生成式人工智能的全球法律规制：制度替代认知、分类锁定与互认路径

摘要

中华智问知识发生器 · sdeuniverses.com · 2026/7/24 13:34:00

生成式人工智能的快速发展正在全球范围内引发一场深刻的法域监管竞赛。然而，当欧盟、美国、中国等主要经济体各自建立起独立的监管框架后，一个比技术风险本身更根本的规制困境开始浮现：这些框架之间不仅在AI的法律定义上存在分歧，更令人担忧的是，它们各自正在不可逆地加深对自己所选路径的制度依赖。本文提出，全球AI治理的真正难题并非缺乏统一的定义，而在于各法域在遭遇“法律认知失败”——即现行法律体系无法将AI这一“本体论上不稳定的对象”装入既有概念格子（如产品、服务、表达）——之后，不约而同地启动了一个“制度替代认知”的程序。各法域基于其既有的、在AI出现之前就已成熟的监管机器、法律工具和行政能力，对AI做出了“就当它是什么”的强制性分类。欧盟调用其产品安全监管机器将其归类为“高风险产品”，美国以联邦贸易委员会法案为依托将其视为“商业行为”，中国则借力其网络内容审查与算法治理体系将其界定为“深度合成服务”。这些选择并非源于对AI更准确的认知，而是源于各自可用的制度资源不同。更为关键的是，每个分类答案一旦被制度化，便会围绕其生长出专业群体、利益结构和行政惯性，形成巨大的沉没成本，使得退出任何一个答案的代价日益高昂，最终导致全球AI法律分类的分歧被各自的累积制度成本锁定。本文的诊断是，这一“分类成本不对称锁定”的机制，正在催生一种此前未被系统识别的新型规制风险——法律本体论套利：企业通过微调技术架构和使用条款，使其模型在不同法域表现为不同的法律对象以规避监管。基于这一理论诊断，本文提出了一套三层次的、递进式的规制工程方案：首先，通过建立“AI法律分类的规则起源地分析方法”使各法域的分类选择透明化；其次，构建“AI合规要求差分矩阵”并推动基于“功能等价”的跨法域互认，以降低企业重复合规成本；最后，引入“法律本体论审计”机制，从资本市场端对企业的跨法域法律身份矛盾形成自下而上的收敛压力。本文为破解全球AI治理的碎片化困局提供了一个可操作的理论框架与政策工具集。

——

全球文库校准后的学术定位

本文与既有AI风险分类、跨境监管和风险管理框架的差异，不在于再次罗列风险，而在于指出：当固定分类替代持续认知时，制度会把新风险强制翻译成旧对象。修订稿据此收缩了“全新理论”的表述，把贡献限定为“分类锁定—认知替代—互认更新”的机制链。

比较范围：SDE站内全文、arXiv、SSRN、PubMed/PMC及开放期刊；检索截止2026年7月24日。

1 引言

1.1 问题的提出

以ChatGPT、Midjourney、Stable Diffusion等为代表的生成式人工智能（Generative AI）技术的涌现，正在深刻重塑内容生产、知识工作与创意产业的面貌。其能够根据用户指令自主生成文本、图像、代码乃至音乐的能力，使得“非人创作”和“无明确人类行为者的损害”成为法律体系必须面对的日常现实。从纽约艺术家诉Stability AI侵犯版权，到深度伪造视频对政治选举的干扰，再到AI驱动的算法歧视在招聘和信贷领域引发的集体诉讼，生成式AI的法律风险议题已经从学术前沿迅速进入公共政策的核心议程。

面对这一技术冲击，全球主要法域在2023至2026年间开启了一场前所未有的立法与监管竞赛。欧盟率先通过了《人工智能法案》（AI Act），构建了基于“风险等级”的分类监管框架；美国联邦贸易委员会（FTC）启动了多起针对AI企业的执法调查，以“不公平或欺骗性行为”的法律路径将AI纳入既有消费者保护体系；中国则发布了《生成式人工智能服务管理暂行办法》，将其嵌入已运行多年的算法备案与内容安全审查机制。表面上看，各法域反应迅速，立法与执法活动频繁，似乎标志着人类正“管住”AI。

然而，一个更深层次的困境正在这些密集的规制活动之下悄然成形。当一家同时在欧盟、美国和中国运营业务的AI企业，需要将其同一个模型分别在欧盟合规文件中描述为“高风险AI系统”、在美国的监管沟通中定位为“自动化决策辅助的商业服务”、在中国的备案材料中界定为“生成合成类深度合成服务”时，我们面对的已不仅仅是术语上的分歧。这一现象暴露出一个根本性难题：全球各个法域不仅对“AI在法律上是什么”这个问题给出了不同的答案，更令人担忧的是，这些答案正在其各自的法律制度土壤中生根、生长，并围绕自身形成全新的专业合规产业、行政操作惯例和司法裁判先例。三个监管框架——欧盟的AI法案、美国的FTC执法判例体系、中国的算法治理基础设施——不是三个“暂定方案”，而是三套正在各自内部生长出自我维持生态系统的制度生命体。当我们讨论“全球AI治理”时，我们实际上是在讨论如何在三套（且未来可能更多套）日益不可逆的制度基础设施之间建立某种秩序。

这一困境的实质，是本文将要揭示并理论化的一个核心命题：生成式AI法律风险的真正危机，不是技术风险清单的长度或立法速度的快慢，而是在各法域对AI进行规制时，其自身的法律认知系统遭遇了分类失败，随后启动的一个“制度替代认知”的程序，正在将全球AI治理推向一个“分类成本不对称锁定”的僵局，并催生了“法律本体论套利”这一前所未有的结构性风险。 任何未能正面处理这一“锁定”和“套利”机制的全球治理方案，都将停留在愿望清单的层面。

1.2 研究问题与论证结构

基于上述观察，本文试图回答三个环环相扣的研究问题：第一，为什么全球各法域对生成式AI的法律分类存在如此深刻且看似不可调和的分歧？这些分歧的根源究竟是各法域对AI的认知不同，还是另有更深层的制度性动因？第二，这一分歧格局正在制造何种超越单项技术风险（如版权侵犯、隐私泄露）之上的新型规制风险？第三，面对这一已陷入“锁定”状态的碎片化格局，是否存在一条不要求任何法域放弃其核心监管主权、不推翻其已作巨额投资的制度基础设施，却能够有效管理碎片化后果的可行规制路径？

本文将按以下结构展开论证。第二节将通过系统的文献综述，梳理既有研究在AI法律分类、跨国规制协调以及制度路径依赖等领域的贡献与盲区，精准定位本文的理论生长点。第三节构建本文的核心理论框架，系统阐述“法律认知失败”、“制度替代认知”与“分类成本不对称锁定”三个概念及其相互关系，并将这一框架与演化经济学中的路径依赖理论和国际税收治理中的“竞争性编码”概念进行实质咬合，奠定跨学科分析的基础。第四节阐明方法论，说明本文如何运用概念分析、制度比较与工程化推演的方法，从理论诊断走向规制方案设计。第五节是本文的核心，分三个分论点逐一展开：首先诊断各法域对AI法律分类的操作性起源（规则起源地分析）；其次揭示分类分歧如何被制度成本锁定，并催生“法律本体论套利”这一新型风险；最后提出一套三层次递进的规制工程方案——从分类透明化（话语层），到功能等价互认（制度层），再到法律本体论审计（市场层），并设计其证伪条件与可操作试点路径。第五节还将包含一个专门的“反驳预期与回应”部分，主动回应可能的批评。第六节结论部分将凝练本文的核心发现，系统陈述其理论贡献与实践政策含义，诚实地讨论研究局限，并开出3-5个可供后续研究独立发展的研究纲领。

2 文献综述

本文的理论定位位于三个学术脉络的交汇处：关于AI法律分类与人格的法理学争论、关于全球规制协调的国际法与治理研究、以及关于制度演化与路径依赖的社会科学理论。本节将对这三个脉络进行评述，指出现有研究的贡献与盲区，从而为本文的理论框架奠定文献基础。

2.1 AI的法律分类与人格地位：法理学争论

关于人工智能在法律体系中应被如何归类的讨论，自20世纪80年代第一批AI系统商业化以来便从未停止。早期的争论往往集中于一个看似根本性的问题：AI是物还是人？基于“物”的定位，AI仅是受产权规则规制的工具，其产生的任何损害均可通过产品责任法追溯至制造者或运营者（Calo， 2015）。而基于“人”的类比则主张，当AI系统展现出足够的自主性时，可借鉴公司法人的拟制逻辑，赋予其特殊的法律人格，以解决其作为独立行为主体的责任承担问题（Abbott， 2020）。Abbott（2020）在其颇具影响力的著作中提出“非人决策者”（nonhuman decision-maker）这一概念，认为面对完全自主的系统，公司法中的“刺破公司面纱”原

则——即穿透表面实体追究最终控制者责任——可以为AI提供一种归责模型。Scherer（2016）则从实用主义出发，主张为AI量身定制一种新的法人形式——他称之为“电子人”（electronic person）——以便在法庭上为AI的受害者提供可诉的实体。

然而，生成式AI的出现使得这一争论显露出其深层局限。上述文献共享一个隐含前提：AI可以被放入法律坐标系中的某个单一、稳定的坐标点。无论是“产品”、“动物”、“雇员”还是“法人”，其分析框架都预设了一个可以在司法过程中完成“识别-归类”操作的对象。生成式AI的独特属性——其输出具有不可预测的创造性、其应用场景跨越几乎所有法律部门、其同一模型在不同场景下表现出截然不同的法律特征——恰恰在这个元操作上构成了挑战：它是一个拒绝单一分类的对象（Pasquale，2015）。当一个生成式AI模型既可作为聊天机器人提供“服务”，又可作为图像生成工具产出版权意义上的“作品”，还可作为招聘筛选工具作出雇佣相关的“决策”时，法律体系面临的不是“该把它划入哪个格子”的选择题，而是“它的性质取决于它被如何使用”这种分类学上的相对主义困境。这一困境暴露了既有文献的第一个关键盲区：它们未能将“分类能力的丧失”本身作为一个独立问题进行理论化。

此外，另一个较少被法律文献严肃对待，但在技术哲学和监管研究中反复出现的视角，是代码与架构本身的规范力。Lessig（2006）在其经典论述中指出，在数字空间中，“代码即法律”（Code is law）——技术架构以其不可协商的刚性，实现着与法律规则等量齐观的规制效果。根据这一理论，AI模型内部的算法架构、训练数据集的构成、提示词（prompt）的处理逻辑，本身就是一套隐性规则系统。但Lessig的分析主要面向单法域内的网络空间治理，未能处理当同一段代码（同一AI模型）须同时满足多个法域不同的架构规范时所产生的矛盾。这一矛盾在生成式AI时代变得无比尖锐：一个在技术上保持一致的模型，可能因不同法域要求添加不同的过滤模块或免责声明接口，而在架构层面就被塑造成不同的“法律对象”。本文将对Lessig的架构规范论进行跨国扩展，分析各法域对AI模型的架构要求如何参与了“制度替代认知”的竞争性编码过程。

2.2 全球AI规制合作与跨国协调：国际法与治理研究

面对全球AI规制的碎片化，国际组织与学术界提出了多种合作与协调方案。世界经合组织（OECD）于2019年通过了首个政府间AI原则，强调“包容性增长、可持续发展与福祉”、“以人为本的价值观与公平”、“透明度与可解释性”、“强健性、安全与安保”以及“可归责性”（OECD, 2019）。联合国教科文组织（UNESCO）则于2021年通过了更全面的《人工智能伦理问题建议书》，试图构建全球性的AI伦理标准。这些努力体现了以“统一原则”、“全球公约”为导向的治理路径，其隐含的预设是：各法域能够并且应当在“AI是什么”以及“我们应当如何治理它”这两个问题上达成基本共识，并以此为基础，逐级向下构建操作层面的协调机制。

另有学者从国际机制设计中寻求启示。Crootof（2019）提出，可将对致命性自主武器系统的治理经验迁移至民用AI，通过制定统一的技术阈值和行为标准来界定被规制的对象。这种思路在瑞典AI委员会向欧盟提交的“通过共同技术标准实现协调”的建议中得到了呼应。在国际金融监管领域，巴塞尔银行监管委员会通过《有效银行监管核心原则》在规范层面上统一全球银行监管标准的成功经验（Pistor，2019），也常被引为AI全球协调的潜在范本。

然而，上述方案在面对本文引言中所描述的“三套制度生态系统日益不可逆”的困境时，其解释力和可操作性均显不足。它们的共同盲区在于：假定分歧仅仅是暂时的、可被理性沟通弥合的认知差异，而忽略了分歧背后深刻的制度性根源。Pistor（2019）在其《资本的法则》一书中提出的“资产编码理论”在此具有启发性。她论证道，法律并非被动地“发现”资产，而是通过一套法律编码机制，将某些东西“构成”为可被持有、抵押和交易的资产。这一理论揭示了法律的构成性力量，但它依然预设了单法域、单套编码系统的场景。当不同法域争相运用各自的法律编码系统去构成同一个技术对象时——比如，欧盟法将某个AI模型构成为“高风险产品”，美国法将其构成为“一项受到欺诈指控的商业行为”——会发生什么？Pistor的理论未能处理这种编码竞争带来的系统性后果。

在国际税收治理领域，我们却观察到了一个与本议题高度同构的挑战及其应对经验。跨国公司将利润转移至低税率法域以规避在高税率法域的纳税义务，这一“税基侵蚀与利润转移”（BEPS）问题本质上也是企业利用各法域税法规则的差异进行套利。经济合作与发展组织（OECD）推动的BEPS项目，并未去追求全球税法的

统一，而是通过建立一套多层工具——国别报告（CbCR）、相互协商程序（MAP）、预约定价安排（APA）等——来管理不一致的后果，即“我们同意不必统一税率，但同意设立规则来追踪和遏制在税率差异中的人为套利”。本文将在理论框架中系统借鉴这一路径，特别是其“功能等价”而非“文字统一”的互认逻辑。

2.3 制度演化与路径依赖：社会科学理论资源

要理解为何各法域对AI的分类分歧没有随时间弥合、反而日益加深，法律与社会科学中关于制度演化和路径依赖的理论提供了关键的分析资源。

制度经济学鼻祖North（1990）在其开创性著作中指出，制度一旦确立，便会因其显著的“报酬递增”（increasing returns）效应而逐步自我强化。制度框架会催生与之匹配的互补性组织、专业化技能以及与之相适配的认知模式，从而产生巨大的制度退出成本。Pierson（2000）将这一分析应用于政治学领域，指出现代政治制度中存在强烈的“正反馈”（positive feedback）机制：既有的制度安排会为它们自身创造出持续的支持者和既得利益者，使得制度变迁变得异常困难。

将此视角转向AI规制领域，我们可以识别出一个高度清晰的路径依赖现象。欧盟在数十年的消费者保护实践中，已建立起一套以“产品安全”为核心的强大监管机器：从化学品注册、评估、许可与限制法规（REACH）实施化工品监管，到《医疗器械法规》（MDR）监管起搏器与手术机器人，这套机器的操作逻辑——“风险等级划分→合格评定程序→CE标志→市场准入”——已经深深嵌入欧盟委员会的行政架构、成员国的市场监管部门、以及无数企业的合规部门和法律服务产业中。当AI出现时，将这部机器对准AI，不是一种在与其他替代方案（如“AI是服务”、“AI是表达”）平等角逐后的最优选择，而是一条路径阻力最小、制度边际成本最低、并且能保护该机器大量既有专业人力资本投入的方案。同样，美国FTC走“不公平或欺骗性行为”的规制路线，不是因为其理论上最优，而是因为自1914年成立以来，FTC Act第5条的宽泛授权已经使其能够灵活应对从石油公司虚假广告到社交媒体数据隐私的一系列新兴商业行为，它所需做的只是将AI纳入“商业行为”这个所有接受FTC管辖的实体都熟悉的解释框架内。中国的“深度合成服务”路径则完美嵌入了其自《网络安全法》生效以来搭建的、覆盖算法、内容、个人信息的全套治理体系，该体系已具备以算法备案和安全评估为主要抓手的成熟操作流程。

既有文献的盲区在于，它们多将这些路径依赖看作是各法域内部的、孤立的现象，而未能看到当不同法域各自独立的路径依赖在国际层面相互遭遇和碰撞时产生的独特后果。本文正欲填补这一理论空白，提出“分类成本不对称锁定”这一概念，以描述当多个法域各自被锁入不同路径后，其相互间的分歧为何非但不会趋同，反而会持续加深，并催生出新的套利空间这一跨国层面的系统性效应。

综上所述，现有研究在三脉文献各自的边界内取得了重要进展，但都未能充分捕捉和解释全球AI规制碎片化困境的一个核心机制：法律认知上的系统性失败如何通过各法域的制度惯性，转化为“制度替代认知”的竞争，并最终被各自的累积制度成本锁定。本文将通过构建一个跨法理学、国际治理与制度演化理论的分析框架，来填补这一空白。

3 理论框架

基于本文引言中的观察与文献综述诊断出的理论空白，本节将系统构建一个由三个核心概念组成的理论框架。该框架旨在为理解生成式AI全球规制困境的深层机制提供一套可操作的概念工具，并为我们诊断这种困境、设计解锁路径提供理论基础。

3.1 核心概念定义与相互关系

（1）法律认知失败（Legal Cognitive Failure）。 法律体系在处理外部世界时，依赖一套根本性的元操作：对任何进入其管辖范围的新对象执行一次初始的、通常是决定性的分类。这个操作将所有对象归入既有的基础法律概念格子——是“物”（产品、器物、自然力）、“行为”（言论、商业活动、侵权行为）还是“人”（自然人、法人），因为后续所适用的整套规则链，都取决于这个第一步的分类。这一法律认知程序在数千年的人类历史中运转有效，因为新出现的事物总能在既有网格中找到位置，或者能被类推适用。但当生成式AI出现时，法律体系首次遭遇了一个拒绝被单一归类的“本体论不稳定对象”。究其原因在于，同一生成

式AI模型在不同使用场景下，可同时满足互斥的法律分类判准：在生成一幅画时，它同时既是用户指令的执行工具（服务的特征），又是一件具有独创性的智力成果的作者功能替代品（创作的特征）；在提供投资建议时，它既是自动化商业服务（行为的特征），又是包含着开发者决策意志的规则应用者（准产品的特征）。法律体系并无现成的元规则来回答，“当一个对象同时满足了多个互斥的初级分类判准时，应将哪一分类定为优先”。这一分类元操作的失能，便是本文所定义的“法律认知失败”。它并非指法律条文不够完善，而是指法律体系最基本的认知扫描器在面对这种全新类别对象时发生了边界失能。

（2）制度替代认知（Institutional Substitute for Cognition）。 法律作为一个社会系统，无法在面临急迫的规制需求时宣布“我不知道如何分类，所以我暂时无法工作”。当认知失败发生、而规制行动又不可逃避时，法律体系会启动一套替代程序：不再追问“这个对象从根本上是什么”（这已无法被回答），而是基于自己已然具备的、独立于该对象而存在的制度资源，给出一个强制性的、可立即投入操作的答案。这个答案回答的不是“AI是什么”，而是“我们手上现有的哪套监管机器，最能覆盖当前可见的AI风险”。我们观察到，在回应生成式AI的挑战时，欧盟、美国和中国各自调动了其用数十年时间、斥巨资搭建和优化的不同监管机器——欧盟的产品安全监管机器、美国FTC的市场不正当行为监管机器、中国的网信内容与算法治理机器。这三台机器不仅是三套制度，更是三种不同的、且早已高度模式化的“审视与干预逻辑”。这种“因为我有这把锤子，所以我把AI看作钉子”的制度性应激，就是“制度替代认知”。它是法律体系在认知失败后的必然操作，其核心特征是：强制性分类结果的可预测性，主要不来自AI的内在属性，而来自各法域既有监管机器的工具特性。

（3）分类成本不对称锁定（Asymmetric Lock-in of Classification Costs）。 当每个主要法域都通过“制度替代认知”为AI选定了一个法律身份，并围绕该身份开始构建合规体系、培训执法官员、催生法务咨询产业以及积累司法判例先例时，就形成了一个正反馈的锁定过程。欧盟批准的AI法案，已催生了新的AI合规官员职位、AI合格评定机构以及企业内部为应对AI审计而跨部门组建的合规团队。美国FTC的每一次AI相关执法案的和解或判决，都在为律所、企业法务和合规咨询公司创造新的、可商品化的知识产品（“FTC对AI商业行为的最新执法趋势”简报）。中国的算法备案体系同样正在围绕“深度合成服务”这一标签，催生出一个完整的、从备案代理到安全评估报告的本地化合规产业。这些投入构成了无法收回的沉没成本。本文用“分类成本不对称锁定”来描述这一更深层的状态：主要法域之间都不认同对方的分类，但退出自身分类的代价极其高昂（已培训的人员被废弃、已投资的系统要重建、已形成的行政程序和司法先例面临动荡），而“忍受分歧继续前行”的成本则相对较低且可被转嫁给跨境运营的企业。这种成本结构的严重不对称，使得妥协与趋同的动力远远弱于路径自我强化的惯性。于是，全球AI法律分类的分歧，便从一种暂时的认知差异，演变成了被各自累积的制度成本锁定的持久的、甚至自我强化的僵局。

3.2 命题与假设

上述三个概念构成了一个环环相扣的因果链条，可被明确表述为以下四个关联性命题：

命题1（认知失败命题）： 法律体系在处理生成式AI这一本体论不稳定对象时，其初始分类元操作会发生系统性失能，导致无法产生单一、一致的法律分类。

命题2（制度替代命题）： 在命题1所述的认知失败发生后，各法域为满足急迫的规制需求，将启动制度替代认知程序，其给出的强制性分类答案的可预测性，主要取决于该法域既有、独立、成熟的监管机器的操作逻辑，而非AI的固有技术属性。

命题3（锁定与套利命题）： 命题2产生的多个互斥的制度化分类，会因各自的累积制度成本而被锁定，这种被锁定的分歧格局将催生“法律本体论套利”这一新型规制风险：企业可通过微调其非核心技术参数、调整使用条款与数据策略，使其模型在不同法域被法律构成为不同的对象，从而系统性地规避本应承担的合规责任。

命题4（解锁路径命题）： 突破命题3所述僵局的可行路径，不是要求任何法域废弃其已投入的制度基础设施（即进行“撤锁”），而是建立一套管理不一致性后果的国际机制——其核心是推动合规动作的“功能等价互认”与“法律本体论审计”，在不要求统一定义的前提下，大幅压缩套利空间并降低重复合规成本。

3.3 跨学科理论资源的实质咬合

为夯实上述框架的理论基础，本文有选择性地与三个学科的理论资源进行实质对接，并显式说明其有效性与借用边界。

第一，借用演化经济学中的路径依赖与递增报酬理论。 North（1990）和Pierson（2000）提出的理论，精准地解释了“制度为何一旦确立便难以改变”。本文将这一理论从解释单法域内的制度变迁，拓展至解释多个路径依赖制度在国际层面相互碰撞、且成本结构不对称时的动态。其有效性在于，欧盟AI法、美国FTC判例体系和中国算法治理体系的发展过程，均呈现出强烈的内部递增报酬特征。但借用有其边界：传统路径依赖理论通常分析的是某项制度在单一环境中的演化，它较少处理多个已经锁定的制度之间如何博弈与互动的问题。本文正是在此边界处作出了理论延伸。

第二，借鉴国际税收治理中“BEPS项目”的制度设计逻辑。 OECD/G20主导的BEPS项目，其核心智慧在于放弃了统一全球税率的幻想，转而设计一套使各国税收规则“透明化”、“互认化”和“可争议化”的工具箱（如国别报告、相互协商程序）。本文提出的规制方案，其精神内核与BEPS完全一致：将目标从“统一定义”转向“管理不一致的后果”。这一类比的有效性在于，国际税收与AI治理同为在高度全球化背景下、面对多法域竞相利用不同规则编码同一对象以获取竞争优势的困境。但核心差异也必须被正视：税收套利的本质是对“利润”这一单一维度的数字在不同法域法规下的再分配；而AI治理互认涉及的则是数据质量标准、基本权利保护、算法透明度等多元且不可直接通约的治理目标。本文在提议功能等价互认时，必须引入一个维度拆解与逐项比对的程序，以应对这种多维性（见下文5.3.2节）。

第三，引入技术哲学与社会系统论中关于“代码即法律”的讨论。 Lessig（2006）的论断——技术架构具有规范力——在本文框架中被操作化：当一个AI模型为满足不同法域的法律要求而在架构上进行变体时（例如，为进入欧盟市场而嵌入更强的提示词过滤模块，为满足中国法规而接入特定的内容安全API），它就在技术上被塑造成了具有不同“法律本体论可塑性”的实体。本文的“法律本体论套利”概念，其操作机理正是企业通过操纵这种可塑性来最大化规避合规成本。此处借用的有效性在于它揭示了“法律规制”与“技术架构”之间双向的、构成性的互动。其边界在于，Lessig的分析主要是在单法域主权空间内展开，本文将其扩展到跨国、多法域兼容的语境中，揭示了当多个主权体分别试图用其法律“编码”同一套技术架构时产生的冲突与套利。

4 方法论

本文是一项旨在进行理论建构与规制方案设计的规范性政策研究，主要采用概念分析、制度比较与工程化推演相结合的研究方法。

4.1 研究设计

本文的研究设计遵循“诊断-建构-检验”的三段逻辑。在诊断层面，我们通过对欧盟《人工智能法案》、美国联邦贸易委员会（FTC）截至2025年的公开执法文件、以及中国《生成式人工智能服务管理暂行办法》及其配套的算法备案规定等一手法律文本和政策文件进行精细的解读与比较，提炼出各法域对生成式AI进行法律分类的独特操作逻辑。同时，借助对Pistor（2019）的资产编码理论、Abbott（2020）的非人决策者理论、Lessig（2006）的代码即法律理论等经典法律与社会理论的批判性对话，确认既有分析框架在解释当前困境时的盲区。在建构层面，本文创新性地提出“法律认知失败-制度替代认知-分类成本不对称锁定”这一理论链条，并以此为基石，推演出一套三层次递进式的规制工程方案。在检验层面，本文为所提出的理论命题和规制方案设置了明确的、可操作的证伪条件，以使其区别于纯粹的价值倡导。

4.2 分析策略

概念分析。 本文的核心分析动作是对“分类”这一法律元操作的解构。我们追问的不是“AI应被分到哪一类”，而是“分类这个动作本身在遭遇AI时发生了什么”。通过对欧盟的“风险等级逻辑”、美国的“商业行为纠偏逻辑”和中国的“服务接入管理逻辑”进行解构，我们揭示出，决定最终分类结果的关键自变量并非AI

的技术本质，而是各法域在AI出现之前便已高度成熟的监管机器的内在操作程序。这一策略使我们得以从表面的话语分歧深入到深层的制度惯性层面。

制度比较。 我们对三种制度生态的“分类可溯源性”进行了结构化的比较。遵循密尔（Mill）的求异法和求同法逻辑，我们既分析三个法域的分类为何不同（源于其监管机器差异），又分析它们在面对认知失败时采取“强制性分类”这一反应上的同构性。为进一步深化比较，我们还引入了国际税收治理（BEPS）和跨国产品标准互认（IECEE体系）作为参照案例，通过跨领域的制度类比与差异辨析，来检验和修正本文提出的规制方案的普适性与特定边界。

工程化推演与证伪设计。 由于本文提出的规制方案（如差分矩阵、功能等价互认协议）尚处于理论构建阶段，不具备事后评估的实证条件，我们采用了“工程化推演”的方法：方案设计的每一步（拆解合规单元、设计互认流程、选定试点法域、构建审计指标）都必须处理具体的实施障碍，并预设解决路径。这项研究策略的一个关键特征是其可证伪性。本文严格遵循波普尔（Popper）的科学发现逻辑，为整个理论框架和规制方案设置了一个2×2矩阵的证伪条件（详见5.5节），清晰指明了理论在何种经验观察下会被证明为伪，从而将其锚定在可检验的学术研究范畴内。

4.3 方法局限的预先承认

本文承认其方法论存在以下几项固有局限。第一，本文属于理论建构与规范性研究，其论证力量主要来源于逻辑严密性与制度剖析的深度，而非大样本的统计推断。因此，本文的结论具有理论上的普遍性诉求，但其对具体国家或案件的应用可能需要根据当地的制度细节进行调适。第二，本研究依赖的文本材料（特别是美国FTC的执法文件）具有一定的时效性，政策和判例的后续演变可能超出本文的预测范围。第三，在跨学科概念的借用中，尽管我们已经力图精确界定其使用边界（如3.3节所述），但从“概念借用”到在不同学科语境下完全“操作化”之间仍存在距离，这部分工作需要在后续的实证研究中由对应领域的专家完成。

5 分析与讨论

本章是本文的核心论证部分。我们将分三个步骤展开。首先，我们执行“规则起源地分析”，实证性地解构欧盟、美国和中国三个法域的分类选择，论证“制度替代认知”这一核心命题。其次，我们诊断这一碎片化格局的后果，揭示“法律本体论套利”这一新型规制风险的生成机制与具体形态。最后，我们提出一套三层次递进的规制工程方案，详细阐述其每步操作、内在逻辑与潜在障碍，并设计其证伪条件。

5.1 诊断：制度替代认知与规则起源地分析

在任何关于全球AI治理的严肃讨论中，第一步都不应是追问“哪个法域的分类更正确”，而应是追问“它们为何会给出如此不同的分类”。这需要我们穿透“概念分歧”的表面，进入“规则起源地”的深处——去考察这些分类答案是从哪台制度机器里被“生产”出来的。

5.1.1 欧盟：“高风险AI系统”与产品安全机器的扩展

欧盟将生成式AI（特别是GPAI，即通用人工智能）纳入《人工智能法案》的规制框架，并最终将其定位为一种须接受风险分级的“产品”。这一选择的逻辑，只有追溯到欧盟自20世纪70年代以来逐步建立起的“新方法”（New Approach）产品监管体系才能理解。该体系的核心是一条铁律：任何进入欧盟单一市场的工业产品，只要落入相关指令或法规的覆盖范围，都必须证明其符合欧盟协调标准下的“基本健康与安全要求”，并加贴CE标志。

这套机器已成功运行数十年，其监管DNA——风险等级分类、合格评定、市场监督、事后严重事故警报系统——拥有跨行业的普适（例如，从儿童玩具到医疗设备的监管均遵循此逻辑）。因此，当欧盟立法者面对如何规制AI的难题时，这套机器的“存在”便提供了巨大的引力场。将AI看作产品，意味着可以立即借用一整套无需从头设计的制度基础设施：它明确划分了制造商（AI开发者）、进口商（将AI引入欧盟市场者）和分销商（如部署API服务的企业）的产业链责任；它提供了多种合格评定模式（从自我声明到第三方机构介入），以灵活适应不同风险级别；它赋予了成员国市场监管机构明确的执法权力。2023年底围绕基础模型的监管争议

——是将其完全按高风险系统要求逐案审计，还是仅施加有限透明度义务——本质上是一场关于这台机器应向一个新对象延伸到何种程度的内部调试，而从未威胁到“AI是产品”这一由机器本身所预定的分类框架。由此，我们看到的不是欧盟对AI本质的认知更清晰，而是它对自身产品安全机器的运用更娴熟。

5.1.2 美国：“商业行为”与联邦贸易委员会的灵活应对

与欧盟不同，美国国会至今未通过一部全面的联邦AI法律。然而，联邦层面的监管并未缺席，联邦贸易委员会（FTC）成为实际的AI监管主力。FTC依据的是其1914年成立时被赋予的、《联邦贸易委员会法》第5(a)条下针对“不公平或欺骗性行为或做法”的宽泛管辖权。

这条路径的选择，同样具有深厚的制度根源。美国政治体系中的权力分立与对“监管过度”的天然警惕，使其极难像欧盟那样建立一套跨行业的、预防性的产品上市前审批制度。但FTC的存在提供了一个高度灵活的执法工具：它不需要立法机关为AI专门定义一个新类别，仅需证明某一AI企业的行为（例如，未经充分告知就收集用户数据训练模型、夸大AI诊断疾病的能力、在广告中使用AI生成虚假名人背书）对消费者构成了实质性的、消费者无法合理避免的、且无法被其他利益所抵消的损害，或是企业作出了与事实不符的陈述。OpenAI、Google等公司近期收到的关于其AI模型可能违反消费者保护规则的问询函，以及FTC对亚马逊Alexa语音助手涉嫌在未获监护人同意下收集儿童数据而开出的罚单，均揭示出美国AI规制的底层逻辑：AI不是一个独立的“东西”，而是一种企业在市场上展开的“行为”。决定其是否合规的标准，不取决于其技术架构的内在风险等级，而取决于企业的公开声明与实际交付之间的一致性，以及其行为是否构成对消费者的剥削。这一逻辑根植于FTC一个多世纪以来在市场监管领域积累的判例法和执法惯性，是制度替代认知的又一明证。

5.1.3 中国：“深度合成服务”与网信治理体系的嵌入

中国对生成式AI的规制，通过《互联网信息服务深度合成管理规定》和《生成式人工智能服务管理暂行办法》等法规，创造性地将其界定为“生成合成类深度合成服务”。这一概念在外观上有别于欧美的分类，但其制度起源同样清晰可溯。

中国的选择是将其无缝嵌入自《网络安全法》（2017）施行以来所构建的、以网络信息内容生态和算法治理为核心的监管体系。该体系的操作枢纽是“服务提供者”这一抓总责任主体，其核心治理工具包括：算法备案（要求将算法的核心技术逻辑、应用场景、数据来源向监管机构进行登记）、安全评估（在上线前和重大更新后进行特定的安全审查）、内容审核（对生成和传播的法律禁止内容进行实时监控和处置）。这个成熟的、覆盖全网的“网信治理机器”，天然地要求将任何新的网络现象“翻译”为其可识别的格式——即一种“服务”类型。因此，生成式AI在法律上被构成为“深度合成”这一具体服务类别（与“网络直播”、“网络支付”等并列），用以接入算法备案与内容安全监管这台既有机器。这一路径的优点在于其启动速度和执行的确定性，它无需进行根本性的法理辩论或立法突破，只需发布部门规章即可将新生事物纳入成熟的行政监管链条中。

上述三域的比较清晰地揭示出：全球AI法律分类的分歧，其根源并非三域的立法者或法学家对AI的认知产生了巧合般的哲学分歧，而是因为他们在面对同一个“认知失败”的困境时，被迫拿起了距离自己最近、用得最顺手的制度工具。这些工具分别来自产品安全监管、消费者保护执法和网络内容治理这三个迥异的传统。

5.2 诊断：分类锁定、本体论套利与新型规制风险

当上节所述的“制度替代认知”完成，并开始其各自的制度化进程后，一个更深刻的风险格局随之显现。本节旨在分析这个格局的两个核心特征：分类成本的不对称锁定与法律本体论套利。

5.2.1 分类成本不对称锁定的形成

正如路径依赖理论所预测，制度一旦建立，其自我强化机制便开始运转。欧盟AI法案的出台，已经催生了一个全新的、以“高风险”与“非高风险”为分界线的合规咨询市场。律师事务所、会计师事务所和专业科技认证公司正在投入巨资培养能解读AI法案条款、并为企业设计从模型开发阶段开始的“合规设计”流程的专业

人员。这种人力资本和知识产品的投资是高度专用性的——它主要用于在“产品安全”这个框架内进行论证和操作。

美国FTC的AI执法案例，则在律所界催生了“FTC AI执法”这一高度专业化的精品法律服务线。律师们通过解读FTC与各家企业达成的不公开和解协议、研究FTC委员的公开演讲和媒体采访，来逆推FTC在当前阶段对AI“欺骗性行为”的具体判准，并将之转化为供客户参考的“内部红线”。这些知识同样被高度封装在“市场行为”的解释框架内。

中国的算法备案与安全评估制度，也催生了一批本土的、深谙监管口径的技术合规服务商。他们的专业能力体现在能够精准地将企业的模型架构和数据处理流程，翻译、包装为监管部门在形式审查中认可的标准语言和格式。

这三个生态系统看似互不相干，但它们在两个关键点上分享了同一个动力学：第一，每个生态系统都在进行大量且不可回收的专用性资产投资；第二，这些投资都使得该生态圈内的专业人士和机构成为维护“本域分类”的忠实既得利益者。由此，退出任何一个分类框架，不仅要面临形式上的法律修改，更意味着对应行业数以亿计的专用知识和商业合同面临突然贬值。面对这种级别的沉没成本，各法域理性地选择了“将就”在既有分歧中。这便是“分类成本不对称锁定”的实质：维持现状的成本（虽在增加但可被转嫁）远低于退出当前路径的成本（一次性的、摧毁性的巨大损失）。

5.2.2 法律本体论套利：一个新型系统性风险的涌现

在这种被锁定的碎片化格局下，一种在单一法域内不存在、也未被现行任何国际贸易规则所覆盖的新型系统性风险悄然浮现：法律本体论套利。

其运作机理如下。一家跨国运营的生成式AI企业，其核心的大语言模型本身是一套统一的技术实体。但为了合规，它必须在三套表格上填上三个不同的“法律名字”。在欧盟，它向合格评定机构提交的文件中，其模型被描述为一种“可能存在风险的技术产品”；在美国，其总法律顾问回复FTC调查函时，将其描述为“一项提供给商业伙伴和消费者的自动化辅助服务”；在中国，其提交的算法备案登记表中，它被描述为“一款生成合成类的深度合成互联网信息服务”。这三个描述，映射的不是三个不同的技术本体，而是三套监管机器的不同形态要求。

精明的企业很快学会了对“法律身份”的主动管理。这不再是消极地遵从法律，而是积极地利用法律编码的差异来优化自身的法律处境。一个典型策略是“法律身份的分场景切换”：当面临版权侵权诉讼时，企业在美国法庭上可能辩称“AI只是一个辅助创作的工具，最终内容由用户创造并负责”（强调工具/服务属性以规避直接侵权），而在欧盟数据监管机构前可能辩称“AI是一个高度自主的模型，其输出具有不可预测性，开发者无法完全预见其具体内容”（强调其独立的、类似自然力量的产品属性，以在GDPr下争取责任豁免或履行举证困难的抗辩）。

另一个策略是“跨法域的合规成本最小化”。当欧盟AI法案要求对高风险模型进行深入、昂贵的第三方审计时，一家企业可能选择在技术上略微调整其向欧盟用户提供的模型版本（例如，禁用某些功能或添加额外过滤层），使其在技术上“降级”为一个非高风险的普通模型，同时在美国和中国维持其原版功能。这种“为法律而拆分技术架构”的行为，本身并不违法，但它巧妙地将原本统一的监管对象，分割成了数个依附于法域边界的法律存在。这使得任何一个单一法域的监管效力都遭到削弱。

法律本体论套利是比单项技术风险（如隐私泄露）更根本的制度性风险，因为它侵蚀的是法律的可预测性这一基础性制度公共品。它使得监管套利不再依赖于寻找地理上的“法外之地”或钻条文漏洞，而是通过在法域之间切换对象的“本体论定义”来实现。这对传统上建立在对稳定事实进行认定基础上的法律裁判体系构成了前所未有的挑战。

5.3 方案建构：一个三层次递进的规制工程体系

面对上述由“制度替代认知”与“分类锁定”共同催生的僵局和套利风险，本节旨在设计一套切实可行的解锁方案。本方案的设计哲学是现实主义的：我们不试图追求一个统一的、全球性的AI定义，不要求任何法域

推翻其已投入巨大成本的既有基础设施，而是专注于建立一套管理不一致性后果的国际机制。本方案分为三个按内在逻辑递进的层次：话语层的透明化、制度层的互认、以及市场层的收敛。

5.3.1 第一层（话语层）：规则起源地分析与分类谱系透明化

打破“谁的分类最正确”这一徒劳争论的第一步，是让各方看清楚自己的分类选择是“如何”以及“从何”做出的。这是我们提出的“规则起源地分析方法”（Rule Provenance Analysis）的核心任务。其操作是一个标准化的三问框架，对于一个法域给出的任何AI法律分类，我们追问：

1. 溯源问题： 构成这一分类的法律框架背后，所调用的核心监管权力（产品上市前审批、市场行为事后纠偏、内容安全审核等），在人工智能出现之前，主要是用来规制什么对象的？（例如，药品、消费品、网络直播、银行资本充足率）。

2. 转化问题： 这套监管权力的标准操作程序（如：企业自我声明、第三方合格评定、政府直接许可），在被应用到人工智能上时，对其输入接口和审查标准做了哪些具体的修改？修改的部分是发生在程序的输入端（定义“什么需要被监管”）还是输出端（判断“怎样算通过”）。

3. 历史验证问题： 该法域在过去十年内，是否曾用同一套监管机器和相似的修改逻辑，成功覆盖过另一个全新的“新对象”？（例如，美国FTC将其监管从虚假广告成功延伸至早期的数据隐私侵犯；中国网信治理机器从网络新闻覆盖到网络直播）。

应用此三问框架，可以生成一个法域的“可用分类器谱系”。这个谱系的分析结果会清晰地揭示：欧盟对AI的“风险等级”分类，是其既有的、运作精密的“产品内外部风险管理体系”的一个新应用场景，而非对它的一次孤立的哲学识别。同样，中国的“深度合成服务”分类，是其“互联网信息服务生态治理体系”按既有接口处理一个新成员。

这一层工作的价值在于实现国际话语框架的一个关键位移：将讨论从“谁对AI的分类更准确”，转向“谁的分类机器对AI的核心风险覆盖得更全面、盲区更少”。当一个分类选择被透明化地解析为“这是我们工具箱里最好的锤子，因此这个新东西暂时被当作钉子来处理”时，国际争论的焦点便能从无法解决的“AI的本体论”之争，转向可被客观评估的“监管工具的覆盖力与效率”之争。这一透明化过程本身便会暴露各法域监管机器在覆盖AI风险上的盲区，从而为下一阶段的“功能等价”互认奠定认知基础。

5.3.2 第二层（制度层）：合规动作差分矩阵与功能等价互认

在各方理解彼此“分类器”的起源和逻辑后，第二步是启动实际的制度协调。此步的核心机制是建立“AI合规要求差分矩阵”（Differential AI Compliance Matrix），并在此基础上开展“功能等价互认”（Functional Equivalence Mutual Recognition）。其核心原则是：不要求法域A承认法域B对AI的分类是正确的，只要求法域A承认，法域B要求企业就某个具体合规动作完成的步骤，在保护功能上等价于法域A自己要求的步骤。

(1) 拆解：从“大分类”到“最小合规单元”。 差分矩阵的第一步，是打破“高风险系统监管 vs. 商业行为监管 vs. 深度合成服务管理”这种整体化、不可比的大概念，将不同法域的合规要求暴力拆解为最小可操作单元。这些单元是具体的、去语境化的合规动作，包括但不限于：

* 数据治理类： 训练数据来源的透明度与文档化要求、受版权保护数据的过滤机制、数据中偏见与歧视性的统计评估义务。

* 透明度类： 向终端用户标识其为AI生成内容的强制标记义务、向B端客户说明模型核心性能局限的技术文件披露要求。

* 风险管理类： 在部署前对特定高风险场景（如雇佣、信贷）进行歧视性影响评估的义务、对模型进行对抗性安全测试（红队攻击）的内部规程要求。

* 问责与接口类： 建立使用者投诉与纠正机制的义务、对监管机构进行系统性风险事故的强制性报告义务、保存可追溯其输出结果日志的技术能力要求。

(2) 比对与互认：从“文字统一”到“功能等价”。 在完成单元拆解后，发起国之间通过组建由技术专家和法律比较学者组成的技术工作组，逐项比对各自单元的功能。以“训练数据透明度”为例，法域A可能要求企业进行“内部自我声明并保存证据备查”，而法域B要求“委托独立第三方进行审计并公开审计摘要”。尽管程序形式不同，但技术工作组可以认定，这两套程序在“确保数据来源可追溯、防止非法数据用于训练”这一核心治理目标上，达到了功能等价。一旦认定成立，便可建立互认。这意味着，一个AI企业如果在法域B完成了一次合格的第三方数据审计，当它进入法域A的市场时，可以将此审计报告作为其履行“训练数据透明度”义务的合规证明提交给A的监管机构，而无需再按A的程序从头做一遍内部自声明和文件归档。

(3) 试点启动策略：利用既有互信，聚焦低争议单元。 功能等价互认的政治可行性，取决于能否找到一个低启动成本、高成功概率的试点。我们的具体建议是，在拥有类似数据保护“充分性认定”互认基础的欧盟和日本之间，启动首个试点。在它们现有的《日欧数字伙伴关系》框架下，增设一个为期18个月的“AI合规功能等价互认试点项目”。

试点核心： 聚焦于最少政治争议的“使用者透明度”单元（即如何向用户标识AI身份）。双方技术工作组就这一单个领域，比对各自的合规要求（如日本可能要求前置说明，欧盟AI法要求明确的水印或标签），认证其功能等价性，并达成正式的互认安排。

成功标志： 在18个月试点期内，通过欧盟-日本互认通道进入任一方市场的AI企业，没有因其在本国完成的合规动作不符合对方标准而被对方监管机构以其本国的透明度规则为由予以处罚。一旦试点成功，其在操作层面的可行性便得到实证证明，可以向其他低政治争议的单元（如训练数据文档化），以及向与欧盟有充分性认定的其他法域（如韩国）进行扩展。

这个方案之所以可行，是因为它规避了触及根基的政治障碍：它不要求任何法域修改其法律条文中的AI定义，不要求欧盟放弃其“产品安全”的预防性逻辑，也不要求美国或任何一个亚洲法域全盘移植GDPR式的数据保护标准。它仅仅要求监管机构在行政审查的“输入端”增加一个选项——“接受同等功能的证明”。这将全球AI治理矛盾，从不可妥协的主权立法之争，转化为了可逐步解决的技术标准与行政效率对话。

5.3.3 第三层（市场层）：法律本体论审计与资本市场压力

互认协议的推进面临一个核心难题：即使政府间达成共识，其扩展速度和覆盖范围的确定性依然脆弱。我们需要一种自下而上的、基于市场力量的驱动机制，与自上而下的政府间合作形成合力。这便是第三层——“法律本体论审计”（Legal Ontology Audit）的作用。

审计的目标与逻辑： 法律本体论审计的核心目标，不是替任何法域的监管机构执法，而是为资本市场提供关于一家AI企业面临的“跨法域法律身份矛盾风险”的可信信息。它的审计对象，是同一个AI企业在不同法域向监管机构、客户和投资者所作的官方陈述之间的一致性。

审计的三个核心问题：

1. 矛盾识别问题： 企业在法域A（如欧盟）的合规文件中对同一AI模型的性能局限描述，与在法域B（如美国）的商业宣传文案或对监管机构的陈述之间，是否存在实质性矛盾？（例如，在欧盟描述模型“不能用于对个人有法律或重大影响的决策”以规避高风险分类，同时在美国向投资客户宣称“本模型能高效准确筛选求职者”）。

2. 归因分析问题： 若发现矛盾，审计师需判断这是源于不同法域表格格式的必然技术性差异（格式不一致），还是企业为了在不同法域实现监管套利的策略性行为（实质性矛盾）。

3. 内控评估问题： 该企业内部是否建立了独立的、跨法域的法律身份一致性审查职能，以主动管理和防范这种矛盾风险？如果没有，这一缺失是否已被明确告知公司的董事会、审计委员会以及潜在的投资者？

资本市场压力的传导机制： 当国际知名的会计或咨询机构（如“四大”）在为企业提供技术合规或ESG（环境、社会及治理）鉴证服务时，将“跨法域AI法律身份一致性”作为一个新的评估模块纳入，便会产生强大的市场规训效应。一家被标注为“高身份矛盾风险”的AI企业，将面临几方面的直接压力：首先，其承保董事及高管责任险（D&O保险）的保险公司可能因法律风险不确定性增加而提高保费或除外相关风险；其次，在

并购交易中，买方的技术尽调和法律尽调团队会将此审计结果作为关键风险项，影响交易的估值和交割条件；最后，大型的机构投资者（如养老基金）在ESG投资原则下，更可能避开此类存在“治理”（Governance）隐患的企业。

这种来自资本市场的压力，其精妙之处在于它完全绕开了冗长的多边条约谈判。它构建了一个直接的商业激励：企业为了获得更低的融资成本、更顺利的并购退出和更稳定的估值，会自动产生收敛其跨法域法律身份的动机。它们会主动游说各法域监管机构采纳功能等价的互认标准，因为一个清晰、可互认的规则体系能减少其“身份矛盾审计”不合格的风险。这最终会形成一个自下而上的、由企业和资本共同驱动的收敛闭环，倒逼全球AI规制向可操作化的互认方向发展。

5.4 反驳预期与回应

本文提出的理论框架和规制方案，可能面临以下有力的学术批评，本节预先予以回应。

反驳意见一：本文过度强调了法律分类的困难，而简化了技术风险的评估。AI治理的首要问题仍然是防范技术滥用带来的实质损害（如歧视、虚假信息），过度聚焦于跨国法律的“分类游戏”可能会错失规制的时间窗口。

回应： 此批评具有重要警示。我们完全同意，任何全球协调机制都不应削弱对公民基本权利和实质损害的即时保护。但本文的论点并非“忽略实质风险”，恰恰相反，是“实质风险因为法律分类的制度僵局而无法被有效规制”。例如，一个在法域A被认定为“需严格控制以避免就业歧视”的高风险招聘AI，因为被法域B归类为“普通商业效率软件”，在法域B使用同样产品的雇主便可能逃脱其在法域A需承担的责任（如使用前须进行偏见审计）。法律本体论套利恰恰是放大、而非缩小了实质技术损害。我们的方案旨在通过降低重复合规成本和消除套利空间，使得法域A能在不丢失其高标准保护的前提下，更有效地将其保护效力扩展至跨境运营的企业。

反驳意见二：“制度替代认知”理论带有制度决定论色彩，忽略了政治意愿、公民社会运动、突发灾难性事故等外部力量打破制度锁定的可能性。

回应： 制度不是一成不变的铁笼，本研究亦非历史决定论。我们承认，一次足以震动全球公众的严重AI事故、或一场跨国的消费者集体诉讼，完全可能成为打破现有制度均衡的“外部冲击”。但一个负责任的政策理论不能将希望寄托于不可预测的灾难。本文所做的工作，是在常态政治的边界下，寻找一种即使在没有外部冲击时也能逐步向前演化、降低风险的路径。我们的方案本身就设计了一种“自下而上的触发机制”——法律本体论审计——其效力正是通过市场中的“微型冲击”（审计不合格导致融资或收购受阻）来不断产生改变的压力，这正是对制度惰性的主动破局，而非被动等待。

反驳意见三：功能等价互认在操作上极易被企业利用，成为“逐底竞争”的工具，即企业总是寻求承认那个标准最低的法域的合规动作为“功能等价”，从而在实质上拉低全球的保护水平。

回应： 这是一个切中要害的批评。确实，在国际贸易的实践中，相互承认原则（Mutual Recognition）有被滥用于放松监管的前科。为了防止这种情况，本文的方案内置了针对性的防御机制。首先，“差分矩阵”的构建本身就是一个使保护水平透明化的过程，低标准的合规单元在矩阵中将一览无余，使其难以在技术上被论证为与高标准“等价”。其次，“功能等价”的认定不是由寻求套利的企业单方面完成的，而是由双边指定的、包含公共利益代表的技术工作组通过严谨的比对做出的，其决策过程有透明度要求。更重要的是，最高标准的熔断机制是“诉诸法庭”：如果一个法域依据互认协议认可的B国合规动作进入A国市场后，其在A国造成了实质性损害，A国的受害者仍然有权在A国的侵权法体系下提起诉讼。此时，A国法院将有机会在个案中审查：B国那套已被协议认定为“功能等价”的程序，在实际操作中是否真的达到了A国可接受的审慎水平。这个事后司法审查的保留权，是对“逐底竞争”的终极威慑。

5.5 可证伪性：理论的检验设计

为使本文的理论超越纯粹的叙事解释，成为一个可被检验的学术命题，我们在此系统设计其证伪条件。本文的核心理论可被归结为一个可检验的假设：在主要法域已建立AI监管体系后，它们之间法律分类的分歧，将

主要由其各自的制度替代路径的累积成本锁定，而非由对AI认知的深入而趋同。其直接后果是，企业将利用这种被锁定的分歧进行法律本体论套利。

5.5.1 最强竞争解释与本理论面对的质疑

与本理论构成最直接竞争的，是我们可以称之为“认知趋同论”的观点。该观点认为，当前的分歧仅仅是AI发展早期的暂时现象。随着全球学术界、立法者和公众对AI技术本质及其社会影响的理解不断加深，一个关于如何监管AI的“最佳实践”共识会自然浮现，各国法律将围绕这个共同认知而逐步趋同。政策制定者和国际组织的主流话语大体反映了此种乐观态度，他们将当前的碎片化视为一个通向最终和谐的、暂时的学习阶段。

此外，另一个强有力的竞争解释是“文化决定论”，认为各法域在隐私、言论自由、集体与个人关系等方面的深层文化价值观差异，是导致其在如何监管具有颠覆性潜力的AI上必然分道扬镳的根本原因，制度选择和成本锁定不过是为这些深层文化价值披上的外衣。

本理论并不否认认知进步和文化价值会影响法律选择，但它提出了一个强有力的机制性替代解释：制度本身的重量（沉没成本与专业生态）才是使分歧锁定的最主要、最直接的力量。

5.5.2 研究设计与检验方法：一个2×2矩阵

为在这两种解释之间做出裁决，我们设计一个以未来3-5年为观察窗口的2×2研究设计。

第一轴是“功能等价互认的启动法域”：一边是从已有互认基础的法域对启动（例如，欧盟-日本或欧盟-韩国，它们之间已有数据保护互认的先例）；另一边是尝试在尚无互认基础、甚至存在地缘政治信任赤字的法域对（如欧盟-中国，或美国-中国的一个具体议题合作）推动同等工作。第二轴是“互认推动的主体性质”：一边是政府间协议驱动（自上而下）；另一边是企业合规联盟驱动（自下而上），即AI产业界通过联合提出一个可同时被双方接受的第三方技术标准，来游说双方政府接受。

由此得到四个方格。本文理论的判别格，是“已有互认基础的法域对 × 企业驱动”。

* 本理论的预测（检验假设H1）： “制度替代认知”理论预测，即便是在已有互信的法域间，只要设计方案是要求监管机构承认“功能等价”，不触动其法定的AI定义，企业驱动的方案在降低政府间谈判成本的情况下，有相当大的成功概率。因为监管机构拒绝它的成本（对抗商业友好形象、无实质性主权损失）高于接受它的成本。因此，我们预测在欧盟-日本这类法域对之间，一个企业驱动的互认试点在18个月内被正式开启是大概率事件。

* 竞争解释的预测（零假设H0）： “文化决定论”会预测，即使在欧日之间，深层的文化差异最终也会阻止任何有实质意义的互认。“认知趋同论”则会预测，若没有对AI的统一定义达成共识，任何互认都是空中楼阁，因此企业驱动的方案会被政府以缺乏坚实的共同认知基础而驳回。

检验的执行方法：

我们采用过程追踪和差别法（Method of Difference）结合的方法。在判别格中，我们选取欧盟-日本作为实验组，选取欧盟-中国（或美国-中国）的一个同类企业倡议作为对照组。我们系统地追踪和比较：在2025-2027年间，面向这两对法域发起的、聚焦于AI治理某一具体单元（如使用者透明度、或训练数据文档化）的、由企业联盟提出的自下而上的互认倡议，其各自的命运如何。我们收集的数据包括：相关商业联合会的公开倡议书和立场文件、双方政府监管机构的官方回应和咨询文件、以及参与企业的内部合规文件（如果可得）。

* 何种结果将证实本理论？ 如果观察到，欧盟-日本的企业倡议在经历了初期技术细节的磨合后，最终被双方政府接受并进入试点阶段；而同期向EU-中国提出的同类倡议却因“缺乏共同价值观和理解基础”而被搁置或拒绝，那么便强有力地支持了H1。这表明，阻碍互认的核心障碍并非“对AI的共同认知”这一抽象目标，而是制度信任与政治关系等前置于AI问题本身的既有条件，这与“制度替代认知”理论中“用既有机器去工作”的逻辑一脉相承。

* 何种结果将证伪本理论？ 如果观察到，即使是在欧盟-日本这对有长期互信合作基础的伙伴之间，其监管机构也明确拒绝了所有企业驱动的、不涉及修改法律定义的功能等价互认提议，且拒绝的核心理由是无法在缺乏对“AI”这一核心对象的统一定义下进行有效的行政协作，那么本理论关于“功能等价互认可以绕开定义分歧”的核心机制就被严重削弱。情况将可能如竞争理论所言，统一定义仍然是协调的第一步。这将迫使我们修正理论，承认在某些情况下，“制度替代”的程序在没有触及认知共识之前无法自发收敛。

这个清晰的、包含实验组和对照组的检验设计，使本文的理论脱离了自我循环论证，为其在未来被证实或证伪提供了明确的操作路径。这一可证伪性是本文作为严肃学术产品的核心分量。

6 结论

本文完成了一项从理论诊断到规制方案建构的系统性工程，为理解和破解生成式人工智能全球治理的碎片化困局提供了全新的分析框架与政策工具集。

6.1 核心发现

本文的核心发现可凝缩为以下四个层层递进的判断：

第一，全球AI定义分歧的根源，并非对AI的认知不同，而是各法域在遭遇法律认知失败后，不约而同地启动了“制度替代认知”程序。 欧盟、美国和中国对AI的分类，不是对同一个问题的不同正确回答，而是他们各自基于手中可用的、在AI出现之前就已成熟的监管机器（产品安全、市场执法、网络治理），对“我们该把AI交给哪个已有部门去管”这个问题的强制性、操作性回答。

第二，这些被制度替代认知产出的分类，正因各自的巨大制度累积成本而被不对称地锁定在分歧中。 围绕每个分类选择的专用知识、产业链条和既得利益集团，使得退出任何一个分类的成本都远超忍受分歧的成本，导致趋同的动力远弱于路径依赖的惯性。

第三，这一被锁定的碎片化格局，正催生一种前所未有的结构性风险——法律本体论套利。 精明的企业正学会通过微调技术架构和法律声明，使其同一AI模型在不同法域被构成为不同的法律对象，以系统性规避全局合规成本。这已不仅是单一的技术风险，而是对法律可预测性这一根本制度公共品的侵蚀。

第四，解锁这一僵局的可行路径，不是奢求一个全球统一的AI定义，而是转向管理不一致性的后果。 一套由“规则起源地分析（透明化话语）”、“功能等价互认（降低制度成本）”和“法律本体论审计（形成市场压力）”组成的三层次递进方案，能够在尊重各法域现有制度主权的前提下，创建连接不同监管生态系统的互操作接口，并自下而上地驱动规制收敛。

6.2 理论贡献

本文的理论贡献主要体现在三个层面。

其一，在法理学领域，本文超越了“AI是物还是人”的经典争论，将分析焦点从被分类的客体转向了分类动作本身，首次系统性地理论化了法律体系在面对本体论不稳定对象时的“认知失败”现象，并揭示了其“制度替代”的必然操作机制。

其二，在全球治理与国际法领域，本文对主流“统一公约论”进行了根本性的批判与发展。本研究借鉴了演化经济学的路径依赖概念，创造性地提出了“分类成本不对称锁定”这一理论，将全球治理碎片化从一种“缺少协调”的认知现象，重新构框为一种由内在制度动力学驱动的、难以自动复原的“结构僵局”。

其三，本文打通了法学、制度经济学与国际税收治理等学科界限，论证了“功能等价互认”可以被创造性地从产品标准和税务协调领域，迁移应用到AI治理这一“多重法律编码竞争”的全新场域。特别是，本文提炼的“法律本体论套利”概念，为法学和金融合规研究开辟了一个新的交叉课题，即法律身份本身如何成为一种可被经营和套利的资产。

6.3 实践与政策含义

本文为从事AI治理的政策制定者、国际组织技术官员和企业的法律合规管理者，提供了以下几项具体且可操作的应用指引：

1. 为跨国政策对话提供新方法。 参与AI政策国际对话的官员和专家，可立即将本文所建议的“规则起源地分析三问”纳入双边和多边会谈的议程。在讨论具体分歧前，先让各方完成对自己和他方分类的溯源性分析，有助于将对话的性质从“维护己方定义的优越性”转向“识别彼此监管机器的覆盖盲区与衔接点”。

2. 为国际组织提供可启动的双边试点蓝图。 经济合作与发展组织（OECD）的AI政策观察站、或全球人工智能伙伴关系（GPAI）等国际机制，可采纳本文提出的“合规动作差分矩阵与功能等价互认试点”方案。特别是，可以在日欧数字伙伴关系或美欧贸易和技术委员会（TTC）等既有高级别对话机制下，启动关于“使用者透明度”这一具体单元的互认试点项目，为其配备18个月的观察期和明确的效果评估指标。

3. 为跨国企业和法律合规产业提供新的战略视角。 对于生成式AI的开发者和部署者，本文的诊断意味着，其法律和合规团队必须从“逐国合规”（country-by-country compliance）的被动模式，转向系统性的“跨法域法律身份管理”。本文提出的“法律本体论审计”框架可被企业直接内化为其内部合规审计的核心流程，以识别和管理跨法域声明的矛盾风险。对于来自四大会计师事务所或大型律所的专业服务提供者，“跨法域AI法律本体论审计”极有可能成为一个新的、高附加值的专业服务品类。本文呼吁有远见的先行者开始这一领域的知识储备和人才建设。

4. 为各国监管机构提供反套利的思维工具。 本文的分析警示，监管机构在起草和执行AI法规时，必须引入一个此前不曾存在的思维维度：即评估其单边规制行为在通过与该国企业互动的其他法域的规制环境交叉折射后，可能产生的下游效应和套利机会。在批准涉及AI的跨境并购、或在处理受害者的跨境损害诉讼时，应考虑对涉案AI企业的跨法域法律身份陈述进行审查。

6.4 研究局限

作为一项以理论建构为主的研究，本文存在以下局限，其本身也是未来研究的重要方向。第一，本文提出了“法律本体论套利”这一概念，并对其机理进行了理论剖析，但受制于当前数据可得性，尚未对“套利”行为在企业个体层面的具体发生频率、规模及危害进行定量测量。第二，“功能等价互认”方案在技术操作层面的细节，如具体合规单元等价性判准的起草，需要技术标准专家和工程师的深度介入，本文仅完成了其制度设计层面的顶层架构。第三，本文的分析主要基于截至2025年上半年的欧盟、美国、中国三套监管体系。其他重要法域（如英国、新加坡、加拿大）的规制创新路径，以及印度、巴西等新兴经济体正在成形的方案，尚未纳入系统比较，这限制了本文结论的全球普适性。

6.5 未来研究方向

本文的研究开启了几个可持续生长的跨学科研究纲领，可供后续研究独立发展：

1. 比较法律生态学：AI分类器的制度谱系数据库。 未来的研究可以拓展本文的规则起源地分析，系统性地收集、编码和比较10-15个主要法域（包括日本、韩国、英国、加拿大、新加坡、印度等）的AI法律治理框架，建立一个公开的跨国“可用分类器谱系”数据库。这使研究者可以运用定性比较分析（QCA）或大样本的定量方法来检验本文的核心假设：一个法域对AI法律分类的选择，与其既有监管机器的结构和管辖范围之间存在显著的相关性，而与其法学界对AI本体论学术讨论的共识之间不存在显著相关性。

2. 功能等价互认的试点后评估与行为经济学分析。 在第一个功能等价互认试点（如在欧-日之间）成功运行18-24个月后，应当进行一项严格的事后评估研究。这一研究应聚焦于参与互认试点的AI企业，并使用双重差分法（DID）等计量经济学工具，对比它们在试点前后的合规成本变化、新市场的进入速度、以及被其国内或国外监管机构处罚的频率变化，并与一组未参与试点的同类公司进行比较。此研究能从实证上验证“功能等价”路径的成本节省与风险控制效果。

3. 法律本体论套利的实证发现与资本市场定价研究。 后续的研究可着力为“法律本体论套利”提供实证依据。方法之一是法律语言学的判例分析：系统检索并分析全球主要司法辖区涉及同一AI公司的跨法域诉讼材料，寻找企业在不同法域或不同案件中对同一模型性能做出自相矛盾陈述的案例。方法之二是事件研究法：当一家主要会计或评级机构宣布将“跨法域AI法律身份一致性”纳入其评估体系时，可以观察资本市场是否对已知存在身份矛盾的AI企业进行重新估值（股价负向波动），以检验市场是否已经开始为这种新型法律风险定价。

4. 互认矩阵的动态更新机制与全球行政法研究。 本文遗留的一个关键未解问题是“互认矩阵的维护权”。当AI技术发生阶跃式变化，使得当前已被认定等价的合规单元不足以覆盖新出现的风险时，谁来触发、评估和完成互认矩阵的更新？这涉及到一种全新的全球行政法编制权归属问题：它应当由初始的技术工作组、受到损害的非政府组织、还是各国监管机构本身来启动？后续研究可以从全球行政法（Global Administrative Law）的视角，深入探讨“合规互认矩阵”的治理架构、合法性来源和问责机制，为其设计一个兼具效率与公信力的长期治理框架。

5. 制度的可逆性研究：法律锁定的成本度量与解锁条件。 本文提出了“分类成本不对称锁定”的理论概念，但尚未能精确度量“锁定”的强度。未来在政治学和法经济学的交叉领域，可以产生一个重要研究议程：开发一套指标体系来度量一个法域在被锁定于某个AI法律分类上的不可逆程度。这可以包括立法修改所需的议会票决门槛、已有行政程序的不可逆程度（例如，是否能以行政指令撤销）、法律从业者专业教育的路径依赖程度、以及该分类锁定的经济受益者（合规产业）的政治影响力等。建立这样的度量体系，可以帮助预测各法域在未来的国际协调中是成为“变革的支持者”还是“顽固的现状维护者”，这将是全球AI治理政治学的一个重大突破。

编辑核验参考文献

M. Hacker et al., A Legal Risk Taxonomy for Generative Artificial Intelligence, 2024.

C. Hacker et al., Report of the 1st Workshop on Generative AI and Law, 2023.

European Union, Regulation (EU) 2024/1689 (Artificial Intelligence Act).

OECD, Recommendation of the Council on Artificial Intelligence.

NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023.