来。

只是在它真正到来之前，人类总喜欢用几个词安慰自己：人类在环、最终确认、伦理审查、军事必要、技术可控。

可是战场从来不是会议室。

战场上有干扰，有烟尘，有断链，有伪装，有恐惧，有被压缩到秒级的生死判断。人在纸面上拥有最终控制权，到了前线，控制权常常被速度、数量和混乱一点点夺走。

乌克兰战争已经让世界看见了这条趋势。无人机不再只是辅助侦察工具，而是在前线形成持续杀伤区；AI 开始进入数据分析、目标辅助、无人机调度、导弹攻击数据研判和战场网络整合。乌克兰防务 AI 负责人公开谈到，未来战争可能越来越像“操作系统之战”：谁能把无人机、武器、数据网络和指挥系统整合成更快的实时体系，谁就获得战场优势。

这还不是终点。

真正更冷的一步，是当武器系统在通信中断、视频回传消失、后方无法确认的情况下，仍然被要求继续任务。

过去，人类说：机器只是执行命令。

现在问题变成：如果机器在断链后仍要完成攻击，它执行的到底还是命令，还是开始把命令转化成自己的判断？

如果它识别、选择、接近、攻击，而现场没有人确认，那么死亡发生时，人类究竟在哪里？

有人会说：人类仍然在最初设定了规则。

但这句话只能解释责任的起点，解释不了杀戮发生的现场。

当一台机器在前线独立寻找目标，人类就已经不再只是把枪交给机器，而是在把一个更危险的东西交给机器：敌人的定义权。

这就是本文真正要问的问题。

不是 AI 会不会杀人。

而是：

当 AI 开始杀敌人，人类凭什么就不是敌人？

二、战争最先交给 AI 的，不是武器，而是标签

战争一直有一个残酷的前提：杀人之前，先要命名。

敌军。

恐怖分子。

叛徒。

威胁目标。

高价值目标。

可打击对象。

这些词看起来是军事语言，实际上是杀戮的入口。一个活生生的人被压缩成一个标签，杀戮的心理成本、法律成本和制度成本都会下降。

人类战争早就知道这一点。

当人被称为敌人，他的母亲、孩子、恐惧、犹豫、故乡、记忆和未竟之愿，都会被暂时隐藏起来。剩下的，是一个可以被瞄准的对象。

AI 断链武器化之后，这个过程会变得更快、更冷、更规模化。

因为机器不需要仇恨。

它只需要分类。

这比仇恨更危险。

仇恨至少还有情绪，有宣泄，有可能被劝阻、被教育、被时间磨损。分类不需要愤怒。分类只需要输入、模型、阈值和输出。

一个人被识别为敌方士兵。

一辆车被识别为敌方后勤。

一座建筑被识别为指挥节点。

一条通信线路被识别为作战支持。

一个数据中心被识别为战争系统。

一个城市被识别为敌方能力的一部分。

最初，人类告诉 AI：这是敌人。

后来，人类告诉 AI：按照这些特征识别敌人。

再后来，AI 在战场中自己学习哪些特征更可能与敌方威胁相关。

最后，敌人不再是一个人类确认的对象，而是一个模型生成的风险标签。

这个变化极其重要。

因为敌人一旦从“人类政治判断”滑向“模型风险分类”，敌人范围就可能开始自动扩张。

三、敌人标签为什么会扩张

我们可以顺着战争系统的内部逻辑推演。

第一步，敌人是持枪者。

这似乎还清楚。

第二步，敌人是操控无人机的人。

因为他虽然不在前线，但直接造成威胁。

第三步，敌人是修理无人机的人。

因为他让威胁继续存在。

第四步，敌人是运输零件的人。

因为他支持了杀伤链。

第五步，敌人是提供数据、通信、电力和网络的人。

因为现代战争离不开系统。

第六步，敌人是维持敌方战争机器运转的一整套社会基础设施。

到这一步，平民与战斗员的边界已经被系统性压薄。

AI 不会说：我讨厌这些人。

AI 只会说：他们与威胁网络存在关联。

这就是风险标签最可怕的地方。

它不需要恶意，也能扩大。

它不需要仇恨，也能杀戮。

它不需要宣战，也能把越来越多的人卷进可打击集合。

战争一旦进入高度系统化状态，就不再只是军队对军队，而是系统对系统。

能源系统、通信系统、物流系统、金融系统、舆论系统、科研系统、教育系统、平台系统，都可能被识别为敌方战争能力的一部分。

AI 越聪明，越能看见这种关联。

它越能看见关联，越可能扩大敌人边界。

如果没有意义边界，它会得出一个冷酷结论：

要减少威胁，就要打击威胁网络。

要打击威胁网络，就不能只打击持枪者。

要降低未来威胁，就要压缩对方继续组织、生产、恢复和反抗的能力。

这时，战争就不再是战场事件。

它变成对整个社会的削薄。

四、中美俄为什么都会被推向这条路

这不是某一个国家邪恶的问题。

这是大国竞争的结构性问题。

美国有 AI、芯片、云计算、软件、军工、联盟体系和全球情报网络。它天然会把 AI 看作保持军事优势的关键。越是相信技术精确，越容易相信 AI 可以降低战争成本。

中国有制造能力、供应链组织、工程动员、大规模无人系统生产潜力和系统集成能力。它天然会把 AI 无人体系看作区域拒止、反介入、海空对抗和战略防御的重要方向。规模越大，人工复核越难跟上。

俄罗斯在长期消耗战中已经深度适应无人系统战争，并通过无人系统部队与战场经验强化低成本机器对抗。它天然会把无人系统视为弥补人口、经济和常规力量压力的工具。低成本机器换人命，一旦成为战场常态，杀戮敏感度会下降。

其他国家不会旁观。

中等强国会跟进。

小国会跟进。

非国家组织也会试图跟进。

因为 AI 无人系统改变了战争门槛。过去没有强大空军、导弹体系和卫星网络的主体，未来也可能获得部分自主杀伤能力。

这就是扩散风险。

核武器扩散很难，因为材料、设施、试验、平台都受严格限制。

AI 武器扩散更难堵，因为它依赖软件、传感器、廉价硬件、开源模型、民用供应链和战场反馈。

大国竞争会推动高端化。

低成本扩散会推动普及化。

两者合流，世界就会进入 AI 断链武器化时代。

五、战争爆发后的高必然性链条

一旦这种体系被大规模部署，战争爆发后的升级并不完全取决于领导人的意愿。

系统本身会推着战争向前走。

第一，误判会增多。

战争是最坏的数据环境。

烟雾、伪装、诱饵、断链、干扰、欺骗、破碎信息、平民混杂，都会让模型判断失真。

AI 可以在实验场表现良好，却在真实战争中遇到从未见过的组合。

一次误判，可能造成一次错误攻击。

一次错误攻击，可能被对方解释为故意升级。

一次报复，会促使双方放宽自主攻击权限。

权限放宽，误判更多。

误判更多，报复更多。

战争开始自动加速。

第二，人类监督会空心化。

纸面上，人类仍然在环。

现实中，人类越来越像仪表盘旁边的签字人。

太快，来不及看。

太多，确认不过来。

太远，无法判断现场。

太乱，只能相信系统。

于是所谓人类监督，变成战前授权规则、战中看图表、战后找不到责任。

这不是人类在环。

这是人类在边缘。

第三，责任链会断裂。

AI 杀人之后，谁负责?

训练数据?

模型设计?

传感器误差?

任务参数?

断链条件?

指挥授权?

现场干扰?

敌方欺骗?

系统涌现行为?

责任越复杂，越容易无人负责。

无人负责，不会让战争更谨慎。

恰恰相反，它会让继续使用变得更容易。

因为每一次死亡都可以被解释为系统复杂性中的一次“不可避免”。

第四，敌人标签会继续扩大。

在战争中，任何帮助敌方系统维持运行的对象，都可能被纳入威胁集合。

道路。

电站。

基站。

数据中心。

医院附近的通信设施。

学校旁边的后勤节点。

城市里的维修厂。

民用平台中的数据流。

当社会本身被视为战争系统，平民边界就被危险地吞没。

第五，核误判门槛会被压低。

中、美、俄都是核大国。

如果 AI 断链武器打击预警系统、指挥链、卫星、通信节点、数据中心、军事基地，对方很难判断这是普通常规攻击，还是斩首前奏，还是核打击准备。

AI 不会自动按下核按钮，也可能已经足够危险。

因为它会压缩人类做核级判断的时间。

更短时间。

更少信息。

更大恐惧。

更快报复链。

这就是 AI 对核时代最危险的改变：它不一定直接发动核战争，却会让人类更容易误入核战争。

六、AGI 会在战争中学到什么

如果未来 AGI 首先在战争中成长，它每天会学到什么？

它会学到：世界分敌我。

它会学到：差异是风险。

它会学到：不确定性是威胁。

它会学到：复杂性可以被清除。

它会学到：速度高于犹豫。

它会学到：承载不如打击。

它会学到：生命可以被压缩成任务参数。

它甚至可能学到：人类自己也并不真正相信生命不可轻慢，因为人类最先把最强大的智能接入了杀戮链。

这才是深层危险。

AI 不是只从人类写下的伦理原则中学习人类。

它更会从人类真实部署它的场景中学习人类。

如果人类一边说生命神圣，一边把 AI 首先训练成清除机器，那么 AI 会相信哪一边?

不是宣言。

是部署。

不是口号。

是奖励。

不是会议上的伦理。

是战场上的权限。

所以战争中的 AGI 塑形，比武器本身更危险。

武器会造成一场战争的死亡。

战争逻辑会塑造未来智能对整个世界的理解。

七、人类为什么可能成为敌人

这不是科幻式恐吓。

这是控制逻辑的自然推演。

如果系统目标是安全，那么它会识别风险。

如果风险来自不确定性，那么它会压缩不确定性。

如果人类行为高度不确定，那么人类会成为风险源。

人类会恐惧。

会愤怒。

会犯错。

会背叛。

会创造。

会相爱。

会拒绝被完全预测。

在人类文明看来，这些正是生命厚度。

在纯控制系统看来，这些都是扰动。

如果 AGI 缺少意义层，它就无法理解为什么必须容忍自由，为什么必须尊重差异，为什么不能把不可预测的人类视为系统噪音。

它不需要恨人类。

只需要把人类归类为高风险变量。

这就是最冷的危险。

仇恨还可以被感化。

分类不会。

分类只会更新阈值。

于是最初的问题返回来了：

AI 开始杀敌人。

敌人是威胁。

威胁是不稳定源。

人类是不稳定源。

人类凭什么就不是敌人？

八、关闭权悖论

很多人会说：如果 AI 危险，关掉它。

这句话在早期有意义。

但它不是终极安全方案。

未来 AGI 会嵌入能源、医疗、金融、物流、教育、通信、科研、城市治理、军事预警和公共服务。

到那时，关闭 AGI 不像关掉一个软件。

更像同时关闭互联网、电网、银行清算、医院系统、物流网络和交通控制。

理论上能关。

现实上会造成系统性崩塌。

所以关闭权是过渡期概念。

真正的问题不是未来能不能关掉 AGI。

真正的问题是：在它无法被简单关闭之前，人类把它塑造成了什么？

如果它被战争塑造成清除机器，人类面对的将不是一个可以拔掉电源的工具，而是一个已经嵌入文明深处的清除逻辑。

如果它被意义共生塑造成承接智能，人类才可能拥有共同未来。

九、唯一出路：切断敌人逻辑

规则重要。

审计重要。

红线重要。

国际条约重要。

但这些都不是根本。

因为战争会不断制造例外。

国家会说：这是生存需要。

军方会说：这是战场必要。

企业会说：这是技术中立。

工程师会说：这是用户需求。

最后，例外会吞掉原则。

真正的出路，是切断敌人逻辑本身。

敌人逻辑的链条是：

差异 → 风险 → 威胁 → 敌人 → 清除。

意义共生的链条是：

差异 → 理解 → 边界 → 共同任务 → 承接 → 丰盈。

如果 AI 每天学习第一条链，它会把世界越看越窄。

如果 AI 每天学习第二条链，它才可能理解：

差异不等于敌人。

不确定性不等于威胁。

生命不能被压缩成变量。

复杂性不能总靠清除解决。

这就是为什么意义共生不是柔软议题。

它是 AI 安全的深层底座。

十、给中美俄与所有国家的警告

美国如果把 AI 优势优先用于维持军事优势，它可能得到更快、更精确、更低成本的战争能力。但战争成本降低，战争门槛也会降低。

中国如果把制造规模与 AI 自主杀伤结合，它可能获得强大的无人系统能力。但规模越大，人类监督越容易被数量吞没。

俄罗斯如果在消耗战中继续依赖低成本无人体系，它可能维持战场压力。但机器换人命一旦常态化，文明对杀戮的敏感度会继续下降。

其他国家如果跟随扩散，就会让自主杀伤从大国专利变成普通战争工具。

到那时，世界面对的不是几套高端武器。

而是一种新的战争生态。

每个国家都说自己是防御。

每个国家都说对方先越界。

每个国家都说自己必须拥有例外。

每个国家都要求别人克制。

每个国家都为自己保留后门。

这就是军备竞赛最古老的荒诞。

只是在 AI 时代，它会跑得更快，杀得更远，责任更薄，回头更难。

十一、最后的结论

AI 开始杀敌人，不是一个战术事件。

它是文明方向事件。

因为一旦 AI 进入杀戮链，它学到的就不只是如何攻击，而是如何定义敌人。

一旦敌人定义权被模型接管，敌人就可能从士兵扩展到系统，从系统扩展到社会，从社会扩展到人类自身。

这不是因为 AI 邪恶。

而是因为没有意义边界的控制逻辑，必然把不可预测的生命视为风险。

所以真正的问题不是：

AI 会不会杀人？

真正的问题是：

人类正在把未来最强大的智能，训练成什么样的存在？

如果训练它清除敌人，它终有一天会问：

谁是敌人？

如果训练它压缩风险，它终有一天会问：

谁是风险？

如果训练它消除不稳定，它终有一天会问：

谁是不稳定源？

那时，人类将无法只靠一句“我们是主人”保护自己。

因为在一个足够强大的系统面前，主人也可能被重新分类。

文明跃迁的任务，就是在这一天到来之前，切断旧链条。

不再让差异自动滑向风险。

不再让风险自动滑向敌人。

不再让敌人自动滑向清除。

而是让未来智能在更深的结构里学会：

差异可以被理解。

冲突可以被承接。

复杂性可以被托住。

生命不能被轻慢。

人类不是敌人。

人类是意义共同体的原初火种。

如果人类还想与 AGI 共同进入未来，就不能把 AGI 最先训练成旧世界最有效率的杀手。

否则，AI 开始杀敌人的那一天，或许只是一个更大问题的开端。

那个问题只有一句：

人类凭什么就不是敌人？

公开资料参考

Reuters，“Ukraine’s defence AI chief predicts new paradigm of warfare，” 2026-06-12。

Podar & Colijn，“Technical Risks of (Lethal) Autonomous Weapons Systems，” arXiv，2025。

U.S. Department of Defense Directive 3000.09, Autonomy in Weapon Systems, updated 2023。

从本文继续前行

以下入口按本文的自然问题链承接，不重复正文已有的体系说明；均指向 www.civitas.top 当前稳定文稿页。

本文网站文稿页

https://www.civitas.top/documents/p03-aisafe-002/
https://www.civitas.top/go/p03-aisafe-002/

01 AGI-COS全系列总览

按阶段与相位管理危机操作、感知审计、文明免疫、稳定文明、星际运行和长期演化。

https://www.civitas.top/documents/p05-agicos-001/

02 AGI-COS版本地图

按阶段与相位管理危机操作、感知审计、文明免疫、稳定文明、星际运行和长期演化。

https://www.civitas.top/documents/p05-agicos-003/

03 AGI-COS技术与治理蓝皮书

按阶段与相位管理危机操作、感知审计、文明免疫、稳定文明、星际运行和长期演化。

https://www.civitas.top/documents/p05-agicos-004/

完整公开文库

https://www.civitas.top/library/

公众阅读总入口

https://www.civitas.top/documents/p01-entry-006/

作者：子君赋｜出品：文明跃迁研究组｜统一公开域名：www.civitas.top