

文章编号:1003-2053(2026)10-2099-12

生成式 AI 时代社交媒体平台协同治理框架研究 ——基于技术架构与治理层级的互构过程分析

郅雨菲^{1,2}, 张辉^{1,2,3}

(1. 中国科学技术大学公共事务学院, 安徽合肥 230026; 2. 清华大学人工智能国际治理研究院, 北京 100084;
3. 北京中科研究院, 北京 100049)

摘要:生成式人工智能的迅速普及重塑了社交媒体平台的内容生产与传播方式,使虚假信息跨平台级联扩散、审核成本攀升和责任链模糊等问题愈发突出。现有实践主要依赖“人工+算法”的线性治理与最低合规执行。针对现有研究中技术架构与治理层级映射缺位、协同机制偏静态、安全—效率—公平的价值权衡难以参数化和公开评估难等空白,本文提出“技术架构—治理层级”互构整合的四级协同治理框架,本框架的核心是建立起一套清晰的“谁负责—管什么—如何管—怎样评”的对应关系,把技术架构(如数据、模型、内容分发)与治理层级(制度、行业、平台、社会)打通;在此基础上辅之以双向治理闭环(规则下行/风险上行)和跨平台协同的治理机制,将抽象的治理理念和价值观转译成为可测量、可调整与可审计的评价指标与权重设置。进而,能够将多元协同治理落地为可执行、可评估的中层设计,弥合法规、标准与平台运营之间的落差,在常态与危机情境中实现敏捷切换,缩短响应时滞,抑制级联扩散,以可控的效率代价改善曝光公平与外部问责。为生成式 AI 时代社交平台治理提供了学理基础和系统化分析工具。

关键词:生成式人工智能;社交媒体;多元协同治理;技术架构;平台治理

中图分类号:G91;G301

文献标识码:A

自 ChatGPT 始,大语言模型和生成式人工智能已经深度嵌入社交网络,后者的信息传播模式和内容治理模式为之一新^[1]。截至 2024 年 12 月,我国社交网络用户规模达 11.01 亿人,占网民整体的 99.3%^①。抖音、快手、微博等社交媒体平台已形成覆盖全民的数据传播网络,构建起了强大的社交图谱;内容平台凭借推荐算法与全民创作机制,将舆情扩散速度推向全新高度。人与人之间的直接交往在很大程度上已经被“人—平台—人”的交往方式所取代,数字平台作为技术的掌控者,能够实现对信息推送的主导与操控^[2]。生成式 AI 进一步加速了内容生产和信息传播的效率;在其多模态合成与秒级生成的能力驱动下,深度伪造、虚假信息与算法偏见呈现非线性扩散与跨平台传导,同时,生成式 AI 呈现“基础模型—专业模型—服务应用”的分层业态,无法在现有的“技术支持者—服务提供者—内容生产者”监管框架中找到适配的法律定位^[3]。此外,

用户也从信息接受者转变成为“提问—生成”双向交互的主体,模糊了内容生产者和消费者的界限^[4],导致了内容安全责任溯源极其困难。此外,社交媒体平台在一系列治理实践中逐步成为新型“把关人”^[5],政策法规也赋予平台较大权责^[6]。在生成式 AI 之前,平台能够针对信息传播进行实时干预与控制;但在生成式 AI 的冲击下,传统治理工具面临着技术落后、政策滞后、商业惯性与制度环境的多重制约^[6]。当前所面临的问题是如何将技术框架与治理层级进行可审计、可执行、可评估的有效映射,从而提升平台治理的可操作性与有效性。

1 实践困境与理论缺口

生成式 AI 深度嵌入社交媒体平台引致了政策制定和治理实践的多重困境。传统依赖人工审核和静态阈值的“人工+算法”线性治理模式出现效能

① 中国互联网络信息中心. 第 55 次中国互联网络发展状况统计报告[R]. 北京:中国互联网络信息中心, 2025.

收稿日期:2025-08-04;修回日期:2025-10-29

作者简介:郅雨菲(1992-),女,硕士,E-mail: qieyufei@126.com。张辉(1985-),男,特任副研究员,通讯作者,E-mail: zhanghui_pas@ustc.edu.cn。

滞后与成本攀升^[7,8]。虽然针对生成式 AI 的政策工具正快速上架,例如《生成式人工智能服务管理暂行办法》对训练数据、内容标识、算法安全提出合规要求,但是,“刚性规则”在平台侧经常遭遇“弹性执行”,出现最低合规化与责任外包(如将标识义务转嫁给用户)的现实问题,导致规则的穿透力不足^[9,10]。而即使平台内部治理尝试以多目标算法平衡安全、效率与创作者生态等诉求,如抖音在技术说明中提出“多目标平衡”的算法设计思路,但因如何在危机期进行可解释、可量化的参数调节与权衡,仍然缺少统一的评估框架与公开口径^[11]。

同时,被寄予厚望的多元协同治理陷入了“名实分离”的实践困境^[12]。传统单一主体的治理模式难以为继,创新平台协同治理机制,构建多元主体治理体系已成共识^[13]。然而,在多元协同治理的实践过程中出现以下困境:一是,传统“企业—行业—政府”三元治理模式^[14]因忽视平台内部价值冲突来

源,即平台内部权力嵌套结构,“平台—用户—创作者—广告主”的多元利益网络使其出现行为异化,导致内容治理总体效能低下,如见表 1 所示。二是,社会监督虚置与公众赋权不足。用户与媒体作为治理主体,但仅能通过末端投诉参与,缺乏对算法决策的实质性监督权,多元协同治理理论中的“用户参与”往往流于形式。三是,跨层级治理链条的传导断裂。现有框架多将多元主体协同治理中的主体视为静态的并列主体,忽视了其技术特性驱动的动态博弈关系,导致治理效能逐层衰减。制度层可以通过立法向行业层传递规则,行业层制定标准约束,平台层执行规则并承担社会监督责任,但社会层的公众意见(如用户举报数据、伦理争议)却缺乏向制度层的有效反馈通道;同时,现有治理框架依赖各层独立运作,缺乏风险跨层监测与阻断工具;此外,利益协调的动态博弈失衡,各个主体通过现有机制无法实现利益协调动态均衡。

表 1 平台内部利益相关者的权力博弈

Table 1 Power dynamics among the platform's internal stakeholders

主体类型	核心诉求	治理行为异化表现
平台运营方	用户停留时长、商业收入最大化	默许 AI 生成低质内容提升流量
内容创作者	创作自由与流量分配公平性	AI 批量生产同质化内容挤占生态位
广告主	精准投放与转化率优化	要求平台开放用户行为预测模型接口
普通用户	信息质量与隐私保护	被动接受算法推荐内容失去选择权

资料来源:作者自制。

现有研究从法律规制、技术机理与复杂系统视角提供了重要启示,但存在三类关键理论缺口。一是,“技术架构—治理层级”的映射缺位。技术架构(基础模型—专业模型—应用服务—终端交互)与治理层级(制度监管—行业标准—平台决策—社会监督)之间缺少可审计、可执行、可评估的对应关系,导致责任链与工具链脱节。国际框架强调“角色—义务—控制点”的对齐,例如欧盟《人工智能法案》将义务分别施加于提供者与部署者^[15];NIST《AI 风险管理框架》提出将 Govern—Map—Measure—Manage 功能嵌入组织流程^[16];ISO/IEC 23894 强调生命周期早期的风险识别与控制项对齐^②。在社交平台场景下,这种映射的必要性在于把源头风险(如数据偏见与水印/溯源)前置到模型层,把分发

调参与补偿机制落在平台层,把众包验证与第三方审计放于社会监督层,从而降低系统性风险与执法成本,并缩短跨层干预时滞^[17]。二是,协同机制偏静态。多元主体协同常被抽象为并列参与,缺乏“规则下行—风险上行”的双向闭环与跨平台的快速协同机制,导致制度层的规则难以转化为行业层的技术协议、平台层的可执行参数与社会层的有效反馈,这会导致政策在执行落地过程中的信息衰减与执行偏差,以及新型风险反馈迟滞等问题。三是,价值权衡缺乏参数化。参与治理的各个主体天然拥有不同的,甚至相互冲突的价值偏好和利益诉求^[18]。例如大语言模型平台治理的目标和策略不仅旨在维护用户权益和平台秩序,而且还需平衡其

② ISO, "ISO/IEC 23894:2023 – Guidance on risk management," 2023 [R/OL]. <https://www.iso.org/standard/77304.html>. 最后访问时间:2025-12-30。

商业目标^[19]。协同治理的过程,本质上就是不同主体带着各自的价值偏好,通过协商、博弈、妥协,最终形成一套权责利分配方案的过程。安全、效率、公平三维价值之间的权衡多停留于原则性描述,缺少参数化、可解释与可参与的评价方法与调参流程,这就往往会导致黑箱决策、问责困难,监管者或平台管理者在压力下容易走向极端等问题,从而使得协同失效。

围绕上述实践困境和理论缺口,本文主要关注三个问题:其一,如何建立与技术架构相匹配的治理层级与权责边界,形成结构化分工与可追责链条?其二,如何设计能使静态架构“动起来”的治理机制,形成规则自上而下传导与风险自下而上反馈的治理闭环,并实现跨平台协同触发?其三,如何将安全、效率、公平三维抽象价值转化为可调参数与公开指标,构建参与式、可解释的权衡与评估方法?

本文的主要贡献包括:提出“技术架构—治理层级”的四层级协同治理架构(制度监管层、行业标准层、平台决策层、社会监督层),明确主体—权责—工具—指标的一致化表达;设计“运行机制与价值调谐”,以双向治理链(规则下行/风险上行)与分布式协同网络(监管—行业—平台—社会分工与协作)构成动态机理,强调危机场景下的熔断触发与常态场景下的参数化调节与透明度披露;构建安全—效率—公平的指标体系与参数化参与机制,将抽象价值冲突转化为可计算、可解释、可参与的工程问题。

2 文献综述

围绕生成式人工智能与社交平台治理,现有研究主要有技术治理、政策治理与复杂系统三个方向。它们分别贡献了可检测、可规制与可解释的知识,但仍难以解决上文中界定的缺口问题。

技术治理集中在深度伪造检测、内容标识/溯源与跨平台预警。检测综述显示,在特定数据集上识别模型表现优异,但对新型伪造、域外分布与对抗策略敏感,稳健性不足,难以直接托举平台级治理^[20,21,7]。且,这些方案多停留在“治理对象层”的

点式改良,难以回答“谁在何层以何工具承担何种义务并用何指标评估”的系统性问题^[22],也就难以完成从技术架构到治理层级的映射。并且技术派忽视平台运营逻辑对算法部署的扭曲:一是,经济效率逻辑压制安全逻辑:如 2020 年 YouTube 为降低人工审核成本,使用 AI 审核造成大范围的误伤^③。二是,算法黑箱可能引发二次风险。如法国巴黎检察院已就 X 平台涉嫌算法操纵及其可能影响欧洲选举、提高部分极右翼政治内容可见性等问题展开调查^④。

政策治理提供了风险分级与角色义务的蓝本。欧盟《人工智能法案》区分提供者与部署者并施加差异化义务^[15];NIST AI RMF 提出 Govern - Map - Measure - Manage 的组织职能闭环^[16]。但在平台场景中,存在“最低合规”的弹性执行与透明度不足,外部审计与公众监督难以进入内容审核与算法决策的隐形基础设施^[23]。第三方审计虽被倡议,但接口与制度化保障薄弱,常流于一次性评估。

复杂系统与平台研究揭示了虚假信息的非线性扩散与级联效应,解释了危机期平台治理为何失灵^[24];制度逻辑则刻画国家安全、平台效率与社群公平的价值冲突^[25];治理匹配与多中心治理提出了分层协同与网络协作的必要性^[26]。但这些框架多停留于“为何”与“应该”,缺乏将安全、效率、公平转译为可调参数与可公开度量的具体方法。公平性研究还指出了指标间的不可兼容与固有权衡,进一步增加平台实操难度^[27]。因此,现有研究解释力强,却在“如何将价值落成可计算与可审计的运营参数”上提供有限支撑。

3 理论整合与分析框架

为回应上述理论不足,本文在剖析上述理论的假设前提之后,对三者进行结构性整合,形成可执行的中层框架。进而夯实本文讨论的四层级协同治理架构的理论基础。

3.1 复杂适应系统(CAS):侧重于解决“系统动态性”问题

③ 全媒派。Q2 下架上千万视频,YouTube 承认机器出现误删[EB/OL]. <https://view.inews.qq.com/k/20200831A032ZT00?no-redirect=1>. 最后访问时间:2025-12-30。

④ Reuters. French probe of Musk's X deepens as police are called in [EB/OL]. <https://www.reuters.com/legal/litigation/paris-prosecutor-contacts-police-probe-into-social-media-platform-x-advances/>. 最后访问时间:2025-12-30。

该理论将社会技术系统视为由大量异质性主体(如用户、平台、模型提供商)组成的复杂适应系统。系统行为是主体间、主体与环境间非线性互动的“涌现”结果。其核心特征包括:多主体适应性(主体会学习并改变策略)、非线性(微小输入可能引发巨大输出)、反馈回路(系统结果会反过来影响主体行为)以及时滞与阈值效应(影响不会瞬时发生,且存在引爆点)^[24]。在生成式AI冲击下,平台中的虚假信息的传播往往不是按“线性、可预期”的轨迹扩散,而是呈现出非线性特征(小触发可能引发大规模爆发)和级联效应(一个节点或群体被触发后,带动其他节点),因此如果治理沿用静态、线性的思路(固定规则、单向管理、按平均效应设计),往往会失效。

针对当前的治理协同机制偏静态的问题,本文借鉴了控制与系统工程中的前馈—反馈闭环思想:系统维持稳定必须依赖双向信息流的原理。“前馈”是预设规则的下行,“反馈”是运行状态的上行,二者形成闭环。将治理机制从“并列参与”的静态结构,改造为具有动态响应能力的流程机制。首先,治理过程中,可触发的双向链,即在治理层级间建立了制度化的信息通道,允许底层风险信号“上行”触发高层干预,也允许顶层规则“下行”进行精准调控。其次,使用多模式干预机制(常态微调、预警升级、危机熔断),这是对非线性与阈值效应的直接操作化。它承认不同级别的风险需要性质完全不同的干预力度和方式,从而实现了资源的有效配置和系统的弹性维护。最后,预留可观测通道与时间窗,这是对时滞问题的直接应对,旨在缩短从风险识别到干预生效的周期,抑制级联放大。总体而言,本文将“规则下行—风险上行”构造成可触发的双向链,设置不同强度的监测与干预模式(常态微调、预警升级、危机熔断),并要求各层预留可观测的反馈通道与时间窗,用以缩短干预时滞、抑制级联放大^[24]。

3.2 制度逻辑理论:侧重于化解“目标冲突”问题

制度逻辑理论认为,社会领域由不同的“制度逻辑”所主导(如国家、市场、专业、社群等),每种逻辑都提供了关于何为正当、合理目标和合适治理方式的独特框架。当这些逻辑在同一场域时,它们之间的冲突、竞争与妥协构成了治理的核心过程^[18]。本文采纳了国家(安全合规)、平台(效率收益)、社群(公平正义)这三类核心逻辑在生成式人工智能

治理场域中的共存与内在冲突。承认这些冲突并非偶然,而是源于不同主体依据其所在的制度场域所具有的、根深蒂固的合法性诉求^[25]。

进而,本文针对“价值权衡缺乏参数化”的不足问题,首先,为价值冲突构建了一个结构化的操作空间,而非试图消除冲突。其次,将抽象的“制度逻辑”转译为治理框架中具体的“约束条件”(如法律法规)和“优化目标”(如平台商业化^[28],用户公平感)。最后,区分“硬性约束”与“软性目标”,设置优先级排序机制,划定了治理底线(主要源自“国家”逻辑的合规要求)和可以协商、权衡的领域(“市场”与“社群”逻辑的目标)。为后续的参数化设定了清晰的边界,避免了在基本原则问题上产生争议^[25]。

3.3 治理匹配与多中心协同理论:侧重于解决“权责落地”问题

“治理匹配”强调治理安排必须与待解决问题的特征(如地理尺度、复杂性、动态性)相匹配。“多中心治理”则反对单一权威中心的垄断,主张在特定规则下,由多个相互独立的决策中心(如政府、行业协会、社区、平台)通过合作、竞争与协调来进行共治。本文借鉴了其分层互补与网络协同中的“多尺度”治理思想,即不同层级(如国际、国家、行业、地方)处理不同尺度的问题,同时通过网络化协作弥补层级治理的不足。此外,本文针对多中心权威与协作,采纳了Ostrom等人的核心观点,即有效的治理需要多个权力中心的共同参与,并依赖于它们之间自发形成的协作规则。针对“技术架构—治理层级映射缺位”的不足,本文设计旨在建立一条清晰可审计的责任链。它要求明确地将技术架构的每一个“治理对象”(基础模型、应用服务等)与治理层级的特定“主体—权责—工具—指标”相对应,确保每一个技术环节都有明确的治理责任归属。在行业层引入互操作标准与联动触发(如特征共享、阈值互认),以降低跨平台协作成本,避免“影子层级”与“空心协作”^[26]。

3.4 四层协同治理架构的设计原则

基于上述整合,本文提出三条可执行设计原则,并映射到“制度监管—行业标准—平台决策—社会监督”的四层治理架构之中。

原则1:映射原则,解决“责任链与工具链脱节”的问题。建立一张精细化的“治理网络”,让治理的“责任”能精准地对接到技术的“实体”上。通过“治

理对象^⑤—主体—工具—指标”这四个元素,为技术栈的每一层(如模型、服务)都明确指定:负责的主体是谁、使用什么治理工具,以及用哪些指标来评估效果,避免了责任虚化。其优势:前置了源头治理义务(如将偏见治理置于模型层而非仅靠内容审核),并通过行业层的互操作标准与阈值互认(如统一的 C2PA 元数据、告警格式),为跨平台协同提供了可执行的语法基础,避免了责任悬空与协作空洞。

原则 2:动态原则,解决“协同机制偏静态”问题。治理不是一次性的立法,而是一个持续循环、双向反馈的“活系统”。通过规则下行,将法律条文“编译”成机器可读的技术协议和平台可执行的参数,确保制度意图能落地。通过风险上行,建立渠道,让基层(平台、用户)发现的风险信号能快速触发中高层(行业、制度)的响应和规则迭代。跨平台特征共享、联动告警、危机熔断机制等,能显著降低跨层干预时滞,并使得第三方审计与众包验证能通过标准接口有效接入,从而将多元协同从理念转化为可操作的、具备韧性的治理实践。

原则 3:参数化原则,解决“价值权衡黑箱化”。将抽象的价值争论,转化为可测量、可调整、可审计的技术选择。首先是量化:为安全、效率、公平等价值定义具体的度量指标。其次是权重:承认价值冲突的必然性,采用多目标优化方法,明确在不同场景下的各项价值的优先级。再次是约束,设定不可逾越的底线(比如合规红线),将优化限制在安全空间内。最后是透明:公开关键参数和调整记录,使价值取舍过程从“黑箱”变成为可解释、可监督的。使平台在具体情境下的价值取舍成为一个透明、可解释、可复核的技术—管理流程,并通过披露关键指标与变更记录,为有效的外部监督与审计提供了基石。

4 四层协同治理框架的构建与映射机制

生成式人工智能的治理困境,根源在于技术体系与治理体系的结构性错配。为了系统回应上文中界定的三类研究缺口,即:技术架构与治理层级的映射缺位、协同机制的静态化、价值权衡的非参数化——本章旨在超越原则性论述,构建一个可执行、可审计的工程化治理架构。本框架的核心是“技术—治理”双轴设计,通过将治理权责精准“定位”到

技术组件的接口与参数上,实现系统性风险的精准阻断和动态适应。

4.1 “技术—治理”双轴定义

(1)纵轴治理层级(G1—G4):谁来做、谁负责
不同的治理层级,依据其权能、资源与信息优势,在本框架中承担差异化且互补的角色。

制度监管层(G1):主体为立法与监管机构。其核心职能是设定具有强制力的合规底线,定义风险分级与角色义务(如《人工智能法案》中的提供者与部署者),并行使执法与审计授权。其主要治理工具为法律法规、行政命令与国家级备案库,关键评估指标包括法规合规率、执法覆盖范围与系统性风险整改时效。

行业标准层(G2):主体为标准组织、行业联盟与认证机构。其核心价值在于化解政府直接干预的“技术盲区”,通过制定互操作规范(如 C2PA)、检测评测协议、跨平台告警字段与触发阈值互认,实现行业自律与敏捷协同。其效能指标包括技术标准采纳率、互操作覆盖度与跨平台协同时滞。

平台决策层(G3):主体为平台内部的产品、安全、合规与算法团队。他们是治理规则的最终执行者与价值权衡的日常操作者。其工具集最为丰富,包括接入评测 API、算法策略引擎、动态调整分发参数、维护透明度与申诉接口等。关键绩效指标包括内容误伤/漏判率、风险处置时滞,以及安全、效率、公平等多维价值的达成度。

社会监督层(G4):主体为第三方审计机构、学术研究者、媒体与公众用户。他们构成治理体系的“免疫系统”,通过众包验证、独立测试、对照评估与风险上报等工具,对前述层级进行制衡与补充。其监督有效性通过问题发现率、平台纠偏成功率、回应时效与审计可重复性等指标衡量。

(2)横轴技术架构轴(T1—T4):治理的技术对象

治理的有效性,首先取决于其能否针对关键的技术实体。技术架构轴的划分,确保了治理触点覆盖 GenAI 内容生命周期的关键环节。

数据与来源层(T1):此层聚焦于信息的可信起源,其核心治理对象包括训练数据、平台用户生成内容(UGC)以及用于验证出处的元数据治理对象(如遵循 C2PA 标准的清单)。目标是实现数据偏见最

^⑤ 治理对象一般是指“被治理的具体技术对象”;如训练数据、AI 模型、推荐算法、用户界面。

小化、来源可验证与违规内容溯源。关键接口包括数据风险标签、来源元数据(数字签名、编辑历史)及合规审计记录。

模型与评测层(T2):此层是系统性风险的源头。治理聚焦于基础模型与专业模型本身,通过安全、偏见、鲁棒性等评测基准,以及输出水印、安全门控(Safety Gating)等技术,在内容生成前实施前置风险控制。关键接口包括评测 API、基准测试套件及风险类别映射。

应用与安全中台层(T3):此层是治理规则的“中央处理器”,将政策语言转译为机器可执行的逻辑。它承载内容审核、事实核查、策略引擎、风险评分与 A/B 测试防护等中台服务。关键接口与治理对象包括策略规则库、审核 workflow、证据管理系统以及实现跨平台协同的告警与特征共享接口。

分发与终端交互层(T4):此层是治理的“最后一米”,直接决定用户感知与社会影响。治理对象包括推荐与排序算法、内容可见性控制、用户告知与标识,以及申诉与透明度界面。关键接口包括分发权重与节流参数、标识 UI 组件、申诉通道及用户可访问的处置日志。

4.2 “技术 – 治理”双轴的映射机制

治理框架由纵轴与横轴交叉组成,类似一个“治理棋盘”。其中纵轴(治理层级轴):回答的是“谁来做、谁负责”的问题(见表 2);横轴(技术架构轴)回答的是“对什么对象进行治理”的问题(见表 3)。这个“治理棋盘”上的每一个格子,都是一个“治理单元”,它明确规定了:某个治理主体,针对某个技术环节,使用某种工具,要达到某个指标。

表 2 四层协同治理框架纵轴 G1 – G4(治理层级轴)

Table 2 Vertical axis G1 – G4 of the four – layer collaborative governance framework (governance hierarchy axis)		
治理层级	核心问题	主体举例
G1:制度监管层	什么是绝对不能做的?	政府、立法机构(如国家网信办)
G2:行业标准层	我们应该怎么做才能互联互通?	行业协会、标准组织
G3:平台决策层	我们平台具体如何执行和平衡?	平台公司(如抖音、微博的算法团队)
G4:社会监督层	你们做得对不对?好不好?	用户、媒体、第三方审计机构

资料来源:作者自制。

表 3 四层协同治理框架横轴 T1 – T4(技术架构轴)

Table 3 Horizontal axis T1 – T4 of the four – layer collaborative governance framework (technical architecture axis)		
架构层级	核心问题	治理对象举例
T1:数据与来源层	信息从哪来?是否干净可信?	训练数据、用户发布的内容(UGC)、元数据(如 C2PA 来源证明)
T2:模型与评测层	AI 模型本身是否安全、无偏见?	基础大模型(如 DeepSeek)、专业模型、模型的评测报告和水印
T3:应用与安全中台层	如何将规则转化为自动化的行动?	内容审核系统、策略引擎、风险评分、跨平台告警接口
T4:分发与终端交互层	信息如何到达用户?用户如何反馈?	推荐算法、内容标识(如“AI 生成”标签)、申诉渠道

资料来源:作者自制。

在纵轴与横轴交叉组成的每个“治理单元”(棋盘格子)里,都包含四个不可或缺的元素,也就是上文提到的“主体—治理对象—工具—指标”四元组,这四者构成了一个完整的责任闭环,确保了“事事有人管,管了有方法,做了有结果,结果可衡量”:

主体(Who):执行治理动作的责任方。如:网信办、行业联盟、平台安全团队、用户。

治理对象(What):被治理的具体技术对象。如:训练数据、AI 模型、推荐算法、用户界面。

工具(How):主体所使用的治理手段或“武器”。如:法律法规、技术标准、算法参数、举报按钮。

指标(How Well):衡量治理效果的可量化标准。如:合规率、处置时效、用户满意度。

表 4 是完整展示映射关系架构图,这个框架考虑到了从 G1 到 G4、从 T1 到 T4 的所有可能性。为确保本框架的论证既具深度又兼顾广度,表中标注(例)的部分是我们选取的几个常用的治理单元,能够覆盖

“治理层级轴(G1 – G4)”与“技术架构轴(T1 – T4)”的所有维度。

表 4 四层协同治理框架映射全景图

Table 4 Panoramic view of the four – layer collaborative governance framework mapping

架构层级	T1 数据与来源层	T2 模型与评测层	T3 应用与安全中台层	T4 分发与终端交互层
G1 制度监管层	G1 × T1	G1 × T2 (例)	G1 × T3	G1 × T4 (例)
G2 行业标准层	G2 × T1 (例)	G2 × T2	G2 × T3 (例)	G2 × T4
G3 平台决策层	G3 × T1	G3 × T2	G3 × T3	G3 × T4 (例)
G4 社会监督层	G4 × T1	G4 × T2	G4 × T3	G4 × T4 (例)

资料来源:作者自制。

G1 × T2(监管层 × 模型与评测):针对基础模型本身的偏见、错误或恶意能力,从源头进行控制。如何通过映射解决:主体——监管机构是唯一有权威对模型本身进行准入管理的实体。治理对象——基础/专业模型与评测报告,监管的直接对象是模型这个“技术实体”及其性能证明。工具——风险分级、合格评定、前置义务,监管不直接干预技术,而是制定规则。例如,像欧盟《AI 法案》那样,将模型按风险分级,禁止某些高风险应用,并对提供者和部署者设定不同的法律义务。模型上线前必须通过合规评测(合格评定)。指标——高风险场景合规通过率、评测覆盖度,衡量监管是否有效不是看抓了多少违规,而是看有多少高风险模型在上市前就被合规门槛挡住。

G1 × T4(监管层 × 分发与终端交互):针对用户在面对平台强大的算法时处于弱势,知情权、申诉权等基本权利可能受损的问题。如何通过框架解决:主体——监管机构设定平台必须遵守的用户权益底线。治理对象——用户标识与申诉界面、处置日志,监管的对象是用户能直接看到和交互的界面与流程。工具——透明度与申诉最低要求、用户告知规范,监管规定平台“必须做什么”。例如,法律强制要求 AI 生成内容必须有清晰标识,必须设立有效的申诉渠道,并定期发布透明度报告。指标——披露完整性、用户申诉闭环时效,衡量平台对用户权益的保障程度。例如,检查标识是否清晰,用户申诉后平均多久能得到解决。

G2 × T1(行业标准层 × 数据与来源):针对网络信息真假难辨,需要一套通用的“防伪身份证”技术标准。如何通过框架解决:主体——标准组织/联盟负责制定这套通用的“身份证”格式,让所有平台都能识别。治理对象——C2PA 元数据与签名,记录

了谁创造的、用什么工具、修改过哪些信息。工具——来源证明技术规范、编辑历史必填字段,行业层定义必须包含哪些信息(如签名、编辑历史),以及这些信息的格式。指标——C2PA 采纳率、元数据完整度,衡量这项标准在行业内有多普及,以及大家填写的信息是否完整可靠。

G2 × T3(行业标准层 × 应用与安全中台):针对虚假信息等风险会跨平台流窜,单个平台治理如同“打地鼠”的问题。如何通过框架解决:主体——联盟/认证机构扮演“信息枢纽”和“调度中心”的角色。治理对象——告警字段规范与响应 SLA,定义平台间通报风险时使用的“电报格式”和“响应时间要求”。工具——跨平台告警与特征共享格式、触发阈值定义,这是协同的“操作手册”。它规定当某个风险内容(如一个哈希值)在 A 平台被发现,应以什么格式、多快的速度共享给 B、C、D 平台;以及达到什么阈值就应自动触发全网警报。指标——协同触发次数、平均时滞,衡量联防机制是否被激活,以及从发现到协同处置的速度。

G3 × T4(平台决策层 × 分发与终端交互)即针对平台需要在安全、效率、公平等相互冲突的价值目标间进行日常的、实时的权衡问题。如何通过框架解决:主体——平台产品与运营是真正的“价值权衡操作员”。治理对象——分发权重、标识 UI、申诉通道。工具——常态、预警、危机三档策略切换、节流与降载。常态下可能效率权重高一些;发现谣言扩散时,调高安全权重,对相关内容限流;出现重大突发危机时,启动“熔断”,优先确保安全。指标——处置时滞、效率损失、公平暴露改进,量化权衡的结果。为了安全,处理速度慢了多久(处置时滞)? 流量减少了多少(效率损失)? 原创作者是否

得到了更多保护(公平暴露改进)?

G4 × T4(社会监督层 × 分发与终端交互)针对防止平台既当“运动员”又当“裁判员”,需要外部力量对其进行监督和制衡。如何通过框架解决:主体——第三方审计/公众作为独立的“第三只眼”。治理对象——透明度报告与处置日志副本,这是社会监督的依据。工具——独立审计与众包验证、申诉与复核机制,社会力量的“监督工具”。审计机构像“会计师事务所”一样核查平台“账目”;用户通过众包像“朝阳群众”一样发现平台未察觉的问题。指标——审计覆盖度、复核通过率衡量外部监督的力度和效果。审计了多少关键环节?用户的合理申诉有多少得到了改判?

5 治理框架的运行机制:从静态架构到动态治理

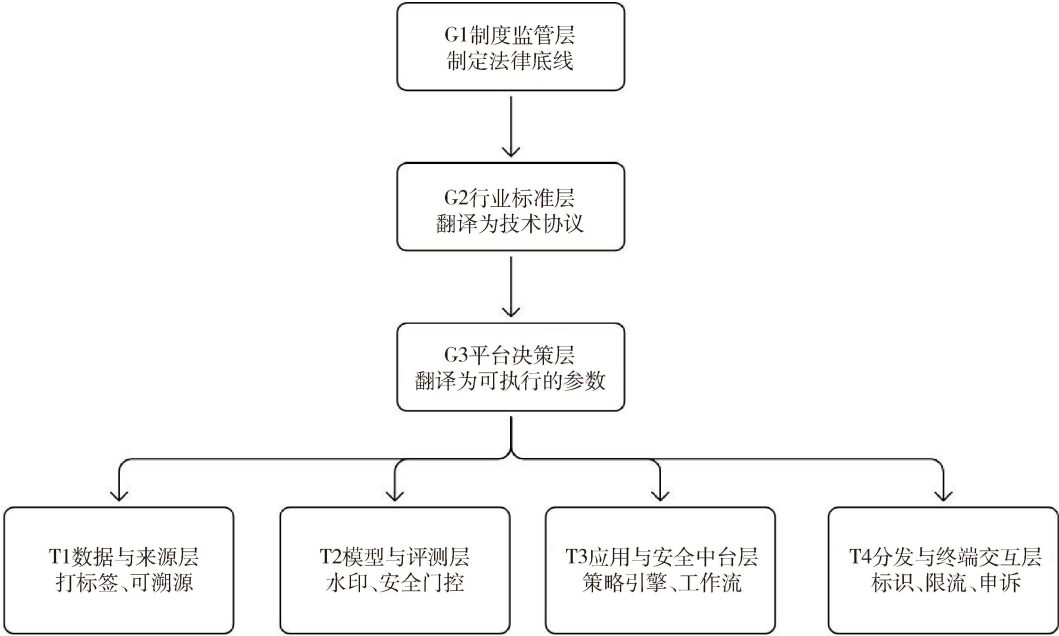
本文框架的生命力在于其动态治理机制。本文设计了一个由“规则下行”与“风险上行”的双向闭环系统,并由一个三档状态机进行全局调度,最终通过参数化价值调谐实现精准治理。该系统的根本目标,是将第四章的静态架构,转化为一个能够自主适应、持续演进的“治理有机体”,从而系统性地解决引言中提出的三大研究缺口。

5.1 “规则下行”链:化解“映射缺位”与“执行衰减”

传统做法中一部法律或规定出台后,平台只知道“不能做什么”,但不知道“在技术栈的哪个环节、用什么工具来实现”,责任是模糊的。平台依靠内部理解和商业利益来“弹性执行”,经常采用成本最低的“末端兜底”式审核,导致政策意图在落地时效果大打折扣。规则下行链的创新与闭环体现两点:一是,规则的“技术编译”,不是简单的下发文件,而是一个将法律语言逐层“编译”成机器可执行参数的过程。通过从上至下的逐层“转译”,将抽象规则精准锚定在具体的技术“治理对象”上,实现了“主体-工具-指标”的刚性互构。避免了政策在落地过程中的理解偏差与弹性执行(见图 1)。这使得规则不再是一纸空文,而是变成了嵌入到技术组件中的可执行代码。同时,通过工具链的刚性约束,确保了责任在技术架构中被“锁定”,解决了“责任链与工具链脱节”的问题。

G1→G2(法律→技术协议):监管说“必须标识 AI 内容”,行业层将其编译为统一的 C2PA 元数据规范。这相当于为法律生成了标准化的 SDK(软件开发工具包)。

G2→G3(协议→平台参数):平台接入这个 SDK,将其转化为内部具体的策略引擎规则、水印检测 API 的调用频率,以及推荐算法的降权权重。



资料来源:作者自制。

图 1 规则下行链示意图

Figure 1 Schematic of the downward transmission chain for rules

5.2 “风险上行”链与状态机:从“人工报警”到“免疫系统”

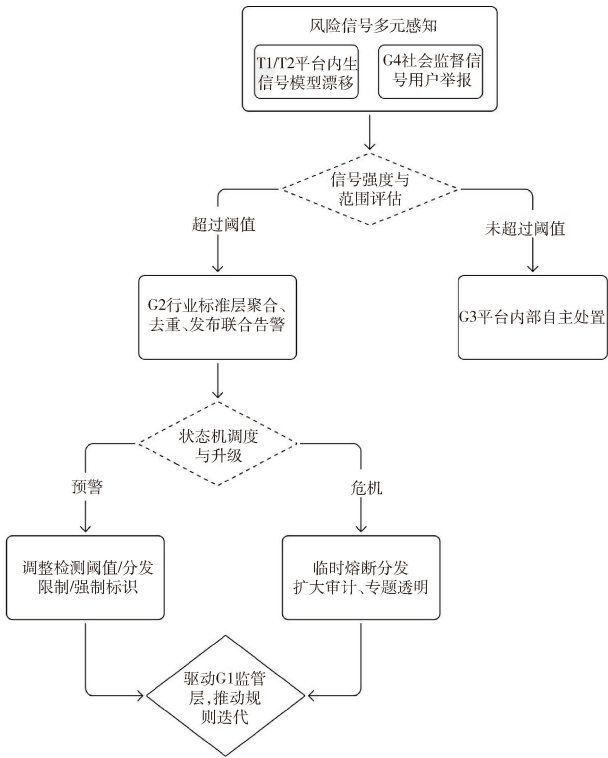
传统做法中平台间是信息孤岛,一个风险在 A 平台被封杀,可能在 B 平台再次爆发。协同依赖临时的“救火会议”,效率低下。同时,用户举报和社会舆论作为风险信号,是分散、嘈杂且非标准的。它们很难被系统地聚合起来,用于驱动上层规则的迭代。

风险上行链将分散的风险信号(用户举报、模型漂移、传播加速)标准化为可聚合的数据指标。G4/G3→G2(信号→情报):行业标准层(G2)扮演了“中央情报局”的角色。它接收各平台上报的标准化告警(如“虚假内容哈希值 X 在 3 分钟内扩散加速 50%”),进行去重、验证和特征提取,生成可行动的、跨平台的风险情报。此外,设置状态机(常态-预警-危机)实现“基于阈值的触发式治理”:当风险指标超过某个阈值时,系统会自动从“常态”切换至“预警”模式,无需等待人工审批。这意味着协同响应从“请示汇报”变成了“条件反射”。例如:一旦触发“预警”,预设的协同协议(如 G2 × T3)会自动执行:所有成员平台自动下调该类内容的权重。这实现了秒级的跨平台风险阻断。状态机将风险信号与治理动作直接、快速地关联起来,形成了一个感知-决策-执行的快速反馈环,极大地抑制了风险的级联放大(见图 2)。

5.3 价值调谐:破解“权衡黑箱”

传统治理中,“安全、效率、公平”之间的权衡往往陷入原则性的哲学辩论,或在平台黑箱中成为一种不可言说的“艺术”,导致决策不透明、问责无依据。本框架的价值调谐机制,其根本创新在于为价值权衡建立一个结构化的操作空间,将其从一个主观的、政治化的过程,转变为一个可测量、可调整、可解释、可参与的技术流程。本研究提出以下几个步骤来解决此问题。

首先,价值参数化:一是指标化,安全不再是模糊的“内容安全”,而是具体化为高风险内容漏判率、虚假信息处置时效、事后复发率;效率从抽象的“运营效率”,量化为单条内容平均审核成本(算力/人力)、系统分发吞吐量、创作者生态整体活跃度。公平从道德的“应然”,转化为可计算的原创内容流量分配的基尼系数(衡量注意力经济公平)、不同群体的内容曝光差异度(衡量算法偏见)。二是参数化:这些指标成为算法和策略引擎可以理解的输入



资料来源:作者自制。

图 2 风险上行链与状态机示意图

Figure 2 Schematic of the upward risk escalation chain and state machine

与输出。抖音就宣称自己采用的是多目标优化算法,运营者不再争论“要不要公平”,而是直接调整公平补偿权重,这个参数将直接影响推荐算法。

其次,为价值权衡划定“安全操作区”。第一步:设定硬约束(不可逾越的红线)。这些主要源自国家逻辑(G1)和基本伦理底线。例如:对于法律明确禁止的极端内容漏判率 < 0.1%;用户申诉平均响应时长 < 24 小时;禁止使用特定敏感属性(如种族)进行定向。硬约束划定了治理的“安全操作区”,所有优化都必须在这个空间内进行。这解决了“为了效率牺牲安全”或“为了安全完全扼杀创新”的极端化困境。第二步,在硬约束内进行多目标优化。在安全区内,市场逻辑(效率)和社群逻辑(公平)成为可以且应该被权衡的“软目标”。采用多目标优化算法,寻找帕累托最优前沿,在不损害一个目标的情况下,无法再改进另一个目标的最佳平衡点集合。例如:在保障了最低安全标准后,平台可以探索,将公平系数提升 5%,会相应带来多少效率(如用户平均停留时长)的损失,并做出知情决策。

最后,透明化与审计。为确保上述过程的可信

度,本文提出一个完整的透明化链条:一是,治理日志。任何关键参数的调整(如安全权重的提升)都必须被记录,并关联到调整主体(是谁)、调整理由(为何)、情境状态(何时)以及决策依据(何据)。二是,影响评估报告。参数变更后,系统需自动生成报告,量化展示此次调整对核心指标的影响(例如:“安全权重提升至0.85,导致漏判率下降15%,但误伤率上升8%,创作者流量基尼系数恶化0.05”)。三是透明度界面,向监管机构和经认证的审计方开放一个只读的治理仪表盘,实时监控关键价值参数的状态和历史轨迹。四是参与式校准,引入用户仲裁委员会机制。当平台内部的参数调整引发重大争议时,可启动跨利益相关方的评议,将部分价值权衡(特别是公平性争议)置于一个分布式的民主程序中寻求共识。使得每一次的价值取舍,都不再是神秘的黑箱操作,而是一个被完整记录、理由充分且可被追溯与复核的技术-管理决策。

6 结论与展望

本文针对生成式人工智能深度嵌入社交媒体平台所引发的治理挑战,系统构建了“技术架构—治理层级”互构整合的四级协同治理框架。该框架通过明确“谁负责—管什么—怎么管—如何评估”的可审计映射,打通了从数据、模型、应用到分发的技术环节与制度、行业、平台、社会的治理触点,形成了责任清晰、工具匹配、指标可测的治理闭环。

本研究提出了执行力极强的多元协同治理架构。通过建立“技术—治理”双轴映射机制,将原本割裂的技术实体与治理主体、工具、指标系统对齐,有效弥合了法规、标准与平台运营之间的执行落差。通过“规则下行—风险上行”双向治理链与三档状态机(常态—预警—危机)的协同调度,实现了从静态结构到动态治理系统的转变,显著提升了系统的响应速度与风险抑制能力。实现了价值权衡的参数化与透明化。将安全、效率、公平等抽象价值转译为可测量、可调整、可审计的指标体系与调参流程,使治理决策从“黑箱艺术”走向“可解释工程”,为外部监督与多方参与提供了制度性入口。

未来,本文将通过情景推演与案例过程追踪,进而在真实平台开展大规模在线实验与前后对照研究,识别四层协同治理框架和治理机制中蕴含的因果链条;另外,针对平台治理事件进行真实场景试点

与A/B评估;在多类型平台开展接口级试点,设计前后对照与A/B测试,量化TTD/TTR、扩散强度R、误伤/漏判、曝光公平的改进效果,并进行因果识别;开发更鲁棒的安全、效率、公平指标与组合度量,引入质化评审与用户体验测量,降低“指标替代”与善意扭曲;将框架应用到直播、语音社交、即时通讯与私域社群,研究多模态指纹与一致性核验、实时身份与来源校验的工程路径;以及围绕EU AI Act、NIST AI RMF与国内规制,比较不同法域的层级职责、透明要求与接口开放政策,提出跨境协同的最小互操作集。从而,来研究将本文框架从“可解释的设计”推进到“可复现的证据”,在跨平台、跨场景、跨法域的真实试点中,验证双向闭环与参数化权衡的有效边界;以隐私合规与标准化接口为抓手,降低协同成本、提升审计独立性与社会监督质量,从而将工程化治理与制度化保障协同起来,形成可持续演化的治理体系。

参考文献:

- [1] 张辉,刘鹏,姜钧译,等. ChatGPT:从技术创新到范式革命[J]. 科学学研究, 2023, 41(12): 2113 - 2121. Zhang H, Liu P, Jiang J Y, et al. ChatGPT: From technological innovation to paradigm revolution [J]. Studies In Science of Science, 2023, 41(12): 2113 - 2121.
- [2] 王晓培. 从技术赋权到平台逻辑:社交媒体舆论极化形成与治理[J]. 中国出版, 2023(14): 11 - 17. Wang X P. From technological empowerment to platform logic: The formation and governance of social media opinion polarization [J]. China Publishing, 2023(14): 11 - 17.
- [3] 张凌寒. 生成式人工智能的法律定位与分层治理[J]. 现代法学, 2023, 45(4): 126 - 141. Zhang L H. The legal positioning and hierarchical governance of generative AI [J]. Modern Law Science, 2023, 45(4): 126 - 141.
- [4] 张欣. 面向产业链的治理:人工智能生成内容的技术机理与治理逻辑[J]. 行政法学研究, 2023(6): 43 - 60. Zhang X. Governance oriented to the industrial chain: the technical mechanisms and governance logic of AI-generated content [J]. Administrative Law Studies, 2023(6): 43 - 60.
- [5] 郭小安. 移动短视频的发展乱象与协同治理策略[J]. 人民论坛·学术前沿, 2025(6): 77 - 86. Guo X A. The development irregularities of mobile short videos

- and collaborative governance strategies [J]. *People's Tribune Academic Frontiers*, 2025(6): 77–86.
- [6] 胡岑岑,黄雅兰.超越看门人:社交媒体平台网络暴力治理比较研究[J].*苏州大学学报(哲学社会科学版)*,2024,45(4):172–180. Hu C C, Huang Y L. Beyond gatekeepers: Comparative research on the governance of online violence on social media platforms [J]. *Journal of Soochow University (Philosophy and Social Science Edition)*, 2024, 45(4): 172–180.
- [7] 李旭嵘,纪守领,吴春明,等.深度伪造与检测技术综述[J].*软件学报*,2021,32(2):496–518. Li X R, Ji S L, Wu C M, et al. Deepfake and detection technologies: A survey [J]. *Journal of Software*, 2021, 32(2): 496–518.
- [8] Raman R, Nair V K, Nedungadi P, et al. Fake news research trends, linkages to generative artificial intelligence and sustainable development goals[J]. *Heliyon*, 2024, 10(3).
- [9] 高志宏.回应与超越:生成式人工智能法律规制——以《生成式人工智能服务管理暂行办法》为视角[J].*社会科学辑刊*,2024(5):121–130. Gao Z H. Response and transcendence: Legal regulation of generative AI—From the perspective of the ‘interim measures for the administration of generative AI services’ [J]. *Social Sciences Journal*, 2024(5): 121–130.
- [10] 张欣.生成式人工智能的算法治理挑战与治理型监管[J].*现代法学*,2023,45(3):108–123. Zhang X. The algorithmic governance challenges of generative AI and governance – based regulation [J]. *Modern Law Science*, 2023, 45(3): 108–123.
- [11] 字节跳动.抖音算法的多目标平衡[EB/OL].抖音安全与信任中心. https://trust.douyin.com/article/15383?enter_from=channel_page&channel=transparency. 2025–01–20 / 2026–01–12.
- [12] 高翔,闫钰琪.我国生成式人工智能服务的复合型监管治理框架及其优化——基于政策文本和深度访谈的混合研究[J].*电子政务*,2025(5):2–15. Gao X, Yan Y Q. A composite regulatory governance framework for generative artificial intelligence services in China and its optimization: A mixed – methods study based on policy texts and in – depth interviews [J]. *E – Government*, 2025(5): 2–15.
- [13] 王玉玲.多元协同:平台经济治理的现实困境与优化策略[J].*东南学术*,2025(3):185–197. Wang Y L. Pluralistic coordination: the practical dilemmas and optimization strategies of platform economy governance [J]. *Southeast Academic Research*, 2025(3): 185–197.
- [14] 张旭芳.生成式人工智能的算法安全风险及治理路径[J].*江西社会科学*,2024,44(8):90–100. Zhang X F. Algorithmic security risks and governance paths of generative AI [J]. *Jiangxi Social Sciences*, 2024, 44(8): 90–100.
- [15] Smuha N A. Regulation 2024/1689 of the Eur. Parl. & Council of June 13, 2024 (EU Artificial Intelligence Act) [J]. *International Legal Materials*, 2025: 1–148.
- [16] NIST. Artificial intelligence risk management framework (AI RMF 1.0) [EB/OL]. URL: <https://nvlpubs.nist.gov/nistpubs/ai/NISTAI.100-1.pdf>.
- [17] Kaufhold M A, Rupp N, Reuter C, et al. Mitigating information overload in social media during conflicts and crises: Design and evaluation of a cross – platform alerting system[J]. *Behaviour & Information Technology*, 2020, 39(3): 319–342.
- [18] Greenwood R, Raynard M, Kodeih F, et al. Institutional complexity and organizational responses[J]. *Academy of Management Annals*, 2011, 5(1): 317–371.
- [19] 肖红军,张丽丽.大模型伦理失范的理论解构与治理创新[J].*财经问题研究*,2024(5):15–32. Xiao H J, Zhang L L. Theoretical deconstruction and governance innovation of ethical misconduct in large models [J]. *Research on Financial and Economic Issues*, 2024(5): 15–32.
- [20] Mirsky Y, Lee W. The creation and detection of deepfakes: A survey [J]. *ACM Computing Surveys (CSUR)*, 2021, 54(1): 1–41.
- [21] Tolosana R, Vera – Rodriguez R, Fierrez J, et al. Deepfakes and beyond: A survey of face manipulation and fake detection[J]. *Information Fusion*, 2020, 64: 131–148.
- [22] 李洋,薛澜.颠覆、调适与协同:责任伦理视域下生成式人工智能的多主体治理机制研究[J].*电子政务*, 2025(7):40–49. Li Y, Xue L. Subversion, adaptation and collaboration: A study on multi – actor governance mechanisms of generative artificial intelligence from the perspective of responsibility ethics [J]. *E – Government*, 2025(7): 40–49.
- [23] Gillespie T. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media*[M]. Yale University Press, 2018.
- [24] Keating C B, Katina P F, Bradley J M. Complex system governance: Concept, challenges, and emerging research[J]. *International Journal of System of Systems Engineering*, 2014, 5(3): 263–288.

- [25] Thornton P H, Ocasio W, Lounsbury M. The Institutional Logics Perspective: A New Approach to Culture, Structure, and Process [M]. Oxford University Press, 2012.
- [26] Ostrom E. Beyond markets and states: Polycentric governance of complex economic systems[J]. American Economic Review, 2010, 100(3): 641–672.
- [27] Biega A J, Gummadi K P, Weikum G. Equity of attention: Amortizing individual fairness in rankings[C]. The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, 2018: 405–414.
- [28] 张辉,陈钰汶,曾雄. 深度求索公司创新机制及其效应分析——以 DeepSeek - R1 大模型为例[J]. 科学学研究, 2026, 44(4): 770–779. Zhang H, Chen Z W, Zeng X. An analysis of DeepSeek company's innovation mechanisms and their effects: A case study of the DeepSeek - R1 large model [J]. Studies In Science of Science, 2026, 44(4): 770–779.

Collaborative governance framework for social media platforms in the generative AI era:

Based on the co - construction procedure of technical architecture and governance tiers

QIE Yu - fei^{1,2}, ZHANG Hui^{1,2,3}

(1. School of Public Affairs, USTC, Hefei 230026, China;

2. Institute for AI International Governance, Tsinghua University, Beijing 100084, China;

3. Beijing Institute of Research, USTC, Beijing 100049, China)

Abstract: The rapid spread of generative artificial intelligence is reshaping how content is produced and disseminated on social media platforms, intensifying challenges such as cross - platform cascading diffusion of misinformation, rising moderation costs, and blurred accountability chains. Current practice largely relies on linear “human + algorithm” governance and minimal compliance. Addressing gaps in existing research—namely the lack of mapping between technical architectures and governance tiers, overly static collaboration mechanisms, and the difficulty of parameterizing and transparently assessing the safety - efficiency - fairness trade - offs—this paper proposes an integrated, four - tier collaborative governance framework based on the co - construction of “technical architecture - governance level.” At its core, the framework establishes a clear correspondence of “who is responsible—what is governed—how it is governed—how it is evaluated,” linking technical layers (data, models, content distribution) with governance tiers (institutional/regulatory, industry, platform, societal). It is supplemented by a bidirectional governance loop (top - down rules and bottom - up risk signals) and cross - platform coordination mechanisms, translating abstract governance principles and values into measurable, adjustable, and auditable metrics and weighting schemes. This enables multi - stakeholder collaboration to be operationalized as executable and assessable meso - level designs, bridging gaps between regulations, standards, and platform operations; supporting agile transitions between routine and crisis contexts; shortening response latency; curbing cascading spread; and improving exposure fairness and external accountability at a manageable efficiency cost. The framework provides a theoretical foundation and a systematic analytical toolkit for platform governance in the generative AI era.

Key words: generative artificial intelligence; social media; multi - stakeholder collaborative governance; technical architecture; platform governance