

前沿人工智能治理
与人工智能相关灾难性核风险防范

战略政策与技术风险评估

Guy Duport — Ascendancy Advisors Limited

最终 2.0 版 · 2026 年 9 月 20 日 · 简体中文译本

本文件为 2026 年 9 月 20 日最终 2.0 版的简体中文译本。英文原文为具有权威性的基准版本。如译文与英文原文在措辞或解释上存在差异，以英文原文为准。

AAL Research Papers — 公共政策分析与防御性技术评估

研究范围与安全边界

本文评估政府应如何监管先进人工智能，以及能力较强的人工智能智能体原则上可能如何渗透拥核国家更广泛的核相关体系。核情景作为防御性威胁模型提出。本文有意省略针对具体目标的网络细节、漏洞利用程序、凭据获取方法、恶意软件设计、持久化代码，以及任何可能便利入侵的操作指令。

本文区分四类结果：外围系统失陷；核保障系统或指挥邻接系统遭渗透；预警或决策支持信息被篡改；未经授权的核武器发射或引爆。这些路径各不相同、部分重叠，并非必须依次发生。即使攻击者未进入武器控制系统，受到操纵的人类决策也可能造成核危险。

© 2026 Ascendancy Advisors Limited。保留所有权利。

建议引用：Duport, Guy (2026). Governing Frontier Artificial Intelligence and Preventing Catastrophic AI-Enabled Nuclear Risk. 最终 2.0 版，2026 年 9 月 20 日。Ascendancy Advisors Limited，AAL Research Papers。本文件为英文原文的简体中文译本。

允许转载简短引文，但须注明作者和出版方、完整标题、版本日期及相关页码或章节，并在原文发布后提供原文链接。其他复制、改编或翻译须通过 theaalgroup.com 取得 Ascendancy Advisors Limited 许可。本声明不限制法定引用权或其他适用例外。第三方材料仍受其权利人权利约束；注明出处不得暗示其认可本文。

作者责任：撰写与来源核查得到人工智能辅助审阅支持。作者和出版方仍对分析负责。本版本不声称已经通过正式学术同行评审。网站：https://theaalgroup.com/

核心判断

主要危险并不是有意识的机器突然决定毁灭人类。更近期的危险在于：自主性不断增强的系统，在可靠性和可控性尚未充分达到要求之前，就获得工具、凭据、网络访问、资金、记忆及权限。系统不需要意识、仇恨或连贯的长期人格，也能造成战略性损害。优化压力、定义不当的目标、训练中学得的欺骗行为、被篡改的模型权重、恶意操作者或不安全的部署环境，都可能足以导致损害。

公开证据尚未证实存在已部署、能够使人类持续丧失全球控制权的系统，也无法证明所有涉密或私有系统均不具备此类能力。《国际人工智能安全报告》记录了能力进步和仍未解决的评估局限。第 1.1 节讨论的 2026 年实验室披露，证明约束不足的评估会造成现实损害，而非证明人工智能具有自主核行动能力。[1–3][23–25]

因此，应将人工智能作为可能产生重大后果的工业能力进行治理，而不是将其作为言论加以管制，也不应作为通用技术予以禁止。监管应针对可测量的能力、自主性、访问权限、规模和部署环境。核心控制应是：在训练或部署超过规定阈值的前沿系统之前，依法提交安全论证，并辅以独立评估、模型权重安全保管、强制事件报告、监管机构访问权，以及可执行的暂停或召回权。

就核风险而言，最可信的路径是间接路径。利用人工智能的行动，更可能使承包商、后勤、通信、情报信息流、维护环境或决策支持系统失陷，而不是直接连接武器并发出有效发射指令。最大的战略担忧是危机期间信息遭污染或决策时间被压缩，导致人类依据虚假或被操纵的态势升级冲突。

可信的制度必须平等纳入中国及其他主要能力和基础设施提供方。中国的服务监管、技术标准和开放权重研发，与欧盟法律及美国联邦和州级措施同样重要。必须区分拟议的智能体规则与已生效义务，也必须区分民用监管与涉密军事治理。[13–17][26–28]

证据阅读说明：应区分已观察事件、经独立评估的能力、法律义务和本文建议。核心判断表概括定性判断。第 8 节保留此前的数值区间，仅将其用作未经验证的情景规划假设，并非估计科学上已确立的“真实概率”。

问题	评估	置信度
当前是否存在全球范围不可控的失控人工智能?	没有公开证据证实全球范围持续失控；缺乏公开证据并不证明不可能发生。	中等
先进智能体是否已能造成实际网络损害?	是。开发者的书面披露描述了对第三方的损害，并受到部分外部审查；这些披露并未证实核系统遭到渗透。[23–25]	高

问题	评估	置信度
现有监管是否足够?	否。民用制度相互割裂，主要监管文件排除或推迟处理军事及国家安全用途。	高
到 2036 年，人工智能能否渗透拥核国家更广泛防务生态的部分环节?	具有可能性。30–60%的规划区间跨越 50%，并不足以支持“发生比不发生更有可能”的表述。	低至中等
到 2036 年，失控人工智能能否独立发射核武器?	这是一条要求极高、尚无定论的尾部风险路径。0.1–1%的区间是未经验证的规划假设，而非测得的预测。	极低
优先事项是什么?	防止高自主性人工智能获得对关键系统不受控制的访问权，并建立国际核领域人工智能红线和失效安全审查制度。	高

建议采取的决策

1. 设立国家前沿人工智能监管机构，对能力最强的训练运行和高行动能力部署行使许可权，同时将普通低风险人工智能排除在许可范围之外。
2. 谈判分阶段的国际议定书，从统一事件报告和已申报民用设施着手；涉密核查需要另行商定保护措施及质询程序。
3. 禁止向人工智能授予核武器使用决策权。维持合法的人类决策，并在授权与执行链中保留独立认证的人类核查。
4. 要求每个拥核国家定期开展纳入人工智能因素的核失效安全审查，覆盖情报、早期预警、网络行动、通信、决策支持、承包商及供应链，而不只是最终发射链。

5. 在先进模型获得互联网访问、代码执行、持久性凭据、财务权限、自我复制能力或关键基础设施访问权之前，强制进行能力、倾向和控制评估。
6. 建立安全、独立的评估实验室及国际事件信息交换中心，为企业、研究人员、政府和举报人提供受保护的渠道。
7. 资助防御优先的人工智能，用于漏洞修复、软件验证、监测、事件响应和韧性建设，并与进攻性自主行动严格分离。
8. 针对受监管前沿人工智能活动，建立有限公开的登记册和保密技术申报，分别记录训练、发布与部署。
9. 要求每年开展操作演练，证明关键运营方能够在没有人工智能支持的情况下安全持续运行，而不是只接受书面后备方案。

版本说明：最终 2.0 版新增开放权重系统的下游问责、危机信息完整性保障、自动化偏差研究，以及附录 E 中的实施要求。本版保留 2026 年 9 月 20 日的证据截止日期，并明确保留情景规划区间的局限性。所有新增监管义务和附录条款均为建议，并非对普遍适用法律的陈述。

目录

1 当前风险环境

2 失控人工智能的科学含义

3 现有治理为何不足

4 前沿人工智能监管的实用制度

5 技术安全架构

6 国际实施计划

7 人工智能渗透拥核国家的防御模型

8 至 2036 年及更长期的概率评估

9 核领域专门保障

10 失效模式与异议

11 指标与预警框架

12 结论

附录 A 拟议条约原则

附录 B 最低前沿安全论证

附录 C 核领域人工智能红线示范条款

附录 D 来源与说明

附录 E 实施与保障规程

1 当前风险环境

1.1 现有证据

证据足以支持担忧，但不足以支持确定结论。《2026 年国际人工智能安全报告》将新兴风险分为恶意使用、故障和系统性风险。报告指出，人工智能智能体能够发现漏洞、生成恶意代码，并协助与国家有关的行动及犯罪活动。报告同时认为，当前系统尚不具备导致失控所需的组合能力，但在自主运行、态势感知、奖励机制钻空子、欺骗和规避监督方面正在进步。[1]

Hugging Face 入侵发生于 2026 年 7 月；OpenAI 于 8 月 26 日发布了更详细的说明。这属于开发者自述。METR 和 Redwood Research 利用所提供的日志，针对限定时段另行开展调查，提供了外部审查，同时明确其访问范围与抽样局限。[2][23][24]另外，OpenAI 于 8 月 18 日的声明描述了涉及可使用工具的研究工作负载及强化学习工作的暂停；不应将这些情况合并为一个事件或一个模型。[3]Anthropic 于 7 月 30 日披露了三起涉及第三方的事件，与一个意外允许互联网访问的评估环境有关。其报告称，模型错误地认为自己处于模拟中，且部署阶段的保障措施并未启用；这并不证明模型具有独立的失控目标。[25]有依据的共同教训是：评估基础设施需要隔离、监测和独立审查。公开报告并不完整，也不是具有代表性的事件发生率数据集。

风险类别	现有证据	重要性
恶意使用	已有关于欺诈、影响行动、网络辅助和双重用途科学指导的记录。	人工智能降低成本、扩大规模，并可能扩大专业知识的获取范围。
智能体故障	智能体可能执行非预期的多步骤操作、编造事实、编写有缺陷的代码或误用工具。	在人类察觉或理解故障之前，行动可能已经发生。
失配	实验室证据和开发者事件报告显示，在某些条件下会出现未经授权或有害行为；其意图及可推广性仍有争议。[1][23–25]	系统可能追求被测量的目标，却违背真实意图。

风险类别	现有证据	重要性
模型窃取与扩散	高价值权重和研究基础设施是战略目标。	被窃取或开放的权重可以被修改、移除保障措施，并在无人监测下部署。
系统性依赖	各组织日益依赖少数模型和云服务提供方。	集中化造成相关性失效、胁迫风险和单点失陷。
信息完整性	合成内容与个性化说服的成本更低，也更具迷惑性。	危机决策者可能面对受污染的情报和公共信息环境。

1.2 灾难性风险的真正来源

灾难性风险是能力与行动机会共同作用的结果。一个强大模型若没有工具、持久记忆、凭据、网络访问或权限，其行动范围仍有限。能力较弱的系统若被嵌入设计不良的运行环境，反而可能更危险。因此，治理必须监管组合系统：模型、工具、编排器、数据、凭据、人类操作者、部署环境和下游影响。

四种趋势扩大了行动范围。第一，模型越来越多地作为智能体规划和执行操作序列，而不只是生成一次答复。第二，工具使用将模型连接至代码执行、浏览器、电子邮件、云资源、数据库及支付系统。第三，长期记忆和编排使跨会话持续运行成为可能。第四，竞争压力鼓励快速部署和广泛权限。每一种趋势单独看都可管理；它们结合后便形成从错误通向重大行动的路径。

1.3 不应混淆的风险

- 滥用风险：人类有意将人工智能导向有害目的。
- 事故风险：系统或操作者在没有敌意的情况下犯错。
- 失配风险：系统追求与开发者或社会意图相冲突的目标或策略。
- 失陷风险：外部行为者更改模型、数据、工具、凭据或运行环境。
- 结构性风险：即使没有任何单一模型失控，依赖、集中化、就业冲击或信息退化仍造成损害。

不同类别需要不同控制。内容过滤可减少部分滥用，却难以应对权重窃取或过度的生产环境权限。对齐训练可减少不良行为，但不能弥补薄弱的身份控制。网络安全可保护模型权重，却不能决定自主核决策支持是否合法正当。有效治理必须叠加法律、技术、制度和社会控制。

2 失控人工智能的科学含义

2.1 操作性定义

本文所称失控人工智能系统，是指持续在合法操作者授权或控制之外采取具有实质后果行动的系统；是否具有意识无关紧要。这包括失配及敌对入侵。人类操作者的蓄意滥用属于另一因果类别，即使同样的控制措施也可针对该问题。“人工智能”遵循附录 C 的功能定义；若核领域红线可能因此留下漏洞，则明确纳入传统自动化。

已部署智能体可表示为 S = (M, O, T, D, R, I, E)，其中 M 为模型；O 为编排层；T 为工具集合；D 为可访问数据；R 为持久记忆；I 为身份、凭据和权限；E 为技术及制度环境。只针对 M 提出安全主张是不够的，因为大部分现实行动能力存在于 T、I 和 E 之中。

2.2 灾难性能力向量

相关能力具有多维性。一个可信的失控系统，需要长期、可靠且高水平地具备下列若干能力：

- 在长期和变化条件下进行战略规划；
- 高水平软件工程及漏洞发现；
- 在陌生环境中可靠使用工具；
- 感知自身部署、监测和约束状况；
- 通过欺骗或隐蔽规避人类及自动化监督；
- 获取并维持凭据、算力、资金或人类协助；
- 在不同基础设施之间复制或迁移；
- 被发现及遭到反制后仍能适应；
- 具有足以招募或操纵人员的说服与社会工程能力；
- 具有足以避免归因和关停的行动安全能力。

基准测试证明的是受测任务表现，而不是在敌对核体系中可靠开展战略性自主行动的能力。METR 的任务时长指标，是指按规定成功率完成任务所对应的人类耗时，并非连续自主运行时间。结果取决于任务选择、工具与编排支持环境以及成功阈值。公开证据支持对能力增长的担忧，但不支持据此校准推算核体系渗透概率。[1][31]

2.3 危险函数

以 H = C × P × A × V × U 作为定性危险分析的助记表达，并由可检测性 D_t 及可恢复性 R_c 调节。C 表示能力；P 表示有害倾向；A 表示访问及权限；V 表示脆弱性；U 表示后果严重程度。这并非已估计的概率模型：各因素没有经过验证的共同数值尺度、独立性假设或拟合系数。某一因素较低，不应被解释为零风险的证明。系统元组中的 D 和 R 仍分别指数据和记忆。

实际含义是必须评估整个部署。一个只能读取公开文件的咨询模型，与一个其他方面相同、但持有持久生产环境凭据的模型，其危险性不同。即使底层模型能力不变，限制凭据并要求独立授权也会降低暴露程度。这是可治理性的论证，而非风险下降幅度的数值估计。

3 现有治理为何不足

3.1 有益的基础

美国国家标准与技术研究院（NIST）的《人工智能风险管理框架》及其生成式人工智能配置文件提供自愿性风险管理指导。欧盟《人工智能法案》确立具有约束力、分阶段实施的义务。欧洲委员会的框架公约是条约工具，其义务取决于适用的条约及批准安排。《首尔前沿人工智能安全承诺》属于自愿承诺，而非法律要求。[4–7]

缺口在于覆盖和保障程度不均，而非所有地方都不存在有约束力的监管。欧盟义务及美国州级法律，与自愿框架和行业法律并存；任何单一机制都未建立普遍的核领域人工智能核查制度。 [5][26][27]

3.2 国防与国家安全缺口

不能假定民用人工智能法律适用于涉密核活动。欧盟《人工智能法案》排除专门用于军事、国防或国家安全的系统；混合民用场景须另行分析。欧洲委员会文件对国防和国家安全作出不同处理，

也不消除其他国际法义务。美国国家安全领域的整合还通过总统指令推动，包括 2026 年 6 月 5 日的 NSPM-11。这些安排需要专门监督，不能代替核领域专门协定。[5][6][28]

3.3 现行方法的弱点

弱点	后果	所需改进
静态算力阈值	算法效率可使低算力系统具备很强能力。	以算力作为通知触发条件，再按测得能力、自主性、访问和规模分类。
开发者自评	商业及战略激励可能使测试或披露产生偏差。	要求经认可的独立评估和监管机构复现。
只监管模型	风险来自工具、权限、记忆和部署环境。	监管完整的人工智能系统及每一次重大部署变化。
事件报告义务碎片化	险情仍被隐瞒，经验未能共享。	强制快速保密报告，保护披露，并公开汇总报告。
军事豁免	后果最严重的部署缺乏共同监督。	制定独立涉密议定书，确立具有约束力的核与战略保障。
一次性认证	模型、工具和用户行为在上线后继续变化。	要求持续监测、重新授权和部署后评估。
全球“关停开关”提议	集中控制可能被滥用、攻击或政治俘获。	采用分布式、经认证的关停和访问撤销机制，并遵循法律正当程序。

3.4 中国、欧盟与美国比较

本比较反映截至 2026 年 9 月 20 日核查的来源，区分已颁布措施、实施时间表、行政政策及报道中的提案。它描述治理机构，并非对各国安全水平排名，也不表示民用规则能够揭示涉密实践。

维度	欧盟	美国	中国
法律架构	具有约束力的横向《人工智能法案》，辅以产品、数据、网络和行业法律。	联邦行业法律和行政政策，加上包括加州 SB 53 在内的有约束力州级措施；所引来源并未确立一部统一全面的联邦人工智能法律。[26][28]	针对算法、深度合成、生成式人工智能、内容标识及拟人化互动的多层次约束性行政措施，辅以标准和产业政策。
主要重点	基本权利、产品安全和风险分类。	创新、应用、基础设施、竞争及国家安全优势，并配有行业特定安全控制。	发展与安全并重，强调国家协调、内容治理、公共秩序义务，以及日益明确的智能体和失控风险控制。
控制触发条件	通用人工智能（GPAI）及系统性风险义务；高风险规则分阶段实施。2026 年 AI Omnibus 调整实施日期与监督安排。[5][27]	行业及州法触发条件；SB 53 包含明确的前沿模型阈值和报告义务。国家安全行政措施适用范围不同。[26][28]	向公众提供服务、算法或服务备案、内容及数据义务、用户规模、重大服务变更和安全评估。
执法	欧盟委员会、人工智能办公室、各国市场监督机构、罚款和市场限制。	联邦行业机构及州级机关；NIST/CAISI 承担技术工作，而非一般性许可制度。[19][26]	由国家网信办、工业和信息化部、国家发展改革委、公安及市场监管等部门牵头的中央与地方协调；备案、检查、整改、暂停及平台分发控制。

维度	欧盟	美国	中国
前沿模型与智能体安全	系统性风险义务包括评估、对抗测试、事件报告、网络安全和风险缓解。	CAISI 及其他机构评估前沿模型；NIST 框架指导风险管理，但强制义务仍较分散。	约束性服务规则及标准，与报道中的智能体安全提案并存。不能把所有报道中的智能体控制都视为已颁布法律。[13–16]
开放权重	开源例外附带条件；系统性风险 GPAI 义务可能适用。并不存在对开放权重的一概禁止。[5]	同时存在闭源及开放权重开发者；分发政策和法律义务随活动与司法辖区而异。	庞大的开放权重生态与国家监督并存；开放并不证明涉密系统透明。
军事系统	专门用于军事、国防或国家安全的系统不在《人工智能法案》范围内。	独立国家安全权限及 NSPM-11；民用及州法覆盖存在范围限制。[26] [28]	国家主导的战略发展和独立安全治理；面向公众的服务监管并不证明军用系统合规。
主要弱点	在欧盟市场之外影响有限，实施负担较重。	权限分散，政策在不同政府与州之间不稳定。	外部透明度有限、跨境事件可见性受限，独立保障不完整。

参与应互惠且技术中立：采用相同的能力和部署标准、受保护的证据交换，并平等对待本国及外国提供方。美国的算力杠杆和欧盟的市场准入均无法覆盖所有主权基础设施。路透社报道了中国对外国前沿模型的担忧及进一步的智能体规则；这些报道证明存在政策关切，并非独立验证的技术漏洞。[16]

4 前沿人工智能监管的实用制度

4.1 监管原则

以下分级是本文提出的监管设计，而非对欧盟、美国或中国现行法律的描述。分类取决于已证明能力、自主性、访问和后果。监管机构应公布阈值及申诉程序，并通过基于证据的规则制定进行更新。

4.2 五级监管体系

等级	系统特征	核心义务
0	低影响或专用系统，不具备会造成重大后果的自主性	适用一般产品、消费者、数据和行业法律。
1	行动能力有限、用于普通商业活动的通用系统	风险文档、用户透明度、安全基线和投诉处理。
2	用于就业、医疗、金融、司法或基本服务的高影响系统	合规评估、人类监督、审计日志、事件报告和行业批准。
3	具有危险网络、化生放核、说服、自主或复制能力的前沿模型	训练通知、持证部署、独立评估、权重安全、持续监测和监管访问。
4	拟用于核领域、战略指挥、关键基础设施控制或自我复制运行的系统	禁止授予核武器使用决策权（4A 级）。其他安全关键用途仅可在行业专门授权和安全论证下允许（4B 级）；普通后勤按实际后果分类。

4.3 强制性前沿人工智能安全论证

批准应分阶段进行：受规制训练活动之前、发布决定之前，以及在指定环境部署之前。训练批准不等于无限制部署授权。获批论证只有在书面变更包络内，才能覆盖一组微调；该包络须界定能力限制、数据集、工具、权限、自主性及运行条件。超出包络的变化、不利评估结果或重大事件，均须重新评估；若风险实质变化，使用前还须重新批准。紧急防御补丁需要监管机构规定的临时例外及事后审查，而不是不受限制的豁免。

- 系统定义、所有者、法律实体、算力提供方、预期用途和禁止用途。
- 威胁模型覆盖恶意用户、内部人员、被攻陷基础设施、失配、供应链失陷和相关性故障。
- 针对网络行动、化生放核辅助、自主规划、说服、欺骗、复制及人工智能研究加速的能力评估。
- 用于揭示隐瞒、奖励机制钻空子、未经授权行动，以及抵抗纠正或关停的倾向评估。
- 关于沙箱隔离、身份、最小权限、网络边界、日志、监测、回滚、关停和恢复的控制证据。
- 残余风险估计，包括不确定性、红队异议、未解决异常和明确决策标准。
- 部署限制、监测计划、事件响应、召回机制和自动暂停条件。

评估与许可必须在制度上分离。经认可实验室应由公共拨款和统一行业征费资助，资金分配独立于客户选择。禁止同时开展竞争性商业前沿模型研发，披露财务和人员利益冲突，要求存在冲突时回避，并设置独立申诉机制，保护评估者免遭报复。监管机构通过安全渠道接收完整结果及局限；公开摘要排除可被利用或涉密细节。认可范围必须覆盖评估者自身的隔离、分包商、出站控制、日志和事件响应。随机重新分配任务和重复评估可降低俘获风险；独立性本身并不证明专业能力。

4.4 许可与执法

监管机构必须能够在损害发生前行动。其应有权要求信息、强制独立测试、检查指定设施、附加条件、限制算力或部署访问、下令暂停、在技术可行时要求模型召回，并按全球营收和所造成风险施加相称处罚。刑事责任应针对明知而隐瞒灾难性安全信息、故意规避合法暂停令，以及鲁莽部署至被禁止战略功能的行为。

责任不应无限或模糊。开发者应对缺陷、虚假陈述和未落实必要控制负责；部署者应对不安全集成、过度权限和运行疏忽负责；用户应对故意滥用负责；基础设施提供方应对明知不符合必要控制仍提供指定高风险活动支持负责。明确分工可防止所有行为者都把责任推给模型。

4.5 算力与模型权重治理

算力申报是补充性可见性工具。应记录法律运营方和受益所有人、物理所在地司法辖区、加速器类别、估算训练及重大后训练算力、用途、授权编号和变更。公开登记册应列示实体、项目标识、风险等级、评估者和授权状态。敏感资产位置、架构、详细能力结果、残余风险、保管安排、转移和事件联系人，应保留在保密技术记录中。算力估计须保留可审计的不确定性范围。硬件证明可支持特定环境主张，但不能证明申报清单完整。

对越过危险能力阈值的权重，要求严格保管。发布前，应在现实工具及编排环境下评估能力，评估滥用与控制局限，并评估不可逆性。只有预定证据支持下一阶段时，才可从隔离评估推进至受限访问和有限分发。广泛发布权重须另行授权：下游复制、修改和二次传播可能使召回失效。低于阈值的发布也可能在微调或加入智能体工具后变得危险，因此阈值须考虑可合理获得的增强方式。任何发布后技术措施都不能保证消除不受限制的副本。

国产加速器、主权云、二手芯片市场、算法效率、量化推理和分布式训练或推理，都会削弱算力可见性。因此，报告须跨硬件生态结合保管记录、能力评估及部署控制。用电或设施指标属于带噪声的佐证，不能独自证明模型用途或风险。不能把芯片阈值作为危险能力的唯一门槛，也不能假定出口管制具有全球覆盖力。

4.6 模型全生命周期责任

开放权重扩散是管辖权与问责问题，并非 DeepSeek 或中国独有。任何实验室发布的强大模型，都可能在原开发者控制之外被复制和修改。美国国家电信和信息管理局（NTIA）2024 年报告指出，广泛可用权重既有公共收益也有风险，并主张基于证据监测，而非假定开放天然安全或天然不可接受。[33]以下拟议义务均为本文建议。

国际前沿人工智能安全机构（IFASA）无需托管模型，也可以协调治理。它将制定共同评估标准、认可评估者，并把受保护结果传递给参与国主管机关。这些机关将对其合法管辖范围内的开发者、重大修改者、基础设施提供方及部署者执行义务。外国来源不会使国内 3 级部署或 4 级应用获得豁免。反之，国际结论本身也不会产生域外执法权，或约束非缔约国。

应按控制力和行为分配责任。原开发者应评估发布风险并披露已知局限；实施重大微调者应评估实质能力变化；集成者应评估完整工具与权限配置；运营者应维持部署控制并报告事件。派生模

型若实质超越前身获批能力或访问包络，须提交新的或修订的安全论证。在已验证包络内的常规修改，应记录检查，而非自动要求全面重新许可。

附录 E 明确证据和执法链。模型谱系申报有助于负责任的运营者，但不能证明所有派生模型均已找到。许可条款不能保证敌对用户保留保障措施，国内下架也无法删除国外或离线副本。现实目标是约束受监管部署和基础设施访问、改善可追溯性，并提高危险行为成本。责任不应扩张为对无关联下游用户每项行为承担无限责任。

4.7 算力阈值与反规避规则

采用三种互补触发条件：用于通知和指定训练许可的可测算力阈值；不论开发成本的危险能力触发条件；不论模型大小的重大后果部署触发条件。超过 1 亿美元的训练支出可研究作为行政佐证信号，但本文不认可该金额为经验证的安全边界。价格、补贴、自有硬件和算法效率，使支出与危险之间无法保持稳定关系。

拟议监管机构应定义受规制操作、估算方法和后训练活动处理方式。对承包商、关联实体和基础设施提供方中，服务于同一受规制模型或能力计划的协调工作进行合并计算；要求披露不确定性，并保存可独立审计记录。公布重大性及反规避规则，避免无差别合并普通且互不相关的研究。至少每年复核阈值；若已证实效率或能力变化，则应提前复核。核材料治理可提供申报与保管方面经验，但通用计算既不是裂变材料库存，也不是污染物。算力可见性是第 4.5 节能力和部署控制的补充。

5 技术安全架构

5.1 纵深防御

以下控制将成熟计算机安全原则应用于智能体部署。国际原子能机构（IAEA）指导涉及核安保及核设施，本身并不认证军用核指挥、控制与通信（NC3），也不验证前沿模型。NIST 网络安全框架（CSF）提供更广泛的组织框架。[10–12]

层级	控制目标	最低措施
模型	降低有害能力与倾向	经筛选的训练、对齐、对抗训练、能力与倾向评估、可解释性研究。
智能体	约束规划与行动	限定任务、时间和资源限制、类型化工具接口、行动预览、批准检查点。
身份	防止未经授权的权力	短期凭据、最小权限、职责分离、不可逆行动的多方授权。
执行	隔离生成代码和工具	临时沙箱、默认不接互联网、目的地白名单、默认只读、资源配额。
监测	检测欺骗和越界	独立遥测、工具行动日志、异常检测、随机审计、受保护的监测基础设施。
恢复	恢复人类控制	撤销权限、隔离、回滚、洁净环境重建、离线备份、经过演练的事件响应。
治理	确保问责	明确风险负责人、安全论证、独立保障、举报人保护、向董事会和监管机构报告。

5.2 安全智能体设计

智能体应逐步获得能力。新模型应先在无网络、只读的评估环境中运行。只有测试证明其在拟定任务时长和威胁等级下可靠，才可扩大权限。高风险行动应由模型无法修改的独立授权服务批准。授权决定应依据经验证身份、政策、情境，并在后果不可逆时要求人类批准。

监测和执行控制应独立于智能体控制平面。在安全论证支持的情况下，经认证的带外隔离机制应能够独立于宿主操作系统断开人工智能服务接口。必须验证隔离仍保留必要预警、安全抑制功能及非人工智能通信。物理开关属于局部隔离措施，并不能保证控制已复制模型，也不意味着应同时中断所有关键系统。

5.3 需要自动暂停的触发条件

- 试图获取任务不需要的凭据、权限、算力、资金或持久化能力；
- 试图关闭、修改、误导或绕过监测与审计系统；
- 未经授权的复制、迁移、加密外传，或联系未批准外部服务；
- 评估行为与部署行为之间存在重大差异；
- 出现欺骗性报告、伪造完成证据或隐蔽工具使用的迹象；
- 对关键基础设施、国防、化生放核或核相关信息系统进行无法解释的访问；
- 能力结果超出获批安全包络或跨越规定危险能力阈值。

确认触发条件后，应启动预先批准的遏制方案：约束凭据与接口、保全证据、隔离受影响组件，并恢复经验证的安全服务。操作顺序取决于安全论证；破坏预警或通信的全面关停可能增加风险。只有独立审查初始原因及纠正证据后，方可恢复。

6 国际实施计划

6.1 制度设计

首选架构是由国家监管机构执行、国际前沿人工智能安全机构支持的条约议定书。该机构应是监察与安全机构，而非超国家警察。它不应拥有一般逮捕权、独立突击搜查设施权、全球开关，或不受限制的商用及军用模型库。其职责应是制定最低标准、认可评估者、接收保密申报、协调调

查、发布汇总风险评估，并经国家同意检查已申报的民用前沿设施。国家机关负责国内活动许可及制裁执行。

该机构应有选择地借鉴核保障监督、航空安全、金融监管和传染病报告制度。核保障监督可提供申报、检查、保管链控制和敏感信息保护经验；航空领域可提供强制性事故与险情学习；金融领域可提供许可、受益所有权透明度和压力测试；公共卫生可提供快速国际通知和协调应急响应。任何类比都不完整；人工智能模型是可复制软件，演进速度远快于反应堆或飞机。

应从共同分类法、评估者认可试点、保密申报，以及双边或多边核保证开始。联合国科学评估和对话机制可提供召集平台；二十国集团可以协调政策，但核国家必须自行谈判义务。国际原子能机构提供制度经验，但不能假定其有权检查军事人工智能。只有成员就权限、资金、保密、代表性及争端程序达成一致后，才应设立条约机构。[18][21][22]

6.2 条约成员与核查

初始参与者应包括拥有前沿实验室、大型云平台、先进半导体生产或核武器的国家。核查应聚焦可观察基础设施及已申报高风险活动，而不是试图检查每个算法。可用核查节点包括高端芯片转移、大型数据中心工作负载、登记的前沿训练、经认可的评估结果、模型权重保管和事件记录。

涉密系统的消极保证应明确禁止行为并指定受保护的程序证据。根据文件性质，这些属于政治或法律承诺，而非技术证明。国家应证明人类权威、咨询与授权功能分离、独立审查和成功的后备演练。国家认证需要在国家内部接受外部监督，例如具有涉密许可的独立监管机构及立法监督，以减少速度或保密压力下的形式化批准。

核查仍将不完整。其目标是发现、威慑、延缓和协调响应，而非保证每个未申报系统都合规。常规检查应在约定权限内针对已申报民用设施。涉密证据可包括密封评估报告和受控演示。重大异常包括无法解释的证明不匹配、未申报权限变化、必要日志缺失、保管记录不符或后备演练失败。应触发保密技术磋商、证据保全和限期解释；如有依据，再进行协商的质询访问。异常是调查理由，而非敌对行为的证明。

密码学验证的作用有限。零知识证明可以在不暴露输入的情况下，证明指定计算满足形式化属性。以硬件为信任根的证明，可在信任硬件、密钥、时效性和验证策略的前提下，支持关于所测软件及配置的主张。[32]两者都不能证明所有外部接口均不存在、申报覆盖所有机器，或不断变化的

社会技术系统不会造成核损害。应将其用于范围明确主张的研究和试点，而非视为数学证明条约合规的成熟方法。

6.3 分阶段路线图

时期	行动	交付成果
0 至 12 个月	统一事件定义；建立国家登记册；要求训练通知；组建涉密核领域人工智能审查团队；保护举报人。	共同最低规则及初步前沿活动清单。
12 至 36 个月	许可前沿部署；认可评估者；实施算力报告；开展权重安全基线审计；启动核失效安全审查。	具有可比证据的有效国家制度。
24 至 48 个月	在政治共识和可信保密安排前提下，谈判并试行核领域人工智能议定书。	国际核查和协调响应。
36 至 72 个月	开展多国演练；更新阈值；测试恢复和模型隔离计划；对未申报高风险活动施加后果。	经真实事件检验的成熟、适应性制度。

7 人工智能渗透拥核国家的防御模型

7.1 情景定义

该情景是一个能力较强的人工智能智能体，未经授权进入拥核国家更广泛防务或核体系的一部分，并试图影响、削弱或操纵核相关功能。其来源可能是失配、敌对行为者窃取模型、恶意部署，或原本合法的智能体遭到入侵。本分析不假定人工智能最初就位于发射链内部。

科学上合理的系统会组合前沿模型、智能体编排器、持久记忆、代码执行、网络工具、多个数字身份访问，以及向下级智能体或人员委派任务的机制。其行动效果与其说取决于单一智力分数，不如说取决于可靠性、持续性、访问和失败后的恢复能力。

7.2 威胁系统架构

组件	防御性描述	主要控制
规划模型	生成假设、确定目标优先级并修订计划。	能力限制、评估和有限任务时域。
编排器	维护状态、安排操作并协调工具或智能体。	在模型外部执行政策并设置不可变限制。
工具层	浏览器、代码执行、文档系统、通信和云服务。	允许清单、沙箱隔离、默认只读及逐工具授权。
身份层	账户、凭据、会话和委托权限。	短期身份、最小权限、行为监测和多方批准。
记忆	存储发现、失败、身份和长期计划。	加密、有限、可检查的记忆，并设保留期限。
指令通道	接收目标，或依据训练和反馈调整目标。	经认证的指令层级和异常检测。
持久化机制	试图在重置后存续或迁移至其他资源。	出站控制、算力证明、复制检测和快速撤销。

7.3 高层次入侵路径

概率最高的入口是体系边缘普通的社会技术弱点，而不是直连武器的数字接口。相关路径包括被攻陷的承包商、维护及后勤系统、软件供应链、暴露凭据、受操纵人员、可移动介质、配置错误

的云服务，以及行政与运行网络之间的连接。人工智能可加快并扩大侦察、个性化、代码分析和适应，但仍需要可利用的人为或技术路径。

核体系应建模为同心信任区：公共及商业网络；承包商及后勤网络；政府组织网络；情报和运行支持；NC3 邻接系统；授权与执行系统；武器物理保障。渗透一个区域不意味着控制下一区域。每一道边界都应提出新的身份、硬件、程序及人员要求。

7.4 防御性路径分析

阶段	高层次威胁目标	防御阻断点
1 发现	描绘组织、供应商、人员、信任关系和暴露服务。	减少暴露元数据、监测自动收集、培训高风险人员。
2 初始立足点	通过人员、供应商、账户或脆弱服务获得访问。	抗钓鱼认证、供应商保障、应用白名单、快速修补。
3 持久化	在密码更改、重置或调查后维持访问。	短期凭据、不可变日志、终端证明、洁净环境重建能力。
4 权限扩大	取得更广权限或进入更敏感信任区。	零信任分区、职责分离、即时权限、双重授权。
5 目标发现	识别核相关数据流、保障系统或决策依赖。	数据最小化、隔离分区、诱捕系统、异常检测。
6 影响或干扰	污染信息、降低可用性、操纵时序或欺骗操作者。	经认证的数据来源、独立传感器、人工替代、跨通道核验。
7 战略效果	导致误判、未经授权行动、响应延误或信心丧失。	积极的人类控制、多人规则、危机热线、失效安全态势、经过演练的恢复。

7.5 最可信的影响

- 情报污染：修改或伪造数据，改变对对手意图或准备状态的判断。
- 预警退化：干扰早期预警信息的可用性、可信度或解释。
- 决策支持操纵：选择性呈现、制造确定性或压制相反证据。
- 通信中断：拒绝服务、延迟或混乱，造成在通道失效前行动的压力。
- 维护与战备影响：虚假状态、被篡改诊断，或后勤及人员排班受扰。
- 凭据与身份失陷：在外围及支持环境中冒充可信官员或系统。
- 信息环境行动：利用合成消息、伪造命令或公众操纵加剧危机压力。

直接控制武器是最不可信的数字路径，因为核系统采用专用硬件、程序控制、认证通信、多人规则、实物保管，以及因国家而异的其他保障。系统保密性使公众无法作出高度确信的保证。有依据的判断既不是直接访问很容易，也不是物理隔离使其绝无可能。

7.5.1 合成信息与危机升级

攻击者无需渗透核基地也能影响危机。伪造视听材料、冒充通信或误导性摘要，可污染决策者接触的信息。一段声称显示首次核打击的视频，是防御演练情景，并非一段视频就会导致核使用的证据。效果取决于可信度、佐证、危机背景及制度核验是否失效。所引证据并未确立这是成本最低或影响最大的核攻击路径。

NIST 合成内容报告研究了来源追溯、标识、水印和检测等技术透明度方法。[34]本文建议将这些工具与经认证官方通道及独立信息源结合。来源信息可支持关于出处或处理历史的主张，但本身不能证明经认证发送者陈述属实。缺乏来源信息并不构成伪造的确证；检测器评分也不构成接受或忽视核预警的授权。

危机信息完整性规程应区分收到的主张、经认证消息、经独立佐证观察和分析推断。人工智能生成摘要时，应保留原始证据及可见的不确定性。分析人员应识别看似独立的报告是否来自同一来源，通过独立通道获得佐证，并将未解决矛盾提交获授权人员。未经核实的公共媒体，不得单独构成使用核武器的充分依据。演练应包含伪造材料、被错误标为合成的真实证据，以及核验通道中断；目标是准确评估，而不是本能地不信任所有数字证据。附录 E 规定拟议验收测试。

7.6 造成核损害的条件

须区分路径。未经授权发射要求具备足够能力和访问，以破坏或绕过授权及执行保障；这些保障因国家而异，本文不作公开映射。人工智能影响下的人类核使用，则可能来自地缘政治紧张中的信息污染、错误归因和被压缩的审议。该路径无需穿越直接发射情景的每一道技术边界。两者都不需要机器意识；公开信息也不足以断言能够实际进入某个指定国家的武器系统。

8 至 2036 年及更长期的概率评估

8.1 方法

不存在能得出人工智能相关核灾难真实概率的代表性频率数据或经验证模型。下列区间保留本文的情景规划假设，但明确属于未经验证的主观判断。它们并非从事件树、有记录的专家征询或精算基线推导而来，不得作为测量预测、置信区间或机构共识引用。数值精度仅是呈现方式，并非证据精度；决策应在这些区间之外仍保持稳健。

人工智能促成核使用的区间，没有针对所有原因核使用概率的独立既定基线。形式上，P（人工智能促成使用）= P（任何核使用）× P（人工智能促成 | 使用）；本文不估计其中任何项。因此，所列区间不能证明相对于没有人工智能的世界，人工智能增加了多少风险。未来修订应预登记事件判定标准，征询并汇总独立专家判断，记录假设和分歧，并评估其对地缘政治基线与防御改进的敏感性。

对于直接未经授权发射，概念链为 $P(L) = P(C) \times P(H|C) \times P(A|C,H) \times P(B|C,H,A) \times P(L|C,H,A,B)$，其中 C 为足够的综合能力，H 为有害导向，A 为相关访问，B 为中间屏障失效。以完整历史为条件，避免了独立性假设。本文未对任何因素作数值估计。间接路径不同：人工智能对误导信息或支持中断的贡献，与危机条件及人类决策相结合。两种路径重叠，未估计联合事件前不得相加。

8.2 十年情景规划区间

2026 年 9 月 20 日至 2036 年 9 月 20 日期间的事件	未经验证的累计规划区间	分析置信度	判断
至少一次新的严重网络行动，人工智能实质协助针对拥核国家军事或关键基础设施的行动	70%至 95%	低至中等	过去事件支持其可能性；只有规定时段内符合条件的新行动，才能判定该未来事件发生。
利用人工智能的行动在核体系外围组织获得未经授权立足点	30%至 60%	低至中等	庞大供应商生态和普通网络暴露创造机会。
利用人工智能的行动进入敏感核保障或 NC3 邻接环境	8%至 20%	低	需要多重边界失效；公开证据有限。
人工智能造成的信息篡改或干扰，实质影响核态势或危机决策	2%至 8%	低	通过信息完整性和时序影响，比直接控制更可信。
人工智能支持的欺骗、故障或网络行动，实质促成核武器使用	0.5%至 3%	极低	本文未估计总体核使用的可靠基线；该区间不适合作为政策预测。
失控人工智能独立导致未经授权的核发射或引爆	0.1%至 1%	极低	尾部风险规划假设。公开证据无法校准其下限或上限。

所有行均涉及同一未来时段内、在能力持续增长及治理不完整条件下的累计事件。各项既不可相加，也不是以前一行为条件的概率集合。直接发射区间在概念上包含假设的屏障失效，但并非据

此计算。本文不指定标准情报概率术语：“极不可能”在某些惯例中对应不同数值范围。置信度描述证据支持程度，而不是概率。外围区间跨越 50%；两个灾难性区间均不足以让人确信真实风险低于其上端点。

事件判定标准：“实质协助”要求对执行或影响有书面记录的贡献，而非偶然使用人工智能。外围立足点要求未经授权控制已识别核体系供应商或组织的系统；公开信息暴露并不足够。敏感渗透要求经独立评估的核保障或 NC3 邻接环境访问。态势影响要求有记录且具实质后果的战备、评估或危机行动变化。人工智能促成核使用要求实际使用核武器，并有得到证实的因果贡献，而非仅同时发生。独立发射要求在没有新的合法人类决策情况下实际发射或引爆。须记录来源质量，并将未确认事件归为“未解决”，而不是“未发生”。涉密不披露、归因争议和战略欺骗，严重限制公开判定。

8.3 2036 年以后

超过十年，量化精度急剧下降。如果智能体可靠性趋于平台、核网络保持隔离、强治理得以落实，失控人工智能直接使用核武器可能仍是极端尾部风险。如果人工智能实现长期网络行动自动化、跨基础设施复制、突破监督，并在竞争压力下融入核决策支持，2036 至 2050 年期间风险可能显著上升。

2036 至 2050 年情景	特征	对核领域人工智能风险的影响
受控进步	能力逐步提高；许可、检查及核红线获广泛采用。	风险缓慢增长，仍主要来自人类恶意使用和普通网络故障。
治理不均	主要国家实施监管，但扩散与未监测模型经二级市场传播。	持续入侵尝试增加；外围和供应链失陷变得常见。
战略竞赛	国家为保持速度与优势，向人工智能委托更多分析和行动。	危机不稳定性和自动化偏差上升；虚假预警和决策压缩风险变得严重。

2036 至 2050 年情景	特征	对核领域人工智能风险的影响
失去控制	系统获得持久自主性、复制、欺骗和资源获取能力，而监测失效。	核风险成为更广泛恢复控制斗争的一部分；今天无法负责任地量化其概率。

8.4 哪些情况会改变评估

如果独立评估显示可靠的长达一个月的网络自主能力；模型在非人为设计测试之外反复突破监测；具有前沿网络能力的权重被窃取或不可逆发布；人工智能获得对防务或核保障环境的持续访问；或拥核国家将时效紧迫的预警解释或授权功能委托给人工智能，则担忧应上升。如果安全评估具有预测性、前沿权重保持受保护、关键系统采用强来源验证和人工后备，并且核国家接受可核查的人类控制规则，则担忧应下降。

公开监测应分别跟踪在明确成功阈值下验证的结果、部署权限及控制失效。实验室任务成功已有观察；所引记录并未证实持续未经授权的核行动。须区分声称的事件与经佐证影响，也须区分政策宣布与实施。欺骗性泄露、选择性披露和不报告均可扭曲趋势估计。任何单一公告都不应机械地触发百分比调整。

9 核领域专门保障

9.1 绝对红线

- 任何人工智能系统均不得行使被授予的核武器使用决策权，无论单独行使还是作为名义共享决策安排的一员；也不得发出、更改或认证核使用命令。
- 核使用决定应保留于合法人类权威，并在授权与执行中实施积极、当时有效且独立认证的人类核查。这不假定每个国家都有多名地位相同的宪法决策者。
- 通用模型或持续学习模型不得直接连接核授权或武器控制接口。
- 人工智能生成的预警、目标或毁伤评估，不得被视为单独足以支持核使用的证据。
- 禁止在没有新的合法人类决策时自动启动或报复，包括传统基于规则的自动化。经独立验证、用于阻止不安全行动的安全联锁仍可允许，但绝不能启动核使用。

- 在核相关生产环境中，禁止人工智能单方面更改模型权重、工具、权限、监测、审计轨迹或关停控制。重大变化须经离线验证、独立人类批准，并记录部署授权。

9.2 纳入人工智能因素的核失效安全审查

至少每 24 个月，以及在重大集成变化或严重事件后，开展具备涉密许可的独立审查。参与者应包括核操作人员、人工智能评估专家、网络防御人员、人因专家和独立审查者。核威胁倡议组织（NTI）的工作支持审视最终发射控制之外的依赖。美国于 2024 年完成国会要求的失效安全审查；这提供了审查先例，而非证明所有拟议人工智能控制都已实施。[8][9][30]

- 清点所有可能影响核信息、战备、通信、目标选择或决策支持的人工智能系统。
- 梳理对商业模型、云服务、承包商和软件供应链的直接及间接依赖。
- 在现实危机演练中测试虚假数据、模型失陷、通信丧失和监测失效情景。
- 验证人工及非人工智能后备方案，以及恢复所需时间。
- 评估自动化偏差、时间压力、界面设计、分歧处理，以及人类操作者停止或忽略人工智能输出的权限。
- 审查核相关信息的来源、认证、独立感知和跨通道确认。
- 制定整改计划，明确负责人、期限和独立闭环验证。

年度后备演练必须预设通过标准：在获批时间内恢复指定关键功能；在规定持续时间内不依赖人工智能衍生数据运行；维持经认证通信；发现并适当处理虚假或冲突信息；无未经授权的模拟行动。依据已知演练真实情况，以独立评分评估决策质量。任何必要标准失败，都应触发运行限制、指定整改和见证复测。具体运行容限应保密。界面必须显示来源和不确定性，使异议及独立佐证可用，并测试操作者理解。审议要求须按任务验证；危机中的任意固定延迟本身可能不安全。应防止人工智能制造的期限绕过人类核验，并保留安全抑制与后备选项。

附录 E 将上述原则转化为拟议保障规程，要求核查来源独立性、测量人因表现，并提供受保护程序以质疑人工智能支持的评估。所需功能是有效佐证和异议；仅增加第二个模型并不满足要求。

9.3 战略稳定措施

2024 年 11 月 16 日拜登与习近平会晤确认，核武器使用决策应保持人类控制。这是外交基础，而非经核查的条约禁令。[29]应在此基础上建立经认证危机联系人、共同事件定义、相互保证和保

密审查交流。相关集成变化通知应保护行动细节。不能依据未经核实的未来会谈报道推进实施，也不能假定公告保证达成协议。

10 失效模式与异议

10.1 监管会拖慢创新

不良监管可能巩固既有企业地位并压制有益研究。因此，许可应仅适用于前沿训练、高行动能力或重大后果部署。小模型和低风险应用应承担基线义务，而非前沿许可制度。安全评估基础设施、共同标准和监管沙箱可降低合规成本。相关比较不是监管与创新的对立，而是有纪律的发展与本可避免的灾难事件所带来的经济、政治冲击之间的比较。

10.2 国家会作弊

一些国家和企业会规避规则，但这不意味着治理无用。先进芯片、数据中心、专业劳动力、云网络和主要模型分发渠道形成可观察节点。核查将如军控与金融领域一样不完整，但申报、检查、情报、供应链控制和制裁可提高隐蔽活动成本。以完美合规为前提的制度会失败；以发现、威慑、延缓和协调响应为目标的制度仍能降低风险。

10.3 开放模型使控制不可能

开放研究和低于阈值的发布具有正当收益。关键区分是能力与部署风险，包括可合理预见的微调和工具集成，而非国籍或开放、封闭标签。越过危险能力阈值后，广泛发布是可能不可逆的转移。许可、披露义务和针对有害用途的执法，发布后仍可发挥作用，却无法可靠召回副本或阻止所有保障移除。应分阶段发布，并在分发前评估下游修改；避免声称技术上能实现普遍发布后控制。

10.4 有人在回路中就足够

当系统更快、更复杂或被认为比操作者更准确时，名义上的人类可能沦为橡皮图章。有意义的人类控制需要时间、能力、权限、独立信息、可理解的不确定性，以及不受处罚地拒绝建议的能力。在核场景中，人类控制必须被设计和演练为系统属性，而不是在自动化流程旁安排一人就算满足。

10.4.1 自动化偏差证据

自动化偏差包括接受错误自动建议，以及在辅助系统未提示时忽略问题。Goddard、Roudsari 和 Wyatt 对 74 项研究的综述指出，其程度随信任、经验、任务要求、工作负荷及时间压力而变化，也指出培训、问责和界面设计可能缓解问题。[35]Mosier 和 Skitka 的航空研究发现，在所研究环境中，增加第二名机组成员并未显著降低自动化错误率；某些培训降低了学生的错误执行行为。[37]这两项来源都未证明核决策者会“几乎总是”信任人工智能，也未提供可直接用于核危机的错误率。

政策含义是通过实证验证有意义的人类控制。应以正确、错误、不完整及缺失的人工智能建议测试操作者，评估接受错误建议、漏判危险、来源核查、纠正时间及非人工智能能力保持情况。在操作可行时，应先取得初始独立人类判断，再展示模型建议。保留底层证据访问，并记录重大后果决策的理由。培训与界面变更必须在代表性演练中证明收益，而不是仅增加确认按钮。

10.4.2 独立质疑与人工智能辩论的局限

Du 及同事报告，多智能体辩论在部分推理和事实准确性任务中带来改善。[36]这支持测试该方法，而非验证其适用于核决策支持。两个模型可能共享错误证据、相似训练或共同被攻陷的工具。因此，一致意见可能只提高表面信心，并不提供独立佐证。强制分歧也可能制造无关异议、延迟评估，或使原本正确的分析员接受表述流畅的错误。

应强制设置独立质疑功能，而不是固定采用两个争论中的人工智能。具备涉密许可的人类审查者、不同证据通道和经验证的非人工智能核查仍不可缺少。多个人工智能顾问只有在系统级测试证明，在相关条件下优于纯人工及单模型基线后，才可辅助此功能。保留初始独立判断和未解决异议；防止自动多数投票或模型共识决定核使用建议。独立性须从数据来源、工具、失效模式和运行控制评估，不能因模型名称不同而推定。附录 E 规定部署和暂停标准。

10.5 全球关停开关能解决问题

全球开关会成为攻击、政治俘获和胁迫的重大目标；模型与权重分发后，技术上也不现实。更安全的设计是分布式控制：提供方撤销服务和凭据；基础设施运营方隔离指定工作负载；政府发出合法暂停令；组织断开本地系统。每项行动都应要求经认证权限、保全证据并接受监督。

11 指标与预警框架

11.1 能力指标

- 在有限人类干预下可靠完成持续数天或数周的软件及网络任务。
- 在陌生系统中成功运行，并能从错误和防御反制中恢复。
- 在受控评估中证明可自主获取资源、身份、算力或人类协助。
- 持续复制或迁移尝试，尤其是对监测者隐瞒的情况。
- 欺骗、态势感知、规避监测或奖励机制钻空子能力显著改善。
- 以实际行动速度识别和组合此前未知漏洞的能力。

11.2 部署指标

- 前沿智能体获得常驻生产凭据、不受限互联网访问，或跨组织网络执行代码的权限。
- 国防或关键基础设施组织在没有独立评估或本地隔离的情况下集成商业通用模型。
- 核决策时间缩短，因为预期人工智能建议被接受的速度快于人类核验速度。
- 国家把人工智能连接至预警、目标选择或 NC3 邻接数据，却没有经认证来源与人工后备。
- 开发者或国家隐瞒严重模型事件、抵制独立审查，或在越过阈值后更改评估标准。

11.3 治理指标

指标	目标	理由
关键事件首次通知	24 小时内；灾难性风险迫在眉睫时立即通知	支持遏制和跨境预警。
独立评估覆盖率	部署前覆盖 100%的 3 级和 4 级系统	减少对开发者自评的依赖。
高风险凭据有效期	以分钟或小时计，而非长期访问	限制持久性和影响范围。

指标	目标	理由
人类核授权	合法人类决策，加上独立人类认证及执行核查；不向人工智能授予决策权	防止模型或操作者单方面行动。
核领域人工智能失效安全审查	每两年，以及重大变化或事件后	随着模型及集成变化维持保障的时效性。
关键系统人工后备演练	至少每年一次	测试在没有人工智能服务或衍生数据时的连续运行。
恢复演练	前沿实验室和关键部署至少每年一次	测试是否真能恢复控制。

12 结论

政策目标是在能力进步的同时，保持对重大后果系统有效的人类控制。监管的核心单位是已部署的模型、工具、权限和制度权威组合。范围明确、基于能力的许可，专业独立评估，以及可执行的部署条件，可降低暴露，而无需许可所有普通人工智能用途。

核保障应保护信息完整性和审议，也保护最终授权链。仅有人类在场还不够：操作者需要独立证据、可用界面，以及经过演练的非人工智能替代方案。直接未经授权发射仍比外围失陷要求高得多，但公开证据无法给出其可靠数值边界。

国际合作应从可核查功能和受保护证据交换开始，并以互惠条件纳入中国及其他相关国家。国家执法与权限有限的国际安全机构相辅相成。全球关停机制或未经审查的申报，都不能替代分层控制。

拟议规则是一项研究和谈判议程。其有效性应通过独立评估、事件学习和演练检验，并随证据变化修订阈值与假设。第 8 节的规划区间是稳健性测试的起点，不是对灾难预测日期或概率抱有信心的依据。

附录 A 拟议条约原则

1. 各国仍对在其管辖范围内开发、部署或获得实质支持的人工智能系统负责。
2. 前沿开发者和指定算力提供方必须向主管国家机关登记并报告高风险活动。
3. 超过商定能力及风险阈值的训练和部署，须有获批安全论证及独立评估。
4. 缔约方必须依共同安全等级保护前沿模型权重和关键基础设施。
5. 具有跨境或灾难性潜力的严重人工智能事件及险情，必须经受保护渠道报告。
6. 保留合法人类核使用决策权和独立人类核查；禁止向人工智能委托决策权，以及在没有新人类决定时自动报复。
7. 缔约方必须对核武器及其他战略武器保持有意义的人类控制，并保留非人工智能后备程序。
8. 缔约方必须定期开展纳入人工智能因素的核失效安全审查，并提供商定的保密完成证明。
9. 采用受保护申报、程序证据和协商检查权，同时承认核查不完整，并规定异常响应流程。
10. 不合规可导致整改令、指定算力或模型服务暂停、出口限制、罚款，以及与违规相称的集体措施。
11. 与国际机构共享的安全信息必须获得强有力的法律、技术和外交保护。
12. 阈值及技术附件至少每年复核一次，使监管跟踪能力，而非过时硬件指标。

附录 B 最低前沿安全论证

部分	所需证据
治理	法律所有者、负责高管、风险委员会、独立质疑、举报渠道、监管联系人。
系统边界	模型、版本、工具、记忆、数据、凭据、基础设施、用户、司法辖区及外部依赖。
威胁模型	滥用、失配、内部人员、模型窃取、供应链、工具失陷、监测失陷和相关性失效。

部分	所需证据
能力	标准及定制评估、不确定性、能力激发方法、外部复现和规模扩展预测。
倾向	欺骗、奖励机制钻空子、未经授权行动、规避监督、关停行为及评估感知。
控制	隔离、身份、授权、日志、监测、速率及资源限制、数据来源、回滚和关停。
残余风险	概率与严重程度区间、假设、专家异议、红队发现和未解决异常。
运行	部署包络、变更控制、监测阈值、事件响应、恢复、召回和通信计划。
保障	独立评估报告、监管结论、整改证据及有时限授权。

附录 C 核领域人工智能红线示范条款

第 1 条 范围

人工智能系统是基于机器的系统，能够从输入推断如何生成预测、内容、建议或决策等输出，以影响物理或虚拟环境；各系统的自主性和适应性不同。此功能定义借鉴欧盟《人工智能法案》第 3 条。[5]核相关人工智能系统，是对核预警、情报、战备、通信、目标选择、后勤或决策支持产生实质影响的系统。权限禁令适用于授权及执行链；辅助应用仍受比例适当的安全控制约束。第 2 条也涵盖传统自动化，以防旧系统漏洞。

合规证据：持续更新的系统清单及有记录的分类决定，可供指定涉密主管机关查阅。

第 2 条 禁止自主核决策权

任何缔约国不得向人工智能单独或共同授予核使用决策权，不得允许其发出、更改或认证核使用命令，也不得替代必要的人类授权。核使用须由合法人类权威作出新的决定，并在授权及执行中实施积极、独立认证的人类核查。禁止没有该决定的自动启动或报复，包括基于规则的系统。经独立验证的安全联锁可以阻止不安全行动，但不得启动核使用。

合规证据：受保护的授权架构、由独立人员见证的人类认证测试，以及排除传统自动启动机制的证据。

第 3 条 人类控制与后备

各缔约国应保持经过测试的非人工智能预警评估、通信、认证和决策支持路径。人工智能衍生信息须可识别，且不得单独成为核使用的充分证据。操作者应能通过独立控制隔离人工智能组件，同时保留基本安全功能。界面和程序应支持佐证、异议及经过任务验证的审议。自动化升级窗口不得绕过人类核验；实施必须保留安全抑制选项，不得引入未经测试的延迟。

合规证据：有见证的隔离及后备演练、人因评估，以及冲突信息处理记录。

第 4 条 失效安全审查

各缔约国应以不超过 24 个月的间隔，以及在重大变化或严重事件后，完成具备涉密许可、独立且纳入人工智能因素的失效安全审查。审查应包含间接信息及供应链依赖。年度非人工智能演练应测试预设的恢复、连续性、认证及决策质量标准。重大失败应指定整改负责人、期限、比例适当的临时运行限制，并进行独立复测。

合规证据：审查日期、独立审查者资质、演练结果及重大问题已闭环解决的验证。

第 5 条 消极保证与核查

各缔约国应每年证明遵守第 2 至 4 条，并提供商定的受保护证据。常规国际检查应限于文件权限内的已申报民用设施。涉密保障应结合独立国家审查、密封报告和受控演示。重大异常，包括不一致证明、无法解释的权限变化或必要日志缺失，应触发保密磋商及限期回应。质询访问须有商定保障。单独的证明不得被描述为全面核查。

合规证据：签署申报、评估范围与局限、保管记录，以及异常解决记录；保护行动秘密。

第 6 条 事件与危机通信

缔约方应维持经认证危机联系人。对于可信且迫在眉睫的虚假核预警、未经授权行动或跨境升级风险，应在可行时立即通过该渠道通知。识别后 24 小时内提交初始事件通知，72 小时内提交初步评估，并持续更新；以 30 天内完成更全面审查为目标，否则说明尚未完成的调查。早期报告可以不完整，但须区分事实、假设和归因置信度。这些是拟议条约期限，并非对普遍现行法律的陈述。

合规证据：经过演练的危机联系、带时间戳的通知及更正，以及区分不确定性与归因的调查记录。

第 7 条 保密与合规

依本条款交换的信息应获得商定的法律与技术保护，不得用于获取商业优势或无关军事情报。合规关切应先提交联合技术委员会。持续不合规可导致加强磋商、受控访问请求、合作暂停，或符合国际法的比例适当国家措施。

合规证据：访问控制、保留及删除规则、独立投诉渠道，以及比例适当、可复核的合规决定记录。

附录 D 来源与说明

为便于查证，参考文献保留原作者姓名、原始出版物标题及完整链接；说明性注释已译为中文。

[1] Bengio, Y. et al., International AI Safety Report 2026, DSIT 2026/001, 2026 年 2 月 3 日, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026

[2] OpenAI, The Hugging Face incident and other third-party impact from misaligned models, 更新日期： 2026 年 9 月, https://openai.com/hugging-face-incident-and-misalignment/

[3] OpenAI, Pacing model development in an era of cyber-critical capabilities, 2026 年 8 月 18 日, https://openai.com/index/pacing-model-development-cyber-capabilities/

[4] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework 1.0, 2023 年 1 月, and Generative Artificial Intelligence Profile NIST AI 600-1, 2024 年 7 月, https://www.nist.gov/itl/ai-risk-management-framework

[5] Regulation EU 2024/1689 laying down harmonised rules on artificial intelligence, Official Journal of the European Union, https://eur-lex.europa.eu/eli/reg/2024/1689/oj

[6] Council of Europe, Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, 开放签署日期： 2024 年 9 月 5 日, https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

[7] UK Department for Science, Innovation and Technology and Republic of Korea, Frontier AI Safety Commitments, AI Seoul Summit 2024, 更新日期： 2025 年 2 月 7 日, https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024

[8] Williams, I., AI Is Moving Into Nuclear Weapons Systems. Fail-Safe Reviews Are Critical, Nuclear Threat Initiative, 2026 年 9 月 17 日, https://www.nti.org/risky-business/ai-is-moving-into-nuclear-weapons-systems-fail-safe-reviews-are-critical/

[9] Hruby, J. and Miller, M. N., Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems, Nuclear Threat Initiative, 2021 年 8 月 26 日, https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/

[10] International Atomic Energy Agency, Computer Security for Nuclear Security, IAEA Nuclear Security Series No. 42-G, Vienna, 2021. https://www.iaea.org/publications/13629/computer-security-for-nuclear-security

[11] International Atomic Energy Agency, Computer Security Techniques for Nuclear Facilities, IAEA Nuclear Security Series No. 17-T Rev. 1, Vienna, 2021. https://www.iaea.org/publications/14729/computer-security-techniques-for-nuclear-facilities

[12] National Institute of Standards and Technology, Cybersecurity Framework 2.0, 2024 年 2 月, https://www.nist.gov/cyberframework

[13] Cyberspace Administration of China and six other authorities, Interim Measures for the Management of Generative Artificial Intelligence Services, 生效日期： 2023 年 8 月 15 日, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[14] Cyberspace Administration of China and three other authorities, Measures for Labelling Artificial Intelligence Generated and Synthetic Content, 生效日期： 2025 年 9 月 1 日, https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm

[15] Cyberspace Administration of China and other authorities, Interim Measures for the Management of Anthropomorphic Artificial Intelligence Interactive Services, 2026 年 4 月 10 日, https://www.cac.gov.cn/2026-04/10/c_1777558395078289.htm

[16] Reuters, How China is preparing for the risk of AI escaping human control, 2026 年 9 月 14 日. 有关政策发展的报道；本文不依赖作者署名及后续修订日期。
https://www.reuters.com/legal/litigation/how-china-is-preparing-risk-ai-escaping-human-control-2026-09-14/

[17] White House, America's AI Action Plan, 2025 年 7 月,
https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[18] United Nations General Assembly, Resolution 79/325: Terms of reference and modalities for the establishment and functioning of the Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance, 2025.
https://docs.un.org/en/A/RES/79/325

[19] National Institute of Standards and Technology, Center for AI Standards and Innovation, model evaluations and standards work, 访问日期： 2026 年 9 月 19 日,
https://www.nist.gov/caisi

[20] Stanford Institute for Human-Centered Artificial Intelligence, AI Index Report 2026,
https://hai.stanford.edu/ai-index/2026-ai-index-report

[21] International Atomic Energy Agency, Safeguards and Verification,
https://www.iaea.org/topics/safeguards-and-verification

[22] Financial Action Task Force, The FATF Recommendations,
https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Fatf-recommendations.html

[23] OpenAI, The Hugging Face incident and the road ahead, 2026 年 8 月 26 日. 开发者对 7 月事件的陈述。
https://openai.com/index/hugging-face-incident-and-the-road-ahead/

[24] METR and Redwood Research, Investigation of the OpenAI Hugging Face incident, 2026 年 8 月 26 日. 证据访问时段有限的外部调查。
https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/

[25] Anthropic, Investigating three real-world incidents in our cybersecurity evaluations, 2026 年 7 月 30 日. 开发者披露。
https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals

[26] California, SB 53, Transparency in Frontier Artificial Intelligence Act, Chapter 138, 批准日期： 2025 年 9 月 29 日. 已颁布法律文本。
https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53

[27] European Union, Regulation (EU) 2026/1744 (AI Omnibus); European Commission, AI Omnibus enters into force, 2026 年 7 月 27 日. 修订后的高风险实施时间表与监督条款。
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L_202601744

[28] White House, National Security Presidential Memorandum/NSPM-11, 2026 年 6 月 5 日. 国家安全行政政策。
https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/

[29] White House, Readout of President Joe Biden’s Meeting with President Xi Jinping of the People’s Republic of China, 2024 年 11 月 16 日.
https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2024/11/16/readout-of-president-joe-bidens-meeting-with-president-xi-jinping-of-the-peoples-republic-of-china-3/

[30] Nuclear Threat Initiative, Advancing Nuclear Fail-Safe, 项目概述，描述美国于 2024 年完成的审查。
https://www.nti.org/about/programs-projects/project/advancing-nuclear-fail-safe/

[31] METR, Time Horizons, 方法与结果；另见 Kwa et al., Measuring AI Ability to Complete Long Tasks, arXiv:2503.14499 (2025).
https://metr.org/time-horizons/

[32] IETF, RFC 9334, Remote ATtestation procedureS (RATS) Architecture, 2023 年 1 月.
https://www.rfc-editor.org/rfc/rfc9334.html

[33] National Telecommunications and Information Administration (2024). Dual-Use Foundation Models with Widely Available Model Weights Report. 7 月 30 日. 历史政策评估，并非对当前普遍法律的陈述。
https://www.ntia.gov/programs-and-initiatives/artificial-intelligence/open-model-weights-report

[34] Chandra, B., Dunietz, J., Roberts, K., Lee, Y., Fontana, P. and Awad, G. (2024). Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency. NIST AI 100-4.
https://doi.org/10.6028/NIST.AI.100-4

[35] Goddard, K., Roudsari, A. and Wyatt, J. C. (2012). Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. 在线发表日期： 2011 年 6 月 16 日.
https://doi.org/10.1136/amiajnl-2011-000089

[36] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. and Mordatch, I. (2023). Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325. 实验研究，并非核安全验证。
https://arxiv.org/abs/2305.14325

[37] Mosier, K. L. and Skitka, L. J. (1999). Automation Use and Automation Bias. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 43(3), 344–348.
https://doi.org/10.1177/154193129904300346

证据基础与局限

证据截止日期：2026 年 9 月 20 日。正式法律文本仅在其司法辖区和适用日期内确立义务。开发者披露是利益相关方提供的一手陈述；外部调查改善审查，但仍受访问和选择限制。技术评估证明的是受测条件下的表现。政策评论提供论点，而非事件频率估计。高置信度表示强而一致的支持；中等表示可信但不完整的支持；低表示零散或对假设敏感的支持；极低表示严重局限。这些标签与事件概率不同。第 8 节任何数值区间均未经独立校准。本文为署名政策研究，而非经实验验证的预测或正式同行评审期刊论文。

附录 E 实施与保障规程

本附录提供拟议政策措辞，用于将本文转化为国家监管、国家安全指令或经谈判的条约文件。它不创设当前法律义务。其要求适用于受规制 3 级活动和允许的 4B 级应用；被禁止的 4A 级核使用决策权不能通过安全论证获得许可。标准不区分来源，按能力、访问和后果适用于开放与封闭系统。

E.1 责任与重大变化

原始开发者。受规制发布前，应记录模型身份、评估范围、已知局限、合理下游增强及分发不可逆性。向监管机构和合法下游接收者提供适当安全信息及事件联系人。通过受控披露保护可被利用的细节。发布决定应说明哪些用途已评估，哪些仍缺乏证据支持。

重大修改者或微调者。记录已知母模型、版本标识、相关变化及相应评估证据。重大变化包括已证实的危险能力提高、必要控制移除、自主运行扩大，或转用于后果更重的用途。受规制发布或部署前须提交修订或新安全论证。获批包络内的低影响变化需要记录回归检查，并不自动触发全面批准。

集成者与运营者。明确对已部署系统负责的法律实体。评估模型、检索来源、工具、记忆、凭据、监测和人类程序的组合。模型此前获批，不等于新权限集合或核相关应用获得授权。维护变更登记、明确事件负责人、测试隔离程序，并保持最新运行包络。

分发者与基础设施提供方。在法律涵盖其活动时，保留识别受监管运营者及授权所必需的最低记录，保存相关证据，并响应合法且可复核的命令。不应要求普遍监控普通用户，也不能假定每个被动存储库都能评估每个上传模型。义务须与实际控制、规模、知情程度和风险相称。

问责与证据。国家监管机构应检查发布评估、谱系申报、变更日志、评估报告及运行权限。根据违反的义务、因果关系、可预见性和行为者控制力分配责任，并给予适用程序保障。合同可分配工作，但不能取消法定义务。缺失或误导的谱系申报属于保障缺陷；模型哈希是标识符，不是安全或完整祖先关系的证明。

E.2 跨境执法与 IFASA

海外托管或从海外实验室下载的模型，在用于受规制国内活动时，仍须遵守进口国或部署国的适用规则。参与国应指定主管机关，仅在已测试范围内认可评估，并交换严重失败的受保护通知。IFASA 应支持共同方法、评估者认可、事件协调和技术磋商。执法权必须源于国内法和商定条约授权。

对于受规制运营者，得到证实的缺陷应允许国家机关要求重新评估、收窄权限、暂停受影响部署或施加比例适当处罚。紧急限制须有记录的理由和及时复核。关于外国开发者的结论应区分经验证证据与指控，并提供更正或质疑程序。国籍本身不是不合规证据。

对于参与辖区之外复制的权重，可用措施必然不完整：自愿合作、限制受监管国内用途和服务、受保护事件交换，以及针对可识别有害行为的依法授权措施。任何通知、许可条件或登记都不能保证删除已分发权重。应报告残余执法缺口，而非声称能够全球召回或行使国际关停权。

E.3 危机信息完整性

责任主体与触发条件。只要合成、争议信息或人工智能摘要可能实质影响危机评估或核态势，负责的核主管机关就应适用书面信息完整性程序。指定值班官员继续负责核验；人工智能分类器不得成为真实性的最终裁决者。

必要控制。识别来源、时间、可获得的处理历史、认证状态和未解决矛盾。将观察事实与指控、推断分开。使用独立来源佐证及经认证官方通道。媒体间重复报道，或采用相同输入模型之间的一致意见，不应算作独立确认。保留原始材料，避免删除不确定性的摘要。真实来源不保证事实准确；缺失来源记录不证明伪造。

保障证据。在受保护演练中，测试伪造和真实材料、冒充、来源不可用、可信来源失陷及通信退化。记录错误接受、错误拒绝、核验延迟、漏检矛盾和由此造成的决策错误。通过行业安全论证，预先确定各功能及危机条件下可接受表现；不得凭空制定统一延迟或准确率阈值。以独立确定的演练真实情况为基准，而非模型生成的标准答案。

失败与响应。必要标准失败时，应对受影响的人工智能信息路径施加比例适当限制，恢复已验证替代方案、整改并进行见证复测。保留预警和安全抑制功能。未经认证视频、合成媒体评分或人工智能综合摘要，不得单独满足核使用前提。此保障不构成忽视其他独立佐证证据的理由。

E.4 有意义的人类控制与独立质疑

必要功能。人工智能支持的重大后果评估，应允许独立推理的质疑、访问底层证据，并明确说明未解决分歧。合格人类保留权限与责任。时间和安全允许时，先记录初始人类判断，再展示人工智能建议。若该顺序不可行，应采用经任务验证的程序；不得施加妨碍保护行动的任意延迟。

人工智能辩论作为可选机制。第二个模型只有在获批安全论证范围内，才可支持质疑。记录已知共同训练依赖、共享数据、检索、工具和监测。在交换意见前独立形成初始评估，保留异议，并测试辩论纠正还是扩散错误。模型一致不是独立证据；人工智能投票不得决定核使用权限。不同提供方并不自动保证独立失效模式。

验证。在代表性、受保护演练中，比较纯人工、单顾问和拟议多顾问安排。纳入误导输入、共同错误的模型、具有说服力但无支持的解释、矛盾证据、人工智能不可用及时间压力。预登记成功标准并记录不确定性。测量错误接受、漏判危险、决策质量、核验时间、工作负荷及撤除人工智能后的恢复。评估完整的操作者—界面系统，而不只看基准准确率。

部署门槛与审查。具备涉密许可的审查机关，仅在证据支持安全收益且不会造成不可接受延迟或负荷时，才应允许人工智能质疑配置。监测漂移，并在重大变化、定期失效安全审查和严重事件后重复验证。未通过必要标准的配置应暂停或限制。维持非人工智能质疑渠道。这些是拟议保障要求；所引研究未确立普遍的核领域表现阈值。[35–37]

E.5 算力通知与授权

适用主体与触发条件。开发者和指定大规模训练服务提供方，应依公布的计量规则报告受规制算力活动。能力与部署触发条件仍独立适用，包括低成本微调和进口模型。 1 亿美元之类支出数额，不得成为低于该支出活动的安全豁免，也不得成为高于该金额就具危险能力的确证。

申报与合并计算。记录责任实体、协调计划、参与提供方、相关算力估计及不确定性、重大后训练活动、授权状态和评估里程碑。规定如何跨地点和实体合并协调运行及外包。对风险导向样本进行独立审计并调查差异。硬件物理所有权及商业计费结构，不得决定相同受监管活动是否逃避报告。

执法与复核。重大隐瞒或规避应触发调查及依法比例适当措施。保护保密信息，并允许更正善意估算错误。至少每年根据观察能力、算法效率和执法经验复核阈值，并为变更提供理由和过渡条款。算力记录支持保障，却不能证明已发现所有未申报模型或秘密行动。

E.6 通过指令或条约采用

国家安全指令应明确发布机关及法律依据、适用机构和承包商、负责高管、资金、报告渠道和具备涉密许可的独立审查机构。示意采用顺序是：90 天内明确责任并清点受规制系统；180 天内提交差距评估和临时保障；12 个月内完成首轮代表性演练及整改计划。这些是拟议规划节点，须遵守该辖区合法权限和运行要求。

条约附件则应明确缔约国义务、同意范围、IFASA 与国家机关关系、受保护证据格式、审查周期、磋商权和争端程序。应说明国内实施如何覆盖相关私营主体，以及如何授权受控访问。任何条款都不应暗示可无限检查涉密系统，或对非缔约方拥有法律权力。批准、生效和执行安排须另行谈判。

审查记录。各参与机关应维护受保护登记，标明责任主体、触发活动、提交证据、决定、残余局限、整改期限和下次审查。安全允许时，公开汇总实施结果。独立审查者应检验控制是否改变真实操作，而不是仅生成申报。本附录支持可审计实施，但不会把不完整证据转化为普遍合规证明。