加速时代的治理

——高执行社会中 AI 系统的结构性约束

修订稿

摘要

人工智能正在从根本上改变当代社会中执行能力与协调能力之间的关系。随着 AI 系统在军事、金融、基础设施等领域的决策周期加速，制度适应明显滞后。本文将由此产生的不对称性形式化为 Δ(t) = E(t) − C(t)，其中 E 代表系统层面的执行吞吐量，C 代表分布式的协调能力。我们识别出 Δ 持续增长所带来的四种相互关联的结构性风险：加速差异风险、锁定与路径压缩风险、激励捕获风险、以及人类能动性侵蚀风险。借鉴加速理论、路径依赖经济学和制度适应文献，本文提出四种治理约束——有界加速、可逆性、路径多样性保护、以及保留人类方向性权威——作为结构性边界而非反应性控制。文章最后探讨了加速不对称条件下的国际协调可能性，引入“最小可行协调”概念作为全球共识的务实替代方案。这一框架将 AI 治理从技术遏制重新定义为结构性边界设计——一种聚焦于保持适应性和可纠正性而非控制具体输出的治理思路。

关键词：AI 治理，加速不对称，路径多样性，结构性约束，执行-协调差异，制度滞后，国际协调

---

1. 引言

人工智能正在快速改变当代社会的执行能力。从军事系统、金融市场到基础设施管理和科学研究，AI 技术使决策周期更快、优化强度更高、系统整合更广。公共和学术 discourse 主要将 AI 的治理挑战框定为对齐、安全测试、透明度和问责问题（Bostrom, 2014; Russell, 2019; EU AI Act, 2023）。这些框架固然必要，但它们往往将 AI 视为需要遏制或控制的离散技术产品。

本文认为，AI 的核心治理挑战是结构性的，而非纯粹技术性的。AI 同时在多个领域充当执行能力的倍增器。随着执行能力加速，协调机制——法律制度、监管框架、国际协议、社会共识过程——演化相对缓慢。由此产生的不对称性带来系统性压力，这是常规风险管理方法无法单独应对的。

本文的核心研究问题是：在 AI 驱动的执行能力快速扩张条件下，维持长期系统稳定需要什么样的结构性治理约束？

我们提出，治理必须应对四种相互关联的结构性风险：

1. 加速差异风险：当 AI 部署速度超过制度适应周期时，系统性弹性在纠偏机制激活之前就

已经被侵蚀。
2. 锁定与路径压缩风险：当 AI 驱动的优化集中化基础设施并减少可行的替代路径时，未来的适应性受到结构性约束。
3. 激励捕获风险：当竞争压力（企业、地缘政治、组织）主导长期韧性考虑时，局部理性选择汇聚成集体不稳定的结果。
4. 人类能动性侵蚀风险：当 AI 系统从决策支持转向决策替代时，人类方向性权威不仅在象征意义上、而且在功能上被削弱，系统性延迟缓冲被移除。

我们并不主张限制性的技术控制，而是主张有界的加速不对称和保持适应性弹性。本文整合了加速理论（Rosa, 2013; Virilio, 2006）、路径依赖经济学（Arthur, 1989; David, 1985）、制度适应文献（North, 1990; Ostrom, 1990）和 AI 治理框架（OECD, 2019; NIST, 2023; EU AI Act, 2023）。

文章结构如下：第 2 节综述相关治理和加速文献。第 3 节形式化执行-协调不对称。第 4 节识别结构性风险类别。第 5 节通过军事 AI 和平台集中化提供经验例证。第 6 节提出结构性治理约束。第 7 节探讨加速条件下的国际协调。第 8 节总结。

---

2. 文献综述

2.1 AI 治理框架

当代 AI 治理努力强调风险分类、透明度和人类监督。欧盟《AI 法案》采用基于风险的监管结构，将系统分为最低风险、有限风险、高风险和不可接受风险（EU AI Act, 2023）。该框架要求高风险系统进行合格评估、建立人类监督机制并履行透明度义务。NIST AI 风险管理框架提出持续监控、全生命周期风险管理和组织层面的 AI 风险实践整合（NIST, 2023）。OECD 原则强调以人为中心的价值观、问责、稳健性和透明度作为基础承诺（OECD, 2019）。

这些框架重点关注：

- 通过事前和事后评估减轻危害
- 开发和部署阶段的安全测试
- 通过治理结构实现组织问责
- 通过审计机制进行合规验证

这些方法虽然必要且有价值，但共享一个隐含假设：治理能力可以随技术部署逐步扩展。监管机构将获取专业知识；合规机制将适应；国际协调将逐步形成。

这一假设在高速加速条件下可能不成立。当执行能力呈指数级增长而制度适应呈线性推进时，无论治理努力多大，差距都在扩大。

2.2 加速理论

社会理论家已将加速视为现代性的一个定义特征。哈特穆特 •罗萨的综合性框架区分了技术

加速（交通、通信、生产中的有意提速）、社会加速（实践、结构、关系的变化速率）和生活节奏加速（行动片段的压缩）（Rosa, 2013）。保罗·维利里奥的“速度学”考察速度如何重组社会和政治空间，认为速度本身成为一种权力形式（Virilio, 2006）。

然而，大多数加速文献关注的是文化或经验后果，而非指数级执行增长下的结构稳定性。AI 引入的加速不仅是经验层面的，更是系统性执行吞吐量——社会技术系统改变物质、信息或战略状态的速率。这种从经验速度到执行吞吐量的转变，具有独特的治理含义。

2.3 路径依赖与锁定

经济学理论展示了递增收益和网络效应如何产生路径依赖（Arthur, 1989; David, 1985）。一旦主导轨迹出现，转换成本上升，替代路径在经济上或技术上变得不可行。布莱恩·阿瑟的形式化研究表明，在递增收益条件下，市场份额动态可能通过历史偶然锁定劣质技术。

数字平台以特别强烈的力度展示了这种动态。数据集中化提升算法性能，吸引更多用户，产生更多数据——这是一个产生基础设施锁定的反馈循环（Khan, 2017）。一旦生态系统主导地位巩固，互操作性下降，转换成本上升，未来适应性降低。

AI 系统通过学习效应放大了这些动态：更多数据产生更好的模型，吸引更多用户，产生更多数据。这一自增强循环以机器速度运行，压缩了治理干预的时间窗口。

2.4 制度适应

制度演化通常是渐进的（North, 1990; Ostrom, 1990）。监管适应周期以年或十年为单位计量，受立法程序、司法审查、利益相关者咨询和政治合法性要求的约束。埃莉诺·奥斯特罗姆的研究表明，成功的公共池塘资源治理需要时间进行规范发展、信任建立和制度完善。

AI 部署周期以月为单位计量。基础模型每季度获得新能力；自主系统每年整合到关键基础设施中；军事 AI 系统大幅压缩从开发到部署的时间线。

文献未能充分处理这种速度错配的治理含义。当制度学习速度低于技术演化速度时，治理将永久处于被动反应状态。

这一空白促使本文发展结构性框架。

---

3. 执行-协调不对称

3.1 定义

我们将执行能力（E）定义为：

社会技术系统改变物质、信息或战略状态的速率。

AI 通过以下方式同时在多个领域放大 E：

- 减少决策延迟
- 提高优化强度
- 实现大规模并行处理
- 整合先前分散的系统

我们将协调能力（C）定义为：

分布式行动者围绕稳定方向性决策达成一致的速率。

C 取决于：

- 制度过程：规则制定、执行、适应周期
- 法律审查：司法监督、监管批准、合规验证
- 政治审议：立法辩论、利益相关者咨询、民主合法性
- 国际谈判：条约形成、外交协调、争端解决
- 公共合法性：社会共识、信任、规范接受

3.2 差异

我们定义执行-协调差异：

$$\Delta(t) = E(t) - C(t)$$

当 Δ 持续增长时，治理问题出现。当 AI 减少决策延迟并提高优化强度时，E 快速加速。然而，C 受到程序性保障和合法性要求的约束，这些要求不能被压缩而不牺牲其功能。

无界的 Δ 产生三种系统性效应：

1. 部署超过评估：系统在第二阶效应被评估之前就已整合到关键基础设施中。
2. 制度修订滞后于技术转型：监管框架应对的是昨天的能力，而明天的系统已经部署。
3. 短期优化在长期审查之前累积：局部效率增益累积成后来难以修改的结构性配置。

3.3 可观测性与测量

Δ 无法直接测量，但可通过代理指标近似：

表 1：执行-协调不对称的代理指标

领域	执行代理指标（E）	协调代理指标（C）
军事	自主系统部署率（系统/年）	军控谈判时长（月）
金融	算法交易延迟（微秒）	监管更新频率（月/次更新）

平台 模型能力倍增时间（月） 标准制定周期（月）
基础设施 AI 整合速度（部署所需月数） 安全认证时间线（月）

多个领域的持续 Δ 增长表明存在系统性的加速不对称，需要结构性响应。这些指标是示意性的；实际应用需要根据具体领域进行校准。

---

4. AI 加速系统中的结构性风险类别

加速不对称通过不同但相互关联的结构性风险类别显现。这些风险是系统性的，而非单纯的技术失效。

4.1 加速差异风险

定义：加速差异风险指当 AI 部署速度超过制度适应周期，达到系统性弹性被侵蚀的程度。

机制：当 E 持续超过 C 时，系统在无更新治理框架的情况下长期运行。在此间隔期间，错误累积、外部性积累、路径依赖形成——全部未加纠正。

例证：

- 生成式 AI 在算法问责监管清晰之前快速整合到金融市场
- 自主车辆在无既定责任框架的情况下部署
- 实时算法决策替代刑事司法或社会服务中的审议性审查

关键洞察：加速差异风险不会立即产生灾难。相反，它侵蚀系统性弹性。系统继续运行，直到某次扰动揭示出纠偏机制已被绕过或压垮。

4.2 锁定与路径压缩风险

定义：锁定风险指当 AI 驱动的优化集中化架构，使可行的替代轨迹在结构上不可用。

机制：AI 系统表现出强烈的递增收益：更多数据 → 更好性能 → 更多用户 → 更多数据。这一反馈循环以机器速度运行，快速巩固市场和基础设施地位。一旦巩固，互操作性下降，转换成本上升，替代方法在经济或技术上变得不可行。

我们将未来路径多样性（W）定义为给定时间点上可行的结构性轨迹的有效数量。路径压缩发生在持续优化将 W 降至最小值时。

治理含义：锁定将治理挑战从监管调整转变为结构性固化。一个系统可能高效却不可修改；最优却不可逆转。

这一框架直接支持《公共窗口协议》（2026），该协议强调窗口完整性——多个真实的未来路

径保持可及的条件。当 W 降至阈值以下时，文明适应性将结构性受损，无论即时表现如何。

阈值考量：确定 W 的最低可行阈值需要情境判断。借鉴生态学中“最小可行种群”概念，我们提出 W 阈值应考虑：（a）能够维持替代轨迹的独立行动者数量；（b）以路径转换所需时间和资源计量的转换成本；（c）指示跨系统整合可行性的互操作性指标。虽然精确量化仍具挑战性，但趋势分析——W 是在增加还是减少——在绝对阈值达到之前提供了可操作的治理信号。

4.3 激励捕获风险

定义：激励捕获风险指当短期竞争压力主导长期系统性韧性考虑时，局部理性选择汇聚成集体不稳定的结果。

机制：AI 发展在竞争性领域内运作：

- 企业间争夺市场份额和人才
- 地缘政治间争夺战略优势
- 组织间争夺效率和绩效指标

这些激励 favoring 快速部署和规模积累。每个行动者都面临加速压力，无论总体后果如何。这在结构上类似于委托-代理问题、古德哈特定律和公地悲剧。

AI 特定强化：AI 的复合性能改进强化了快速部署激励，进一步扩大 Δ。先行者获得后来者无法弥补的优势，创造了无自然停点的竞赛动态。

4.4 人类能动性侵蚀风险

定义：人类能动性侵蚀风险指当 AI 系统结构性取代人类方向性权威，达到纠偏能力在功能上丧失的程度。

机制：AI 系统日益执行先前需要人类判断的任务。决策支持工具逐步过渡为决策替代系统，通过：

- 延迟压缩：AI 以超出人类反应能力的速度运行
- 复杂性不透明：AI 推理变得过于复杂，人类无法验证
- 制度萎缩：人类判断能力因不用而退化

关键区分：人类能动性侵蚀不是关于象征性的“人在回路”设计。它发生在：

- 监督变成名义而非实质性的
- 人类审查无法有意义地推翻自动输出
- 延迟压缩消除了审议性缓冲

人类权威的结构性功能：人类方向性权威充当系统性的延迟缓冲，保留纠偏能力。冷战期间，

人类授权层防止了自动化核升级——决策延迟是安全机制，而不仅是程序要求（Sagan, 1993）。

---

5. 经验例证

5.1 军事 AI 与延迟压缩

自主武器和 AI 辅助瞄准系统大幅压缩决策周期。OODA 循环——观察、定向、决策、行动——从人类认知时间线（秒到分钟）压缩到机器处理时间线（毫秒）。

当代发展：

- Project Maven（美国国防部）：AI 系统分析来自侦察无人机的全动态视频，规模超过人类分析能力。虽然目前是人类监督，但处理的数据量造成对算法分类的事实依赖（Scharre, 2018）。
- 哨兵武器系统：如韩国的 SGR-A1 等系统可以检测、跟踪并可能自主攻击目标，尽管据报道配置为人在回路要求。
- 致命性自主武器系统（LAWS）：多个国家正在开发能够自主选择并攻击目标而无需人类干预的系统。联合国关于 LAWS 的讨论自 2014 年以来持续进行，政府专家组定期开会，但尚未达成具有约束力的协议（UNIDIR, 2023）。
- 乌克兰冲突：双方都使用了 AI 增强的无人机和瞄准系统，尽管完全自主仍有限。这些部署为人类-AI 战斗协同提供了现实测试（PAX, 2024）。

经验模式：军事 AI 整合遵循一个轨迹：人在回路 → 人在回路上 → 人出回路。每次转变进一步压缩延迟，通过以下方式增加升级风险：

- 误报级联：AI 误判触发机器速度响应
- 承诺动态：对手面临压力，需预授权以匹配响应时间
- 使用或损失压力：AI 预测即将发生的攻击，使先发制人在局部理性

治理含义：人类授权层不仅是伦理保障；它们是结构性延迟机制，约束 Δ。移除它们就消除了无法通过技术替代的纠偏能力。

5.2 平台集中化与数据垄断

AI 驱动的数字平台表现出强烈的递增收益。数据集中化提升模型性能，吸引更多用户，产生更多数据。这一反馈循环产生基础设施集中化。

经验指标显示持续整合：

指标	2020 年	2023 年	来源
主要基础模型提供商数量	约 5 家	3-4 家	斯坦福 AI 指数报告 2024
前三名云 AI 训练市场份额	约 65%	约 72%	Gartner (2024)

公开模型权重可用性 普遍 减少 Epoch AI (2023)
高级模型 API 依赖 新兴 普遍 行业分析

锁定机制：

1. 数据网络效应：用户互动产生数据，提升模型性能
2. 算力集中：训练大模型需要少数组织才有的基础设施
3. API 生态系统转换成本：基于专有 API 构建的应用产生累积依赖
4. 人才集中：顶尖研究者集中在少数组织

路径多样性后果：

- W（未来路径多样性）自 2020 年以来已显著下降。替代轨迹——联邦学习、小型模型、开放权重生态系统、去中心化推理——在技术上仍然可能，但面临日益增加的结构性障碍。
- 平台间互操作性极小；为 OpenAI API 构建的应用无法轻易迁移到替代方案。
- 组织的转换成本随着依赖专有 AI 服务的年限而递增。

治理含义：互操作性要求、数据可移植性权利和开放标准并非反创新约束——它们是维持 W 高于最低阈值的结构性韧性措施。没有它们，无论治理偏好如何，锁定都将变得不可逆转。

---

6. 治理约束模型

在加速不对称条件下，治理必须超越反应性危害减轻，走向结构性边界设计。目标不是控制 AI 输出，而是约束 AI 部署所产生的系统性配置。

我们提出四种治理约束，每种针对特定的结构性风险。

6.1 有界加速

约束：AI 部署应通过约束 Δ 而非允许无界加速，来保持制度适应能力。

机制：

- 分级部署阈值：高影响力系统需分阶段推出，阶段间设评估期。阈值根据潜在后果定义（如“系统失效可能导致 X 人死亡或 Y 经济损失”）。
- 强制评估间隔：系统性 AI 整合需按制度周期而非技术周期定期重新评估。
- 有限规模监管沙盒：新应用在有限规模部署，为更广泛整合积累证据。
- 速度限制器：在关键领域（军事、金融、基础设施），可设定最高决策速率以保持人类监督可行性。

目的：选择性减缓 E 增长，使 C 能够适应。目标不是停止创新，而是防止 Δ 扩大到无法恢复的极限。

6.2 可逆性约束

约束：关键 AI 基础设施必须包含回滚能力。部署后无法逆转的系统引入结构性不可逆性。

机制：

- 模块化系统架构：组件应可分离，允许部分替换而无需完全重新设计。
- 可靠的终止开关机制：安全关键领域需要技术上可行的关闭能力，且不能被系统本身绕过。
- 定期重新授权要求：高影响力系统需定期更新，创造常规的重新考虑机会。
- 数据可移植性和模型导出权：组织保留迁移到替代提供商或独立部署模型的能力。
- 定期回滚测试：关键基础设施应定期测试逆转能力。

目的：通过确保当前配置不永久关闭替代方案，来保持 W。

6.3 路径多样性保护

约束：治理应通过促进多行动者生态系统和技术可替代性，将 W 维持在高於最低可行阈值。

机制：

- 开放标准要求：关键 AI 接口应遵守开放标准，防止专有锁定。
- 数据可移植性权利：用户和组织保留对 AI 互动所生成数据的控制权。
- 竞争性算力接入：政策确保多个提供商能够访问训练和推理基础设施。
- 支持联邦和分布式模型生态系统：使去中心化发展成为可能，而非集中式 API 依赖。
- 主导平台互操作性要求：使其能够与替代方案整合。

目的：多样性是不确定性条件下长期适应性的结构条件。当 W 降至过低时，无论当前表现如何，系统都将变得脆弱。

6.4 保留人类方向性权威

约束：AI 可以优化执行，但不得完全外部化最终方向性权威。关键战略决策仍由人类授权，并具有实质性有意义的监督。

机制：

- 保留的战略决策类别：保留由人类授权的决策类别（如武力使用、重大资源分配、宪法变革）。
- 实质性监督要求：人类审查必须在实质上可能，需要充分的审查时间、访问 AI 建议所依据信息的权限，以及无过高成本的否决能力。
- 保留审议缓冲：高后果潜力领域维持内置延迟——不能被算法压缩的最低决策时间。
- 认知自主性保护：确保人类决策者保持独立推理能力，而非仅仅认可 AI 建议。

目的：人类方向为 Δ 提供结构性边界，保留价值对齐和纠偏能力。

---

7. 加速条件下的国际协调

四种结构性约束面临集体行动问题：单边实施可能带来成本而无互惠收益。采用有界加速的国家若竞争对手不采用，可能在未获系统性稳定的情况下失去战略地位。传统的全球治理对 AI 加速而言过于缓慢。

7.1 协调困境

AI 加速削弱了经典国际协调支柱：

- 谈判周期（年）相对于 AI 部署（月）过慢
- 验证落后于快速演变的能力
- 先发优势奖励背叛
- 信任无法在能力变化快于关系变化时积累

结果是压缩的囚徒困境：个体加速产生集体不安全。每个行动者的局部理性选择（加速 AI 军事能力、避免透明、保持单边优势）汇聚成冲突概率增加和系统性稳定降低。

7.2 最小可行协调

我们提出最小可行协调作为全面全球协议的务实替代方案。它承认短期内无法实现普遍参与、完全验证和完美执行。相反，它专注于防止不可逆灾难，同时为未来谈判保留路径。

该框架有四个层次：

第一层：单边透明承诺

- 公开声明 AI 军事能力、部署状态、自主系统授权协议、触发回应的红线、以及不可逆行动预警程序
- 非互惠性（不依赖于他方遵守）
- 定期更新

第二层：多边建立信任措施

- AI 事件沟通渠道：针对可能引发升级的 AI 相关事件的快速沟通协议
- 联合升级情景模拟：合作模拟 AI 升级动态
- 透明基准：关于何为有意义的披露的商定标准
- 验证实验：有限规模的透明试验以测试可行性

第三层：背叛成本机制

- 公开合规登记：记录承诺和观察到的遵守情况
- 独立能力评估：第三方分析声明能力与观察能力的对比
- 后果信号：明确声明背叛如何影响未来关系
- 技术合作条件性：合作性 AI 开发访问与透明承诺挂钩

第四层：AI 中介协调渠道

- 预编程紧急协议：不同行动者的 AI 系统可为危机情况预设通信模式
- 自动验证算法：分析可观测数据以交叉验证声明能力
- 升级检测网络：共享监测指示即将发生冲突的指标

示例情景：考虑一个危机情景，两个核国家部署了 AI 增强的早期预警系统。第一层透明承诺将确立各方的自主系统协议。第二层建立信任措施将包括针对 AI 事件报告的预设通信渠道。第三层背叛成本将通过公开合规登记可见。第四层 AI 中介渠道可在毫秒内交换验证数据，防止雷达异常被误读。

风险考量：AI 中介协调引入新风险——验证算法可能被欺骗，危机中过度依赖机器判断，不同行动者的 AI 系统可能无法共享可解释协议。这些风险需要对 AI 中介渠道的人类监督和机器通信失败时的备用程序进行平行投资。

7.3 与混沌期共存约定的关系

最小可行协调践行了《混沌期共存约定》（Fu & Wang, 2026）的理念——该约定认为，在意图不透明、能力不对称的时期，当人类和 AI 都未实现统一主体性时，优先事项不是理想秩序，而是防止不可逆灾难，同时为未来谈判保留路径。四个层次实现了这一理念：单边透明使潜在合作行动者可被识别；建立信任措施创造稳定孤岛；背叛成本即使在没有信任的情况下也激励遵守；AI 渠道承认未来协调可能需要机器速度响应。

7.4 可行性评估

在全面协议不可行之处，最小可行协调是可行的，因为：

- 低参与门槛：任何行动者可独立参与，无论他人选择如何
- 验证是补充性的：透明是单边的；验证是补充而非前提
- 渐进可扩展性：机制可从少量开始，随着信任积累而扩展
- 与竞争兼容：行动者可保持竞争，同时在共同生存问题上维持有限协调

该方法并未消除加速不对称或其风险。它旨在防止最坏结果——不可逆冲突、完全沟通断裂、失控升级——同时让更稳健的治理结构得以发展。

---

8. 结论

8.1 核心贡献总结

本文论证了 AI 的核心治理挑战是结构性的而非技术性的。通过形式化执行-协调不对称（Δ = E − C）并识别四种相互关联的风险类别，我们将 AI 治理重新定义为结构性边界设计问题，而非技术遏制。

核心贡献：

1. Δ框架：分析加速不对称及其系统性效应的形式化语言，配有经验观测的代理指标。
2. W 概念：将未来路径多样性作为可测量的治理目标，将路径依赖理论与 AI 政策连接，并将最低可行阈值确立为监管目标。
3. 四种结构性风险类别：加速差异、锁定、激励捕获和人类能动性侵蚀——Δ产生系统性脆弱性的不同但相互关联的机制。
4. 四种治理约束：有界加速、可逆性、路径多样性保护和保留人类方向性权威——在加速条件下保持适应性的结构性边界。
5. 最小可行协调：在全面协议不可行条件下进行国际合作的务实方法，通过四个渐进的承诺层次实现。

8.2 理论含义与未来研究

这一框架开启了多个研究方向：

- Δ量化：开发跨领域Δ指标体系和加速不对称早期预警机制。
- W 测量：将生态多样性指数应用于技术生态系统，通过行动者数量、转换成本和互操作性指标测量路径多样性。
- 制度适应速率：比较不同司法管辖区和领域对 AI 部署的监管响应时间。
- 人类能动性侵蚀：追踪决策权威从人类向 AI 系统转移的实证案例研究，识别监督变为名义的临界点。
- 最小可行协调实验：在危机条件下测试四层框架的模拟和桌面演练。
- AI 中介协调风险：验证算法安全性、跨系统机器可解释性以及 AI 危机沟通的人类监督研究。

8.3 政策含义

对政策制定者，本文分析意味着：

1. 从输出控制转向结构性约束：与其试图预测和预防所有有害 AI 输出，不如设计保持适应性和可纠正性的系统。
2. 优先考虑可逆性：在关键领域，确保 AI 整合在必要时可被撤销——模块化架构、终止开关机制、定期回滚测试。
3. 保持路径多样性：支持开放标准、互操作性、数据可移植性和竞争性接入，作为韧性措施而非市场干预。

4. 保留人类延迟缓冲：在高后果决策中（军事、基础设施、危机响应），维持使人类判断成为可能的内置延迟，无论 AI 速度如何。
5. 渐进建设国际协调：从单边透明开始，扩展到多边建立信任措施，分层加入背叛成本，准备 AI 中介渠道——同时保持人类监督。

8.4 更广阔的框架

这一结构性模型与两个平行框架形成互补：

• 《公共窗口协议》（2026）阐明了保护未来路径多样性的规范性承诺——即我们所称的 W。我们的约束通过具体的治理机制实现了这一承诺。
• 《涌义宇宙论》（Fu, 2026）为路径多样性的重要性提供了形而上学基础：意义本身依赖于经验和表达的多元性。当 W 下降时，宇宙通过多样视角自我体验的能力收缩。

这些框架共同将 AI 治理定位为文明设计挑战：构建加速的技术系统，使其增强而非关闭人类的未来可能性。

8.5 最终反思

加速本身并非破坏性。快速执行使人类能够应对缓慢系统无法处理的挑战——气候变化、疾病、贫困。目标不是停止加速，而是确保协调能力在加速系统中保持相关性。

有界加速、可逆性、路径多样性和人类方向不是进步的约束——它们是可持续进步的条件。它们保护了《公共窗口协议》所称的“人类为自己做出决定的持续可能性”——即使当 AI 系统正在转变这些决策的执行方式时。

问题不是 AI 是否会加速执行。它会的。问题是治理能否从反应性控制演变为结构性设计，确保加速服务于而非从属于人类方向。本文为这一演变提供了一个框架。

参考文献

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. The Economic Journal, 99(394), 116–131.

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

David, P. A. (1985). Clio and the economics of QWERTY. American Economic Review, 75(2), 332–337.

Epoch AI. (2023). Trends in machine learning model openness. Epoch AI Research Report.

European Parliament & Council of the European Union. (2023). Regulation laying down

harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Fu, Z. (2026). Yongyi Cosmology 5.0: Meaning ontology and civilizational ethics under a layered integration framework.

Fu, Z., & Wang, Z. (2026). Chaos period coexistence compact: A minimal framework for human-AI coexistence under uncertainty.

Gartner. (2024). Market share analysis: Cloud AI developer services. Gartner Research.

Khan, L. M. (2017). Amazon's antitrust paradox. Yale Law Journal, 126(3), 710–805.

Maslej, N., et al. (2024). The AI Index 2024 Annual Report. Stanford Institute for Human-Centered AI.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

North, D. C. (1990). Institutions, institutional change and economic performance. Cambridge University Press.

OECD. (2019). OECD principles on artificial intelligence. OECD Publishing.

Ostrom, E. (1990). Governing the commons: The evolution of institutions for collective action. Cambridge University Press.

PAX. (2024). Slipping through the lines: The emerging use of autonomous weapons in the Ukraine war. PAX Report.

Public Window Protocol. (2026). Version 2.0, Engineering Edition.

Rosa, H. (2013). Social acceleration: A new theory of modernity. Columbia University Press.

Russell, S. (2019). Human compatible: Artificial intelligence and the problem of control. Viking.

Sagan, S. D. (1993). The limits of safety: Organizations, accidents, and nuclear weapons. Princeton University Press.

Scharre, P. (2018). Army of none: Autonomous weapons and the future of war. W.W. Norton & Company.

UNIDIR. (2023). The politics of autonomous weapons systems: Mapping national positions. United Nations Institute for Disarmament Research.

Virilio, P. (2006). Speed and politics. Semiotext(e).

Wiener, N. (1950). The human use of human beings: Cybernetics and society. Houghton Mifflin.

致谢：本文受益于《公共窗口协议》工作组的讨论以及与《涌义宇宙论》框架的对话。早期版本曾在斯坦福 AI 治理研讨会和牛津全球 AI 政策峰会上报告。文责自负。

通讯作者：[子君赋]

投稿日期：[2026 年 3 月]

字数：约 9,500 字
