

AGI 时代的国际关系与战略安全重新对账

——智能竞速、威慑、危机稳定与共同存续边界

Reckoning with International Relations and Strategic Security in the Age of AGI

Intelligence Competition, Deterrence, Crisis Stability, and the Boundary of Shared Survival

国际关系／战略安全独立外发专业深稿 V1.2.1

作者：子君赋 | 出品：文明跃迁研究组

证据截点：2026 年 8 月 26 日

摘要

高级人工智能进入国际安全，真正改变的并不只是武器性能，而是国家如何看见威胁、形成判断、传递信号、配置预授权、组织联盟并决定行动。国际关系与战略安全因此不能被“AI 是否安全”这一模型级问题替代。一个系统在技术上更可靠，并不自动意味着相互竞争的国家会更稳定；一方缩短决策链以获得优势，在特定条件下可能促使对方加速、增加预授权或缩短核验、沟通和降级窗口，也可能引发退让、独立核验、监督、拒用或共同护栏。反过来，更好的感知、模拟、预警、防御和危机沟通也可能降低误判，扩大纠错时间。AI 的战略后果不是单向技术宿命，而取决于能力特征、组织采用、相互预期、联盟结构、战略姿态、法律制度与对手反应的共同作用。

本文从国际关系与战略研究已经形成的安全困境、威慑、危机稳定、联盟、信号、误判、军备竞赛与组织过程传统出发，区分四类论证身份：A 类为较稳定的历史、制度与现实部署事实；B 类为 AI 究竟增强还是削弱威慑稳定、提高透明度还是增加不确定性、增强人还是推动自动化升级等尚未收敛的竞争解释；C 类为“AI 能力→战略部署→组织流程→决策时间→对手预期→行为调整→危机稳定→共同存续风险”等跨域结构判断；F 类为作者公开承担的规范承诺。本文不把 A 类事实直接写成战争必然，也不把共同存续写成世界主权的自动理由。

文章的核心判断是：AGI 时代国际安全最需要防止的，不是某个国家拥有更强智能本身，而是各方在信息不完全、相互猜疑和竞争压力下，把“更快、更自动、更不可落后”逐步变成共同默认，并在特定条件下使本来用于增加安全的能力共同压缩纠错空间；这种结果必须被经验检验，不能由单方性能或采用事实自动推出。战略稳定因此必须重新从单方性能转向互动条件；所谓“人在环中”也必须进一步检验人是否真正拥有足够信息、时间、权限与组织支持去改变结果。共同存续在本文中只作为任何战略竞争不能失守的 F 类底线：它不取消国家正当安全利益，不要求单方面解除防御，不预设世界政府，也不替国际法、伦理学、政治学或计算机科学作专业裁决。

关键词：AGI 时代；国际关系；战略安全；人工智能军事应用；威慑；危机稳定；安全困境；战略自主系统；决策时间；共同存续

Abstract

The strategic significance of advanced artificial intelligence does not lie only in improved weapon performance. AI is entering the processes by which states perceive threats, fuse intelligence, generate options, allocate authority, signal resolve, coordinate alliances, and act under time pressure. A technically reliable system does not therefore imply a strategically stable interaction. Under some conditions, one actor's attempt to shorten its decision cycle may induce rivals to accelerate, expand preauthorization, or shorten verification, communication, and de-escalation windows; under others, it may prompt restraint, independent verification, oversight, refusal, or reciprocal guardrails. Better sensing, simulation, warning, defensive coordination, and crisis communication may also reduce misperception. The effect of AI on strategic stability is therefore conditional rather than technologically predetermined.

This article builds from the established literatures on the security dilemma, deterrence, crisis bargaining, alliances, signaling, misperception, arms competition, command and control, and organizational failure. It distinguishes established facts and formal institutional materials (A), competing and unsettled explanations (B), cross-domain structural hypotheses requiring empirical validation (C), and the author's explicit normative commitments (F). Its central claim is that the key unit of analysis is strategic interaction, not the AI system alone. Shared survival is treated only as a normative floor: it does not erase legitimate state security interests, authorize unilateral disarmament, create world sovereignty, or substitute for legal and ethical judgment.

Keywords: AGI era; international relations; strategic security; military AI; deterrence; crisis stability; security dilemma; autonomous systems; decision time; shared survival

研究定位、对象边界与证据纪律

本文是一篇国际关系／战略安全专业深稿。它研究国家、联盟和军事组织在相互竞争、信息不完全与高后果条件下，如何因 AI 能力进入情报、预警、决策支持、网络行动、自主系统、指挥控制与基础设施而改变彼此行为。研究对象首先是互动关系，而不是一个孤立模型。模型是否可靠、系统如何保证，由人工智能与计算机科学重新判断；武力是否合法、国际人道法如何适用，由法学与国际法重新判断；公共权力为何正当、主权如何授权，由政治学重新判断；何种伤害可被正当化，由伦理学重新判断。战略研究只在自己的证据范围内回答：这些能力一旦进入相互竞争，会怎样改变威慑、危机稳定、升级控制和共同安全。

全文继续采用 A／B／C／F 四种论证身份。A 类只包括较稳定的历史事实、成熟研究、正式政策与可核验部署状态；B 类包括理论竞争、机制争论与外部效度尚未收敛的经验结果；C 类包括跨域因果连接、未来情景和制度压力测试；F 类是作者公开承担的规范承诺。A/B/C/F 不是证据等级，而是提醒读者一句话正在承担什么论证责任。A 不能自动生成 B、C 或 F；F 也不能伪装成战略科学事实。

身份	本文中的典型内容	能够支持什么	不能自动推出什么
A 较稳定事实/正式材料	经典安全研究、历史危机材料、正式军事 AI 政策、已核验应用范围	支持限定对象、时点和条件下的事实判断	不能自动推出 AI 必然导致战争、某制度必然正当
B 竞争解释	AI 增强稳定/破坏稳定；透明/新不确定性；人机增强/自动化升级	提出可比较机制与经验检验	不能写成战略共同体已收敛的普遍结论
C 跨域结构判断	速度竞争→预授权→核验窗口可能压缩；共享模型→联盟相关性风险	形成可反驳的战略假说和压力测试	不能伪装成已发生的历史事实
F 规范承诺	共同存续不能失守；高不可逆决定须保持责任与纠错条件	公开说明本文为什么关心某些风险与方向	不能替代国际法效力、政治授权或伦理裁决

表 1 A／B／C／F 论证身份与战略研究边界

文明跃迁理论在本文中的最低接口

陌生读者无需预先接受文明跃迁理论。本文只使用一个最低接口：现实按因果发生，不按学科发生；学科分别求真，文明共同承担后果。国际关系与战略安全能够说明国家为何竞争、威慑如何成立、危机为何升级或被控制，却不能因为问题最终涉及文明存续，就取得法律、伦理、技术或政治权力的终审权。文明尺度也不能反向替战略研究预设答案。

本文的 F 类底线是：共同存续是任何战略竞争都不能失守的文明底线。但这句话只为战略选择增加一个不可忽略的后果维度，不把国家安全利益宣布为不正当，也不从“人类整体”直接推出统一世界权力。国家仍必须处理真实威胁、联盟责任与防御需求；如果某项共同安全安排要求一方承担不可接受的单边脆弱性，它在战略上也很难稳定。共同存续与国家现实必须同时显影，而不是彼此取消。

目录

- 一、导论 | 当智能进入战略互动，安全不再是单方能力问题
- 二、成熟专业遗产 | 国际关系与战略安全已经证明到哪里
- 三、现实断裂 | AI 正在改变战略互动的哪些输入条件
- 四、智能竞速与安全困境 | 各自合理为何可能共同更不安全
- 五、威慑 | 更快、更准、更强为什么不自动更稳定
- 六、危机稳定 | 时间压缩、核验窗口与升级控制
- 七、信号、归因与欺骗 | 信息更多不等于不确定性更少
- 八、自主系统与决策支持 | 从系统自治到战略互动
- 九、联盟与互操作 | 共同能力也会生成共同依赖
- 十、AI 基础设施与战略依赖 | 算力、芯片、云、数据与供应链
- 十一、核风险与战略稳定 | 最高不可逆压力测试
- 十二、能力扩散与非国家行为体 | 扩散不等于等量化
- 十三、从军事优势到共同安全 | 哪些 AI 用途可能提高稳定
- 十四、国际治理正在形成什么、尚未形成什么
- 十五、共同存续边界 | 文明尺度进入哪里、在哪里停止
- 十六、什么事实能够改变本文 | 最强反驳、败北条件与研究议程
- 结论 | 当更快成为可能，更慢的核验反而更珍贵
- 续读与共同研究入口
- 附录 A | 来源—主张边界登记
- 附录 B | 战略压力测试矩阵
- 生成式人工智能使用声明
- 参考文献
- 版本说明

一、 导论 | 当智能进入战略互动，安全不再是单方能力问题

安全从来不是一个主体单独拥有的属性。Baldwin 提醒，讨论“安全”必须说明谁的什么价值免于何种威胁、付出何种成本；Herz 与 Jervis 所发展的安全困境则进一步指出，一方为提高自身安全采取的措施，可能在无意中降低他方安全，并诱发反向军备与怀疑。[1-3] 这一成熟遗产在 AI 时代没有失效，反而变得更重要：AI 的战略影响很少由一方性能单独决定，而取决于对手如何解释、如何预期、如何模仿以及如何加速。

联合国大会于 2024 年通过 A/RES/79/239，将军事领域 AI 及其对国际和平与安全的影响作为专门议题处理，明确指出国家正在把 AI 整合进广泛军事应用，并强调国际法在相关 AI 能力生命周期中的适用。[27] 2025 年秘书长报告 A/80/78 汇集成员国、国际组织、红十字国际委员会、研究机构与产业提交，所呈现的机会与挑战远超自主武器本身；2025 年大会 A/RES/80/58 又决定在 2026 年召开开放参加的三日非正式交流。[28-30] 这说明军事 AI 已不再只是技术预测，而成为正式国际安全议程。

但“进入国际议程”不等于“风险已经被证明”。UNIDIR 对军事 AI 的研究特别强调，AI 可参与武器之外的十八类军事任务，涵盖信息处理、规划、后勤和支持功能；技术可行性与现实采用、组织吸收及战略后果仍须分别判断。[25] 因此，本文拒绝两种简化：一是把每一次 AI 军事采用都写成升级必然；二是因为尚无单一决定性证据，就把战略风险当作夸张。真正的研究对象是中间链条。

1.1 本文的核心问题

本文追问：当 AI 持续提高信息处理、感知、预测、决策支持、网络行动与部分自主执行能力，而国家仍在相互竞争、信息不完全并担忧相对优势时，安全困境、威慑、危机稳定、升级控制、联盟互动和共同存续条件会怎样变化？这个问题至少要求区分五个层次：技术能力、组织采用、战略姿态、对手认知、互动结果。任何一层都不能被另一层直接替代。

1.2 三个不能自动转换的等号

观察到的变化	常见自动推断	必须继续证明的桥
系统更快	国家更安全	速度只有在核验、责任、对手反应与降级窗口仍成立时才可能转化为安全
信息更多	误判更少	数据质量、欺骗、模型校准、组织认知和归因决定信息是否真正降低不确定性
武器更自主	威慑更可信	自主性可能提高响应，也可能降低可控性、增加误读和失控担忧

表 2 从技术性能到战略结果之间的三座断桥

二、 成熟专业遗产 | 国际关系与战略安全已经证明到哪里

2.1 安全困境：安全是互动结果

安全困境传统的关键并不是“国家一定敌对”，而是即使行为者主要追求自保，只要意图难以可信传达、攻防姿态难以区分、对未来缺乏信任，防御性选择也可能被对方读成威胁。[2-3] AI 因此首先改变的不是国家“有没有自利”，而是攻防能力是否更易区分、优势变化是否更快、意图与能力之间的信号是否更难被解释。

2.2 战争与危机：信息、承诺和误判仍然关键

Fearon 的理性主义战争解释把私人信息与误报激励、以及承诺问题置于核心位置；Schelling 则把胁迫、威慑和风险操纵置于相互预期中理解。[5,7] Jervis、Allison 与后来的组织研究进一步表明，国家并不是无摩擦的统一理性体：知觉偏差、官僚流程、组织例程和有限信息可以改变危机选择。[8-12] 这些遗产意味着，AI 不会简单“替理性增加算力”。它可能改变信息结构，也可能把旧的组织偏差写入更快流程。

2.3 战略稳定不是一个单一指标

战略稳定至少应区分：和平时期是否持续产生军备竞速压力；危机时是否存在先发或预防行动诱因；冲突发生后是否容易跨域、跨层级升级；威慑威胁是否可信且仍可控制；指挥控制是否能在压力下维持授权和辨认。[13-19] 新兴技术研究反复提示，同一种技术对这些维度可以产生相反影响，甚至在一个阶段稳定、另一个阶段不稳定。[13-15] 因此本文不建立“AI 战略稳定总分”。

维度	核心问题	AI 进入后的主要变量
和平/军备竞赛稳定	是否因担心落后而持续提前部署、扩张军备	采用速度、相对能力、扩散、可见性、攻防区分
危机稳定	危机中是否产生“现在不动以后更糟”的先发压力	预警、决策时间、误报、对手姿态、通信窗口
威慑稳定	威胁与承诺是否可信且可理解	能力、意图、可控性、信号一致性、报复韧性
战时升级稳定	冲突后是否可把行动限制在可控层级	跨域耦合、自动反应、目标识别、组织例程
指挥控制韧性	权威链是否能识别错误、接管、停止和修复	数据、通信、权限、冗余、人机关系

表 3 战略稳定的分维审查

三、现实断裂 | AI 正在改变战略互动的哪些输入条件

AI 并非第一次把技术引入战争。雷达、卫星、核武器、精确制导、网络技术都改变过信息、火力和组织。当前断裂在于：同一类通用 AI 能力可以同时进入情报融合、目标识别、网络防御、后勤、模拟、信息行动、决策支持和自主系统，而且很多能力来自民用市场并可快速软件更新。Horowitz 因此把 AI 更接近“使能型技术”而非单一武器；其影响高度依赖组织吸收和配套制度。[16]

UNIDIR 2023 报告把军事 AI 视野从“自主武器”扩展到广泛任务，NATO 2024 修订战略也把生成式 AI、双用民用方案、数据、算力、能源、人机问责与互操作纳入同一政策框架。[25,36] 2025 年 NATO 年度报告还记录其完成 Palantir Maven Smart System NATO 采购并在演习中测试，用于增强数据分析、决策流程和战略响应。[39] 这些是 A 类采用事实；它们只能证明 AI 正在进入军事组织，不证明其已经改善或恶化战略稳定。

3.1 六个需要重新研究的输入条件

输入条件	可能的能力变化	战略上必须继续追问
速度	信息处理、方案生成、任务执行可加快	核验、授权与沟通是否同步缩短

输入条件	可能的能力变化	战略上必须继续追问
可见性	传感与数据融合可能提高态势感知	对手是否把更强感知视为防御、攻击准备或先发优势
不确定性	模型可量化概率并处理海量数据	数据污染、分布外情境、欺骗和模型错误是否制造新不确定性
自治	部分任务可在更少即时人工输入下完成	授权边界是否清楚，异常时是否仍能接管
扩散	软件和民用能力可跨组织快速传播	能力是否真正等量扩散，还是被算力、数据、人才和组织能力重新分层
双用性	民用基础设施可成为军事能力条件	商业供应链、云、芯片和模型是否进入战略依赖

表 4 AGI 时代战略互动输入条件的现实断裂

3.2 “AGI”不是本文成立的前提

本文所称“AGI 时代”，不是声称某个统一定义的 AGI 已经完成，而是指高级 AI 持续进入高后果决策链，使战略互动开始按“越来越强的智能能力可被普遍调用”的条件重组。真正 AGI 何时出现、是否具备主观体验或主体连续性，均不是本文的战略结论前提。当前系统能否被信任、如何测试、是否具有意识，也分别交由计算机科学、认识论和心灵研究继续判断。

四、智能竞速与安全困境 | 各自合理为何可能共同更不安全

当 AI 被视为经济、情报、网络、无人系统和指挥优势的共同底层能力时，竞争者会产生一个现实激励：即使无法确定“更多 AI”是否真正增加安全，也可能担心“不采用”先带来相对劣势。这个机制与传统安全困境一致，却因软件扩散、民军融合和能力更新速度获得新的时间结构。[3,16]

4.1 竞速不是单纯“谁先获得 AGI”

战略竞速可以发生在多个中间层：更快情报融合、更广传感覆盖、更低成本无人系统、更强网络自动化、更短指挥链、更大算力储备、更快模型更新。若研究只盯着一个“AGI 到达日”，会遗漏现实中已经发生的渐进竞速。相反，把所有 AI 投资都叫“军备竞赛”也会夸张。真正需要观察的是：一方投资是否系统性改变他方威胁认知和部署选择，以及双方是否因此进入相互强化的采用循环。

4.2 攻防属性不能由 AI 标签决定

Jervis 的攻防区分和 Garfinkel、Dafoe 关于攻防平衡随投入规模变化的研究都提示：一项技术究竟偏向进攻还是防御，必须相对于具体任务、地理、对手和投入水平判断。[3,15] AI 可提高网络攻击、侦察和目标搜索，也可提高防空、异常检测、维护和韧性。不能因为“AI 更强”就预设攻势占优，更不能从抽象算力规模直接推出先发优势。

4.3 B/C 类条件命题：速度竞争何时可能转化为制度竞争

现有证据只支持链条的若干局部环节，不足以证明完整结果已经发生。第一，对手得知 AI 或自动化介入高后果决策，可能改变其对威胁、可信性、事故和责任的判断（P3a，B）；已有实验与认知研究显示，算法来源会影响

响国家安全专家或军事人员的行动意愿、归责和信任，但方向并不单一。[43,45-46] 第二，这种感知可能引出报复、退让、独立核验、追赶、反制、监督、拒用或共同护栏，方向取决于任务、组织与危机条件（P3b，B—C）。对中美军事 AI 认知与自动化核威胁的研究提示，相对优势焦虑、模仿和可信性变化都可能推动回应；但它们不能证明对手必然采取对称加速。[44,47]

第三，只有当一方的流程变化被对手理解为缩短侦测、决定或行动时间，并触发进一步自动化、预授权、警戒或缩短授权程序，而这些回应又反过来改变另一方流程时，双方才可能共同压缩验证、沟通和政治重新进入的保留时间（P3c，C）。第四，只有在二次打击信心低、教义不透明、独立核验弱和危机沟通脆弱等条件下，这种共同压缩才可能提高误判、意外或非预期升级风险（P3d，C）；更好的防御、独立复核、热线、可逆授权和共同护栏也可能扩大纠错时间。[17,48-50]

这里真正重要的不是某个算法跑得多快，而是整个战略互动把“慢一步”定义成多大成本。但本文不假定 AI 必然加速危机，不假定对手对称适应，也不声称这条完整链已经被经验观察。

与本文直接相邻，Yuan（2026）已经讨论美中军事 AI 竞争中的行动—反应压力、采用激励、网络风险、威慑与危机决策。[51] 本文不主张互动式时间压缩已经被经验观察，也不主张首先提出上述宽问题；本文的特定贡献，是将“感知—回应—共同时间预算—危机风险”拆成 P3a—P3d 四个可分别败北的命题，并以（T_V，T_C，T_P）操作化验证、沟通与政治重新进入的共同保留时间，从而把问题从“单方能否更快决定”转向“互惠、多轮适应是否在后果不可逆前缩短纠错时间”。

五、威慑 | 更快、更准、更强为什么不自动更稳定

威慑不是拥有能力本身，而是让对手相信某种行动会带来不可接受后果，同时又让威胁、承诺与限制足够可信。[6-7] Huth 与 Russett 的历史研究说明，威慑成败与具体关系、局部力量条件和承诺结构有关，并不能被一个总体军力指标概括。[6] 因此，AI 进入威慑至少要分开能力、可见性、可信性、可控性和解释。

5.1 AI 可能提高威慑的三条路径

第一，更好的态势感知与预警可能提高报复能力生存性，让对手更难相信突袭可以奏效。第二，模拟和决策支持可能帮助领导者比较更多危机路径，识别误判与替代方案。第三，网络防御、异常检测和无人高风险任务可能降低脆弱性，使一方不必因担心瞬间失去能力而提前行动。[20-24] 这些都是值得认真研究的稳定化可能，而不是技术悲观主义应该忽略的例外。

5.2 AI 也可能削弱威慑的三条路径

同样的感知与分析能力如果被认为能够定位对方的隐蔽力量、瘫痪指挥网络或提高首击有效性，就可能增加“窗口即将关闭”的担忧；自动化响应若被对手认为缺乏政治可控性，也可能使威胁看起来既更可信又更危险；模型输出和数据来源不透明，还可能使一方无法判断对方究竟真的获得新能力，还是只是相信自己获得了新能力。[17-23] 因此，威慑稳定取决于能力与认知共同变化。

AI 变化	可能的稳定效应	可能的不稳定效应
预警更快	更早发现攻击，增强报复韧性	误报更快进入决策，压缩人工复核

AI 变化	可能的稳定效应	可能的不稳定效应
目标识别更准	降低误击，提高防御效率	对手担心隐蔽力量更易被发现，产生脆弱性
自动响应	降低通信中断后的瘫痪风险	可控性下降、误触发或被对手误解为预授权
模拟更多方案	帮助比较后果	模型共同盲点、虚假精确度和自动化依赖
网络 AI 防御	提高系统韧性	同类 AI 也可用于攻击、欺骗和侦察

表 5 同一 AI 能力对威慑的双向可能

六、危机稳定 | 时间压缩、核验窗口与升级控制

危机稳定关注的是：双方在高度紧张时，是否因为担心等待会使自身处境恶化，而产生先行动机。AI 最值得警惕也最容易被夸张的变量是速度。速度本身不是风险；真正的问题是哪些环节被加速、哪些环节没有同步获得更好的证据，以及政治决策是否仍有时间重新进入。

6.1 决策时间不是越短越好

在灾害、防空或网络异常中，更快识别可以扩大纠错窗口；在相互战略竞争中，更快也可能把“等待核验”变成显著成本。Horowitz 把更高操作速度与更低人类控制作为自主武器可能影响危机和战略稳定的两个关键特征。[17] 但 Talmadge 对冷战案例的研究提醒，新技术通常不是升级的独立充分原因，更常是介入既有政治和军事条件的变量。[14] 两者并不矛盾：速度值得被研究，却不能被写成自动升级定律。

6.2 有意义的人类控制首先是一个时间问题

形式上的“人在环中”不够。若人只能在数秒内确认已经被模型完成的问题定义、证据筛选、目标排序和行动建议，他虽然仍然点击按钮，却未必真正改变结果。战略意义上的人类控制至少需要：看到与后果相称的信息；理解不确定性与替代解释；拥有拒绝、延迟、降级或升级复核的权限；组织不以“错过窗口”为由惩罚一切谨慎；关键时刻可以把机器建议重新转成政治判断。DoD 3000.09 和多项国际政治承诺都强调适当人类判断、测试和失效处置，但这些制度要求能否在真实高压环境中有效，仍是经验问题。[37-38]

6.3 决策窗口应被当作战略资源

传统战略研究重视武器、生存性与通信链；AI 时代还需要把“可核验、可质疑、可沟通的时间”作为独立研究对象。最短反应时间不是天然优势。如果更快响应同时消除判断冗余，系统可能在常态下表现优秀，在异常情形下却更脆弱。相反，刻意保留暂停、双重确认、交叉传感和危机沟通可能在局部牺牲速度，却增加整体稳定。这个权衡需要战争推演、组织实验和跨国危机研究，而不能由本文在纸面上决定。

本文把共同时间预算表示为三维向量（ T_V, T_C, T_P ）： T_V 是从出现关键异常或竞争性解释，到完成独立验证、反例生成与授权复核仍可支配的时间； T_C 是各方在行动不可逆前核对意图、纠正归因、传递限制与启动危机沟通仍可支配的时间； T_P 是有权承担后果的政治决策者重新进入、改变机器或组织既定路径仍可支配的时间。三个维度不得合成为“战略纠错总分”，也不得以一个维度的增加兑换另一个维度的失守。

最低相互压缩判据要求观察到双边或多边、至少两轮的流程反馈：一方流程变化触发对手回应，对手回应又反过来改变前者；同时，每一方至少有一个相同维度相对其自身基线出现净减少，且该维度没有被新增的独立验证、有效沟通或政治复核恢复。单方变快、单轮时长下降、纯技术延迟减少或总流程缩短，都不足以证明共同压缩。退让、独立核验、监督、拒用和共同护栏必须作为竞争性结果同时编码。

七、信号、归因与欺骗 | 信息更多不等于不确定性更少

国际危机中的困难从来不只是“缺信息”，还有信息由谁发出、对方是否相信、是否存在欺骗激励，以及行动是否被错误归因。Fearon 的私人信息与误报激励、Jervis 的知觉研究都说明，更多信息并不会自动消除战略不确定性。[5,8] AI 扩大生成、筛选、合成和分析能力后，可能同时改善可见性与制造新的信任问题。

7.1 深度伪造不是唯一问题

合成音视频当然可能影响危机信号，但更深的问题是信息链整体的真实性：传感器是否被欺骗，数据是否被污染，模型是否被提示注入或对抗操纵，来源链是否完整，决策者是否知道模型正在推断而非观察。NATO 修订 AI 战略已把 AI 赋能的信息行动和虚假信息列为关切，并强调测试、数据和人机问责。[36] 这些政策关切只证明风险已被正式识别，不证明每种操纵在现实危机中都有效。

7.2 归因能力提高也可能改变升级结构

网络冲突研究对“匿名性必然摧毁威慑”一直存在反论。Gartzke 与 Lindsay 都指出，网络行动的战略效果取决于目标、情境、情报和与现实军事行动的连接，而不是“网络”这一标签本身。[40-41] AI 可能提高异常检测和归因速度，也可能提高攻击自动化与欺骗复杂度。战略研究必须比较净结果：哪些场景下 AI 使归因更可靠，哪些场景下它只是让错误归因更快。

7.3 信号必须保留“可被对方理解”的条件

威慑和降级都依赖对方能理解行动的含义。如果大量决定由不可见模型和组织内部规则生成，对手即使观察到行动，也可能不知道它代表政治意图、自动响应、误报还是局部指挥行为。AI 治理因此不能只做内部可靠性，还要研究战略可解释性：哪些能力、限制、授权结构和危机沟通信息可以在不泄露关键秘密的前提下被外部理解。

八、自主系统与决策支持 | 从系统自治到战略互动

自主武器是军事 AI 的重要子集，但不是军事 AI 全部。ICRC 把自主武器描述为激活后能够在没有进一步人类干预的情况下选择并对目标施加武力的系统，并持续主张对不可预测系统和针对人的自主攻击建立新的法律约束。[34] 这是一项具有明确机构立场的法律与人道主义主张，不能被本文冒充为国际社会已经完成的统一法律结论。

8.1 自主性不是一个开关

在战略分析中，更有用的做法是沿行动链分解：谁定义目标，谁选择数据，谁建立识别规则，谁允许系统进入哪些地理和时间范围，谁可以修改任务，谁确认攻击，谁可以停止，谁在错误发生后承担修复。系统可以在执

行层高度自主，却在目的和权限层受到严格控制；也可以表面保留人工确认，却在实际上由算法预先决定几乎全部可见选项。

8.2 “责任在人类”仍需组织结构承载

国际政策文件越来越强调人类责任。美国政治宣言要求军事 AI 在负责任的人类指挥控制链中使用，并提出法律审查、明确用途、全生命周期测试与失效防护；REAIM 2026《Pathways to Action》同样把人类判断、风险评估、测试验证和降低误解、误算与非预期升级列为实践方向。[31,38] 这些原则的战略问题是：责任是否在速度、层层外包和供应链依赖中仍然真实存在。

8.3 从战术可靠到战略稳定还隔着对手

一个自主系统可以在限定任务中高度可靠，却仍可能因规模、速度、部署位置或对方想象而改变战略互动。相反，一个技术上不完美的系统若用途透明、范围有限、错误可控且对方不把它视为先发工具，未必造成同样压力。战略研究不能把“技术可靠”直接翻译成“国际稳定”，也不能把“技术存在缺陷”直接翻译成“必然升级”。

行动链层次	首先要回答	专业交接
目的/任务定义	谁决定系统要实现什么军事目的	政治/军事授权；本文只分析其战略互动后果
数据/识别	哪些观察进入系统、怎样形成分类	AI/计算机、认识论；战略研究看误读与信号后果
权限/范围	可在哪些时间、区域、目标类别行动	法学/制度授权；战略研究看对手预期
执行/自治	系统何时可自行采取行动	AI 工程；战略研究看速度、可控性与升级
停止/修复	谁能中止、隔离、回收并解释事故	系统工程/组织；战略研究看危机降级能力

表 6 自主系统进入战略互动前必须分开的行动链

九、联盟与互操作 | 共同能力也会生成共同依赖

联盟研究长期指出，结盟既提供安全，也制造被抛弃与被牵连的双重担忧；联盟内部承诺、能力差异与战略自主不是附属问题。[4] AI 时代，互操作进一步进入模型、数据、云、接口和软件更新层。NATO 修订 AI 战略把“联盟范围内 AI 系统互操作”列为明确目标，并同时强调测试、标准与责任使用。[36]

9.1 互操作的稳定收益

共同数据标准、态势图和决策支持可能提高联盟协调，减少友军误判，缩短信息孤岛，增加共同防御效率。对于跨国防空、海上态势、网络防御和后勤，这些收益具有现实价值。联盟因此有充分理由追求 AI 互操作。

9.2 互操作也可能制造相关性失败

若多个盟国依赖相同模型、云服务、数据管线或供应商，局部漏洞、模型偏差和政策变化可能同步传播。更复杂的是，能力不对称会改变联盟内部责任：技术更强的成员可能事实上提供更多判断基础，技术较弱成员仍在政治上承担行动后果。此时“共同使用”不等于“共同理解”，更不等于“共同授权”。

9.3 联盟 AI 需要两层可见性

第一层是内部：盟国知道系统能力、限制、版本和权限。第二层是外部：对手能够在必要程度上理解某些部署意味着什么，避免把例行互操作升级误读为突然形成的进攻性预授权。透明当然有安全边界，但完全不可解释也有战略代价。具体披露范围必须由国家安全、军控和法律专业进一步决定。

十、AI 基础设施与战略依赖 | 算力、芯片、云、数据与供应链

AI 作为使能型技术的另一特点，是军事能力越来越依赖民用与商业基础设施：芯片、数据中心、云平台、网络、模型供应商、软件生态与能源。NATO 2024 战略明确把民用双用方案、算力和能源需求、产业与学术生态纳入防务 AI 条件。[36] 这使“战略自主”不再只看武器库存，还要看关键数字基础设施能否持续工作、替代和恢复。

10.1 基础设施控制不是自动主权

控制重要云、模型或芯片供应链可以产生显著事实影响力，但它不自动生成合法公共权力。这一点由政治学和法学继续判断。战略安全在这里关心的是依赖如何改变危机脆弱性：供应中断会不会使某国在危机中失去关键能力？商业服务条款、跨境数据和软件更新是否可能成为战略摩擦点？

10.2 打击基础设施可能产生跨域升级

AI 基础设施往往同时服务民用经济和军事组织。针对云、数据中心、通信或芯片供应链的行动，其军事价值与民用后果可能高度耦合。对手也可能无法判断某次网络或物理行动是有限军事措施、广泛经济攻击还是更大升级的准备。这里需要国际关系、网络安全、国际法、基础设施韧性和生态能源专业重新交接，而不能由“战略必要性”一次结算。

十一、核风险与战略稳定 | 最高不可逆压力测试

核领域是本文最严格的压力测试，因为错误后果可能无法修复。SIPRI 2020 以来的系列研究反复分析 AI 进入预警、ISR、指挥控制、无人平台和相关常规系统的双向可能：更好的感知、维护与决策支持可能增强韧性，过早采用、误报、网络攻击、能力误判和军备竞速也可能增加风险。[22-24] 这些研究多以技术—战略综合、区域专家研讨和公开资料为基础，适合作为 B 类与政策分析材料，而不应被写成“AI 已进入核发射决定”的普遍 A 类事实。

11.1 2026 年的证据边界

Saalmann 2026 对中国、印度和美国的公开资料进行系统调查，发现三国关于 AI 与核及核相关系统的公开程度差异显著，并将其工作定位为未来接触和建立信任措施的基线。[24] 这恰好说明当前研究最需要克制：公开材料可以证明一些 AI 相关功能、政策和趋势，却通常不足以重建各国完整 NC3 内部算法、授权和实际核决策流程。本文因此不声称已知任何国家把核使用最终授权交给 AI。

同样重要的是，联合国大会于 2025 年 12 月 1 日通过 A/RES/80/23，专门处理人工智能融入核武器指挥、控制与通信系统的潜在风险。该决议要求在核武器尚未彻底消除的条件下，对整合 AI 技术的相关系统维持人类控制

和监督，并以 118 票赞成、9 票反对、44 票弃权通过。[42] 它证明 AI—核指挥控制通信问题已经进入独立的多边核风险议程，但不能被写成所有核武器国家已经形成一致实践，更不构成新的条约义务。

11.2 核领域最需要保留的不是“人类签字”，而是完整政治判断

在核危机中，真正危险的可能不是系统直接发射，而是它越来越早决定什么被看见、哪些信号被认为可信、哪些选项排在前面、需要多快回应。如果政治领导者只在最后一步选择已经由机器压缩后的极少数选项，形式人类授权仍在，实质判断空间却可能变窄。这个问题不能靠一条抽象“人类控制”口号解决，需要指挥控制研究、组织模拟、危机推演和真实程序审计。

应用环节	潜在收益	主要风险	优先验证方向
预警/态势感知	提高检测与融合速度	误报、对抗欺骗、虚假精确度	不同源交叉核验、保留不确定性
NC3 辅助	提高信息组织和通信效率	共同组件故障、网络渗透、自动化依赖	冗余、隔离、分层授权
常规反力量能力	提高定位与任务规划	对方担心核力量脆弱，先发压力上升	战略对话、限制误读、透明/信任措施
无人/自主平台	提高危险环境生存与持续性	部署意图不清、失控或错误升级	限制任务范围、停止能力、事故沟通
模型推演	扩展方案比较	训练数据偏差、把模型当现实	多模型/人类对抗审查、显式失败条件

表 7 AI—核风险的双向压力测试

十二、能力扩散与非国家行为体 | 扩散不等于等量化

AI 软件的可复制性容易产生一种想象：能力会快速扩散，因此所有行为体很快接近同一水平。Horowitz 关于军事技术扩散的框架和 Goldfarb、Lindsay 对数据与判断互补性的分析提醒，真正军事能力仍依赖组织、数据、人才、硬件、训练、任务理解和整合能力。[16,19] 因而，模型开放或成本下降可能降低部分门槛，却不必然消除国家间与组织间差异。

12.1 非国家行为体的风险必须按任务分解

AI 可能降低信息搜集、代码生成、语言操纵、无人平台规划或目标分析的门槛，但不同任务的现实门槛差异巨大。把“可获得模型”直接写成“获得战略级能力”会制造恐慌；反过来，因为非国家行为体缺少国家级基础设施就忽视某些低成本高影响任务，也会漏掉真实风险。研究应按具体任务、资源、数据和行动链评估，而不是用“AI 民主化”一词结算。

12.2 扩散会反过来改变国家危机决策

当国家无法迅速确认攻击来自谁，或担心非国家行动会被误认为国家行为，归因就成为升级控制的一部分。AI 生成内容、代理网络和自动化攻击工具可能进一步模糊责任链。这里必须与网络安全、情报、法学和社会信息研究重新交接。

十三、从军事优势到共同安全 | 哪些 AI 用途可能提高稳定

如果本文只列风险，它会犯与技术乐观主义相反的同类错误。战略研究真正需要比较的是：哪些 AI 用途能在不制造更大相互不确定性的条件下，增加可验证的防御、预警、沟通和纠错能力。UNODA、REAIM 与多国政策都同时讨论 AI 改善信息处理、人员安全、态势感知、后勤和决策支持的机会。[26,28,31]

13.1 防御性、可辨认、可验证的能力更可能减少困境

安全困境研究提供一个方向：若防御比进攻更有优势，且防御姿态与进攻姿态更容易区分，合作条件通常更好。[3] 因此 AI 的共同安全价值可能不在“所有人都少用 AI”，而在于优先发展并显影那些提高基础设施韧性、事故检测、通信可靠性、灾害响应和有限防御，同时不显著增加先发打击能力的应用。具体哪些用途满足条件必须逐项验证。

13.2 AI 可用于保留“更多选择”，而不只是选择更快

高质量决策支持最有价值的功能之一，可能是主动生成反例、显示不确定性、比较降级路径、提醒信息缺口和保留延迟选项，而不是把一个“最优动作”推到领导者面前。Goldfarb 与 Lindsay 强调预测变便宜会提高判断的相对重要性。[19] 战略系统若只优化响应速度，就可能错用 AI；若用于扩展领导者对不确定性和替代方案的理解，AI 可能扩大而非压缩政治判断。

13.3 军控与信任措施也可以拥有 AI 工具

AI 可以帮助监测公开信息、整理条约数据、识别异常、模拟核查方案和支持危机热线的信息处理。REAIM 2026 也把自愿信任措施、政策与风险评估信息分享、减少误解误算和非预期升级列入行动路径。[31] 这些工具并不能自动生成信任；但如果它们提高可验证性且不把敏感判断集中到不可争议黑箱中，就可能成为稳定化基础设施。

十四、国际治理正在形成什么、尚未形成什么

截至 2026 年 8 月 26 日，军事 AI 治理已经明显从零散倡议进入多层并行阶段，但不存在一项覆盖所有军事 AI 问题、全球统一且全面具有法律约束力的单一制度。准确描述这一状态十分重要：既不能说“国际社会没有规则”，也不能把政治宣言、联合国大会决议、国家政策与条约法写成同一法律效力。

14.1 联合国大会：议题已经正式全球化

A/RES/79/239 将军事领域 AI 及其对国际和平与安全的影响作为专门议题纳入联合国大会决议，强调国际法适用于相关 AI 能力生命周期，并要求秘书长报告。[27] A/80/78 于 2025 年提交，汇集多方观点；A/RES/80/58 随后决定继续在联合国框架内开展 2026 年非正式交流。[28-30] 2026 年 6 月 15 日至 17 日的日内瓦会议已经举行，议题包括术语、能力趋势、生命周期、技术融合、国际合作、规范提案和下一步。[30] 这些是重要制度进展，但大会决议和交流本身不等于新条约。

14.2 LAWS：CCW 谈判仍在进行

CCW 框架下致命自主武器系统政府专家组 2026 年第一阶段会议已于 3 月 2 日至 6 日举行，第二阶段安排在 8 月 31 日至 9 月 4 日；截至本文证据截点，第二阶段尚未开始。[33] 因此，本文不能把仍在协商的“滚动文本”写成已经达成的最终法律规范。2026 年 8 月 25 日，联合国秘书长与 ICRC 主席再次公开呼吁尽快启动具有法律约束力的国际文书谈判，这一呼吁体现强烈规范与政策立场，但仍不是条约已经存在的事实。[35]

14.3 REAIM 与国家/联盟政策：从原则走向操作

REAIM 2026《Pathways to Action》明确承认军事 AI 既有人员保护、平民保护和改善决策的潜在收益，也存在误算、偏见、失控和升级风险，并提出法律审查、风险评估、TEVV、人类参与和信任措施等实践方向。[31] UNIDIR 对 2025 年五场区域咨询的总结同时显示，国家政策、区域优先事项与多方利益相关者关切并不完全相同，这进一步提醒我们：REAIM 更适合被理解为正在形成共享行动空间的开放政治进程，而不是已经收敛的全球规则。[32] NATO、美国等国家与联盟也形成自己的责任使用原则、审查和测试要求。[36–39] 这些实践能否形成跨阵营共享、可验证且不会变成单方竞争劣势的稳定机制，仍是国际关系问题。

治理层	已经形成什么	不能写成什么
UNGA 79/239、80/58	全球政治/制度议程；确认国际法适用并推动报告与交流	联合国大会决议，不等于专门军事 AI 条约
UNGA 80/23 AI—核 C3	专门处理 AI 进入核武器指挥、控制与通信系统的潜在风险，并要求维持人类控制与监督	以 118—9—44 通过；重要政治规范压力，不等于条约义务或全体一致实践
A/80/78 与 2026 非正式交流	汇集国家与多方观点，形成共同问题空间	意见多元，不是单一共识结论
CCW GGE on LAWS	围绕 LAWS 规范与操作框架持续谈判	截至截点仍在协商，无全面最终文书
REAIM Pathways to Action	国家主导、多方参与的政治与实践建议	政治文件；开放支持，不等于普遍法律义务
NATO/国家政策	联盟或国家内部责任 AI、测试、授权和使用规则	适用主体有限，不能代表全球统一规范
ICRC 立场/联合呼吁	人道法解释、禁止与限制主张，推动法律发展	机构立场与政策倡议，不等于现行普遍条约条文

表 8 截至 2026 年 8 月 26 日的军事 AI 治理层级

十五、共同存续边界 | 文明尺度进入哪里、在哪里停止

战略研究处理的是相互竞争如何产生安全或不安全，但并不自行决定“什么样的文明值得存在”。本文在这里才引入 F 类文明底线：在核升级、失控自主武器、大规模关键基础设施打击等可能产生高度不可逆后果的场景中，共同存续必须作为国家局部收益之外的独立后果显影。

15.1 共同存续不取消国家安全

一个国家面临真实威胁时，有充分理由防御自身、维护盟友承诺和保护人民。要求它为了抽象“人类整体”接受单方面不可逆脆弱，并不会自动产生稳定秩序。共同存续更准确的战略含义是：各方的安全安排不能把“所有人都更接近毁灭”当作可忽略外部性；应寻找能让主体长期利益与共同存续条件更相容的机制。

15.2 共同存续不生成世界主权

从“某些风险跨境且可能不可逆”不能直接推出“必须建立单一世界政府”。主权、授权与合法性仍由政治学、政治哲学和法学另行证明。国际关系在这里最多说明：某些外部性无法由单方安全逻辑充分内部化，因此需要协调、信任措施、军控或共同规则；究竟采用条约、多边机制、联盟安排、多中心治理还是其他结构，必须比较现实效果。

15.3 高不可逆决定的最低 F 类方向

本文公开承担四项方向：责任必须可追溯；授权必须明确；关键假设应尽可能可审查；必须保留纠错、终止和降级路径。它们不是国际关系经验事实，也不自动具有法律效力，而是因为不可逆后果使“事后知道错了”可能已经失去意义。任何具体制度仍须接受战略可行性、国际法、国内政治和技术工程的共同验证。

十六、什么事实能够改变本文 | 最强反驳、败北条件与研究议程

16.1 最强反驳一：技术并不是升级的独立原因

这一反驳成立相当大一部分。Talmadge 的历史研究明确对“新技术作为升级的独立、首要驱动”持怀疑，强调技术常是必要但很少充分的介入变量。[14] 本文因此不主张 AI 技术决定战争。若未来研究进一步证明政治意图、既有战略与组织过程几乎解释了全部所谓 AI 效应，AI 变量应被降级，甚至不再需要独立专业稿。

16.2 最强反驳二：AI 可能总体提高危机稳定

完全可能。如果更好的传感、预测、网络防御、模拟与通信长期显著降低误报，扩大领导者可见选项，并没有造成新的先发压力或责任稀释，那么本文关于“速度竞争压缩纠错窗口”的主张必须收缩。Shandler、Gross 与 Shereshevsky 使用高保真军事决策支持系统界面，对 2,015 名以色列现役或退役军人开展两项实验；AI 来源标记在两项研究中均未显著缩短决策时间，第二项研究加入解释性信息后，AI 组与人类分析员组的决策时间仍无显著差异。[45] 这是一项任务层面的重要负结果，足以否定“AI 介入必然使人更快决定”的无条件推断，但不能单独裁决国家间危机的多轮互动。战略安全稿不能因为风险叙事更有传播力就拒绝正向证据、负结果或零结果。

16.3 最强反驳三：人机协同可以稳定解决控制问题

如果长期实战外推、演习、组织实验与事故数据表明，高度自动化系统仍能稳定保持有效接管、清晰责任 and 高质量政治判断，而且人类没有被推到末端确认位置，那么“人在环不等于人有控制”的风险需要重新限域。本文关心的不是维持一个原则，而是控制是否真实。

16.4 最强反驳四：国际治理会比风险扩散更快成熟

当前多边进程确实在加速。若未来形成跨主要军事国家可执行、可核验、能处理技术更新和危机信息的规则体系，并稳定降低误算与竞速压力，那么本文关于治理滞后的判断应修订。相反，如果规则只在低竞争场景有效，也必须精确说明适用边界。

16.5 败北条件

可检验命题	可能推翻/改变本文的事实	本文必须怎样改
P3a 感知改变	实验与跨国研究显示，对手识别 AI 介入或流程变化后，并不系统改变威胁、可信性、事故或责任判断	删除或收缩“感知改变”机制
P3b 回应方向	对手反应无稳定方向，且加速、预授权或追赶并不比退让、核验、监督、拒用或共同护栏更常见	取消对称加速倾向，保留开放的回应集合
P3c 共同时间压缩	双边／多轮数据未观察到流程相互反馈，或 AI 介入未共同压缩验证、沟通和政治重新进入时间	删除或显著收缩“共同时间压缩”命题
P3d 危机风险传导	即使共同时间预算下降，误判、事故或非预期升级风险也未上升，或被独立复核、热线、可逆授权等稳定抵消	收缩风险传导链，提高稳定化机制权重
自动化与责任	长期数据表明更高自动化没有造成责任稀释且接管质量稳定提高	收缩组织控制风险
感知与误判	更强态势感知在多种危机中显著降低误判、先发压力与错误归因	提高稳定效应权重
战略能力差距	AI 扩散后能力差异远低于预期，且不形成新的联盟/国家竞速	收缩相对优势与竞速判断
制度护栏	跨国护栏稳定抵消新风险且不会制造不可接受单边脆弱	弱化治理不足判断
独立事实域	所有关键命题均可被政治学、AI 科学、系统科学和国际法无越权承担	重新审查甚至取消本文独立成篇资格

表 9 本文的败北条件与强制修订规则

16.6 优先研究议程

- P3a | 对手感知：比较人类来源、AI 来源和不同自动化程度的同一信息，检验威胁、可信性、事故、责任与行动意愿是否改变；预注册方向假设，并保留无显著差异。
- P3b | 回应方向：把报复、退让、独立核验、追赶、反制、监督、拒用和共同护栏作为并列结果，检验何种任务、组织、联盟与危机条件推动哪一种回应。
- P3c | 共同时间预算：在双边／多边、多轮战争推演中分别测量 T_V、T_C 和 T_P，观察一方流程变化是否触发对手回应并反向改变前者；报告三维向量，不建立总分，不以单方或单轮变快替代互动证据。

- P3d | 危机后果：在二次打击信心、教义透明度、独立核验、热线与可逆授权等条件间比较共同时间压缩是否提高误判、事故或非预期升级；把反向、负向与零结果纳入强制报告。
- 信号与归因：测试合成内容、数据污染和模型辅助归因在不同情报质量下如何改变领导者置信度与升级判断。
- 联盟相关性：研究共同模型、云和数据接口的失败是否产生跨盟国同步错误，以及多样化和冗余能否缓解。
- 核领域：在不暴露敏感作战细节的前提下，建立 AI 进入预警、NC3 辅助和相关常规系统的透明度与信任措施研究。
- 正向能力：系统识别哪些 AI 应用稳定提高防御、事故检测、危机沟通和降级能力，而不显著提高对手的先发恐惧。
- 制度实验：比较法律审查、TEVV、独立红队、决策暂停、双重确认、模型多样化和跨国事故通报在真实组织中的效果。

结论 | 当更快成为可能，更慢的核验反而更珍贵

AGI 时代的国际关系与战略安全，不应被压缩成“AI 会不会发动战争”。真正的问题发生得更早：国家怎样因 AI 重新看见威胁，怎样因为速度优势改变组织流程，怎样解释对方的新能力，怎样把更多步骤预授权，怎样在联盟中共享模型和责任，又怎样在危机里保留重新判断的时间。

传统战略研究没有因此过时。安全困境、威慑、承诺、信号、误判、联盟、组织事故与升级研究，恰恰提供了避免技术决定论的专业遗产。AI 时代的新任务，是把这些成熟问题放到新的能力条件下重新检验：一项性能提升进入互动后是否仍然是安全提升；一个国家的局部合理选择在聚合后是否仍然增加共同安全。

本文不要求国家放弃竞争，也不把共同存续变成世界主权的许可。它只坚持一条底线：当错误可能不可逆时，速度、效率和能力都不能替人类省略核验、责任、沟通与降级。更强智能真正值得争取的战略价值，不是让战争更快发生，而是让更多误判在变成战争以前被看见。

当每一方都用 AI 让自己“更安全”，共同世界可能变得更不安全。战略研究的责任，是找出这句话在什么条件下成立，也找出它在什么条件下不成立。不成立的条件同样属于答案。

续读与共同研究入口

本文在国际关系／战略安全处只回答：AI 进入国家与联盟互动以后，怎样改变安全困境、威慑、危机稳定、信号、升级与共同存续风险。它不完成模型安全、法律效力、政治合法性、生态承载或文明最终方向的判断。续读因此不是文件堆叠，而是让下一门真正拥有事实责任的专业重新进入。

向上 | 如果继续追问：为什么战略安全最终成为文明问题

进入《AGI 时代的学科重新对账——学科边界、相互作用与共同研究议程》V2.4，以及《AGI 时代人类文明跃迁的机遇、方向、方法与目标实现》文明方向合流深稿 V1.1。前者说明学科为何必须独立、现实因果为何又不能停在学科边界；后者公开承担“这些共同变化值得把人类带向哪里”的方向判断。战略稳定是存续条件之一，却不是文明全部目的。

向下 | 如果只继续国际关系与战略安全本身

继续进入威慑理论、危机历史、军控、安全困境、联盟、情报与误判、核战略、网络冲突、指挥控制、战争推演和跨国危机数据库。优先工作不是再写一篇宏大判断，而是建立可以比较“纯人工／不同 AI 辅助／不同自动化与预授权结构”的真实研究设计。

向左 | 哪些学科必须重新进入

重新进入的专业	为什么不能由本文代替
AI／计算机科学	系统能做什么、怎样失效、如何测试、权限和停止是否可靠
政治学	国家能力、公共权力、主权与授权何以正当
伦理／法学／国际法	武力使用、国际人道法、责任、权利与伤害边界
系统科学	级联、紧耦合、相关性故障和韧性的一般机制
认识论／认知科学	情报不确定性、模型推断、自动化依赖、时间压力与误判
经济／政治经济	军备投入、产业结构、供应链、算力和战略资源
社会／人类学	敌我叙事、信任、身份、动员与战争社会后果

表 10 战略稿的主要学科交接

向右 | 如果不同意本文，应优先从最强反方进入

优先检验四类真正能够改变本文的方向：AI 是否总体增强战略稳定；新技术是否只是既有政治冲突的介入变量而非独立驱动；更强人机协同是否足以解决控制与责任问题；军事 AI 军控是否因为核查困难、相对优势与双用性而难以形成。相反证据不是附录，而是本文继续成立的条件。

向前 | 什么专业工作足以改变本文

新的危机数据、跨国演习、组织实验、战争推演、负结果、事故材料、形式模型、信号实验、制度比较、军控核查工具与红队结果，都可以直接改变本文。Human-COS（高后果选择在不可逆之前进行多主体、多视角后果推演与判断支持的候选系统）若用于战略研究，只能作为候选研究与验证接口，不能反向成为战略事实的证据。

文明未来链 | 避免毁灭之后，人类还希望走向哪里

如果希望进一步看见：当战略安全不再长期吞噬文明全部能力、技术更多用于保护而不是压缩共同未来时，人类的工作、教育、家庭、关系、公共生活和技术协作可能怎样变化，可进入《活的文明》三部曲及五卷书。这里必须保持一句边界：愿景不是证据。《活的文明》用于让读者看见、比较、拒绝或重构可能的新世界，不用于证明本文的任何战略因果。

文明跃迁五卷书：<https://www.civitas.top/documents/p02-five-001/> | 公众总入

口：<https://www.civitas.top/documents/p01-entry-006/> | 完整文库：<https://www.civitas.top/library/>

最有价值的参与，不是赞同，而是足以改变理论的专业工作。

不是为了让更多人读完文明跃迁，而是为了让更多人一起完成文明跃迁。

附录 A | 来源—主张边界登记

来源类型	本文如何使用	明确不能推出
安全困境与威慑经典研究	说明互动、承诺、攻防与信号为何重要	不能直接证明 AI 时代会复制过去结果
新兴技术/AI 战略学术研究	提出速度、自治、组织、攻防与稳定的竞争机制	多数为理论、历史类比或有限案例，不能当未来事实
SIPRI/UNIDIR 研究	提供公开资料整合、区域观点、军事任务与核风险基线	机构研究报告不等于国家实际秘密流程或国际法共识
UNGA/UNODA	证明全球议程、正式决议、报告和会议状态	大会决议/会议不自动构成新条约
UNGA 80/23（AI—核 C3）	证明联合国大会已把 AI 融入核武器指挥、控制与通信系统的潜在风险单列为正式议题，并记录其人类控制/监督主张与表决状态	不能推出核武器国家已有统一实践、形成普遍条约义务，或证明 AI 已经进入核使用最终授权
REAIM/NATO/美国政策	证明国家或联盟责任 AI 原则与治理实践	政治承诺、国家政策不代表全球统一规范
ICRC	证明其人道法解释、政策立场与新规则倡议	ICRC 主张不能写成所有国家已接受的现行条约规则

表 A1 来源与主张强度边界

附录 B | 战略压力测试矩阵

压力事件	最先进入的事实层	战略问题	优先检查
预警模型出现高置信错误	数据/模型/人类校准问题	危机稳定：是否触发先发/报复压力	能否独立复核、延迟、沟通、降级
一方部署更快自动防御	工程提高响应	安全困境：对方如何解释攻防属性	姿态可辨认性与危机通信
联盟共用同一基础模型	互操作提高效率	相关性故障与责任共享	模型多样化、备份、版本可见性
深伪模拟国家领导人命令	信息真实性受损	信号与归因：政治意图是否被误判	来源认证、人工回呼、多渠道核验
AI 辅助常规反力量打击	目标识别更强	核脆弱性感知与跨域升级	战略对话、限制误读、指挥控制韧性
供应商服务中断	商业/技术故障	国家或联盟战略依赖暴露	替代供应、低技术备份、主权边界
非国家行为体使用通用模型	任务门槛下降	归因与报复对象不确定	任务级能力评估、证据链与比例响应
AI 建议显著改善危机决策	正向绩效出现	本文风险主张是否需要收缩	跨案例复现、撤除 AI 后的能力、对手反应

表 B1 八类跨专业战略压力测试

生成式人工智能使用声明

本稿在研究和编写过程中使用生成式人工智能辅助资料与文献线索整理、结构设计、部分文字起草与润色、交叉核查、引用线索核验及版式检查。AI 输出不作为经验数据，不作为权威来源，也不作为独立证据。所有关键事实、文献、正式文件与时点性信息均要求回到可核验来源；研究问题、理论命题、A/B/C/F 身份判断、材料取舍、规范立场、引用采用与最终定稿均由作者审定，作者承担最终学术责任。生成式人工智能不列为作者。正式投稿时，将按目标期刊要求进一步披露所用工具、模型版本、使用日期与具体范围。

参考文献

- [1] Baldwin, David A. "The Concept of Security." *Review of International Studies* 23, no. 1 (1997): 5–26. <https://doi.org/10.1017/S0260210597000053>.
- [2] Herz, John H. "Idealist Internationalism and the Security Dilemma." *World Politics* 2, no. 2 (1950): 157–180. <https://doi.org/10.2307/2009187>.
- [3] Jervis, Robert. "Cooperation under the Security Dilemma." *World Politics* 30, no. 2 (1978): 167–214. <https://doi.org/10.2307/2009958>.
- [4] Snyder, Glenn H. "The Security Dilemma in Alliance Politics." *World Politics* 36, no. 4 (1984): 461–495. <https://doi.org/10.2307/2010183>.
- [5] Fearon, James D. "Rationalist Explanations for War." *International Organization* 49, no. 3 (1995): 379–414. <https://doi.org/10.1017/S0020818300033324>.
- [6] Huth, Paul, and Bruce Russett. "What Makes Deterrence Work? Cases from 1900 to 1980." *World Politics* 36, no. 4 (1984): 496–526. <https://doi.org/10.2307/2010184>.
- [7] Schelling, Thomas C. *Arms and Influence*. New Haven: Yale University Press, 1966.
- [8] Jervis, Robert. *Perception and Misperception in International Politics*. Princeton: Princeton University Press, 1976.
- [9] Allison, Graham T., and Philip Zelikow. *Essence of Decision: Explaining the Cuban Missile Crisis*. 2nd ed. New York: Longman, 1999.
- [10] Posen, Barry R. *Inadvertent Escalation: Conventional War and Nuclear Risks*. Ithaca: Cornell University Press, 1991.
- [11] Sagan, Scott D. *The Limits of Safety: Organizations, Accidents, and Nuclear Weapons*. Princeton: Princeton University Press, 1993.
- [12] Perrow, Charles. *Normal Accidents: Living with High-Risk Technologies*. New York: Basic Books, 1984.
- [13] Sechser, Todd S., Neil Narang, and Caitlin Talmadge. "Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War." *Journal of Strategic Studies* 42, no. 6 (2019): 727–735. <https://doi.org/10.1080/01402390.2019.1626725>.
- [14] Talmadge, Caitlin. "Emerging Technology and Intra-War Escalation Risks: Evidence from the Cold War, Implications for Today." *Journal of Strategic Studies* 42, no. 6 (2019): 864–887. <https://doi.org/10.1080/01402390.2019.1631811>.
- [15] Garfinkel, Ben, and Allan Dafoe. "How Does the Offense-Defense Balance Scale?" *Journal of Strategic Studies* 42, no. 6 (2019): 736–763. <https://doi.org/10.1080/01402390.2019.1631810>.
- [16] Horowitz, Michael C. "Artificial Intelligence, International Competition, and the Balance of Power." *Texas National Security Review* 1, no. 3 (2018): 36–57. <https://doi.org/10.15781/T2639KP49>.
- [17] Horowitz, Michael C. "When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability." *Journal of Strategic Studies* 42, no. 6 (2019): 764–788. <https://doi.org/10.1080/01402390.2019.1621174>.
- [18] Altmann, Jürgen, and Frank Sauer. "Autonomous Weapon Systems and Strategic Stability." *Survival* 59, no. 5 (2017): 117–142. <https://doi.org/10.1080/00396338.2017.1375263>.
- [19] Goldfarb, Avi, and Jon R. Lindsay. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46, no. 3 (Winter 2021/22): 7–50. https://doi.org/10.1162/isec_a_00425.
- [20] Nadibaidze, Anna, and Nicolò Miotto. "The Impact of AI on Strategic Stability is What States Make of It: Comparing US and Russian Discourses." *Journal for Peace and Nuclear Disarmament* 6, no. 1 (2023): 47–67. <https://doi.org/10.1080/25751654.2023.2205552>.
- [21] Erskine, Toni, and Steven E. Miller. "AI and the Decision to Go to War: Future Risks and Opportunities." *Australian Journal of International Affairs* 78, no. 2 (2024): 135–147. <https://doi.org/10.1080/10357718.2024.2349598>.
- [22] Boulanin, Vincent, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson. *Artificial Intelligence, Strategic Stability and Nuclear Risk*. Stockholm: SIPRI, June 2020. <https://www.sipri.org/publications/2020/policy-reports/artificial-intelligence-strategic-stability-and-nuclear-risk>.
- [23] Chernavskikh, Vladislav. *Nuclear Weapons and Artificial Intelligence: Technological Promises and Practical Realities*. Stockholm: SIPRI, September 2024. <https://doi.org/10.55163/VBQX6088>.
- [24] Saalman, Lora. *AI in Chinese, Indian and US Nuclear Postures, Norms and Systems*. Stockholm: SIPRI, February 2026. <https://doi.org/10.55163/QZBL7397>.
- [25] Grand-Clément, Sarah. *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain*. Geneva: UNIDIR, 2023. <https://unidir.org/publication/artificial-intelligence-beyond-weapons-application-and-impact-of-ai-in-the-military-domain/>.
- [26] Unal, Beyza, and Ulysse Richard. *Governance of Artificial Intelligence in the Military Domain*. UNODA Occasional Papers No. 42. New York: United Nations, June 2024. ISBN 978-92-1-003174-5.

- [27] United Nations General Assembly. Resolution 79/239, "Artificial intelligence in the military domain and its implications for international peace and security." Adopted 24 December 2024. A/RES/79/239. <https://docs.un.org/A/RES/79/239>.
- [28] United Nations Secretary-General. Artificial intelligence in the military domain and its implications for international peace and security: Report of the Secretary-General. A/80/78, 5 June 2025. <https://digitallibrary.un.org/record/4086346>.
- [29] United Nations General Assembly. Resolution 80/58, "Artificial intelligence in the military domain and its implications for international peace and security." Adopted 1 December 2025. A/RES/80/58. <https://docs.un.org/A/RES/80/58>.
- [30] United Nations Office for Disarmament Affairs. "Informal exchanges on artificial intelligence in the military domain." Geneva, 15–17 June 2026, convened pursuant to A/RES/80/58. <https://meetings.unoda.org/unoda-stu-meeting/unoda-science-technology-and-international-security-unit-meeting-2026>.
- [31] REAIM 2026. A Coruña Declaration: "REAIM 2026 Pathways to Action." 5 February 2026. <https://www.reaim2026.es/en/REAIM2026/Documents/REAIM%202026%20Pathways%20to%20Action.pdf>.
- [32] Afina, Yasmin. The Global Prism of Military AI Governance: Reflections from the 2025 Regional Consultations on Responsible AI in the Military Domain. Geneva: UNIDIR, 2 February 2026. <https://unidir.org/publication/the-global-prism-of-military-ai-governance-reflections-from-the-2025-regional-consultations-on-responsible-ai-in-the-military-domain/>.
- [33] United Nations Office for Disarmament Affairs. Convention on Certain Conventional Weapons—Group of Governmental Experts on Lethal Autonomous Weapons Systems (2026), first session 2–6 March 2026 and second session scheduled 31 August–4 September 2026. <https://meetings.unoda.org/ccw/-convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2026>.
- [34] International Committee of the Red Cross. Autonomous Weapon Systems and International Humanitarian Law: Selected Issues. Geneva: ICRC, 5 January 2026. <https://www.icrc.org/en/publication/autonomous-weapon-systems-and-international-humanitarian-law-selected-issues>.
- [35] United Nations Secretary-General and International Committee of the Red Cross President. "Renewed call from UN Secretary-General and ICRC President to adopt rules on autonomous weapon systems." Joint appeal, 25 August 2026. <https://www.icrc.org/en/statement/renewed-call-un-secretary-general-and-icrc-president-adopt-rules-autonomous-weapons>.
- [36] NATO. "Summary of NATO's revised Artificial Intelligence (AI) strategy." 10 July 2024. <https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy>.
- [37] U.S. Department of Defense. DoD Directive 3000.09, Autonomy in Weapon Systems. 25 January 2023. <https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DODDIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF>.
- [38] U.S. Department of State. Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy. Latest version, 1 November 2023. Archived official text: <https://2021-2025.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy>.
- [39] NATO. The Secretary General's Annual Report 2025. Brussels: NATO, 2026. <https://www.nato.int/content/dam/nato/webready/documents/publications-and-reports/annual-reports/sgar25-en.pdf>.
- [40] Gartzke, Erik. "The Myth of Cyberwar: Bringing War in Cyberspace Back Down to Earth." *International Security* 38, no. 2 (2013): 41–73. https://doi.org/10.1162/ISEC_a_00136.
- [41] Lindsay, Jon R. "Stuxnet and the Limits of Cyber Warfare." *Security Studies* 22, no. 3 (2013): 365–404. <https://doi.org/10.1080/09636412.2013.816122>.
- [42] United Nations General Assembly. Resolution 80/23, "Possible risks of the integration of artificial intelligence into command, control and communications systems of nuclear weapons." Adopted 1 December 2025. A/RES/80/23. <https://docs.un.org/A/RES/80/23>.
- [43] Horowitz, Michael C., and Erik Lin-Greenberg. "Algorithms and Influence: Artificial Intelligence and Crisis Decision-Making." *International Studies Quarterly* 66, no. 4 (2022): sqac069. <https://doi.org/10.1093/isq/sqac069>.
- [44] Zeng, Jinghan. "The US Factor in Chinese Perceptions of Militarized Artificial Intelligence." *International Affairs* 101, no. 2 (2025): 677–689. <https://doi.org/10.1093/ia/iaae323>.
- [45] Shandler, Ryan, Michael L. Gross, and Yahli Shereshevsky. "Black Box Warfare: Human Judgment and Military Decision-Making in the Age of AI." *Journal of Conflict Resolution*, OnlineFirst (2026). <https://doi.org/10.1177/00220027261463443>.
- [46] Lushenko, Paul. "AI, Trust, and the War Room: Evidence from a Conjoint Experiment in the US Military." *Cambridge Forum on AI: Law and Governance* 1 (2025): e52. <https://doi.org/10.1017/cfl.2025.10019>.
- [47] Schwartz, Joshua A., and Michael C. Horowitz. "Delegating Destruction: Coercive Threats and Automated Nuclear Systems." *International Organization* 80, no. 1 (2026): 179–204. <https://doi.org/10.1017/S0020818325101215>.
- [48] Johnson, James. "Revisiting the 'Stability–Instability Paradox' in AI-Enabled Warfare: A Modern-Day Promethean Tragedy under the Nuclear Shadow?" *Review of International Studies*, First View (2024): 1–19. <https://doi.org/10.1017/S0260210524000767>.
- [49] Zatsepina, Luba. "Waltzing into Uncertainty: AI in Nuclear Decision Making and the Challenge of Divergent Deterrence Logics." *Cambridge Forum on AI: Law and Governance* 1 (2025): e56. <https://doi.org/10.1017/cfl.2025.10021>.
- [50] Johnson, James. "GenAI and Synthetic Foresight at the Brink: The Future of Nuclear Crisis Decision-Making." *European Journal of International Security*, First View (2026): 1–22. <https://doi.org/10.1017/eis.2026.10063>.
- [51] Yuan, Shaoyu. "World and War in the Age of AI: Cyber Risk, Deterrence, and Decision-Making in U.S.-China Competition." *Defence Studies*. Published online 15 August 2026. <https://doi.org/10.1080/14702436.2026.2718345>.

版本说明

版本说明：V1.2.1 为最小引用补丁。在 V1.2 基础上新增 Yuan（2026）参考文献及一段原创性边界说明，明确本文不主张互动式时间压缩已被经验观察，也不主张首先提出行动—反应压力、采用激励、威慑与危机稳定等宽问题；本文的特定贡献继续限定为 P3a—P3d 可分别败北的命题拆分、 (T_V, T_C, T_P) 共同时间预算与相应失败条件。V1.2 的正文主轴、专业对象、A/B/C/F 边界、共同存续 F 类定位及跨学科交接关系均保持不变。本次升版不改变理论主体，不提升证据身份，不构成英文外发稿或投稿稿。V1.2.1 为当前 Canonical 专业研究基线。