doi. 10. 3724/1005-0566. 20260102

全球开源人工智能博弈中的中国监管方案：经验借鉴与自主构建

廖慧姣，张韬略
（同济大学法学院，上海　200092）

摘　要：我国正处于开源人工智能各方博弈和外部条件快速变化的窗口期，如何通过理性的监管方案设计准确把握，将直接决定中国能否抓住这一历史性机遇，实现在全球 AI 竞争中的战略突围。本文基于多元主义利益集团和“间断—均衡”等理论，设立了一个开源 AI 监管方案的钟摆模型，以回答不同条件下所采取的应然政策。对欧美监管方案进行诊断发现：美国的监管实践与钟摆模型结果高度吻合，由此实现了多元博弈下的平衡；欧盟的监管实践显著偏离模型结果，而错失了突围的黄金窗口，并陷入“监管越严、差距越大”的困境。我国监管方案应符合防控风险的激励走向，体现“激励主导，风险可控”的特征。基于此，建议应先以统筹化思维完善目前的激励政策，并构建多维度、动态化的开源风险预警体系，以及预设前瞻性的类型化框架，为钟摆摆动预留制度空间。

关键词：开源人工智能；钟摆模型；监管方案；风险与价值博弈
中图分类号：D63；D92　　**文献标识码**：A　　**文章编号**：1005－0566(2026)01－0016－11

China’s regulatory strategy in the global open-source AI landscape: Comparative insights and independent framework construction

LIAO Huijiao, ZHANG Taolue
(*School of Law*, *Tongji University*, *Shanghai* 200092, *China*)

Abstract: China is currently situated in a critical window period characterized by the dynamic interplay of domestic open-source AI stakeholders and rapid shifts in external conditions. The design of a rational regulatory framework that accurately navigates this period will directly determine whether China can seize this historic opportunity to achieve strategic breakthrough in global AI competition. Drawing upon theories such as pluralistic interest group theory and punctuated equilibrium theory, this paper establishes a Pendulum Model for open-source AI regulatory policy, designed to determine the appropriate course of action under varying conditions. Policy diagnosis of Western countries reveals that: The United States’ regulatory practice aligns highly with the predictions of the Pendulum Model, thereby achieving equilibrium amid pluralistic interests. Conversely, the European Union’s regulatory practice significantly deviates from the model’s trajectory, leading the EU to miss a golden window for technological advancement and fall into a predicament where “stricter regulation exacerbates the gap”. The analysis suggests that China’s regulatory approach should adhere to an incentive-driven policy characterized by risk control. This approach embodies the principle of “Incentive Dominance, Risk Controllability”. Specifically, China should: Prioritize the systematic

基金项目：国家社会科学基金项目“人工智能大模型对开源软件著作权保护范式的挑战与应对研究”(25BFX178)。

作者简介：廖慧姣(1998—)，女，江西宜春人，同济大学法学院博士研究生，互联网与人工智能法律研究中心研究员助理，研究方向为人工智能法。通信作者：张韬略。

correction and integration of current incentive policies. Establish a multi-dimensional and dynamic open-source risk early warning system. Design a forward-looking, categorized regulatory framework to reserve institutional space for potential future policy shifts (pendulum swings).

Key words: open-source artificial intelligence; pendulum model; regulatory strategy; risk-value dynamics

当前人工智能(artificial intelligence, AI)竞争格局中,开源生态已以不可阻挡之势重塑产业格局和国际政治形势[1]。值得关注的是,我国已借助本土开源 AI 的优势[2]在全球 AI 军备赛中强势突围。然而,开源本身作为一种普惠性技术,其开放精神在鼓励协同创新、降低准入门槛等方面意义重大,但也由于其两用性(dual-use)的特点,衍生出滥用、技术泄漏等风险隐患[3]。因此,开源 AI 的监管方案设定绝非简单的激励或限制问题,而是涉及社会安全、产业发展及国家战略等多维度利弊权衡的复杂博弈过程。在这一背景下,开源 AI 的监管问题已成为全球 AI 治理的前沿议题,逐渐受到各国立法者[4]及学界的关注。现有研究普遍关注到开源 AI 特殊性所带来的监管难题,并提出如"利益平衡—风险防范—协同治理"[5]、"发展导向、国际合作和敏捷影响三大原则"[6]等宏观治理框架,但未能揭示不同外部条件下,不同风险与价值背后博弈主体及其互动关系,因而无法为监管方案在动态演进过程中的应然走向提供分析依据。

为充分透视动态关系下开源 AI 监管方案制定的理想轨迹,本文通过综合借鉴多元主义利益集团理论、"间断—均衡"理论及 PEST 等交叉学科理论方法,尝试构建一个理想监管方案的钟摆模型,旨在回答不同条件下对开源 AI 应采取的应然方案。目前,我国正处于开源 AI 产业和技术快速变化的政策窗口期,其监管方案的布局将直接决定中国能否实现在全球 AI 竞争中的战略突围。然而,我国尚未对开源 AI 的基本监管立场及具体监管路径形成共识。基于此,本文在全面梳理全球开源 AI 博弈格局的基础上,以钟摆模型诊断欧美所施行的监管方案,通过借鉴美国"精准校摆"的成功经验与欧盟"战略自缚"的失败教训,进一步提出适合中国国情的开源 AI 监管方案。

一、开源 AI 监管政策的钟摆模型理论与构建

既有文献分别从"劳动—资本—公共产品"[7]、"公地治理理论"[8]、"伦理—技术"[9]及"价值—风险"[10]等路径探索如何制定中国的开源 AI 监管方案,但在分析范式上存在两大局限:一是侧重静态假设,难以预测政策为何转变;二是单维分析,聚焦特定维度而忽视多因素的交互关系。为拓展动态多维分析视角,本文构建的钟摆模型旨在通过分析不同利益集团的博弈机制与外部变量的影响,确定不同条件下监管机构应采取的合理立场,由此评估现行监管方案的合理性并为监管方案的调整指明方向。相较于既有研究,本文的差异化贡献体现在 3 个层面:一是引入动态机制,模型可分析政策应何时转向和如何演进;二是整合多维视角,在外部条件和内部博弈等多重维度下寻求"适配监管方案";三是设置规范性分析工具,通过对比模型推导的理想方案,实现监管实践的诊断和监管方案的指引功能。

(一)理论基础

根据多元主义利益集团理论,公共政策为不同利益集团之间的一种平衡产物[11],而非单一理性行为者的决策。开源 AI 的监管方案,也并非是某个"全知全能"政府的静态理性设计产物,而是产业、公众及政府等不同利益集团博弈的结果。因此,本文所建立的开源 AI 理想监管方案钟摆模型,其最终的应然走向(既是摆向激励端又是限制端)由这种动态博弈结果决定。

通过系统梳理相关的立法意见①,开源 AI 所涉及的博弈集中在公共安全、国家竞争及产业发展三大核心议题。本文尝试借鉴宏观环境分析中的 PEST 框架②,将上述博弈划分为相互独立的社

① 信息来源:https://www.regulations.gov/document/NTIA-2023-0009-0001/comment.

② 在此框架中,技术(T)本身是分析的起点和动因,而非被分析的影响领域之一,因而不予纳入考察范围。这种做法旨在避免将动因与结果混淆,使分析更聚焦于开源 AI 最为核心的社会经济与治理影响,从而为政策制定者和公众提供一张清晰、全面的认知地图。

会、经济和政治场域。公共政策的调整和转变将由注意力分配驱动[12]，场域的影响权重取决于注意力占比，本文设定了“技术能力、产业生态及地缘政治”三大外部观测变量。由此，三大场域在此权重占比下的加权合力将共同决定最终钟摆模型的走向。

根据“间断—均衡”理论，大多数政策都在很长的时间内保持稳定和渐进演化，若存在某些打破政策均衡状态，引发政策“间接性”变革的外部事件或条件，政策可能会经历快速且根本的变化[13]。在该钟摆下，当某个关键事件发生时，钟摆便进入政策制定的窗口期，发生“标点式”的剧烈摆动。一旦某个政策方向被采纳，它就会产生制度惯性。

（二）钟摆模型的适用范围及运行规律

1. 开源 AI 的定义

考虑到商业实践中主流的开源 AI 模型均为开放权重模型，本文采纳 OSI（open source initiative）对开放权重模型的定义，即指通过允许使用、学习、修改及分享自由的许可证，发布已训练神经网络的最终权重与偏置参数[14]。

2. 三大场域的内部博弈

社会场域涉及公众与监管主体就开源安全的博弈。一方面，开源被公认为可增加模型透明度、向善监管[15]，从而防止模型被滥用或受到攻击[16]。另一方面，开源的无差别访问权限[17]也为恶意滥用敞开了大门[18]，且与模型性能成正比[19]。社会场域主体立场具有条件性与摇摆性：当外部条件平静，无重大安全事件且透明度优势获得认可时，博弈处于中间态；若此时已有可控风险缓解措施，则博弈倾向激励方向；而在重大安全事故、负面案例曝光的安全危机期，钟摆会急速偏向限制方。

经济场域涉及开源利益方与闭源利益方围绕市场利益展开的博弈。开源利益方主张开源极大地促进了 AI 产业的民主化和创新[20]，有利于新入场者和小型玩家跨入 AI 领域[21]。闭源利益方认为，开源方可获取无偿的集体智慧[22]，并借助开源主导力制造假性竞争、实则集中的垄断[19]，损害闭源企业的市场份额。该场域立场取决于两方的力量对比：当开源方占优时，经济场域偏向激励端；当闭源派占优时偏向限制方；当势均力敌时，趋向中间。

政治场域涉及国家政府就标准主导与地缘政治风险间的博弈。开源 AI 可为国家带来超越市场份额的主导“间接权力”[20]，有利于在全球范围内传播其背后的价值观或伦理取向。但开源 AI 同时可能导致先进技术外泄，并对引入国造成后门风险[23]。当本国占据技术和产业优势时，涉及标准主导权和技术外溢的博弈衡量；当本国处于技术弱势端，依赖国外开源模型时，推动钟摆可能摆向限制方向，以防范后门风险。

3. 三大场域的外部变量

技术能力变量决定社会场域的权重。在风险预防理论看来，当技术存在不确定性但有潜在严重的风险时，即使缺乏科学共识，也应采取预防性措施[24]。若开源 AI 的技术水平超乎一定阈值，其不可控风险将急剧放大社会场域的权重，公众安全诉求将压倒其他考量。

产业生态变量决定经济场域的权重。当一个产业在经济上具有关键地位时，公共政策总会习以为常地适用于实业界的必要需要[25]。该变量取决于两方面：一是 AI 产业的生态圈层和规模，规模越大，“监管俘获”能力越强[26]；二是本土产业对开源的依赖程度，权力将在相互依赖关系中分配[27]。即便产业生态不占优势，若本土 AI 发展依托开源，其经济场域影响力仍不可忽视。

地缘政治变量决定政治场域的权重。在大国竞争回归的“后全球化时代”[28]，地缘政治紧张使得议题“安全化”，政治场域影响将超越常规范畴[29]。这一过程由产业升级触发的结构性矛盾奠定基调：中低端产业国家倾向淡化地缘政治色彩，但一旦产业升级进入技术前沿，国家诉求转向供应链安全和标准输出[30]，与守成国利益碰撞。在此基调上，特定事件将使得“咽喉效应”风险[31]变得具体且急迫，从而急剧放大政治场域的影响力。

二、钟摆模型下欧美监管方案的诊断

本节旨在运用钟摆模型推导的理论结果，对欧美开源 AI 监管方案进行诊断，揭示其内在机理，为我国提供经验借鉴。模型诊断遵循以下步骤。①基于钟摆模型，分析不同国家在其不同条件下，

监管方案的理想状态。②梳理欧美实际出台的监管实践及演变轨迹,并与模型的应然结果进行对比:若两者高度吻合,说明该国监管实践较好地平衡了各方利益,则作为正面经验;若出现显著偏离,可透过模型揭示监管实践偏差的原因,可作为反面案例。

本文选取欧盟和美国作为诊断对象的依据如下:美国作为 AI 产业的领先者,与我国"追赶的领先者"在产业生态和竞争格局方面存在一定的相似性,而欧盟早期的产业生态变量及地缘政治变量,与我国具有内在的一致性;从诊断结果来看,美国监管实践的契合及欧盟的背离形成理想对照,可为我国开源 AI 监管方案提供充分的经验素材。

(一)美国:高度吻合模型结果的精准校摆

1. 第一阶段(2022 年下半年到 2023 年中后期)

(1)模型结果:社会安全恐慌下的限制摆动

社会场域在此阶段占据绝对主导权重,且形成高度一致的限制共识。2022—2023 年,AI 技术向人类基准急速靠拢[32]的跃迁趋势促使社会场域下的 AI 安全问题成为众矢之的[33]。2022 年,美国民众对 AI 的正面支持率低至 35%[32]。特别地,2023 年年初,Meta 的 LLaMA 模型权重泄露事件[34],进一步加剧了公众的恐慌。

相比之下,经济场域的权重和影响力远不及社会场域。从外部变量来看,此时的开源模型与先进闭源模型相差甚远[35],尚未达到产业化应用要求,产业对开源的依赖有限。从内部博弈来看,其呈现出明显的闭源倾向格局。以 OpenAI、Anthropic 为主的先进闭源企业获得巨量资金支持[36],并利用其话语权优势,巧妙地将"闭源与负责任""开源等同危险"的叙事逻辑植入监管讨论,使得经济场域偏向限制端。

地缘政治场域虽然已经存在,但并非是主导性议题。由于美国 AI 此时处于一家独大的优势地位,地缘政治担忧尚未被放大。况且,作为技术先进方,美国此时更担心开源 AI 会被中国等国家利用,实现技术"弯道超车"。因此,该场域也偏向限制端。

三大场域的内部倾向呈现出高度一致的限制性取向,形成以社会场域为主导的限制方向合力。根据理论模型,美国在这一阶段的监管政策应呈现安全恐慌下的限制特征:强调开源的安全风险,对模型的开源持审慎态度,要求报备或审查。

(2)监管实践:以"安全"为中心的限制政策

美国在该期间出台的监管政策呈现限制偏向。2023 年,美国参议院举行了多场针对 AI 监管的听证会,开源模型的滥用风险成为核心议题[37]。随后,拜登政府签署的《关于安全、可靠、可信开发和使用 AI 的行政命令》(Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence)要求"任何所有或占有的、具备双用途的基础模型权重信息,以及所采取的物理和网络安全措施都必须报备",并任命 NTIA(National Telecommunications and Information Administration)审查和提出模型开放所带来的政策建议。这些监管实践清晰地反映出美国政府在整体判断上更偏向技术开放的风险防控。

2. 第二阶段(2023 年下半年到 2024 年年底)

(1)模型结果:产业创新主导的中间回摆

社会场域的权重被显著削弱,从限制端回落至中间端。与第一阶段形成鲜明对比,第二阶段的技术能力发展进入常态化轨道(现有模型基本处于 AI 能力量表的中间区间[38]),并未出现恐慌的"超预期跃迁"。公众和监管机构逐渐将 AI 视为"可适应的现实存在",对 AI 持正面态度的比例也上升至 39%[32],博弈结果重新摆向中间。

经济场域权重显著增加,摆向中间端。随着开源模型性能向顶尖闭源模型靠拢[33],美国对开源模型的依赖度大幅提升,以 62% 的使用率位居全球首位[39]。产业生态变量的变化致使经济场域权重显著增加。同时,经济场域的内部博弈发生逆转,从"闭源主导"转向"势均力敌"。Meta 的成功示范[40]及下游应用企业对成本优势的追求[39],促使产业上下游纷纷拥护开源,其政治动员能力显著提升[41]。

政治场域权重随中国的赶超有所提升,内部博弈陷入"标准主导"与"技术外泄"的拉锯战。中国开源模型的追赶[42]奠定了地缘政治的紧张基调,触发了场域内部的激烈博弈:以 Meta 为代表的科技企业主张通过开源强化美国的技术标准主导

地位[43]。而在 2024 年年底,中国利用美国开源模型进行军事用途开发的消息[44]加重了美国政府官员对技术外泄的担忧[45]。美国面临着“被追赶”与“被泄露”的双重危机,博弈结果短期内尚不明显。

虽然场域的权重占比发生大幅变化,但内部博弈结果普遍趋向中间,钟摆模型从“限制端”向“中间地带”摆动。根据理论模型,美国此时的监管政策应呈现产业创新主导的中间端摆向特征:从“如何限制开源”转向“如何平衡开源的风险与收益”,“审慎出台”限制性监管政策。

(2)监管实践:从限制转向观望型政策

美国这一阶段呈现出从“限制”到“观望”的转向。其一,NTIA 于 2024 年 7 月发布《具有双重用途的广泛可用模型权重的基础模型报告》,基于“难以衡量开放模型的风险与收益关系”,美国政府明确应转向事后监管的观望立场[46]。其二,陆续叫停或搁置不利开源的人工智能法案,如旨在通过出口管制限制模型跨境开源的《增强海外关键出口限制的国家框架法案》(ENFORCE Act)及增设“安全护栏和一键关闭”要求的加利福尼亚 SB 1047 法案。其三,从“一刀切监管”转向“差异化监管”。2025 年 1 月,美国商务部发布的《AI 扩散框架》(Framework for Artificial Intelligence Diffusion)已明确将开源权重排除在监管之外,仅限制闭源模型的权重出口行为。

3. 第三阶段(2024 年年底至今)

(1)模型结果:标准主导权驱动的激励摆向

技术发展的持续“常态化”进一步削弱社会场域的权重,其降至历史最低水平,且延续之前的“理性评估”立场,呈现中立态势。由于产业生态变量未发生显著改变,经济场域的权重较为稳定,“开源阵营占优”的力量格局继续维系,博弈继续呈现中间态势。

政治场域的权重却大幅上升,博弈结果指向激励端。中国开源模型的反超及 DeepSeek、Kimi 等咽喉事件的刺激,彻底激活国家结构性矛盾。作为技术守成国,美国希望通过开源模型维持全球 AI 领导地位的战略紧迫性空前提升。此时,“技术泄露”的限制逻辑已经逐渐失效。尽管美国对引入中国开源模型的“后门风险”担忧亦浮现,形成了一股制衡力量,但其紧迫性和影响力尚不足以扭转由“标准主导权”所驱动的强大激励倾向。

政治场域以其绝对主导的激励倾向,推动钟摆模型从“中间端”摆向“激励端”,但其所面临的“后门风险”也同样会被重视。基于模型结果,美国监管方案应呈现以标准主导权为目标的激励摆向特征:出台积极的产业政策,鼓励本土开源企业引领全球技术标准,限制竞争对手的开源模型以防止“后门风险”;积极主导全球开源 AI 治理规则的制定。

(2)监管实践:仅面向本土开源的激励型政策

自 2025 年起,美国施行了一套仅面向本土开源模型的激励型政策,地缘政治成为主导逻辑。其一,激励本土开源模型的发展与出口。2025 年 1 月 23 日,白宫颁布《消除美国 AI 领导地位的障碍》(Removing Barriers to American Leadership in Artificial Intelligence)行政令,“维持和增强美国在 AI 领域的全球主导地位”设为政策首要目标。随后的 DeepSeek[47]及“轻监管促竞争”的听证会[48],均体现了这一目标。2025 年 7 月 23 日,白宫发布《推动美国 AI 技术栈出口》(Promoting the Export of the American AI Technology Stack),希望“通过支持美国原产 AI 技术的全球布局,维持和扩大美国在 AI 领域的领导地位,并降低国际社会对竞争对手所开发 AI 技术的依赖”。同月,《美国 AI 行动计划》(AI Action Plan)亦将“鼓励开源和开放权重 AI”列为加速 AI 创新的关键环节。其二,限制竞争国家的开源模型。美国以安全为由,已禁止多个美国政府机构在联邦设备上使用中国开源 AI[49]。

(二)欧盟:显著偏离模型结果的战略自缚

1. 第一阶段(2022 年下半年到 2023 年上半年)

(1)模型结果:社会安全警惕的限制偏向

与美国情况类似,AI 技术能力的跃迁致使社会场域的安全考虑成为决定政策走向的因素。然而相较于美国公众的极度焦虑,欧盟民众 2022 年对 AI 持正面态度的比例高达 42%[32]。因此,该场域的内部博弈结果虽偏向限制端,但不应过度。

受限于落后的产业发展[50]，经济场域的权重占比较小，但其内部明确指向激励端。欧盟产业内部并不存在闭源集团主导的限制力量，且其擅长利用开源工具进行应用创新。特别是，法国的 Hugging Face 已经成为全球最重要的开源 AI 平台，仅 2023 年上半年托管的模型数量就超过 10 万个[51]。

政治场域的权重处于最低水平，且博弈结果处于中间偏激励端。欧盟处于“追赶者”的竞争态势，尚未出现如美国对外封锁这类引发地缘政治紧张的关键事件。尽管追求“战略自主”与“数字主权”是欧盟的长远目标，但在技术能力与顶尖国家存在显著差距的产业发展初期，政策重心应为避免“错失技术革命”，技术外泄及主权争夺的博弈力量尚未形成，后门风险则相对次要。综上，政治场域的博弈结果处于中间偏激励端。

虽然社会场域占据主导并形成限制方向的合力，但欧盟的经济场域和政治场域却存在向激励端的反向拉力，模型钟摆最终仅呈现微弱的限制偏向。基于模型结果，欧盟监管方案应呈现以下特征：虽有一定限制举措，但已经认识到开源的战略价值，为开源发展留有余地。

（2）监管实践：过度限制的严监管政策

出于对 GPT 等通用目的模型的安全恐慌，2023 年欧盟《人工智能法》理事会提案（General Approach）特别增设了对“通用 AI 系统”的监管条款。这一变动直接导致市面上几乎所有流行的开源模型被纳入监管范围，开源者需满足与闭源模型同等甚至更严格的合规要求，已然远超模型推导的理想政策走向。

2. 第二阶段（2023 年下半年到 2024 年下半年）

（1）模型结果：开源产业主导的激励转向

与美国情况相似，AI 技术能力的常态化发展使得社会场域的权重有所下降，内部博弈逐步从限制端趋向中间端，欧盟民众 2024 年对 AI 持积极态度的比例小幅增长至 45%[32]。

经济场域成为主导力量，并形成高度倾向激励的博弈结果。欧盟本土产业规模和开源依赖度都大幅提升，产业生态变量发生较大变化：Mistral AI 的开源模型有力证明其具备与中美直接竞争的潜力[42]；Hugging Face 以百万级开源规模而成为全球事实上的开源 AI 资源中心[52]，与欧盟形成深度绑定关系。因此，场域内部以 Hugging Face、Mistral AI 等为代表的强大的亲开源产业力量形成，博弈结果高度摆向激励端。

政治场域权重有所上升，且偏向激励端。虽然欧盟的 AI 发展尚未触及引发国家“产业升级结构性矛盾”的临界点[35]，但开源模型的快速追赶也为欧盟迎来了一次战略窗口期。欧盟处于追赶者地位，致使限制开源（如技术泄露或后门风险）的博弈力量微弱。相反，凭借 Mistral AI 的突破和 Hugging Face 的生态枢纽地位，欧盟即便无法在模型性能上全面领先，也可以通过主导开源标准对全球产生影响。因此，此时“主导开源标准”的激励力量应占上风。

在该阶段，经济场域权重占比最大且倾向激励，而政治场域的“主导开源标准”诉求也偏向激励端，钟摆模型应从限制端转向激励端。根据模型结果，欧盟监管方案应呈现以下特征：叫停不利于开源发展的严监管政策，出台激励开源创新的政策以促进本土开源生态发展。

（2）监管实践：纠偏不足的假性激励政策

这一阶段，欧盟对开源 AI 的限制政策进行了表面调整，但纠偏力度远远不够，政策本质仍为限制而非“激励”。其一，开源豁免条款的适用空间有限。欧盟在意识到开源的创新价值后，于《人工智能法》三方会谈草案（Trilogue Draft）中首次增设了开源监管豁免的规定，并最终出台了一套有限的开源豁免条款。但豁免条款的适用范围有限，开源方仍需承担大量合规义务。其二，激励政策的效果有限。欧盟采取了一套旨在“建立欧盟开源大模型”的激励政策，如《促进合法、安全和值得信赖的发展和使用的战略愿景》（A Strategic Vision to Foster the Development and Use of Lawful, Safe and Trustworthy Artificial Intelligence Systems in the European Commission）、《关于促进可信 AI 的创业和创新的交流》（Communication on Boosting Startups and Innovation in Trustworthy AI），但这些带有强烈计划性色彩的政策，并不直接激励开源 AI 的市场发展。

3. 第三阶段（2024 年下半年至今）

（1）模型结果：惯性作用下的激励维持

由于第三阶段的外部变量和内部场域博弈

均未出现显著变化,根据模型理论,在缺乏“关键事件”打破政策均衡状态时,监管方案应延续第二阶段“激励端转向”的趋势,朝着激励端方向保持稳定和渐进演化。因此,欧盟理想的监管方案特征为:鼓励本土开源发展并逐步放松限制。

(2)监管实践:豁免收紧的限制政策

欧盟在这一阶段采取的监管政策,使其监管实践的限制走向愈演愈烈,越来越偏离钟摆的应然轨迹。在欧盟《人工智能法》的执行阶段,欧盟委员会采纳了多位议员所呼吁的“限缩开源豁免条款”的观点[53],以防止诸如 Meta 等巨头公司规避监管。特别地,2025 年 7 月发布的《关于通用 AI 模型提供者义务范围的指南》(Guidelines on the Scope of Obligations for Providers of General-Purpose AI Models Under the AI Act)通过设定严格的适用前提和列举排除例证等形式,大幅限缩了通用 AI 模型的监管豁免空间。即使是欧盟本土流行的开源模型,如 Mistral Large 2 也因存在商业限制条款而难以满足其豁免要求[54]。

(三)诊断与总结

美国的开源 AI 监管实践与钟摆模型推导的应然结果高度吻合,精准回应了每次“关键事件”下场域力量的动态变化,体现了高度理性且适应性极强的“精准校摆”过程。虽然美国政策调整的最终效果尚待时间检验,但其政策制定的内在合理性已经得到充分验证。目前,美国依然保持着 AI 领域的一定领先地位[55],通过精准的政策调控有效平衡了内部场域的多元博弈,并在全球 AI 竞争格局中占据了战略主动权。

相比之下,欧盟的监管实践则在限制道路上陷入越走越远的“战略自缚”,严重偏离钟摆模型推导的应然结果。虽然欧盟在第二阶段曾短暂地进行了激励方向的微调,但其监管惯性最终束缚了调整空间。随后阶段的进一步限制,使得其开源 AI 政策彻底偏离了应然方向。这一偏离已经产生了可观测的不良影响:2025 年欧盟最先进的开源模型性能表现与全球领先水平的差距持续拉大[42],其在全球 AI 生态系统中的影响力不断减弱[56]。欧盟案例构成了一个典型的反向验证,其政策走向偏离模型结果的理性轨迹,不仅让欧洲错失了突围的黄金窗口,更陷入了“监管越严、差距越大”的困境。

三、钟摆模型下我国监管方案的构建

(一)监管实践诊断:符合方向但存有不足

1. 模型结果:防控风险的激励走向

我国的社会场域呈现独特的博弈格局。与欧美国家不同,我国公众对 AI 技术始终保持着高度的积极态度(从 2022 年的 78% 到 2024 年的 83%)[32],但监管机构对潜在风险(技术滥用、安全隐患)保持警惕。因此,无论技术跃迁引发的权重如何变化,在尚无明显风险事件时,我国社会场域的博弈始终处于中间端。

经济场域随着产业生态变化跃升为主导场域,并明确摆向激励端。我国的海量应用场景,致使 AI 的采用率远超其他国家[57],产业基础较大。与此同时,我国已成为全球最先进的开源模型提供者,其性能和生态都超越欧美[56]。由于国内 AI 头部企业普遍采取开源战略,以快速积累声誉并抢占市场(如阿里、DeepSeek 等),内部不存在闭源集团的封锁阻力,因此博弈结果明确倾向激励端。

因地缘政治变量的剧烈波动,政治场域急速上升为次要场域。本土开源 AI 的弯道超车使我国迎来“产业升级”的关键时刻,内在诉求转向“标准制定者”。在政治场域,尽管存在技术泄露和植入后门的限制力量,但由于与美国的技术差距较小(泄漏风险后果有限)且后门风险存在其他解决方法,场域博弈结果仍偏向激励端。

综合各场域的博弈力量,钟摆模型推导的应然走向明确摆向激励端,但需同时为社会及政治场域的双向安全风险提供应对备案。根据模型结果,我国的政策制定应体现“激励主导,风险可控”的特征:一方面,通过强有力的激励政策巩固并扩大既有的开源领先优势,加速构建开源生态护城河;另一方面,对社会与政治场域中的风险因子进行前瞻性的监测与备案,以便在未来的“外部影响因素”来临时,监管机构能够迅速、精准地调整政策,在发展与安全之间实现动态再平衡。

2. 监管实践:缺乏风险关注的冗余激励政策

我国对于开源 AI 的监管实践契合模型的激励大方向。目前我国采取以“应用端”治理为核心的 AI 监管体系,重点规制面向公众的 AI 服务提供者,如基础法律层面的《中华人民共和国个人信息保护法》《中华人民共和国网络安全法》《中华人民共和国数据安全法》,以及针对具体风险场景出台的部门规章,如《生成式人工智能服务管理暂行办法》《人工智能生成合成内容标识办法》等。在这一监管体系下,由于开源模型的发布者通常不直接提供服务,且多在跨境开源平台上发布,开源 AI 事实处于监管的“灰色地带”。与此同时,我国政府在促进开源 AI 发展方面始终保持着积极的政策步调,形成了“中央定调、地方竞跑”的激励格局。随着本土开源生态的繁荣,《第十四届全国人民代表大会第三次会议关于 2024 年国民经济和社会发展计划执行情况与 2025 年国民经济和社会发展计划的决议》中提出要大力支持“构建开源模型体系”,而国务院《关于深入实施“AI +”行动的意见》则明确要求“推动 AI 技术开源可及”。这些中央政策文件的表态标志着开源 AI 已上升为国家科技自立自强的战略抓手。在中央的引领下,各地政府于 2023 年后纷纷将开源 AI 纳入地方规划,推出包括财政奖励(如《北京市关于支持信息软件企业加强 AI 应用服务能力行动方案(2025 年)》)、资金资助(如《广东省推动 AI 与机器人产业创新发展若干政策措施》)及项目引育(如《杭州市 AI 全产业链高质量发展行动计划(2024—2026 年)》)等在内的多项扶持政策。

我国监管实践方向虽符合模型结果,但其对场域内部风险的关注不足及当前激励政策的过度冗余同样蕴含风险。其一,监管体系的“应用端”定位致使我国缺乏对社会场域的安全风险及政治场域的技术外溢与后门植入风险的防控关注。若未来出现相应的关键事件,我国可能无法精准调控监管走向。其二,各地激励政策虽形式多样,但多聚焦于财政补贴、算力支持等“输血式”措施,且呈现同质化的竞争趋势。总而言之,风险防控机制的缺位及地方政策的同质化竞争,易导致激励型方案难以持久推动开源 AI 的健康可持续发展,需进行一定填补。

(二)监管方案指引:秉持“激励主导,风险可控”原则

1. 统筹化完善现行的激励政策

开源 AI 社区和项目的建设需要持续投入大量人力、物力资源,我国目前形成的“中央定调、地方竞跑”的激励格局导致相似项目、同质化平台重复立项,存在资源浪费、标准制定权碎片化等问题。一方面,各地方扶持的开源平台和项目可能存在不一样的标准,不利于形成统一的技术标准和生态规范,无法满足政治场域以“主导标准”为核心的需求。另一方面,各地为政的支持力度有限,难以聚焦具有战略意义的“基础性”开源项目,无法最大限度地适配经济场域下对“创新激励”的需求。在此背景下,我国激励政策机制亟须进行统筹完善。对比来看,欧盟采取一套政府主导的资源整合模式,旨在打造一个本土开源大模型及开源资源共享平台[58]。美国则发挥产业主导的政府赋能优势,通过合作关系[43]和科研资金[59]等手段进行间接激励,并借此制定统一标准。鉴于政府主导路径难以激发市场活力且滞后性较强,并且我国已经形成完备的开源生态,美国以产业为主导的赋能形式更具备借鉴意义。

综上,我国可尝试构建“中央统筹、地方试点”的双层激励政策体系。中央层面可由国家发展改革委、科技部、工业和信息化部联合成立“国家开源 AI 发展协调小组”,统筹全国开源生态的建设。通过政府项目合作等形式(如提供算力支持),促进产业核心开源平台(如魔搭社区)进行升级,并借助参与信息,完成开源许可证、模型接口标准、开源安全评估规范等系列标准的制定。与此同时,由中央财政设立“国家开源 AI 创新基金”,重点资助开源平台建设、核心基础项目突破(如中文通用模型、多模态基础模型、垂直领域专用模型)等,引导地方项目对接中央资源,实现激励成果“由下而上”的汇聚。地方层面则侧重开源 AI 的场景化落地。选择 AI 产业基础较好(如北京、上海等)的城市开展试点工作,由各地方政府科技主管部门牵头,用以探索“开源模型 + 公共服务”的落地方案。具体而言,可优先选择低敏感性的政

务服务及基层治理场景,通过合作、安全审查前置及反馈循环等机制实现场景开放。项目完成后,由国家开源 AI 发展协调小组组织第三方机构对试点城市进行评估,推广优秀的落地经验,并纳入中央政务的 AI 赋能项目中。

2. 构建多维度、动态化的开源风险预警体系

正如前文钟摆模型所揭示的,即便我国当前采取以激励为主的监管立场,仍需同步推进对开源 AI 潜在风险的识别与评估工作,以确保未来标点时刻来临时,监管机构能够快速识别并精准调整监管方案。根据模型结果,我国目前有待完善对模型开放发布的风险评估、国产开源模型的扩散跟踪,以及使用境外开源模型的后门风险监测机制。对比而言,美国已通过产业、学术和政府的三方协同,建立起一套监控开源 AI 风险的评估机制[46]。据此,为保证风险评估的科学性和政策的可接受性,我国开源风险预警体系也必须建立在"政产研"协同的多元视角之上。

具体而言,国家开源 AI 发展协调小组可常设一个开源 AI 风险评估中心,汇集产业(头部开源企业和开源平台)、研究(AI 安全专家)、政府(网信办、工业和信息化部、国家安全部等监管机构)三方主体,统一风险的评估流程、方法与标准。再针对"开源模型滥用的社会安全风险""技术外溢的地缘政治风险""域外模型的后门风险"启动首轮专项风险评估、发布年度报告。评估方法与标准每两年修订一次,形成动态迭代机制。根据风险评估结果,建立"绿、黄、红"三级响应机制,实现风险识别与政策调整的动态联动:若风险等级上升,评估中心需立即召集联席会议、研判形势、制定应对方案,并上报相关部门。

3. 预设前瞻性的类型化框架

如前文所述,我国政策钟摆明确指向激励端,场域中限制开源的博弈力量尚未成熟,并不具备迫切的规制必要,当前对开源的"克制监管"具有合理性。过早干预可能抑制技术创新、削弱产业活力,不利于我国在开源领域"弯道超车"的战略目标。然而,技术、产业及地缘政治形势瞬息万变,一旦出现"技术跃迁""地缘政治矛盾激化"等关键事件,可能触发钟摆模型的"标点"时刻。届时,我国将处于"监管缺位"的被动局面。因此,有必要前瞻性地设计分级分类框架,为未来钟摆的摆动预留制度空间。欧盟采用了一套基于风险等级的类型化开源豁免,但其"严监管"底色易过度限制开源,不宜采纳。美国采取触发式监管的渐进策略,在不监管前提下预设若干触发监管门槛(如技术阈值、特殊适用领域等[46])。该思路符合我国国情,可予以借鉴。

该类型化框架的制定工作应由上述开源 AI 风险评估中心进行,从而充分发挥该中心的功能协同优势,即评估中心可充分利用其在信息调研阶段积累的"数据"和"专业知识",完成分级分类框架的设计、标准制定。例如,针对开源模型滥用的社会安全风险,可将其分为低扩散风险模型(能力有限或受控发布)、中扩散风险模型(具备一定两用性,但发布或风险可控)、高扩散模型(具有高度智能水平,一旦被恶意行为者获取将引发严重后果)。通过积极推广形成的开源类型化框架,有助于我国在全球开源治理规则协商中争取话语权。当相关场域的限制力量成形并影响钟摆走向时,可依据该类型化框架所设立的监管机制,快速、平稳地从"不监管"有序过渡到"分级分类监管"。

四、结语

开源 AI 无疑是推动我国 AI 产业快速发展的关键动因之一,对于加快实现核心技术自主可控、抢占未来产业制高点具有重要战略意义。然而,随之而来的技术泄露、安全和监管、后门侵入等风险同样不可忽视。在当前全球 AI 竞争日趋激烈的背景下,如何在我国独特的开源生态格局下,权衡开源 AI 的利与弊,科学构建兼顾发展与安全的监管体系,将直接关系到我国 AI 产业的创新能力、生态主导权及国家科技安全的整体布局。因此,构建"激励优先,风险可控"的监管方案已成为我国在 AI 开源治理中亟待回应的核心命题。

参考文献:

[1]封帅,薛世锟. 开源人工智能与国际政治变革[J]. 东南亚研究,2025(5):1-16,154.

[2]DeepSeek-AI. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning[EB/OL]. (2025-01-22)[2025-12-09]. https://github.com/deepseek-ai/DeepSeek-R1/blob/main/DeepSeek_R1.pdf.

[3]郑晓龙,李家彤. 人工智能时代的开源与闭源技术模

式探讨[J]. 中国科学院院刊,2025,40(3):459-464.
[4]HARRIS D E. How to regulate unsecured "Open-Source" AI: no exemptions[EB/OL]. (2023-12-04)[2025-12-09]. https://www.techpolicy.press/how-to-regulate-unsecured-opensource-ai-no-exemptions/.
[5]傅宏宇,贾开,彭靖芷. 人工智能开源的价值、风险与生态治理研究[J]. 电子政务,2026(2):58-69.
[6]苏宇,郭雨婷. 人工智能开源生态的法律治理[J]. 宁夏社会科学,2024(5):119-130.
[7]贾开. 经济民主与开源人工智能[J]. 开放时代,2025(5):211-223,11.
[8]王哲,薛澜. 大模型开源创新公地:历史演进、价值逻辑与中国叙事[J]. 探索与争鸣,2025(7):128-137,179,2.
[9]徐进,王珏. 开源人工智能"伦理—技术共构"的治理范式演进[J]. 华东师范大学学报(哲学社会科学版),2025,57(4):27-35,176.
[10]周辉. 开源人工智能模型的法律治理[J]. 上海交通大学学报(哲学社会科学版),2024,32(8):18-33.
[11]LATHAM E. The group basis of politics[M]. New York:Cornell University Press, 1952: 36.
[12]余敏江,李粤昊. 中国共产党探索社会治理现代化的百年历程:以注意力资源配置为视角[J]. 理论探讨,2021(5):12-19,2.
[13]鲍姆加特纳. 美国政治中的议程与不稳定性[M]. 曹堂哲, 文雅, 译. 北京:北京大学出版社, 2011: 9.
[14]OSI. Open weights[EB/OL]. [2025-12-09]. https://opensource.org/ai/open-weights.
[15]SPIRLING A. Why open-source generative AI models are an ethical way forward for science[J]. Nature, 2023, 616(7957): 413.
[16]The Gradient. Why we released grover[EB/OL]. (2019-07-15)[2025-12-09]. https://thegradient.pub/why-we-released-grover/.
[17]SCHNEIER B. Big Tech isn't prepared for A.I.'s next chapter[EB/OL]. (2023-05-30)[2025-12-09]. https://cyber.harvard.edu/story/2023-05/big-tech-isnt-prepared-ais-next-chapter.
[18]IWF. How AI is being abused to create child sexual abuse imagery[R]. Cambridge: Internet Watch Foundation, 2023:6.
[19]SEGER E, DREKSLER N, MOULANGE R, et al. Open-sourcing highly capable foundation models: an evaluation of risks, benefits, and alternative methods for pursuing open-source objectives[EB/OL]. (2023-09-29)[2025-12-09]. https://doi.org/10.48550/arXiv.2311.09227.
[20]SPOHRER J. The role of open-source software in artificial intelligence[J]. AI magazine, 2021, 42(1): 93-94.
[21]Creative Commons. supporting open source and open science in the EU AI act[EB/OL]. (2023-07-26)[2025-12-09]. https://creativecommons.org/2023/07/26/supporting-open-source-and-open-science-in-the-eu-ai-act/.
[22]Medium. Mark Zuckerberg: open source is good business[EB/OL]. (2024-02-02)[2025-12-09]. https://machine-learning-made-simple.medium.com/mark-zuckerberg-open-source-is-good-business-c62828f24bf1.
[23]ZHANG Z, SUN Y H, YANG J X, et al. Be careful when fine-tuning on open-source LLMs: your fine-tuning data could be secretly stolen![EB/OL]. (2025-05-22)[2025-12-09]. https://arxiv.org/abs/2505.15656.
[24]桑斯坦. 恐惧的规则:超越预防原则[M]. 王爱民,译. 北京:北京大学出版社, 2010: 15.
[25]林德布洛姆. 政治与市场:世界的政治—经济制度[M]. 王选舟,译. 上海:上海三联书店, 2016: 275-280.
[26]STIGLER G J. The theory of economic regulation[J]. The bell journal of economics and management science, 1971, 2(1): 3-21.
[27]罗伯特·基欧汉, 约瑟夫·奈. 权力与相互依赖[M]. 4版. 门洪华, 译. 北京:北京大学出版社, 2012: 11.
[28]利夫西. 后全球化时代:世界制造与全球化的未来[M]. 王吉美,房博博,译. 北京:中信出版集团,2021: 252.
[29]BUZAN B, WAVER O, WILDE J. Security: a new framework for analysis[M]. London: Lynne Rienner Publishers, 1998: 23.
[30]吉尔平. 世界政治中的战争与变革[M]. 宋新宁,杜建平, 译. 上海:上海人民出版社, 2007: 144-161.
[31]FARRELL H, NEWMAN A L. Weaponized interdependence: how global economic networks shape state coercion[J]. International security, 2019, 44(1): 42-79.
[32]Stanford Institute for Human-Centered Artificial Intelligence. The 2025 AI index report[R]. Stanford, CA: Stanford University, 2025: 61, 95, 247, 400.
[33]Future of Life. Pause giant AI experiments: an open letter[EB/OL]. (2023-03-22)[2025-12-09]. https://futureoflife.org/open-letter/pause-giant-ai-experiments/.
[34]VINCENT J. Meta's powerful AI language model has leaked online: what happens now?[EB/OL]. (2023-03-08)[2025-12-09]. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse.
[35]Stanford Institute for Human-Centered Artificial Intelligence. The 2024 AI index report[R]. Stanford, CA: Stanford University, 2024: 95.
[36]OpenAI. OpenAI forms exclusive computing partnership with

Microsoft to build new Azure AI supercomputing technologies[EB/OL]. (2019-07-22)[2025-12-09]. https://news.microsoft.com/2019/07/22/openai-forms-exclusive-computing-partnership-with-microsoft-to-build-new-azure-ai-supercomputing-technologies/.
[37] U.S. Senate Committee on the Judiciary. Oversight of A.I.: rules for artificial intelligence[EB/OL]. (2023-05-16)[2025-12-09]. https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-rules-for-artificial-intelligence.
[38] OECD. OECD AI capability indicators technical report[R]. Paris: OECD Publishing, 2024:36.
[39] McKinsey & Company, Patrick J McGovern, Mozilla. Open source technology in the age of AI[R]. New York: McKinsey & Company, 2025: 6, 9.
[40] PATEL D, AHMAD A. Google "We Have No Moat, And Neither Does OpenAI"[EB/OL]. (2023-05-04)[2025-12-09]. https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/.
[41] KUTV. Utah attorney general leads multi-state challenge against federal AI regulation[EB/OL]. (2024-02-06)[2025-12-09]. https://kmyu.tv/news/local/artificial-intelligence-utah-attorney-general-sean-reyes-leads-multi-state-challenge-against-federal-president-joe-bidens-ai-regulation-technology-government.
[42] Artificial Analysis. Open-weights frontier language model intelligence by country over time[DB/OL]. (2025-11-10)[2025-12-09]. https://artificialanalysis.ai/trends#open-weights-frontier-language-model-intelligence-by-country-over-time.
[43] CLEGG N. Open source AI can help America lead in AI and strengthen global security[EB/OL]. (2024-11-04)[2025-12-09]. https://about.fb.com/news/2024/11/open-source-ai-america-global-security/.
[44] Reuters. Exclusive: Chinese researchers develop AI model for military use on back of Meta's Llama[EB/OL]. (2024-11-01)[2025-12-09]. https://www.reuters.com/technology/artificial-intelligence/chinese-researchers-develop-ai-model-military-use-back-metas-llama-2024-11-01/.
[45] Global News. U.S. finalizing rules to curb AI investments in China[EB/OL]. (2024-10-28)[2025-12-09]. https://globalnews.ca/news/10835445/us-ai-rules-investments-China-restrictions/.
[46] NTIA. Dual-use foundation models with widely available model weights report[R]. Washington: NTIA, 2024: 2, 34, 47.
[47] U.S. House Committee on Science, Space, and Technology. DeepSeek: a deep dive[EB/OL]. (2025-04-08)[2025-12-09]. https://docs.house.gov/meetings/SY/SY15/20250408/118111/HMTG-119-SY15-20250408-SD002.pdf.
[48] U.S. Senate Committee on Commerce, Science, and Transportation. Winning the AI race: strengthening U.S. capabilities in computing and innovation[EB/OL]. (2025-05-08)[2025-12-09]. https://www.commerce.senate.gov/2025/5/winning-the-ai-race-strengthening-u-s-capabilities-in-computing-and-innovation_2.
[49] Reuters. US Commerce department bureaus ban China's DeepSeek on government devices, sources say[EB/OL]. (2025-03-18)[2025-12-09]. https://www.reuters.com/technology/artificial-intelligence/us-commerce-department-bureaus-ban-chinas-deepseek-government-devices-sources-2025-03-17/?utm_source=chatgpt.com.
[50] Stanford Institute for Human-Centered Artificial Intelligence. The 2023 AI index report[R]. Stanford, CA: Stanford University, 2023: 189.
[51] WHEATLEY M. Report: AI startup Hugging Face hoping to raise millions from Salesforce and other investors[EB/OL]. (2023-08-22)[2025-12-09]. https://siliconangle.com/2023/08/22/report-ai-startup-hugging-face-hoping-raise-millions-salesforce-investors/.
[52] RONIK. Every hugging face statistics You Need to know (2024)[EB/OL]. (2024-05-01)[2025-12-09]. https://weam.ai/blog/guide/huggingface-statistics/.
[53] CADE Hub. EU lawmakers urge strict definition of open source AI in EU's AI Act[EB/OL]. (2025-04-11)[2025-12-09]. https://cadeproject.org/updates/eu-lawmakers-urge-strict-definition-of-open-source-ai-in-eus-ai-act/.
[54] Mistral AI. Large enough[EB/OL]. (2025-11-08)[2025-12-09]. https://mistral.ai/news/mistral-large-2407.
[55] Artificial Analysis. Leading-models by country[DB/OL]. (2025-11-10)[2025-12-09]. https://artificialanalysis.ai/trends#leading-models-by-country.
[56] ATOM Project. ATOM: American truly open models[DB/OL]. (2025-11-10)[2025-12-09]. https://atomproject.ai.
[57] SAS. Generative AI global research report: strategies for a competitive advantage[R]. North Carolina: SAS Institute Inc., 2024:30.
[58] EuroLLM. Meet EuroLLM Large language model made in Europe built to support all official 24 EU languages[EB/OL]. [2025-12-09]. https://eurollm.io/.
[59] U.S. National Science Foundation. National artificial intelligence research resource pilot[EB/OL]. [2025-12-09]. https://www.nsf.gov/focus-areas/artificial-intelligence/nairr.