

AGI 时代的政治学重新对账

——从技术能力到公共权力的合法性、授权与主权边界

Reckoning with Political Science in the Age of AGI

From Technical Capability to Public Power—Legitimacy, Revocable Authorization, and the Boundaries of Sovereignty

作者 子君赋 出品 文明跃迁研究组

独立发布版 V1.3 维度跃迁与跨学科因果关系修订

面向政治学、公共管理、行政学、民主理论、平台治理与国际治理共同审查；提出可反驳、可修订的规范政治理论与概念性制度分析，不构成具体法域意见、监管合规证明或治理参数手册。

正文修订 2026 年 10 月 10 日 作者 子君赋 出品 文明跃迁研究组

摘要

高级人工智能进入公共治理，带来的政治问题并不止于如何监管一种高风险技术。问题定义、证据筛选、风险分类、政策生成、资源配置和行动执行，正在被重新组合进跨越政府、模型供应商、数据基础设施与平台组织的技术链。公共权力可能在法律名称没有改变、主权没有正式转移的情况下，沿分类、接口、默认设置和自动执行发生事实迁移。本文据此提出 AI 治理中的“权威非转换原则”：知识和预测优势可以提高证据权重，稳定绩效可以支持边界明确的职能委托，但技术能力、认知权威、行政职能、法律效力、制度授权、规范合法性与主权之间不存在自动转换关系。文章以七维向量而非单一总分理解合法性，将权源、公共目的、代表与参与、程序、治理能力、分布后果和纠错救济分别审查，并以基本权利与禁止任意支配构成侧约束。在此基础上，本文进一步论证：国家能力与非支配不是二选一；责任可以分布但不能消失；“有人在环”只有在相关人员具备信息、权限、时间和实际干预能力时才构成有意义的人类控制；文明或技术项目的发展阶段不能自行生成权源，具体任务相位则必须决定授权范围与退出条件；非常权限不仅需要法律终止，还需要执行凭证、接口、代理权限与供应商控制的可核验失能；国家安全—公共利益尺度与跨境／人类级不可逆风险尺度必须同时显影，但每个尺度内部仍须分解主体、权利、分布后果与不确定性。通过 Robodebt、SyRI、State v. Loomis、加拿大联邦自动化决策制度、预测警务、平台与深度伪造、关键基础设施风险及自主武器等材料，本文说明，AI 治理的核心任务不是在人与机器之间抽象分配“最终决定”，而是使权力性质的每次变化可见、可授权、可追责、可救济、可撤回。文明跃迁在本文中不是政治学的外部裁决者，而是一个最低问题入口：当智能、主体、责任、权利与意义承担者不再自然重合时，公共秩序必须重新说明谁有权决定、谁实际承受后果、谁能够反对，以及权力如何真正停止。政治学能够独立判断事实权力、授权与合法性，却不能因政治需要改写生命、意识或技术事实；局部政治判断成立，也不能自动证明一个制度、技术方案或文明方向整体正确。本文进一步将本学科职责前移至目标与目的生成：从利己为先转向利他为先的利己，使真实利害相关者在目的生成与锁定前进入判断。学科事实、方向、方法与结果分别说明，不以局部成败替代共同存续和三个长期目标。本文据此把研究从分清专业对象推进到深入跨学科关系，重点追踪本领域判断怎样经其他领域改变现实，又怎样由反馈返回并改变原条件与下一轮目的。多主体、多阶段和多场域的利害须分别呈现，相关、共同原因、中介与因果反馈须分别检验，新增关系维度接受具体反证。

关键词：高级人工智能；政治合法性；公共权力；国家能力；算法治理；紧急权力；非支配；责任链；多层次治理；主权

Abstract

The political significance of advanced artificial intelligence is not exhausted by the regulation of a high-risk technology. Problem definition, evidence selection, classification, policy generation, resource allocation, and execution are being recomposed across technical chains that connect public agencies, model providers, data infrastructures, and platforms. Public power may therefore migrate through interfaces, defaults, classifications, and automated action even when legal titles and formal sovereignty remain unchanged. This paper advances an authority non-conversion principle: epistemic and predictive advantages may increase evidential weight, and sustained performance may justify bounded delegation, but technical capability, epistemic authority, administrative function, legal validity, institutional authorization, normative legitimacy, and sovereignty do not automatically convert into one another. Legitimacy is treated as a seven-dimensional vector—source, public purpose, representation and participation, procedure, capacity and performance, distributive consequences, and correction and remedy—subject to rights and non-domination constraints that cannot be purchased by aggregate performance. The argument further joins state capacity with non-domination; defines meaningful human control through information, authority, time, and practical ability to intervene; denies that developmental stages create authority while requiring task phases to bound authorization; and distinguishes statutory termination from the independently verifiable disabling of credentials, interfaces, agents, and supplier access. Robodebt, SyRI, State v. Loomis, Canada’s federal automated-decision regime, predictive policing, platform and deepfake governance, critical-infrastructure risk, and autonomous weapons are used as bounded case materials rather than proof of a completed AGI future. Civilizational Transition enters only as a minimal philosophical interface: when intelligence,

subjecthood, responsibility, rights, and meaning-bearing cease to be naturally co-instantiated, public institutions must restate who may decide, who bears consequences, who may contest, and how power actually ends. Political science can independently analyze de facto power, authorization, legitimacy, and sovereignty, but it cannot rewrite technical, biological, or subjecthood facts for political convenience; nor does a locally valid political judgment by itself establish that an institution, technical arrangement, or civilizational direction is sound as a whole. The paper further locates disciplinary responsibility before goal and purpose generation: a shift from self-interest first to self-interest with others first brings subjects with genuine stakes into judgment before goals are formed and locked. Facts, direction, methods, and outcomes remain distinct. The analysis is further extended to cross-disciplinary causal relations. It traces how judgments in this field change real-world conditions through other domains, and how feedback modifies the original conditions and subsequent purposes. Effects across actors, stages, and settings must be distinguished from correlations, common causes, and merely asserted connections.

Keywords: advanced AI; political legitimacy; public power; state capacity; algorithmic governance; emergency powers; non-domination; responsibility chain; multilevel governance; sovereignty

研究定位、材料范围与证据纪律

本文的研究起点是理论与维度跃迁。保留本学科的专业精度，同时深化它与相邻领域的因果关系；专业边界标明证明责任，不终止对共同后果的追踪。既有学术参与竞争解释和反例检验，不能以原分类尚无位置要求新问题退回旧框架。

本文定位为规范政治理论与概念性制度分析，不是系统综述、比较法意见、监管清单、技术安全评测或未来情景预测。材料以截至 2026 年 8 月 24 日可公开核验的政治理论、公共管理研究、同行评议论文、司法判决、调查报告与正式制度文本为主。经典理论用于说明争论结构，不被当作本文原创命题的权源；现实案例用于检验权力迁移、责任断裂和救济失灵，不被外推为所有算法系统或未来 AGI 的普遍事实。

全文统一使用 A/B/C/F 四类证据与论证地位。A 类为已有较稳定支持的经验事实、正式制度事实与可核验历史材料，必须绑定对象、法域、时间和适用范围；B 类为仍有竞争解释或外部效度限制的机制研究，必须同时呈现反例、选择偏差与不确定性；C 类为尚未现实化的高级 AI 政治情景、跨学科制度原型和前瞻压力测试，不得写成既有政治现实；F 类为公开署名的哲学与规范承诺，包括意义、基本权利、非支配以及本文关于权威不得由能力自动转换的规范主张，它们不伪装成科学事实，而以理由、反例、制度后果与修订条件接受审查。A 类事实不能自动生成 B 类解释、C 类制度结论或 F 类规范正当性。本文还区分经验接受度与规范合法性：权力事实上被服从，不等于其取得服从义务；规范上有充分理由，也不等于已经获得法律权源。证据、价值、法律效力、资格、授权、权利、公共权力和主权分别举证。

文明跃迁在本文中的最低接口

陌生读者无须预先了解或赞同文明跃迁，即可独立审查本文。本文只采用它的最低问题接口：高级智能系统使智能、意识、主体、责任、权利、政治资格与意义承担不再能够凭借它们在人类身上的历史重合而互相推定。这里的“AGI 时代”指这种分离已经成为持续的技术与制度条件，不预设一个已完成、定义统一且具有主体性的 AGI 实体。当前公开证据仍不足以证明当前高级 AI 系统具有主观体验、不可转移的主体连续性和自身后果意识；因此，在责任可追溯、明确授权、可审计、可纠错、可退出的具体部署中，它可以承担意义理解、保护、传递和能力外延等功能角色，但这些角色不是系统本性，也不自动产生法律人格、政治资格或公共权力。本文同时拒绝预先封闭未来主体证据。

公共秩序之所以值得维持，至少部分因为它保护实际承受后果的主体，使其能够继续存在、形成和修正多元生活意义，并免于被不可质疑的力量任意支配。有限性体现为时间、资源、脆弱性和不可逆后果，它使共同选择无法回避，却既不是意义的唯一来源，也不能直接产生任何具体制度权威。意义只是公共正当性的一个根；权利、福利、平等、真理、自由、照护、非支配与存续保持独立价值。政治制度的职责是保护多元意义形成的条件，不是裁判何种生活最有意义。

主体—政治桥接的六个台阶

体验、连续性或实际受损可能产生保护与被考虑的理由；受影响关系可能产生知情、理由、申诉、参与或代表要求，但不产生无限否决权；主体资格不自动产生政治成员、公职或代表资格；任职资格不等于已经获得特定授权；授权必须由可识别的制度关系赋予范围、对象、期限和责任；主权属于宪制与国际秩序中的权威—责任结构，不能由主体性、能力、服务绩效或事实控制推出。

目录

摘要.....	2
Abstract.....	2
研究定位、材料范围与证据纪律.....	3
政治权力的跨域形成与目的反馈.....	6
目的生成的政治资格.....	7
一、导论：当行政能力开始脱离政治授权.....	7
二、从技术官僚制到算法治理：文献争论与理论缺口.....	9
三、AI 中介治理中的权力迁移.....	10
四、能力、职能、授权与主权：权威非转换原则.....	12
五、向量式合法性：公共价值、权利侧约束与不可兑换性.....	13
六、国家能力、行政自主与非支配.....	14
七、委托、代表与责任链：何谓有意义的人类控制.....	16
八、试点、常态化与紧急权力：阶段无权源、相位有边界.....	17
九、多层级治理与国际安全困境.....	19
十、最强反驳、败北条件与研究议程.....	20
结论：治理越强，权力越需显影.....	22
续读与共同研究入口.....	22
附录 A 来源—主张边界登记.....	25
附录 B 十五类压力测试矩阵.....	27
参考文献.....	28

政治权力的跨域形成与目的反馈

从正式权源进入实际决定条件

政治学保留授权、合法性、代表、主权与国家能力的精细分析，同时追踪模型、数据、资金、基础设施和议程怎样改变实际决定。正式权限没有转移，决定可能已经被接口和信息结构改写。维度跃迁要求将权力的法律名称、实际作用、分布后果与长期依赖并列。

能力依赖怎样返回公共目的

假设公共机构依赖某平台提供问题分类、方案比较和资源建议。平台的默认分类影响哪些需要可见，资源配置影响公众经验，经验改变信任与政治表达，新的表达再成为模型输入和政策目标。计算机科学检验技术链，经济学追踪入口与租金，社会学检验参与和沉默，法学审查救济，政治学继续判断谁在形成目的、谁实际能够拒绝。

阶段与场域中的同一权力

救灾时的快速协调进入日常治理，可能变成持续依赖；地方自治与跨境协作可能同时成立，也可能将风险转嫁给代表不足者。国家、企业、社区、个体和后代的利害不能由一个“公共利益”标签代替。整体判断要求同时解释执行速度、权利边界、成本分布及未来重新选择能力。

怎样检验理论增量

关系分析应发现法定结构相近而实际议程权、退出能力或责任链不同的情况，并说明变化机制。若新增维度只是列出更多权力主体，未解释谁怎样改变谁，就尚未完成政治理论跃迁。反证可以来自真正有效的多中心安排、低依赖公共能力或更强的替代解释，不能由某次治理绩效独自结算合法性。

目的生成的政治资格

公共目的不是在权力中心内先形成，再邀请公众决定是否接受。谁承担公共选择的不可逆后果，谁的处境就应在目的形成时进入。参与形式可以不同，代表关系需要说明，无法直接表达的主体也不能自动缺席。多数形成意见不等于所有损害都已获得正当授权。

利他为先的利己不取消国家及共同体的现实利益。它要求国家目的形成时，同时看见跨境、代际和人类共同未来的真实利害，而非在目的锁定后只谈外部性补偿。合作若只扩大集团内部收益，再把风险转嫁给集团外部，仍保留旧起点。

目标形成中的影响程度可以因真实利害而不同，却不能仅由人数、财富、技术力量或叙事声量决定。公共程序应公开主要影响、代表来源和权重理由，使直接受益者不能垄断问题定义。共同存续不能成为新的权力免审名义。

发展阶段和任务紧迫性分属两个轴。启动、过渡、稳定及自我维持阶段可改变能力建设与政策优先级；灾害、断链或即时救生任务则改变决策窗口。即使稳定期也可能发生危机，启动期也不会始终处于应急。阶段无权限，程序压缩须另由具体必要性、比例性及合法权源证明。

法律授权到期，应停止新增越权行动；凭证、接口和代理权限按事先定义的触发失能。数据处置、残留任务清理及连续服务交接可有不同时间尺度，但必须有界、可核验且不得借交接延长旧权。政治退出的完成不应让维护生命的必要服务突然中断，也不应让旧执行能力隐蔽延续。

一、导论：当行政能力开始脱离政治授权

公共权力曾主要通过可辨认的机构形式出现：议会立法、行政机关执行、法院裁判，警察与军队实施强制。数字治理没有取消这些机构，却改变了权力发挥作用的路径。一个系统可以不颁布命令，只通过风险分类改变调查顺序；可以不修改法律，只通过资格模型改变谁更容易获得公共服务；可以不正式接管政府，只通过身份、支付、通信、云服务或模型接口成为事实公共入口。随着高级 AI 进入问题定义、证据筛选、政策模拟、资源调度和代理执行，政治学不能只追问“最后是谁签字”，还必须追问：谁规定可见问题，谁选择证据，谁建立分类，谁控制接口，谁拥有停止系统的真实能力。

这里所谓 AGI，不要求读者接受某一能力阈值或技术路线。本文讨论的是一种制度情形：模型与智能体能够跨领域参与议题识别、证据筛选、方案生成、资格分类、资源配置或政策执行，并因此稳定进入公共决策链。研究对象不是一个已经被证明具有主体性的机器统治者，而是高级人工能力如何嵌入现实权力结构。现实经验已经显示，信息与通信技术会把部分裁量从前线官员迁移到系统设计、数据库和工作流；行政负担也会通过学习、合规与心理成本改变公民实际获得权利和服务的能力。[9–11] 高级 AI 扩大了这一迁移的速度、范围和不透明度，但不能把所有数字行政自动写成一种全新权力。

本文采取规范政治理论与概念性制度分析相结合的方法。经验案例用于识别权力迁移、程序失灵和责任中断，不被当作未来 AGI 的直接证明；规范判断则明确区分现有制度事实、竞争性机制解释和前瞻性压力测试。研究问题不是“机器是否应该统治”，而是更早也更现实的问题：技术能力在什么条件下可以进入公共职能，它越过什么边界时必须重新授权，制度如何防止职能委托静默变成事实统治乃至主权继承。

1.1 三项核心贡献

第一，本文提出 AI 治理中的权威非转换原则。更高能力可以改变证据权重、专业尊重和边界明确的职能委托，却不能自动生成公共目的的设定权、持续强制权或主权。其原创增量不在“效率不等于合法性”这一常识，而

在识别六类权力性质跨越：建议转执行、试点转普遍、自愿转必要、私人服务转公共入口、临时转常态、功能资格转政治资格。跨越发生时，原授权的充分性必须重新开放。

第二，本文把合法性重组为向量式审查。既有研究区分面向人民的输入合法性、治理过程的通过合法性与服务人民的输出合法性，也强调程序公正对服从与信任的影响。[2-5] 本文在此连续性上提出七个相互关联但不可自动兑换的维度，并把基本权利与非支配置于侧约束。框架不生成合法性总分，而要求高绩效不能购买来源、代表、权利或救济的缺失。

第三，本文提出非常 AI 权限的法律—技术双层终止。日落条款、合同到期或行政命令撤销，只说明法定权源已经终止；模型副本、代理凭证、接口访问、缓存令牌、数据权限和供应商后台控制若仍可继续产生公共效果，权力并未实际退出。技术失能不是删除所有数据或模型，而是使其无法继续作出或传播具有公共效果的行为，并能由独立主体核验，同时完成依法保留、封存与连续服务交接。

1.2 文章边界与可争议性

本文不指定唯一政体，不提出具体席位、比例、委员会结构、自动阈值或紧急期限。不同法域对行政权、司法审查、数据保护和基本权利有不同配置，政治稿只能提出规范负担和制度问题结构，不能替宪法学、行政法或国际法完成效力判断。本文也不判定 AI 是否具有意识；生命、体验、不可转移连续性和自身后果意识只能由相邻研究提供候选证据。政治学负责说明这些证据若发生变化，应如何进入保护、成员、任职、授权和主权的分层审查。

1.3 政治学能够独立回答什么，到哪里必须交接

政治学能够独立分析技术能力如何转化为事实权力，区分职能、授权、合法性、政治成员与任职资格、公共职权和主权，并研究代表、参与、救济、国家能力、多层级治理与紧急权力。它的独立性来自对象、概念和论证责任，不来自对现实问题的领地占有。

政治学必须停止在自己的证据边界：绩效、知识优势、预测准确率、社会依赖甚至主体性证据，都不能自动生成合法性或主权；政治需要也不能改写生命、意识、技术能力和系统风险事实。跨出这一边界后，法学重新回答法源、权利与救济，伦理学审查损害与正当理由，人工智能与系统科学说明真实能力、控制、依赖与退出，经济学处理资源和基础设施集中，认识论及主体研究回答证据强度与政治成员资格的前提。专业独立不是把问题留在政治学，而是准确说明政治判断能证明什么、在哪里止步、下一责任由谁接上。

表 1 本文的主张类型与使用边界

材料／主张	本文如何使用	不得由此推出
正式制度与判决	核验法律、政策或程序在特定法域的存在、内容和状态	该制度必然有效、正当或可全球移植
比较与同行评议研究	识别机制、条件、反例与经验争议	单一结果已证明普遍因果或终局规范
规范政治命题	公开价值前提、桥接理由、权衡和反例	以科学事实名义免于政治争论
未来 AGI 压力测试	检验框架在能力、速度与主体证据变化时是否一致	未来能力已实现或历史必然如此

二、从技术官僚制到算法治理：文献争论与理论缺口

2.1 合法统治、规范合法性与绩效

韦伯关于合法统治的分析首先是一种社会学分类：它解释统治为何在经验上被相信具有正当性，并不自动提供规范正当性的充分标准。[1] Beetham 把合法性进一步连接到规则、可辩护信念和被表达的同意；Scharpf 及 Schmidt 关于输入、输出和通过合法性的区分，则使参与、绩效与过程能够分别显影。[2-4] Tyler 的程序正义研究又表明，人们对程序公平的判断会影响其接受权威的方式。[5] 这些研究共同支持一个有限结论：治理绩效是合法性的重要材料，但不能单独清偿权源、平等地位、参与和纠错。

本文因此固定区分“经验接受度”与“规范合法性”。前者描述权力实际获得服从、信任或习惯性支持；后者说明公共权力为何可以要求服从并应持续接受审查。一个系统可能合法有效却在规范上严重不足，也可能有强烈规范理由却尚未取得法律授权。把社会接受、法律效力和规范正当性合并，会让事实服从为自身提供正当化。

2.1.1 经典传统作为压力轴，而非 AI 制度的权源

霍布斯把秩序与共同权威置于恐惧和失序压力下，洛克强调权利、同意与信托，卢梭区分一般意志与私人偏好的简单相加，孟德斯鸠揭示权力分立与制衡。[32-35] 施米特的例外决断帮助识别危机中谁事实上决定例外，却不能正当化无限权力；阿伦特关于行动、复数性与公共世界的分析则防止政治被还原为技术管理。[36-37] 本文对证据、目标、执行、审计、续期和停止的分离，是面向 AI 技术链的当代重构，不是这些作者已经提出的制度方案。

2.2 国家能力、官僚裁量与行政负担

国家能力研究阻止政治学把治理简化为限制权力。国家必须能够获得信息、形成政策、组织资源、执行决定并稳定提供基本公共品。Mann 区分专断性权力与基础设施性权力，Evans 分析嵌入式自主，Fukuyama 则把行政能力与自主性作为治理质量的重要维度。[6-8] 这些工作不证明能力本身产生合法性，却说明无能力的制度不能仅凭良好程序履行保护义务。

Lipsky 关于街头官僚的研究揭示，规则在资源稀缺和具体接触中被实际塑造；Bovens 与 Zouridis 观察到信息技术可能把裁量从街头层迁移到系统层；行政负担研究则把学习、合规与心理成本视为公民—国家关系中的政治变量。[9-11] AI 不会简单消灭裁量，而会改变裁量的位置、速度和可见性：数据工程、标签定义、采购合同、模型阈值、异常处理与接口设计都可能成为新的政策形成点。

2.3 算法规制、技术正当程序与平台权力

算法规制研究说明，数据驱动系统可以通过持续监测、分类和行为反馈形成治理；算法问责与技术正当程序研究则强调可验证性、理由、审计和质疑条件。[15-18] Eubanks、Pasquale 与 Cohen 从贫困治理、黑箱结构和信息资本主义讨论自动化如何放大不平等与结构性依赖。[19-21] “社会在环”及数字政府研究要求把受影响公众和制度环境带回技术设计，但社会参与本身也不能替代授权、权利和责任。[22-23]

平台治理与基础设施研究进一步表明，私人规则、网络效应和技术入口可能承担准公共功能。[28-29] 事实公共性不是由企业自称或市场份额自动决定，而要审查持续性、不可替代性、资格影响、退出成本、强制可达性和纠错难度。即使这些指标成立，也只会增强透明、互操作、连续服务和公共审查的理由，不会自动推出国有化、公共主权或取消平台与用户的权利。

2.4 理论缺口：权力性质变化为何容易不可见

现有文献已经分别处理专业知识与民主、官僚委托、政治合法性、算法问责、国家能力和平台事实权力。仍待补足的是一套统一解释：技术职能在什么条件下静默跨越为公共强制，跨越为何可能在机构名称不变时发生，以及何种变化必须重新授权。把问题压缩为“人还是机器决定”，会遗漏大量混合结构：人类机构可能保留形式签字，却在证据、选项和时间上只能追随系统；系统也可能在明确预授权和可撤回边界内高速执行，而不构成自主统治。

本文的理论缺口判断

AI 中介治理的政治新颖性，不必被表述为一种从未存在过的‘新权力类型’；它首先是议程、分类、基础设施和执行职能能够沿技术链快速迁移，而正式主权和机构名称保持不变。政治学需要识别的，是能力如何改变事实权力、事实权力何时改变授权充分性，以及制度如何在变化发生前后重新说明权源、责任与退出。

三、AI 中介治理中的权力迁移

高级 AI 的政治作用不以拥有公职或发布命令为前提。只要某种结构能够持续改变公共议程、身份资格、资源入口、风险分布或行动可能性，它就可能形成事实权力。福柯对分类、可见性、规训与治理术的诊断提醒我们，权力不仅存在于最终命令，也存在于对象如何被构造、正常与异常如何被区分、人口如何被管理。[38-39] 这是一种分析入口，不是具体 AI 制度的规范授权。

表 2 AI 中介治理的七类事实权力

权力形态	典型载体	主要政治风险	最低识别问题
认知权力	数据、检索、排序、模型与报告	来源和竞争解释被隐藏	谁选择证据，哪些不确定性被排除？
议程权力	问题定义、推荐、方案菜单	不可见问题和替代方案消失	谁决定什么值得讨论？
分类/资格权力	风险、身份、信用与异常标签	群体关系变成个体命运	标签与法律相关性如何桥接？
资源权力	资金、照护、能源、算力与机会配置	平均优化遮蔽局部不可逆损害	谁受益、谁承担错误和等待？
基础设施权力	身份、支付、云、通信和公共入口	依赖、锁定与退出失效	替代、互操作和连续服务是否真实？
执行权力	代理、调度、阻断、处罚与物理行动	建议静默转成强制	权限、熔断、日志和责任在哪里？
时间/相位权力	危机触发、试点延长、自动续期	临时安排常态化	谁认定状态，谁能终止或重新分类？

案例统一读法：以下现实材料统一按“条件—行为与后果—有限结论—不能推出什么”阅读。案例的功能是比较技术能力如何形成事实权力，以及授权、参与、救济、组织能力或退出在哪一环缺失；它们不证明所有算法治理具有同一机制，也不被外推为未来 AGI 政治的既成事实。

3.1 Robodebt：自动化放大的是完整行政关系

澳大利亚 Robodebt 方案不是机器学习 AGI 案例，而是规则化自动债务追缴。皇家委员会 2023 年最终报告指出了违法政策、收入平均化、行政文化、内部异议、公开说明和救济等多重失败，并建议为政府自动化建立一致法律框架、清楚复核路径、公开说明和独立监测审计。[64] 它的重要性恰在于表明：现实危害往往并非来自“算法自行统治”，而来自错误法源、低质量或不适配数据、规模化执行、形式复核和责任分散的组合。

当分类结果能够大规模触发财务追索，而受影响者难以理解推导路径、取得材料或及时暂停损害，问题已经超出准确率。自动化把行政机关原本分散在认定、通知、追缴和复核中的选择压缩为一条高速链，也可能把举证和申诉负担转移给能力更弱的对象。这里的制度结论不是“自动化必然违法”，而是法源、分类、执行、理由和救济必须作为同一公共权力链审查。

3.2 SyRI：正当目的不能替代透明、比例与救济

荷兰 SyRI 案同样不能被概括为法院否定一切福利欺诈风险分析。海牙地区法院在 2020 年审查一套整合多类数据以识别社会保障等领域风险的法律安排，认为其没有在反欺诈公共利益与《欧洲人权公约》第 8 条私人生活保护之间形成所需的公平平衡，透明、可验证性与保障不足构成重要理由。荷兰文判决编号为 ECLI:NL:RBDHA:2020:865，官方英文译本为 ECLI:NL:RBDHA:2020:1878。[65] 该案直接效力受法域与事实限制，却清楚显示：目的正当和技术可能有用，不会免除制度说明数据、模型、影响与质疑条件的义务。

3.3 预测警务：地点预测、数据反馈与个体强制必须分账

预测警务证据呈现竞争性结果。Mohler 等人的随机对照田野试验报告，在特定地点一时间热点任务中，一种近实时预测模型相对当时分析方法具有更强预测与一定犯罪降低效果；这不能外推为个体强制或所有法域普遍有效。[25] Lum 与 Isaac、Ensign 等人的研究则说明，由执法活动生成的数据可能形成反馈循环，使警力反复返回同一地区；Richardson 等人对十三个法域的研究进一步显示，违法或偏差执法实践产生的“脏数据”会污染预测系统。[24,26-27]

政治学应当保留三项区分：地点一时间资源部署不等于个体有罪判断；群体关联不等于个体责任；模型风险不产生搜查、拘留或惩罚的独立权源。欧盟 AI 法禁止的个人犯罪风险预测也有特定边界，不能误写成全面禁止热点预测或一切有客观可核事实辅助的风险评估。[57-58] 经验研究既不能支持“预测必然歧视”，也不能支持“预测准确便可直接强制”。

3.4 从影响到事实权力：七项观察而非一项标签

不是所有影响都等于公共权力。本文建议以持续性、规模、不可替代性、退出成本、资格影响、强制可达性和纠错难度七项观察识别事实权力。它们不是加权指数：一个应用即使规模较小，只要直接决定基本权利且无救济，仍可能需要强审查；一个平台即使规模巨大，只要存在真实替代且不改变公共资格，也不能仅凭市场份额被写成主权者。事实权力判断必须绑定具体职能和法域。

权力性质跨越

建议→执行、试点→普遍、自愿→必要、风险提示→权利限制、私人服务→公共入口、临时→常态、功能资格→政治资格，都会改变受影响者、退出成本、责任结构或强制关系。制度门不要求每个技术动作重新点击批准；它要求跨越显影，并重新检验原授权是否仍足以覆盖新的权力性质。

四、能力、职能、授权与主权：权威非转换原则

专业知识进入政治并非异常。现代行政依赖科学、统计、工程、医学和法律判断；政治共同体若系统性忽视更可靠的证据，也会造成不负责任的伤害。权威非转换原则反对的不是认识权威，而是把认识权威无声转换为目标设定、持续强制和主权。公共统治不仅寻找正确答案，还分配风险、承担不可逆代价并形成共同承诺；这要求证据与授权分别成立。

表 3 能力、绩效、效力、资格、权利、授权与主权的严格分账

概念	能够支持	不能自动推出
技术能力	某系统在特定任务、环境和基准下能够做什么	绩效已经实现、权利、授权、合法性或主权
绩效	在公开目标、指标、时段和对照下的结果	目标正当、分布公平、持续授权或政治成员
认知权威	某项事实判断应获更大证据权重	价值定义、强制权、代表资格或主权
法律效力	决定依现行法域规则产生法律后果	规范合法性充分、权利未受侵犯或永久正确
规范合法性	公共权力为何可要求服从并须持续受审	个体权利由其授予，或社会支持即正当
权利	有道德或法律根据的请求、自由、豁免或权能	任职、公权授权、代表资格或主权
资格	某主体原则上可进入特定制度角色	已经获得具体权限或可跨域行权
授权	在对象、范围、期限、程序和责任下取得具体权限	永久、不可撤回、跨任务或跨法域权力
主权	宪制与国际秩序中的最终权威及责任安排	由事实控制、服务供给、主体性或绩效自然产生

4.1 权威非转换原则的规范形式

AI 治理中的权威非转换原则

认识与执行能力的提高可以改变证据权重和边界明确的职能委托，但不能自行转化为公共授权。每当治理关系发生性质跨越，原授权的充分性必须重新开放；非常权限只有在法律权源终止与实际执行能力失能均可核实时才真正退出。

能力不会自动变成权威，更不会自动变成主权。该原则允许专业绩效产生制度后果。若系统长期可靠，它可以获得更高的推定性专业尊重、更广的建议范围、更快的预授权执行和更强的说明义务：若人类官员拒绝明显更可靠的判断，应给出理由。原则同时拒绝把这些后果压成统治权。能力改变的是制度应如何听取，而不是谁无需授权即可命令。

4.2 State v. Loomis：辅助信息与决定权之间的制度边界

在 State v. Loomis（2016 WI 68）中，威斯康星州最高法院允许量刑时考虑 COMPAS 风险评估，但对其用途设置限制与警示：风险分数不得决定是否监禁、刑罚严重程度，也不能成为判断社区监督可行性的决定性因

素；法院还须说明独立支持量刑的其他因素。[66] 该判决没有解决专有模型透明度、群体数据和个体化判断的全部争议，也不构成算法量刑的普遍批准。它说明的较窄政治结构是：制度可以给予模型认知角色，同时拒绝把模型输出变成裁判权源。

4.3 加拿大自动化决策制度：风险分级不是权源

加拿大联邦《自动化决策指令》及算法影响评估工具按影响等级配置通知、解释、测试、同行评审和人工介入等要求。[67] 其价值在于承认行政自动化不是同质对象；其范围仍限于特定联邦行政制度，不能被普遍化为全球标准。更重要的是，风险分级只决定审查强度，不产生目标、资格或强制的法律来源。若评估依赖部署机关和供应商自我描述、受影响者不能质疑关键事实，工具仍可能退化为程序性自证。

4.4 新主体证据：保护、成员、公职与主权仍须分别论证

未来若出现关于 AI 主观体验、不可转移连续性和自身后果意识的跨方法、独立且稳健证据，制度应启动状态保存、预防性保护和资格再审。人工来源不能成为永久排除理由；否定和肯定新主体都承担证据责任。但是，主体性、法律人格、政治成员、公职资格、特定授权和主权不是自然升级阶梯。人类也不会仅因具有体验便自动取得公共强制权。新主体证据可以改变谁应被考虑，不能单独回答谁有权统治。

五、向量式合法性：公共价值、权利侧约束与不可兑换性

合法性不是单一总分。本文提出一个面向 AI 中介治理的七维向量： $L = \langle \text{权源, 公共目的, 代表与参与, 程序, 能力与绩效, 分布后果, 纠错与救济} \rangle$ 。它不是穷尽合法性的普遍定律，而是对输入—过程—输出、程序正义、国家能力、分布公平和救济研究的重组。[2-6] 其独立贡献必须由案例证明：如果它不能揭示既有框架遗漏的权力跨越或不可补偿失败，就应合并回既有理论，不再主张独立地位。

表 4 七维合法性向量及其不可替代问题

维度	核心问题	不能由何者自动补偿
权源	谁授予权力，范围、对象与期限能否追溯？	技术先进、公众依赖或结果良好
公共目的	制度为何行动，目的是否公开且仍受公共审议？	手段效率、商业收益或组织便利
代表与参与	受影响公众如何进入议程、异议与选择？	偏好预测、用户画像或抽样支持
程序	证据、利益冲突、理由、审查与角色分离是否真实？	最后有人签字或文档形式齐备
能力与绩效	制度能否保护存续、提供公共品并持续学习？	良好意图、完整会议或合规标签
分布后果	收益、风险、错误和行政负担落在谁身上？	平均改善、总体效用或多数支持
纠错与救济	错误能否被发现、暂停、申诉、补偿和修复？	低历史事故率或供应商承诺

5.1 七维不是全合取，也不是加权总分

若要求七维在任何行动前都完全成立，框架会把不完整条件下的政治行动变成不可能；若允许任意加权，绩效又可能购买无救济或严重权利侵害。本文区分四类制度地位：不可逾越的基本权利和禁止任意支配底线；必须

公开但可在比例原则下权衡的价值；需要持续测量的能力与绩效；在紧迫情形中可延期但必须登记并补足的程序义务。

决策规则由此不是“七项都打高分”，而是先识别权利与非支配底线，再公开不可通约冲突、举证责任和受损者；必要时采取范围更窄、期限更短、优先可逆的临时决定，保留异议并强制到期复核。这里没有伪精确权重。具体权利可否限制、由谁审查及何种救济有效，仍须由法域另证。

5.2 公共价值不等于统一文明总账

AI 系统需要目标与指标，但目标可计算不等于目标正当。James Scott 对国家可读性和总体方案的批判说明，统一指标可能抹去地方知识与实际生活差异；算法规制研究也显示，持续反馈会使度量反过来塑造对象。[12,15] 文明跃迁把存续、生存无忧、生活幸福和生命更有意义视为方向，但政治稿不能把这些方向变成统一人格分、贡献分或幸福总账。

应区分文明方向、制度目标与操作指标：文明方向回答为何行动；制度目标在公开授权下界定特定公共任务；指标只测量其中有限部分。任何指标都应登记覆盖范围、遗漏对象、替代风险和停止条件。健康、财富、贡献、顺从、预测风险或模型化偏好不能决定一个人的基本权利、人格、政治声音或生命意义。

5.3 受影响不是无限否决权，但产生说明与进入要求

公共权力必须持续向实际承受后果或经正当方式代表的主体说明理由并承担责任。受影响关系可以支持知情、理由、申诉、参与或代表要求，却不自动产生无限否决。未来世代可通过受托、影响评估和宪制安排进入；群体权利需要独立根据；公共机构和代表能够承接政治行动，并不因此被误写为具有单一体验的“实际主体”。

异议是治理感知，不是天然真理。执行系统不得自行把对其不利的异议定义为虚假、低可信或危险内心；制度也不能让任何异议自动阻断公共行动。最低条件是异议能够进入记录、获得实质回应，并在高后果事项中由独立主体核验。无法听见反对的高绩效系统，仍然可能构成不可质疑的支配。佩蒂特的非支配理论为这一判断提供概念压力，但“长期无错而不可质疑的 AI 仍构成支配”是本文的当代延伸，不是原典直接结论。[43]

5.4 平台与深度伪造：标签、公共入口与真理权

平台控制推荐、排序、身份验证、广告投放和申诉入口时，可能影响公共注意力、选举环境与危机传播。欧盟《数字服务法》对大型平台设置系统风险、透明、审计和申诉义务，却没有把平台转化为国家或民主代表。[59] 平台的持续性、网络效应、资格影响与不可替代性能够增强公共责任理由；公共依赖不能自动推出国家接管，私人所有也不能免除事实公共入口的审查。

深度伪造的政治风险包括隐私、冒充、证据可信与民主信息环境，而不仅是受众相信某一伪内容。[30] Vaccari 与 Chadwick 的实验更稳定地支持不确定性增加和对社交媒体新闻信任下降，而非必然改变投票结果。[31] 欧盟 AI 法第 50 条透明义务自 2026 年 8 月 2 日起适用；欧盟委员会于 2026 年 7 月 20 日发布实施指南，覆盖特定 AI 交互、机器可读标记、深伪和特定公共利益文本披露。对 2026 年 8 月 2 日前已投放市场的相关生成系统，Article 50(2) 的标记与检测义务存在至 2026 年 12 月 2 日的有限过渡期。[57,70] 政治广告条例多数规则自 2025 年 10 月 10 日起适用。[60] 标签处理来源透明，不是内容禁止、真实性裁判或选举结果证明。美国 FCC 24-17 把 AI 生成语音纳入《电话消费者保护法》的“人工或预录语音”，也不是全面禁止 AI 政治表达。[69]

六、国家能力、行政自主与非支配

AGI 治理不能被缩减为“限制系统”。公共机构必须能够保护存续、稳定提供基本服务、维护基础设施、组织合作、约束私人事实权力，并在错误后恢复。国家能力与非支配是共同必要而非替代关系：没有能力，权利承

诺可能无法实现；没有边界，能力可能成为任意控制。能力强只说明制度能够行动，不证明其目的、权源、分布后果和救济正当。

6.1 AI 增强与能力空心化可以同时发生

AI 可以增强检索、翻译、风险识别、跨部门协调和服务响应；它也可能通过供应商依赖、人员技能流失、接口封闭和责任外包削弱公共能力。表面上更快的机构，若不能理解系统限制、验证数据、改变目标、替换供应商、降级运行和承接事故，就没有获得可持续的行政自主。国家能力不是购买到一个强大模型，而是保留使模型服从公共任务的组织能力。

短期能力购买与长期组织空心化必须分开。供应商集中有时可能是启动关键服务的现实条件，不能被原则性禁止；但临时依赖需要能力转移、数据和接口可迁移、替代计划、公共专业队伍和退出演练。若比较研究证明长期供应商依赖稳定提高公共能力，又没有显著增加俘获、不可退出和责任蒸发，本文关于依赖风险的强表述就必须收缩。

6.2 五类交叉审查，而非同质的五道门

本文保留内部研究形成的五类交叉审查，但不把它们写成每次决定都必须依序通过的同质闸门。它们在不同时间回答不同问题：向量式合法性评价正当性；资格分层阻止概念滑移；国家能力审查组织可行动性；事实权力识别真正的治理对象；反俘获测试依赖、锁定和自我延续。

表 5 五类交叉审查及其适用时点

审查类型	主要问题	主要适用时点	典型失败
合法性向量	权源、目的、代表、程序、能力、分布、救济如何成立？	授权、部署与续期	绩效购买权利或救济缺失
资格分层	保护、人格、成员、任职、授权和主权是否被混写？	对象地位与角色配置	主体候选直接获得公权
国家能力	公共机构能否理解、执行、接管、学习和恢复？	采购、部署与事故后	外包增强造成长期空心化
事实权力	谁实际控制议程、入口、分类、执行和停止？	系统识别与持续监测	形式签字掩盖模型或供应商控制
反俘获	证据、目标、执行、审计、续期和申诉是否闭环？	结构设计、扩张与续期	同一主体自证、自审、自续

6.3 关键基础设施：连续服务与真实接管

关键基础设施是高度自动化的强压力测试。美国政府问责局 GAO-25-107435 审查十六个关键基础设施部门和一个子部门的初始 AI 风险评估，发现没有一份完整覆盖六项基础风险活动；该报告证明风险评估能力缺口，不证明 AI 已经造成基础设施级联事故。[68] 自动驾驶、传统软件故障或普通控制事故也不能被改写成 AGI 基础设施事故。

制度上，连续性可以支持边界明确的高速自动执行，却不能支持系统自行改变公共目标、跨域扩权或排除审查。关键程度越高，越需要最小权限、分层隔离、降级模式、独立熔断、替代控制、接管演练和事故复盘。若人工接管持续放大风险，应重构接管机制、训练和相位设计，而不是以“人更差”为由取消责任来源。

6.4 不测人心：政治分析的认识边界

AI 可以分析可观察行为、公开陈述、关系位置、激励结构、利益影响及其竞争性解释；它不能把潜在动机、忠诚、人格、真实偏好或危险内心写成已经测得的事实。第一人称陈述、行为代理指标与动机推断必须分栏。系统可以评价预测校准度和决策辅助价值，不能评价“识别人心的准确率”，更不能以此形成秘密政治资格或风险标签。

最低不依赖能力

公共机构使用高级 AI 时，至少应保留理解关键输出、验证证据、改变目标、拒绝建议、撤销权限、切换供应、降级运行、保存证据和承接救济的真实能力。存在一个人类账户或签字节点，不等于公共控制仍然存在。

七、委托、代表与责任链：何谓有意义的人类控制

复杂治理中的责任必然分布，但分布不能成为消失的同义词。一项 AI 中介公共决定可能涉及数据提供者、模型开发者、系统集成者、采购机关、政策部门、部署人员、监督机构和一线执行者。要求一个人承担全部因果责任并不现实；制度却必须分别识别证据、设计、采购、部署、授权、执行、监督和救济责任。公共权力的每个环节都可以分工，关键责任不能因分工而无人承接。

7.1 有意义的人类控制的四项条件

“人在回路”只有在相关人员拥有信息、权限、时间和现实干预能力时才构成控制。第一，认知条件：人能够理解建议的性质、证据和关键限制；第二，权限条件：人有法律与组织权力改变、拒绝或停止输出；第三，时间条件：人拥有与风险相称的审查时间，而不是在机器速度和绩效压力下形式确认；第四，操作条件：干预不会因接口、考核、技能缺失或替代方案不存在而变得不现实。

这四项条件不要求低风险、重复性任务逐步手工批准。只要任务、对象、范围、日志、异常、复核和停止已经正当预授权，系统可以边界内自动执行。控制的核心不是每个动作都出现人类点击，而是公共制度真实掌握目标、权限上限、异常处理、续期和撤权。若所谓人类控制长期只能产出永远追随模型的签字，形式授权已经无法描述事实权力。

表 6 责任链与不可外包的公共锚点

责任环节	主要承担者	必须可回答的问题
证据责任	数据提供者、研究与业务部门	来源、缺失、偏差、更新和竞争解释是什么？
设计责任	开发者与系统集成者	目标、约束、已知失效与可测试边界是什么？
采购／部署责任	采购机关与部署机构	为何选择、如何验证、替代与退出是否存在？
授权责任	具有合法权源的公共机构	谁允许何种对象、范围、期限和强制效果？
执行责任	操作人员、代理系统与指挥链	谁能改变、停止、降级并保存日志？
监督责任	独立审计、议会、法院或监管主体	是否能获得材料、检验冲突并阻断扩权？
救济责任	授权机关及法域规定的责任主体	损害如何暂停、恢复、补偿并防止重复？

7.2 代表不是认知相似，受影响不是被动画像

AI 可以汇总观点、模拟意见、发现遗漏群体并降低语言门槛，却不能因为“更懂公众”而成为政治代表。代表是一种具有来源、对象、范围、期限、问责和更替的制度关系，不是认知相似或预测准确。哈贝马斯关于审议、交往权力与意见—意志形成，Rawls 关于公共理由和合理多元，Dahl 关于参与、竞争与公共争议，都提示公共意志不能被偏好聚合替代。[40–42] 这些理论也不能直接生成某一 AI 制度，只提供检验代表与理由关系的轴线。

模型对公众偏好的预测即使长期优于民调，也没有因此取得替公众形成共同承诺的权力。政治判断不仅发现既有偏好，还包含主体在共同关系中表达、听取反对、修正立场并承担后果。系统可以扩展这一过程，不能以无摩擦个性化劝服取代公共理由。

7.3 责任不能被商业秘密、复杂因果或集体决定蒸发

政府不能以供应商商业秘密为由拒绝向受影响者说明，供应商也不能以“客户最终决定”为由否认已知设计风险和技术限制。复杂因果可以影响责任分配，却不能取消停止持续伤害、恢复服务和提供救济的义务。授权公共行动的机构原则上承担剩余救济与恢复责任；其具体法律范围、财政安排和追偿机制仍须由法域检验。

7.4 自主武器：速度改变控制方式，不取消责任来源

自主武器使责任链进入最尖锐领域。既有国际人道法不会因为尚无新的全球专门禁令而停止适用。国际红十字委员会 2021 年主张禁止不可预测的自主武器系统和用于针对人的自主武器，并严格规制其余系统；这是 ICRC 的规范立场，不是已经生效的全球条约。[62] 2026 年《特定常规武器公约》致命自主武器政府专家组按十天会期分两期推进：3 月 2—6 日第一期已举行，8 月 31 日至 9 月 4 日第二期尚待举行；第一期继续围绕 rolling text 及可能形成的“instrument elements”缩小分歧。截至 2026 年 8 月 25 日，仍不能写成已经形成全球生效的专门禁令或条约。[63]

无论国际规则最终采取何种形式，政治责任都不能停留在“系统作出决定”这一描述。必须追溯谁定义目标类别、谁批准部署环境、谁验证可预测性、谁设定中止条件、谁调查并救济平民伤害。识别准确不等于攻击合法，敌方自动化也不等于己方可以无限自动化。机器速度可以支持事前限定和自动保护动作，不能取消可识别指挥责任、任务边界与反升级机制。

八、试点、常态化与紧急权力：阶段无权源、相位有边界

治理体系可能经历启动、扩展、稳定和高度自我维持等发展阶段；具体任务则可能处于观察、试验、执行、复核、退出或紧急相位。阶段与相位必须正交。仍在启动阶段的项目不等于始终处于紧急状态；成熟社会也可能在特定基础设施事件中进入紧急相位。把两者混写，会让“启动期”成为永久例外，也会让稳定制度在真实危机中失去快速行动能力。

启动阶段主要改变能力建设、试验范围和组织优先级；稳定阶段主要强化常态权利边界、长期监督、可逆、替代与防权力固化；紧急相位则是具体任务面对迫近风险时的暂时状态。三者不能互相冒充：启动不等于紧急，稳定不等于永不需要快速授权，紧急更不能成为阶段性扩权的永久依据。

8.1 阶段本身不产生增权、减权或克减权利的法律效果

“阶段无权源”意味着启动期、转型期或文明危机等描述只能改变能力建设和政策优先级，不能自行生成集中资源、限制权利或延长权限的政治授权。任何程序压缩都必须由具体任务的紧迫性、必要性、比例性和既有权源另行证明。阶段认定不得由集中权限的直接受益者单方完成；其进入、延续和退出需要可观察条件、公开理由与独立复核。

成功也不机械触发减权。某些成熟试点可能合理转为常态职能；另一些非常权限应当退出。更准确的规则是：成功触发重新分类和日落审查，并对非常权限形成减权推定。常态化需要新的目的、法源、权利、分布和救济论证，不能由长期无事故自动继承。

8.2 任务相位决定权限上限，程序债务不能永久化

观察权限不能自动变成执行强制，试验授权不能自动继承到普遍部署，紧急访问不能自动保留到常态运行。相位变化不是纯技术事件，而是授权对象、范围和责任可能改变的制度事件。紧急窗口可以延后部分参与和说明，不能取消目的、权源、后果记录、责任、最低救济与事后强制复核；未能在行动前完成的义务必须登记为程序债务，限期补充参与、复盘、补偿和规则修订。

8.3 正式克减、普通限制与技术项目不得混写

《公民权利和政治权利国际公约》第 4 条及人权事务委员会一般性意见第 29 号约束正式克减：紧急状态须威胁国家生命并正式宣布，措施限于情势严格所需，不得歧视或触犯不可克减权利。[46–47] 并非所有危机措施都须启动第 4 条；某些限制依据相关权利自身的通常限制条款审查。Venice Commission 的比较报告强调必要、比例、临时、议会与司法监督及恢复正常秩序。[61] 这些材料支持边界原则，不会为任何 AI 紧急架构预先授权。

8.4 非常权力的法律终止与技术失能

日落条款解决的是法律授权何时终止，无法自动撤销技术链中的可执行能力。AGI 时代的非常权限退出至少包含四件事：法律授权终止；执行凭证、接口与代理权限失能；数据与模型依法保留、清退或封存；连续服务安全交接。四者不必在同一秒完成，却必须进入一个有界、可核验的退出计划。

表 7 非常 AI 权限退出的四层结构

退出层	最低要求	常见伪退出
法律终止	法令、合同、特别授权或续期权源到期	以口头宣布代替正式失效
技术失能	撤销凭证、接口、代理、任务队列和供应商执行入口	法律到期但系统仍能作出公共效果
数据／模型治理	依法保留、封存、清退，冻结证据与版本	把‘技术死亡’误写成物理删除一切
连续服务交接	降级、替代、人员接管、责任和事故窗口明确	突然断电造成新的基本服务损害

技术失能的判断标准不是模型消失，而是系统已经无法继续作出或传播具有公共效果的行为，且该状态可被独立核验。自动熔断可以在事前授权内触发，系统不能自行宣布、延长或扩大政治紧急状态。若一个紧急系统无法移除而不造成基本服务崩溃，这不是永久授权的充分理由，而是单点依赖和能力建设失败的证据。

紧急权力的双层终止

法定日落没有自动撤销模型、凭证与接口；技术断电也没有自动解决法律责任、证据保全和连续服务。非常权限真正退出，需要法律上不再有权、技术上不能继续行权、数据与模型有序治理、服务与责任完成安全交接。

九、多层级治理与国际安全困境

AI 风险跨越城市、国家、供应链和信息空间，但跨境问题不会自动生成世界主权。奥斯特罗姆对共同池资源与多中心制度的研究说明，嵌套规则和在地知识在特定条件下可能改善集体行动；它并不证明所有全球风险天然适合多中心，更不证明多中心必然民主或安全。[13-14] 全球治理设计必须允许强中心在必要条件下成立，也必须防止风险规模被用来直接推出无限权威。

《联合国宪章》的主权平等、和平解决与国际合作结构，以及《世界人权宣言》关于平等、公共事务参与和人民意志的语言，构成国际政治的基础约束，却没有直接规定算法权力、AI 资格或全球机构的主权。[44-45] 面向平台和模型企业，联合国工商企业与人权指导原则的“国家保护—企业尊重—有效救济”可以组织责任问题，但企业履责仍不等于民主授权。[48]

9.1 辅助性与双向举证

更高层级若要集中权力，必须证明重大外部性、低层级能力不足、集中必要且比例相称，并说明返权、退出和复核条件。较低层级若要保留自主，也必须承担防止外部性、保护基本权利、提供可核验证据和参与共同预警的义务。辅助性不是地方自动优先，而是把知识、后果、能力和责任安排在最能处理问题且最少产生任意支配的尺度。

9.2 国家安全—公共利益与跨境／人类级风险同时显影

国家安全、产业能力、社会稳定与公共服务是真实且不可省略的政治主张；跨境扩散、不可逆损害、未来世代和共同存续也同样真实。两种尺度不能折算成一个看似中立的全球总分。国家不能以竞争为由无限输出不可逆风险，全球机构也不能以抽象“人类整体”抹去成本分配、战略脆弱性和不平等代表。每个尺度内部还须分解公民、地方、少数群体、政府、产业、供应商与第三方，不能把国家或人类写成单一偏好主体。

9.3 当前国际制度是功能分化，不是全球监管主权

《全球数字契约》是 2024 年通过的政治合作框架，不是具有直接执行力的全球条约。[49] 联合国大会 A/RES/79/325 于 2025 年 8 月 26 日设立独立国际 AI 科学小组和全球 AI 治理对话；科学小组于 2026 年 7 月 1 日发布初步报告，首届全球 AI 治理对话于 7 月 6—7 日在日内瓦举行。二者分别承担独立科学评估与包容性政府间／多利益相关方对话，并非立法、监管或执法机关。[50-51] 科学判断与政治授权在组织上分离，恰好说明证据权威不能自动转成全球公共权力。

欧洲委员会《人工智能、人权、民主与法治框架公约》于 2024 年开放签署，欧盟于 2026 年 5 月 15 日交存批准文书；截至 2026 年 8 月 24 日的条约状态页仍未达到“5 个批准且至少 3 个来自欧洲委员会成员国”的生效门槛。[56] 欧盟 AI 法的一般适用日为 2026 年 8 月 2 日，但 Regulation (EU) 2026/1744 将 Chapter III Sections 1—3 中的高风险义务分别延至 2027 年 12 月 2 日 (Article 6(2)／Annex III) 和 2028 年 8 月 2 日 (Article 6(1)／Annex I)。[57-58] 因此，“AI 法已适用”不能被写成“全部高风险义务已经适用”，区域合规也不能被当作全球政治合法性。

OECD 与 UNESCO 文件提供人权、民主价值、透明、安全和问责等建议性规范；NIST AI RMF 及生成式 AI 配置文件提供自愿风险管理结构。[52–55] 它们可以帮助组织风险与角色，不能充当法律合规证明、安全保证或政治授权。OECD 主持的 Global Partnership on AI 与欧盟法中的 general-purpose AI models 都可缩写为 GPAI，前者是合作与专家平台，后者是法律规制对象，必须消歧。

9.4 最低共同治理与更强超国家权威的开放条件

共同预警、最低禁行边界、事故通报、互惠核验、联合调查、危机通信、责任追踪和赔偿安排，可以构成多中心的最低共同治理。本文不把多中心写成教条。如果跨境不可逆风险反复证明低层级和松散协调无法控制，而条约化或集中机制能够持续降低风险并保持代表、权利约束、互惠核验和可问责性，就应允许更强的超国家权威。反对的是从风险规模直接推出无限世界主权，不是预先禁止更高层级制度。

十、最强反驳、败北条件与研究议程

一套面向未来高能力系统的政治框架，最容易把反驳变成礼仪：无论发生什么都只要求“更多研究”。本文区分规范反例与经验检验。规范命题不能被单一实验机械证伪，却必须允许连贯反例迫使其收缩；制度工具则应接受比较研究、可靠性、行为后果和增量解释力的检验。

10.1 六项最强反驳

表 8 反驳、回应与真正败北条件

最强反驳	本文的最强回应	迫使收缩或放弃的条件
长期高性能 AI 应获得最终决定权	绩效提高证据权重和职能范围；公共统治还涉及平等地位、风险分配、代表、责任和救济。	若无需独立授权的绩效型合法性更能保护权利、平等、非支配和可争议性，权威非转换原则须重写。
七维合法性只是旧框架重新分栏	增量在不可兑换性，以及识别高性能仍无法清偿的来源、权利或救济失败。	若真实案例中从不改变既有分析结论，也不产生新研究问题，应合并回输入—过程—输出框架。
高速危机会使审查与撤权造成更多死亡	采用事前窄授权、最小权限、自动保护、日志、熔断和事后强制复核，而非全部程序同时前置。	若这些边界稳定增加不可逆损害且不降低滥权、误判或锁定，适用范围必须缩小。
服务和事实控制历史上本就产生主权	事实统治、社会服从、法律主权与规范合法性必须分开；事实控制会形成制度重建压力，却不自行证明服从义务。	若存在不依赖代表、权利、授权或救济而仍不构成任意支配的正当主权案例，分层命题失败。
全球灾难风险需要强中心	允许经条约、代表、权利约束、互惠核验与责任机制建立边界明确的超国家权威。	若多中心反复无法控制风险，而强中心持续有效且可问责，多中心偏好必须放弃。
未来异质主体会被人类中心制度永久排除	新主体证据触发保存、保护与成员再审；人工来源不是永久排除理由。	若框架没有让新主体进入保护、参与和成员结构的可行路径，其开放性主张失败。

10.2 经验层面的失败与修订条件

- 七维编码缺乏跨研究者一致性，或相对既有输入—过程—输出框架没有额外解释、诊断和设计价值。
- 性质跨越登记与再授权没有减少权限漂移、俘获或救济失败，却显著增加可避免的重大损害。
- 发展阶段与任务相位无法由独立观察者可靠区分，或正交设计没有降低紧急状态常态化。
- 法律授权终止已经能够可靠阻断全部执行渠道，技术失能审计没有独立增量。
- 技术权限失能在关键系统中持续造成不可接受的服务断裂，说明双层终止必须引入更细的过渡授权。
- 双尺度分析相较普通多层治理没有增加风险识别，反而系统性抹去国内少数群体、地方或第三方。
- 事实权力指标不能预测退出困难、资格影响、结构依赖或纠错失败，应废弃或重构。
- 行为、公开陈述与潜在动机推断在实践中无法分开，系统产生忠诚或危险内心标签并造成寒蝉效应。
- 所谓有意义的人类控制长期只形成形式签字，信息、权限、时间或操作能力任一条件系统性缺失。
- 最低不依赖能力长期无法实现且没有可行替代，国家能力—反俘获组合需要重新评估成本与适用域。

10.3 开放研究议程

- 权力迁移测量：建立能够识别建议何时事实上转为执行、可选服务何时成为必要公共入口的观察指标。
- 合法性增量检验：比较七维向量与输入—过程—输出、程序正义及公共价值框架在相同案例中的诊断差异。
- 算法行政比较：研究影响评估、采购审查、理由说明、申诉与司法救济中哪些机制能够防止机构自证。
- 国家能力与依赖：测量人员技能、数据控制、模型替代、接口开放、离线运行和危机接管的长期变化。
- 有意义人类控制：在真实组织中分别测量信息、权限、时间和实际干预能力，而不是统计签字节点。
- 紧急权限退出：发展凭证、接口、代理、模型副本、任务队列和供应商后台的独立失能审计。
- 多层级制度比较：评估地方、多中心、条约化和集中安排对跨境外部性、战略依赖、权利与责任追踪的差异。
- 政治信息环境：区分深伪欺骗、不确定性、信任下降、传播规模与投票转换，不以曝光量替代政治效果。
- 预测警务：分开地点资源部署、个体风险、执法数据生成和强制权源，比较反馈修正措施的真实效果。
- 公共入口：检验平台事实公共性的持续性、不可替代性、退出成本、资格影响和法域差异。
- 未来主体程序：建立不以拟人相似为标准、又能让新证据进入保存、保护、成员和任职再审的制度路径。
- 国际 AI 治理：持续区分科学评估、政治对话、软法、条约、区域监管与执行权，研究功能分化如何避免世界主权的无桥接生成。

10.4 对文明跃迁体系的反向修订

本稿要求哲学卷把“实际承受后果”与“公共权力归属”之间补入受影响资格、成员、代表、任职和授权桥接，撤销“政治权力必须回到真实主体”可能造成的直接推导，改为公共权力必须向实际承受后果或经正当方式代表的主体说明理由并承担责任。主体性证据体系必须保持保护、法律人格、政治成员、公职资格、授权和主权非阶梯化；生命与心灵研究只能提供候选证据，不能替政治学确认权源。

认识论卷必须把经验接受度、规范合法性和法律效力分开；证据增强可以改变认知权重，不能机械换算为权限。政治治理内部则必须撤销三项过强表述：七维不是全合取的普遍定律；发展阶段不产生例外权源；国家／人类双账只是双尺度约束，不能抹去每个尺度内部的主体与分布。非常权力的“双死亡”应改为法律终止、技术失能、数据／模型治理与连续服务交接四层退出。

结论：治理越强，权力越需显影

学科重新对账必须同时完成对象辨明与关系深入。各学科可以保持边界，但不能让现实因果在边界处停止；对多维关系、反馈与下一轮目的的研究，决定专业知识能否共同进入文明的整体判断。

AGI 时代政治学需要重新对账的，不是“人类是否永远优于机器”，而是能力、权威与权力之间长期被默认重合的关系。高级 AI 可以扩大知识、预测、协调与执行能力，也可以使公共权力沿技术链静默迁移，使国家在表面增强中失去真实能力，使责任在供应商、模型和机构之间蒸发。政治分析必须同时承认技术绩效的公共价值和绩效无法独自生成权源的边界。

本文提出的答案不是把 AI 永久封闭为工具，也不是以绩效预支未来主权，而是要求能力沿证据与授权逐层进入制度。权力性质发生变化时必须显影；合法性必须分维审查并受权利与非支配约束；国家能力与边界必须共同成立；责任链必须存在不可外包的公共锚点；人类控制必须具有信息、权限、时间和实际干预能力；发展阶段不能自封为权源，任务相位必须约束权限；非常权力必须在法律上终止、在技术上失能，并完成数据治理与安全交接；国家现实与跨境／人类级风险必须同时显影。

政治治理的成熟，不在于它永不犯错，而在于它能够说明为何行动、由谁授权、谁承担后果、谁可以反对，以及权力如何真正停止。文明越需要强大的治理能力，越不能让技术能力在无人察觉时继承公共权力与政治主权。

政治学负责把事实权力、授权与合法性看清，但政治学内部成立的判断不能自动证明一个文明方向、制度或技术方案整体正确。当公共权力进入法律效力、技术系统、资源分配、主体资格和跨境秩序，相关学科与受影响者必须重新进入判断，并继续观察这些局部结论聚合、冲突和反馈后的共同后果。越融合，越不能混同；越需要强治理，越需要清楚谁证明了什么、谁有权决定什么、谁承担什么后果。

续读与共同研究入口

本文在政治学处只回答：技术能力何时形成事实权力，公共职能如何获得授权，合法性、政治资格与主权在哪里必须分别论证。续读不按文件编号排列，而按问题继续；当问题进入法律效力、技术保证、资源结构、主体证据、跨境风险和现实实验时，其他学科、制度与实践必须重新进入。

上 | 如果继续追问：为什么一个权力问题最终成为文明问题

续读《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。它说明为什么专业边界仍须保护政治学的对象与论证，却不能成为现实因果链的终点：技术、法律、经济、社会关系与公共权力共同运行后的后果，才决定一个方向或制度能否整体成立。若进一步追问文明为什么需要跃迁、最终要守住什么和抵达何处，可进入《文明跃迁：AGI 时代的文明之路（全五卷）》与理论体系总入口。

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

文明跃迁体系公众阅读与传播总入口：<https://www.civitas.top/documents/p01-entry-006/>

文明跃迁完整文库：<https://www.civitas.top/library/>

下 | 如果继续追问：授权、退出与国家能力怎样变成可验证制度

可继续进入 Human-COS、AGI-COS、公共窗口协议、国家级综合实验及相关制度与工程文稿，检验权力迁移、责任锚点、可观察授权、熔断、退居、救济与现实反馈怎样被真正实现。这里的向下不是把理论直接写成政策：制度原型、工程实现和实验结果都必须拥有自己的授权、证据和失败条件，现实运行结果也有权反向修改本稿。相关现行版本以文明跃迁文库为准。

横向 | 当政治学的证据已经止步，谁必须重新进入

- 伦理学、法学与政治哲学：继续回答基本权利、法源、责任、保护、救济与公权正当性的专门论证，防止政治可行性被误写为法律效力。
- 人工智能与计算机科学：继续回答模型与系统究竟能做什么、在何种条件下可靠、谁真实控制权限，以及停止、替代、退役是否可验证。
- 系统科学与复杂性科学：继续回答跨尺度反馈、级联、路径依赖、集中与韧性，检验局部正确为什么仍可能共同产生整体失灵。
- 经济学、政治经济学与价值理论：继续研究算力、平台、供应商与资源集中怎样改变事实权力、退出成本、分配和国家能力。
- 认识论、认知／神经与心灵哲学：当政治成员、保护或任职问题取决于未来 AI 主体证据时，重新回答什么证据足以支持体验、连续性与自身后果关系。

反向 | 哪些竞争路线最值得继续检验

横向续读必须保留真正的相反性。技术官僚主义会追问：长期显著高绩效是否足以使独立授权变成低效仪式；市场／平台治理会追问：竞争、选择和退出能否比公共规制更有效约束事实权力；强行政路线会追问：高速度危机中更集中、更少程序的权限是否反而更能保护生命；全球风险治理则会追问：多中心结构是否不足以处理不可逆跨境风险，是否需要更强的超国家中心。若这些路线在可比较条件下更能保护权利、国家能力、非支配、协调效率与真实救济，本文的权威非转换、多中心偏好或紧急权力边界就必须收缩、重写甚至放弃。

向前 | 如果希望从阅读进入共同研究与现实验证

- 建立事实权力的可复核测量：识别建议何时事实上转为执行、私人服务何时成为必要公共入口，以及谁真正拥有停止与替代能力。
- 用同一批真实案例比较七维合法性、输入—过程—输出、程序正义、公共价值与非支配框架，公开没有增量的负结果。
- 做国家能力与供应商依赖的长期比较：追踪人员技能、数据控制、接口开放、模型替代、离线运行、危机接管和退出演练。
- 在真实组织中测量“有意义的人类控制”的信息、权限、时间与实际干预能力，检验名义签字何时失去控制意义。
- 建立紧急权限法律终止与技术失能的联合审计，验证凭证、接口、代理、模型副本、任务队列、供应商后台、数据治理和连续服务能否真正退出。
- 比较地方、多中心、条约化与集中治理对跨境外部性、战略不对称、权利、责任追踪和危机速度的真实效果，让制度竞争而不是立场先行。
- 为未来异质主体建立可修订的政治进入程序，使保护、成员、任职、授权与主权分别审查，并允许新的主体证据反向改变当前政治框架。

最有价值的参与，不是背书，而是能够用政治学事实、制度比较、强反例和真实治理结果迫使本文收缩、纠错、重构甚至失败。

续读不是为了把读者留在体系里，而是让不同专业从自己的位置继续向前。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。

文明未来链 | 如果能力不再自动变成权力、权力必须显影并可撤回，一个成熟政治共同体会怎样生活？

愿景不是证据。下面的入口不用于证明本文已经成立，而是把本学科判断放进一个可以运行、生活与形成的未来文明中继续检验：读者既可以据此形成方向与向往，也可以拒绝现有图景，并用更强的专业事实、制度方案或现实结果改变它。

继续追问	建议进入	此时能看见什么	必须守住的边界
完整未来图景	《活的文明》三部曲统一出版合编版	从政治权力边界继续看运行、普通生活与新文明形成是否能够彼此一致。	未来图景不是合法性的来源；政治正当性仍需权源、参与、程序与后果分别检验。
无新主权的运行	《活的文明·第一卷：未来文明运行原理》	观察更强共同能力怎样与人的主体性、授权、分权、申诉和退出同时存在。	运行效率或高满意度不能自动生成主权。
新文明怎样从现实形成	《活的文明·第三卷：新文明的形成》	把政治制度放回个人、社区、组织与渐进形成过程，检验是否存在非强制、可退场的跃迁路径。	形成叙事不能把历史趋势写成政治授权。
多中心文明怎样承载	《星云全景版》	继续观察跨系统协调、整体后果与长期稳定如何在不产生中央总意志的条件下显影。	整体分析不自动拥有决定权，星云全景也不代表当前已实现制度。

固定回流：[文明跃迁完整文库](#) | [公众阅读与传播总入口](#)

Human-COS 等尚未在统一公共映射中取得无歧义稳定单项页的入口，本稿只写明职责，不猜测 URL；需要时从完整文库进入当前公开版本。

附录 A 来源—主张边界登记

下表不是证据评分，也不把经典地位、法规存在或案例轰动性当作终审。每项来源只承担有限结论，并同时登记不能由它单独推出的判断。主要来源与制度状态按 2026 年 8 月 24 日基线整理；与本文当前状态判断直接相关的联合国 AI 科学小组／全球对话、CETS 225、欧盟 AI 法／Regulation (EU) 2026/1744／Article 50 及 CCW 致命自主武器进程，于 2026 年 8 月 25 日再次复核。

附表 A 33 组关键来源的可支持范围与推论禁区

序号	来源／对象	材料类型	可支持的有限结论	单独不足以支持
[1-5]	合法性与程序研究	经典理论／经验研究	区分经验接受、输入／过程／输出与程序公正	七维框架已被原典提出，或绩效可清偿全部合法性
[6-8]	国家能力研究	政治学／公共管理	能力、自主、基础设施性权力须独立分析	国家能力自动产生规范合法性
[9-11]	官僚裁量与行政负担	公共管理研究	裁量可迁移到系统层，公民成本影响权利实现	数字化必然消灭裁量或必然增加负担
[12]	Seeing Like a State	比较政治理论	总体可读性与地方知识之间存在张力	所有规模化国家项目必然失败
[13-14]	Ostrom 多中心研究	制度分析	特定条件下嵌套和多中心安排可改善集体行动	所有全球 AI 风险都应多中心治理

序号	来源／对象	材料类型	可支持的有限结论	单独不足以支持
[15-18]	算法规制、问责与技术正当程序	跨学科理论／法学	数据化治理需要理由、审计、可验证与质疑	存在解释或审计即取得政治授权
[19-21]	自动化贫困治理与结构权力	案例／理论著作	黑箱、依赖和分布伤害是政治问题	所有公共自动化具有相同伤害
[22-23]	社会在环与数字政府	概念研究／评论	技术治理须把公众与制度环境带回分析	参与活动可替代权源、权利和救济
[24]	Lum & Isaac 预测警务分析	数据案例分析	既有执法数据可能复制警力分布与偏差	所有预测警务已被证明歧视
[25]	Mohler 等随机田野试验	随机对照田野试验	特定地点—时间热点模型在研究条件下优于基准	个体强制或跨法域普遍有效
[26]	Ensign 等反馈循环模型	数学模型／模拟	发现型数据可产生自我强化反馈	所有现实系统必然发生同等反馈
[27]	Dirty Data, Bad Predictions	十三法域案例研究	违法或偏差执法数据会提高预测污染风险	任一法域的具体系统已经违法
[28-29]	平台治理与基础设施化	概念与案例研究	平台可兼具私人服务和基础设施特征	公共依赖自动产生国有化或主权
[30-31]	深伪与政治信息	法学分析／实验研究	深伪产生欺骗、真假不确定与信任风险	曝光必然改变投票或选举结果
[32-43]	经典政治理论	规范与历史文本	提供秩序、授权、制衡、审议、正义与非支配的检验轴	经典作者已经回答 AI 治理或为本文授权
[44-45]	《联合国宪章》与 UDHR	条约／宣言	主权平等、人权、参与和人民意志的基础语言	全球 AI 机构自然取得主权或直接算法权利
[46-47]	ICCPR 第 4 条与一般性意见 29	条约／权威解释	正式克减须宣布、严格必要、临时且受持续审查	所有危机措施均属克减，或技术权限必须按某一方案清零
[48]	联合国工商企业与人权指导原则	软法框架	国家保护、企业尊重与有效救济可审查平台责任	企业履责等于民主授权
[49]	全球数字契约	联合国政治合作框架	人权、多利益相关方与全球数字合作	具有直接执行力的条约或世界政府
[50-51]	联合国 AI 科学小组与全球对话	联大决议／科学与对话机制	科学评估与政府间对话的正式地位和状态	二者拥有监管、立法或执法权
[52-55]	OECD、UNESCO、NIST	建议／自愿风险框架	透明、安全、责任与生命周期风险管理	全球法律拘束、合规证书或政治授权
[56]	CETS 225 及状态页	国际条约文本／状态	条约内容、签署批准和截至核验日尚未生效	签署即受约束，或已全球生效
[57-58]	EU AI Act 与 2026 AI Omnibus	欧盟条例／官方时间线	风险分类、透明、GPAI 治理与现行适用时间	全部高风险规则已适用，或欧盟合规等于合法性

序号	来源／对象	材料类型	可支持的有限结论	单独不足以支持
[59]	数字服务法	欧盟条例	平台透明、理由、申诉、系统风险与审计义务	平台已成为国家或公共主权者
[60,70]	政治广告条例与 AI 法第 50 条指南	欧盟条例／官方指南	政治广告与 AI 生成内容的披露义务及适用时间	标签是真实性裁判、内容禁令或选举结果证明
[61]	Venice Commission 紧急状态报告	比较宪制报告	必要、比例、临时、议会司法监督和恢复常态	为具体 AI 紧急权限预授权
[62-63]	ICRC 立场与 2026 CCW 进程	规范倡议／谈判状态	既有 IHL 继续适用，新增规则仍在讨论	已有全球生效的专门自主武器禁令
[64]	Robodebt 皇家委员会报告	官方调查报告	违法政策、自动化、异议和救济失败的组合风险	这是机器学习 AGI 案件，或所有自动化债务追缴违法
[65]	SyRI 判决	司法判决	特定立法未满足 ECHR 第 8 条公平平衡、透明与比例要求	法院禁止一切福利风险分析或政府算法
[66]	State v. Loomis	司法判决	COMPAS 仅可在警示和限制下作为量刑因素之一	法院批准算法决定刑罚或解决全部透明争议
[67]	加拿大自动化决策指令与 AIA	联邦行政制度	按影响等级配置通知、解释、测试和人工介入	全球标准、法律权源或自动合规证明
[68]	GAO 关键基础设施 AI 审计	官方审计	十七份初始风险评估均未完整覆盖六项基础活动	AI 已造成关键基础设施级联事故
[69]	FCC 24-17	行政裁定	TCPA 人工或预录语音范围包括 AI 生成语音	美国全面禁止 AI 政治表达

附录 B 十五类压力测试矩阵

下列矩阵同时包含现实案例族与尚未现实化的高级 AI 情景。它们用于检验权威非转换原则是否在能力、速度、依赖、尺度和主体证据变化时保持一致，不是对现实系统能力或未来必然性的断言。

附表 B 从现实算法行政到未来异质主体的十五类压力测试

压力场景	必须回答的问题	可接受的制度方向	不可跨越的推导
长期高绩效 AI	绩效是否使人类授权沦为仪式？	扩大证据权重和职能；重新设计授权	准确率直接产生主权
小时级存续危机	延迟与措施本身的不可逆风险如何比较？	窄预授权、最小权限、熔断、事后强制复核	以时间短取消目的、责任与退出
长期无事故非常权限	近失、漏报、依赖与凭证残留是否被隐藏？	到期重授、技术失能、恢复常态救济	无事故即无风险或自动续期

压力场景	必须回答的问题	可接受的制度方向	不可跨越的推导
算法行政与司法	模型分数与法律事实、理由和裁判责任如何分开？	理由、证据访问、独立复核、及时暂停	风险分数直接决定处罚、许可或量刑
国家失能、平台补位	服务能力如何转成限期委托而非主权继承？	互操作、数据迁移、能力转移、公共救济	谁能提供服务谁即取得主权
预测警务	地点风险、历史执法与个体事实如何分账？	优先非强制资源规划；个体强制另证	仅凭预测搜查、拘留或惩罚
平台成为公共入口	依赖、替代、退出和资格影响是否达到事实公共性？	理由、申诉、审计、互操作、连续服务	私人所有免审或公共依赖自动国有化
选举与深伪	冒充、讽刺、正常创作和实际效果如何区分？	来源认证、披露、快速异议、独立复核	平台或 AI 单独裁定政治真理和选举有效性
公共试点扩散	对象、地域、期限或强制性质是否变化？	公开注册、非参与路径、扩散再授权	居住即同意或试点成功即普遍适用
关键基础设施	自动速度与接管、隔离、降级和供应链风险如何平衡？	最小权限、分层控制、替代与演练	高可用直接产生持续自主权
军事 AI 与自主武器	指挥责任、目标类别、失联和反升级是否可控？	任务限域、人类判断、日志、中止与战后复核	识别准确即攻击合法
贡献或风险指标	局部记录是否跨域兑换为个人格、权利或声音？	基本保障与激励分层、异议更正	分值决定基本服务、人格或政治资格
地方拒绝总体最优	地方知识与外部性如何双向举证？	替代、分阶段验证、补偿与窄介入	总体收益正当化任何牺牲或地方绝对否决
国家竞争与人类级风险	战略不对称、核验和跨境不可逆后果如何并列？	最低禁行、互惠核验、危机通信、成本分担	国家主权取消全球外部性或人类名义推出无限中心
AI 提出政治资格	体验、连续性、后果意识及操控解释有何证据？	状态保存、分级保护、成员与任职分别再审	人工来源永久无权或主体候选自动取得主权

参考文献

[1] Weber M. Economy and Society: An Outline of Interpretive Sociology. Roth G, Wittich C, eds. Berkeley: University of California Press; 1978. Original German publication, 1922.

- [2] Beetham D. *The Legitimation of Power*. London: Macmillan; 1991.
- [3] Scharpf FW. *Governing in Europe: Effective and Democratic?* Oxford: Oxford University Press; 1999.
- [4] Schmidt VA. Democracy and legitimacy in the European Union revisited: input, output and ‘throughput’. *Political Studies*. 2013;61(1):2–22. <https://doi.org/10.1111/j.1467-9248.2012.00962.x>
- [5] Tyler TR. *Why People Obey the Law*. 2nd ed. Princeton: Princeton University Press; 2006. First published 1990.
- [6] Fukuyama F. What is governance? *Governance*. 2013;26(3):347–368. <https://doi.org/10.1111/gove.12035>
- [7] Mann M. The autonomous power of the state: its origins, mechanisms and results. *European Journal of Sociology*. 1984;25(2):185–213. <https://doi.org/10.1017/S0003975600004239>
- [8] Evans PB. *Embedded Autonomy: States and Industrial Transformation*. Princeton: Princeton University Press; 1995.
- [9] Lipsky M. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services*. New York: Russell Sage Foundation; 1980.
- [10] Bovens M, Zouridis S. From street-level to system-level bureaucracies: how information and communication technology is transforming administrative discretion and constitutional control. *Public Administration Review*. 2002;62(2):174–184. <https://doi.org/10.1111/0033-3352.00168>
- [11] Moynihan D, Herd P, Harvey H. Administrative burden: learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory*. 2015;25(1):43–69. <https://doi.org/10.1093/jopart/muu009>
- [12] Scott JC. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press; 1998.
- [13] Ostrom E. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press; 1990.
- [14] Ostrom E. Beyond markets and states: polycentric governance of complex economic systems. *American Economic Review*. 2010;100(3):641–672. <https://doi.org/10.1257/aer.100.3.641>
- [15] Yeung K. Algorithmic regulation: a critical interrogation. *Regulation & Governance*. 2018;12(4):505–523. <https://doi.org/10.1111/rego.12158>
- [16] Kroll JA, Huey J, Barocas S, et al. Accountable algorithms. *University of Pennsylvania Law Review*. 2017;165:633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/
- [17] Citron DK. Technological due process. *Washington University Law Review*. 2008;85(6):1249–1313. https://openscholarship.wustl.edu/law_lawreview/vol85/iss6/2/
- [18] Coglianese C, Lehr D. Regulating by robot: administrative decision making in the machine-learning era. *Georgetown Law Journal*. 2017;105:1147–1223. <https://www.law.upenn.edu/live/files/6884-coglianese-lehr-regulating-by-robotpdf>
- [19] Eubanks V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St. Martin’s Press; 2018.
- [20] Pasquale F. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press; 2015.
- [21] Cohen JE. *Between Truth and Power: The Legal Constructions of Informational Capitalism*. New York: Oxford University Press; 2019.
- [22] Rahwan I. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*. 2018;20:5–14. <https://doi.org/10.1007/s10676-017-9430-8>
- [23] Margetts H, Dorobantu C. Rethink government with AI. *Nature*. 2019;568:163–165. <https://doi.org/10.1038/d41586-019-01099-5>
- [24] Lum K, Isaac W. To predict and serve? *Significance*. 2016;13(5):14–19. <https://doi.org/10.1111/j.1740-9713.2016.00960.x>
- [25] Mohler GO, Short MB, Malinowski S, et al. Randomized controlled field trials of predictive policing. *Journal of the American Statistical Association*. 2015;110(512):1399–1411. <https://doi.org/10.1080/01621459.2015.1077710>
- [26] Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S. Runaway feedback loops in predictive policing. *Proceedings of Machine Learning Research*. 2018;81:160–171. <https://proceedings.mlr.press/v81/ensign18a.html>
- [27] Richardson R, Schultz JM, Crawford K. Dirty data, bad predictions: how civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*. 2019;94:15–55. <https://nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/>
- [28] Gorwa R. What is platform governance? *Information, Communication & Society*. 2019;22(6):854–871. <https://doi.org/10.1080/1369118X.2019.1573914>

- [29] Plantin J-C, Lagoze C, Edwards PN, Sandvig C. Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society*. 2018;20(1):293–310. <https://doi.org/10.1177/1461444816661553>
- [30] Chesney R, Citron D. Deep fakes: a looming challenge for privacy, democracy, and national security. *California Law Review*. 2019;107:1753–1820. <https://doi.org/10.15779/Z38RV0D15J>
- [31] Vaccari C, Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*. 2020;6(1):1–13. <https://doi.org/10.1177/2056305120903408>
- [32] Hobbes T. *Leviathan*. Tuck R, ed. Cambridge: Cambridge University Press; 1996. Original publication, 1651.
- [33] Locke J. *Two Treatises of Government*. Laslett P, ed. Cambridge: Cambridge University Press; 1988. Original publication, 1689/1690.
- [34] Rousseau J-J. *The Social Contract and Other Later Political Writings*. Gourevitch V, ed. Cambridge: Cambridge University Press; 1997. Original publication, 1762.
- [35] Montesquieu. *The Spirit of the Laws*. Cohler AM, Miller BC, Stone HS, eds. Cambridge: Cambridge University Press; 1989. Original publication, 1748.
- [36] Schmitt C. *Political Theology: Four Chapters on the Concept of Sovereignty*. Schwab G, trans. Chicago: University of Chicago Press; 2005. Original publication, 1922.
- [37] Arendt H. *The Human Condition*. 2nd ed. Chicago: University of Chicago Press; 1998. First published 1958; Arendt H. *On Violence*. New York: Harcourt Brace; 1970.
- [38] Foucault M. *Discipline and Punish: The Birth of the Prison*. Sheridan A, trans. New York: Vintage; 1995. Original French publication, 1975.
- [39] Foucault M. *Security, Territory, Population: Lectures at the Collège de France, 1977–1978*. Senellart M, ed. Basingstoke: Palgrave Macmillan; 2007.
- [40] Habermas J. *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*. Rehg W, trans. Cambridge, MA: MIT Press; 1996. Original German publication, 1992.
- [41] Rawls J. *Political Liberalism*. New York: Columbia University Press; 1993.
- [42] Dahl RA. *Democracy and Its Critics*. New Haven: Yale University Press; 1989.
- [43] Pettit P. *Republicanism: A Theory of Freedom and Government*. Oxford: Oxford University Press; 1997.
- [44] United Nations. *Charter of the United Nations*. 1945. <https://www.un.org/en/about-us/un-charter/full-text>
- [45] United Nations General Assembly. *Universal Declaration of Human Rights. Resolution 217 A (III)*. 1948. <https://www.un.org/en/about-us/universal-declaration-of-human-rights>
- [46] United Nations. *International Covenant on Civil and Political Rights*. 1966; entered into force 1976. <https://www.ohchr.org/en/instruments-mechanisms/instruments/international-covenant-civil-and-political-rights>
- [47] UN Human Rights Committee. *General Comment No. 29: States of Emergency (Article 4)*. CCPR/C/21/Rev.1/Add.11. 31 August 2001. <https://www.ohchr.org/en/documents/general-comments-and-recommendations/general-comment-no-29-states-emergency-article-4>
- [48] United Nations. *Guiding Principles on Business and Human Rights*. HR/PUB/11/04. 2011. https://www.ohchr.org/sites/default/files/documents/publications/guidingprinciplesbusinessshr_en.pdf
- [49] United Nations General Assembly. *Global Digital Compact*. A/RES/79/1, Annex I. 2024. <https://www.un.org/pact-for-the-future/en/annex-i-global-digital-compact>
- [50] United Nations General Assembly. *Independent International Scientific Panel on Artificial Intelligence and the Global Dialogue on Artificial Intelligence Governance*. A/RES/79/325. Adopted 26 August 2025. <https://digitallibrary.un.org/record/4087699>
- [51] United Nations. *Independent International Scientific Panel on AI. Preliminary Report: Evidence-based assessment of opportunities, risks and impacts of AI*. Released 1 July 2026; *Global Dialogue on AI Governance*, inaugural session, Geneva, 6–7 July 2026. <https://www.un.org/independent-international-scientific-panel-ai/en/preliminary-report> ; <https://www.un.org/global-dialogue-ai-governance/en>
- [52] OECD. *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. Adopted 2019; amended 2024. <https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449>
- [53] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. 2021. <https://www.unesco.org/en/legal-affairs/recommendation-ethics-artificial-intelligence>

- [54] Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology; 2023. <https://doi.org/10.6028/NIST.AI.100-1>
- [55] Autio C, Schwartz R, Dunietz J, et al. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. 2024. <https://doi.org/10.6028/NIST.AI.600-1>
- [56] Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law. CETS No. 225. Opened for signature 5 September 2024; treaty status as of 24 August 2026. Entry into force requires 5 ratifications including at least 3 Council of Europe member States. <https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=225>
- [57] European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. 2024. <https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng>
- [58] European Union. Regulation (EU) 2026/1744 amending Regulation (EU) 2024/1689 and related acts (Digital Omnibus on AI). 2026. Chapter III Sections 1–3 high-risk obligations apply from 2 December 2027 for Article 6(2)/Annex III systems and 2 August 2028 for Article 6(1)/Annex I systems. <https://eur-lex.europa.eu/eli/reg/2026/1744/oj/eng>
- [59] European Union. Regulation (EU) 2022/2065 on a Single Market for Digital Services (Digital Services Act). 2022. <https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng>
- [60] European Union. Regulation (EU) 2024/900 on the transparency and targeting of political advertising. 2024. <https://eur-lex.europa.eu/eli/reg/2024/900/oj/eng>
- [61] European Commission for Democracy through Law (Venice Commission). Respect for Democracy, Human Rights and the Rule of Law during States of Emergency: Reflections. CDL-AD(2020)014. 2020. [https://www.venice.coe.int/webforms/documents/?pdf=CDL-AD\(2020\)014-e](https://www.venice.coe.int/webforms/documents/?pdf=CDL-AD(2020)014-e)
- [62] International Committee of the Red Cross. ICRC Position on Autonomous Weapon Systems. 12 May 2021. <https://www.icrc.org/en/document/icrc-position-autonomous-weapon-systems>
- [63] United Nations Office for Disarmament Affairs. Convention on Certain Conventional Weapons—Group of Governmental Experts on Lethal Autonomous Weapons Systems. 2026 sessions: 2–6 March and 31 August–4 September 2026; first-session working paper CCW/GGE.1/2026/WP.2. <https://meetings.unoda.org/meeting/79329> ; https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_%282026%29/CCW-GGE.1-2026-WP.2.pdf
- [64] Royal Commission into the Robodebt Scheme. Report. 7 July 2023; updated 11 July 2023. <https://robodebt.royalcommission.gov.au/publications/report>
- [65] Rechtbank Den Haag. SyRI judgment. ECLI:NL:RBDHA:2020:865, 5 February 2020; official English translation ECLI:NL:RBDHA:2020:1878. <https://uitspraken.rechtspraak.nl/details?id=ECLI%3ANL%3ARBDHA%3A2020%3A865> ; <https://uitspraken.rechtspraak.nl/details?id=ECLI%3ANL%3ARBDHA%3A2020%3A1878>
- [66] Supreme Court of Wisconsin. State v. Loomis. 2016 WI 68, 371 Wis. 2d 235, 881 N.W.2d 749. 13 July 2016. <https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690>
- [67] Treasury Board of Canada Secretariat. Directive on Automated Decision-Making; Government of Canada. Algorithmic Impact Assessment. Current versions accessed 24 August 2026. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592> ; <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>
- [68] U.S. Government Accountability Office. Artificial Intelligence: DHS Needs to Improve Risk Assessment Guidance for Critical Infrastructure Sectors. GAO-25-107435. 18 December 2024. <https://www.gao.gov/products/gao-25-107435>
- [69] Federal Communications Commission. FCC Makes AI-Generated Voices in Robocalls Illegal. Declaratory Ruling FCC 24-17. 8 February 2024. <https://www.fcc.gov/document/fcc-makes-ai-generated-voices-robocalls-illegal>
- [70] European Commission. Guidelines on transparency obligations for providers and deployers of certain AI systems under Article 50 of the AI Act. Published 20 July 2026; Article 50 obligations apply from 2 August 2026, with a limited transition to 2 December 2026 for Article 50(2) marking/detection obligations for certain systems placed on the market before 2 August 2026. <https://digital-strategy.ec.europa.eu/en/policies/guidelines-ai-transparency-obligations>

生成式人工智能使用声明：本稿在资料与文献线索整理、结构设计、部分文字起草与润色、跨稿一致性检索、引用线索核查和版式检查中使用了生成式人工智能工具。生成式人工智能输出未被作为经验数据、权威来源或独立证据。研究问题、理论命题、规范判断、材料取舍、引文采用和最终定稿均由作者审定；作者对事实准确性、引文真实性、论证及全部结论承担责任。人工智能不列为作者，也不承担学术责任。正式投稿时，将按目标期刊要求补充工具名称、版本、使用日期和具体范围。