作为或然性工具的人工智能及其法律挑战

郑　戈

【内容摘要】生成式人工智能的崛起标志着工具本质的范式转型。传统法律体系建立在确定性工具的技术逻辑之上，预设使用工具的人类行为与人类意图的线性映射关系，由此形成“设计者可控—使用者可责”的归责框架。然而，生成式人工智能作为或然性工具，其自回归生成机制与概率抽样特性颠覆了工具行为的可预见性原则：神经网络的参数空间阻断了设计意图的完整渗透，强化学习的动态调优消解了使用指令的确定性约束，技术应用呈现去中心化扩散效应。这种技术本体论层面的变革，使得传统法律中雇主责任、产品责任与平台责任的三重归责体系遭遇结构性危机。通过解构大语言模型的系统性幻觉机制，可以揭示算法黑箱与人类认知框架的深层张力，进而提出基于过程导向的动态合规路径——算法透明度强化、以人类为中心的审计机制与职业伦理再造，在技术创新与风险防控之间重构规范性平衡。这一探索最终指向法律科学从机械决定论向概率主义范式的认知跃迁，为人工智能时代的制度调适提供理论框架。

【关 键 词】生成式人工智能　或然性工具　算法幻觉　法律主体性　归责体系重构

【作　　者】郑戈，上海交通大学凯原法学院教授。（上海 200030）

问题的提出

生成式人工智能工具对人们构成致命的诱惑。一方面，它使得脑力工作者过去需要连续熬夜才能完成的工作可以在几分钟乃至几秒钟之内完成；另一方面，它那充斥着排比句和修辞的看似有理的生成内容中又埋着无数的坑，未经人工核验就提交出去，会面临声誉乃至法律上的不利后果。正因如此，人们对它的态度也是完全不同的：一些人主张积极地拥抱这种新的工具，甚至认为不会使用这种工具的人在职场上会被淘汰，因此开始为专业人士提供人工智能工具的使用指南；

另一些人则认为在其没有解决“一本正经地胡说八道”问题之前，对这种工具应该敬而远之。笔者基本上倾向于第一种立场，就法律职业而言，人工智能不会完全替代法律人，但会使用人工智能工具的法律人将替代不会使用这种工具的法律人。但笔者也理解持第二种立场的人士的担忧，本文正是针对这种担忧所基于的误解而提出的一种概念化解释方案。简单地说，本文认为：由于生成式人工智能产生的内容真假莫辨、充满幻觉而拒绝使用它，是因为人们习惯于使用既无法带来惊喜也不会带来惊吓的“确定性工具”,而人工智能是一种既可能带来惊喜也可能带来惊吓的“或然性工具”。本文将确定性工具（deterministic tools）定义为遵循线性因果律且具备全映射功能的技术中介系统，其基本特征可归结为三项核心要件：输入输出关系的完全可预测性、操作过程的可溯源性以及结果生成的逻辑必然性。而或然性工具（probabilistic tools）则是一种基于概率论框架构建的非线性技术系统，其基本特征可归结为三项核心要件：输入输出关系的概率性关联、操作过程的涌现性与部分不可溯源性以及结果生成的非必然性与多元可能性。

由于“或然性工具”是本文提出并阐释的主要概念，这里先对它的三个核心要件进行简要的描述。输入输出关系的概率性关联是指或然性工具不遵循线性因果律，而是通过统计学关联（如大语言模型的自回归机制和注意力机制）重构输入与输出的因果关联。输出结果受内在随机性要素（如神经网络的参数采样和温度参数调控）支配，形成非单一确定解空间，而非完全可预测的映射关系。例如，生成式人工智能的相同输入可能导向多模态输出，其关联本质上是基于训练数据的统计分布而非逻辑规则推导。操作过程的涌现性与部分不可溯源性是指操作过程依赖高维参数空间中的概率抽样和强化学习机制（如 Transformer 架构的交叉注意力运算），导致决策路径呈弥散式分布。设计者意图无法穿透神经网络的隐层空间，使用者的指令难以规训模型的概率采样，从而产生系统性幻觉（如虚构内容生成）。这使得操作过程具有涌现特性（知识通过组合创新涌现），但部分环节（如隐层激活路径）因黑箱效应而难以完全追溯。结果生成的非必然性与多元可能性是指或然性工具的结果生成不具备逻辑必然性，而是通过概率分布重组（如潜在空间的语义向量投影）产生或然性结果。输出具有不确定性特质，同一输入可能对应多模态合理输出,各选项间构成非排他性的可能世界集合（如人工智能生成内容时出现的“系统性幻觉”与创造性输出并存）。这突破了传统工具“可控—可责”的认知框架，强调结果的概率阈值而非确定性边界。

在确定性工具与或然性工具二分的概念框架下审视人们对待人工智能的不同态度，其背后的社会心理预期便能清晰浮现出来：对“一本正经地胡说八道”的抱怨来自人们对传统工具的理解和期待，即确定性期待。确定性工具能够做什么完全是由制造和操纵它们的人来决定的，这也包括人工智能技术成熟前的计算机。这种工具的使用方法和功效完全为霍布斯自然哲学中的“制造者知识”（the maker’s knowledge）[①]所把握，不会出人意料。但人工智能不是这样的工具，它能够在你没有明确想法的情况下根据你模糊表述的提示词帮你生成结构清晰、逻辑严谨的内容，能够根据你用自然语言表述的需求生成编程代码，能够根据你的简单描述为你“创作”图像或视频。幻觉当然存在，但这正是它的“创造力”的代价。没有人希望人工智能只能根据明确表述的关键词来完成搜索，或者是仅仅把你录下的语音准确转化成文字。

技术范式从确定性工具向或然性工具的跃迁正在引发法学体系的深层震荡，其结构性突破主要体现在三个维度。其一，传统归责逻辑遭遇系统性解构。当自动驾驶系统基于实时环境感知自主执行避险决策时，产品责任法预设的“设计缺陷—制造瑕疵”二分框架显露出根本性局限——

科技之治：第四次工业革命中司法制度的“变”与“不变”研讨会海报

算法在动态运行中形成的策略调优，已然超越工业时代静态的产品瑕疵认定范式。更值得警醒的是，深度伪造内容在社交平台的指数级传播，彻底暴露了“通知—删除”机制的时空错位：这套诞生于 Web1.0 时代的责任规则，在面对概率生成的内容洪流时，如同试图用陶罐承接瀑布般徒劳。这种困境恰如 19 世纪铁路事故对传统侵权法的冲击，但此次技术革命的烈度远超往昔。其二，法律主体性理论面临本体论危机。欧盟赋予高级机器人“电子人格”的立法尝试，本质上是对罗马法“人格拟制”原理的路径依赖。除了赋予人工智能体以“主体性”的方案外，法学界还有赋予其从属性人格（类比于罗马法中的奴隶）的提议，主张借鉴罗马法对奴隶经商的法律关系概念化，比如 actiones adiecticiae qualitatis（字面意为“追加性质的诉讼”[②]），构建不依赖主体拟制的风险分配模式。[③]这种法律拟制未能触及算法决策的分布式本质——模型输出是训练数据、初始参数、实时交互与随机变量共同编织的拓扑网络，其决策节点如同区块链的分布式账本，不存在传统法律主体所需的意志中枢。将工业时代的责任容器强加于网络化智能体，不仅暴露了法学界对机器学习架构的认知盲区，更折射出法律想象力的时代局限。这令人想起 19 世纪法人人格争议时梅特兰的告诫：法律拟制需警惕将活生生的现实塞进概念标本箱的危险。[④]其三，法学认识论遭遇范式颠覆。法律推理依赖的类比技艺与涵摄技术，在面对概率驱动的决策系统时遭遇解释力危机。当医疗人工智能基于贝叶斯网络给出超越临床指南的治疗方案时，医生陷入双重困境：既无法追溯黑箱中的逻辑路径，也难以用自然语言重构其论证过程。这种认知断裂本质上是两种认识论的冲突——法律人的“理由之治”与算法的“相关性统治”正在知识论层面分道扬镳。这恰似 16 世纪罗马法学者面对英国判例法时的困惑，但此次知识传统的裂痕更深。

本文旨在回答的基本问题是：在生成式人工智能时代，传统法律体系如何应对归责逻辑解构、法律主体性危机和认识论范式断裂的三重挑战，实现从确定性思维向或然性思维、从意志论向控制论、从教义学向系统工程的全面转型？生成式人工智能的勃兴正在动摇传统法律体系的认知论根基，法律赖以为生的“意志—行为—责任”因果链便遭遇了本体论层面的消解。传统归责体系预设的行为可归因性在算法涌现的混沌中失去支点：产品责任法无法锚定动态调优的算法策略，版权制度难以界定概率生成内容的独创性边界，侵权法则在面对自主决策系统的行为不可溯源性时陷入解释困境。更深层的危机在于法律语言的确定性根基被侵蚀，法律文本作为意义载体的稳定性正被概率性语言游戏所瓦解。这种颠覆迫使法学界必须直面三重认识论转型：从追求唯一正解的确定性思维转向容纳多元可能的或然性思维，从强调主观过错的意志论转向关注风险分配的控制论，从静态规则涵摄的教义学范式转向动态过程治理的系统工程。唯有重构法律与技术的认知接口，在算法透明性框架中重建“人在回路”（Human-in-the-loop）的价值锚点，方能避免法律

沦为技术必然性的注脚，守护以人的尊严为尺度的规范秩序。

表 1　技术 / 工具类型学

类型	定义	核心特性	法律属性
确定性工具	遵循线性因果律且具备全映射功能的技术中介系统，其行为可被人类完全预测和控制。输入输出关系严格对应，操作过程可逆	1. 输入输出关系完全可预测性：相同输入必然产生相同输出 2. 操作过程可溯源性：行为路径可被逆向追踪和解析 3. 结果生成逻辑必然性：输出结果由设计意图直接决定，无随机性	1. 归责逻辑：基于“可控—可责”原则，责任可追溯至使用者或设计者（如产品责任法） 2. 法律基础：预设行为与意图的线性映射，归责框架清晰（如雇主责任、产品瑕疵认定） 3. 适用场景：传统侵权法、合同法等静态规则体系
或然性工具	基于概率论框架构建的非线性技术系统，其行为依赖统计关联和随机抽样。输入输出关系非固定，操作过程部分不可控	1. 输入输出关系概率性关联：相同输入可能导向多模态输出，关联基于统计分布 2. 操作过程涌现性与部分不可溯源性：决策路径弥散式分布，隐层空间不可解释（黑箱效应） 3. 结果生成非必然性与多元可能性：输出具有不确定性，同一输入对应多模态合理输出	1. 归责困境：设计者意图无法穿透参数空间，使用者指令难以规训模型，导致“可控—可责”框架坍塌 2. 法律挑战：需重构归责范式（如动态合规、算法透明度），从结果导向转向过程控制 3. 风险特征：系统性幻觉内生于技术本体，法律需容纳不确定性（如概率阈值替代二元因果）

确定性工具与法律中的主客二元界分

自霍布斯在《利维坦》中将国家喻为人工造物、笛卡尔将动物视作自动机械以降，机械论世界观便为现代法律体系铺设了认知论的地基。传统工具观建立在线性因果律与完全可预测性的物理法则之上，从阿基米德杠杆到内燃机车，工具的物理属性与功能边界被牢牢锚定在经典力学框架内。这种确定性映射在法律系统中凝结为“可控—可责”的归责范式：钟表匠对齿轮啮合误差负全责，马车夫对缰绳失控致人伤亡担过失，饲养人对危险动物承担无过错责任。这些规则皆预设工具行为可溯源于人类意志的单向投射。然而生成式人工智能的崛起，正以概率论的计算统计学逻辑解构着法律赖以存续的机械论根基。当大语言模型在自回归生成中涌现出设计者未曾预见的“系统性幻觉”，当自动驾驶系统在强化学习中演化出脱离原始训练框架的决策路径时，传统法律体系中主客二分的工具论预设便遭遇了本体论层面的挑战。

笛卡尔在《方法论》中将动物比作精密的自动机器，“我们知道人的技巧可以做出各式各样的自动机，即自己动作的机器，用的只是几个零件，与动物身上的大量骨骼、肌肉、神经、动脉、静脉等等相比，实在很少很少，所以我们把这个身体看成一台神造的机器”；[⑤]霍布斯则在《利维坦》中将心脏喻为发条、将神经视作游丝：“由于生命只是肢体的一种运动，它的起源在于内部的某些主要部分，那么我们为什么不能说，一切像钟表一样用发条和齿轮运行的‘自动机械结构’也具有人造的生命呢？是否可以说它们的‘心脏’无非就是‘发条’，‘神经’只是一些‘游丝’，而‘关节’不过是一些齿轮，这些零件如创造者所意图的那样，使整体得到活动的呢？”[⑥]——这些思想巨匠的机械隐喻绝非文学想象，而是近代工具理性建构社会秩序的认知基石。传统工具范

式以双重确定性为特征：物理机理遵循既定的因果律（如斧刃锋面与劈砍度的线性关联），功能指向符合设计者意图（如望远镜仅服务于视觉延伸）。这种确定性保障了人与工具之间关系的透明性和单向操纵性——中世纪匠人能依据齿轮参数推算座钟误差，工业化车间可通过工时动作分析预判生产效能。人类社会迄今为止的规范体系（无论是伦理还是法律）在处理人与工具（包括动物和早先的奴隶）的关系时都是以确定性工具作为工具的当然形态，因此谁控制谁负责就成为基本原则。比如，在承揽关系中，承揽人以自有工具完成工作成果（如自带电动葫芦运输建材），原则上由承揽人承担工具使用风险。定作人仅在存在选任过错（如未审查承揽人资质）或指示过错（如要求危险操作方式）时，需按过错比例承担补充责任。在传统雇佣关系中，工具由雇主提供，雇主需确保工具安全性并承担使用风险；而在“带工具雇佣关系”中，若雇员自备工具，雇主仅负安全保障义务（如提供安全操作指引），但工具瑕疵导致的损害可能由雇员分担责任。如工人使用自带受损工具作业受伤，法院可能认定雇主未尽安全检查义务，需承担主责。“可控—可责”的法理框架正是机械论世界观的体现：一如钟表匠对机芯缺陷承担完全责任，工具设计者的过错认定可通过逆向工程精确追踪。

霍布斯的哲学体系以“制造者知识”为核心，主张人类通过主动构建对象（如几何图形或社会契约）获得科学知识（scientia），其统一性体现在两种科学（几何学与公民哲学）之中。他严格区分了基于因果性推论的确定性知识(scientia)和经验性认知(cognitio),前者依赖生成性定义(如“线是点的运动轨迹”)，后者源于感官经验。其中，生成性定义强调通过构造过程（如几何作图或政治契约的订立）揭示事物本质。换句话说，工具的确定性来自人基于知识对它的制造和控制。[⑦] 人也因此要对自己使用工具的行为负责，不存在所谓“工具伦理”（比如当下人们热议的“人工智能伦理”），而只有人使用工具的伦理。

基于“制造者知识”创造出的工具预设运作机制与结果呈现间的严格确定性对应关系，因此是完全可预期、可控制的。由于这类工具是人类按特定意图设计的，为特定目的而使用的，且使用工具的过程完全在特定的人类主体的控制之下，所以主体而非工具是责任的承担者，谁控制谁负责、谁使用谁负责就成为一般原则。在加工承揽、雇主责任、产品责任等特定民事关系中，还需穿透控制或使用的表象，找出对工具性能和相应风险具有信息优势且因此能以较低成本控制风险、预防损害的主体，并由其承担责任。

体现机械论的法哲学将工具视为人类意志的被动延伸，其典型表征可见于《德国民法典》，其中第 823 条和第 826 条确立的过错责任原则（行为人主观故意或过失）与第 833 条对危险动物饲养人设立的严格责任（区分生计性动物的过错推定责任），典型地体现了机械论世界观下侵权归责的规范逻辑——无论是工具操作的可控性预设（通过因果关系链追溯使用者过错），抑或危险动物的风险归因（基于物种危险性的客观可预见性），均建立在工具与动物行为符合同一律确定性假设之上。从机械论法哲学视角审视《中华人民共和国民法典》相关条文，其规范逻辑呈现出工具理性支配下的递进式建构。首先是过错责任原则（第 1165 条），以行为人主观过错为归责核心，预设工具完全受人类意志支配的线性因果链，将损害视为行为人意志缺陷的客观延伸，损害结果被视为行为人主观过错（故意 / 过失）的外化，其归责逻辑依赖于因果关系链的完整性（如第 1169 条教唆帮助责任中“意思联络”对因果关系的强化）。机械论线性因果观还体现在第 1170 条共同危险行为侵权责任承担上，要求“不能确定具体侵权人”时方适用连带责任，反映出对个体行为与结果精确对应的执念。其次是严格责任体系（第 1166 条及相关特别法），通过风险客观化

实现责任扩张，如动物致害责任（第 1245 条）摒弃“生计性 / 奢侈性”区分而直接以饲养人 / 管理人为风险容器，将动物重构为去意志化的危险源，表明立法者更关注风险的社会可预见性而非生物学特性。在责任主体层面，动物管理人责任与建筑物管理人责任（第 1254 条）突破所有权中心主义，揭示了机械论框架下责任主体从“意志控制者”向“风险承担者”的嬗变。第 1249 条遗弃动物致害责任中，原饲养人责任不因丧失占有而中断，切断了“控制力减弱则责任递减”的传统逻辑。这些特点都表明中国民法典在继承“行为—结果”因果链等机械论法律责任原理的同时，引入了以下几个方面的制度创新：首先是风险概念的抽象化，严格责任适用范围从物理危险物（比如第 1237 条民用核设施）扩展至社会性风险（第 1254 条高空抛物）；其次是责任主体的社会角色化，动物管理人、建筑物管理人等非直接控制者被纳入责任网络；再次是突破了损害救济的矫正正义框架，通过引入惩罚性赔偿等措施来标记和阻遏对社会有害的特定行为，比如第 1185 条对故意侵害知识产权的惩罚性赔偿。这种“形式机械论，实质社会化”的特质，揭示出中国民法典在工具理性框架内嵌入了风险分配正义的现代性诉求。这些特质都为迈向应对或然性技术的法律创新准备了初始条件。

如上所述，霍布斯所阐发的“制造者知识”理论，强调工具使用者可通过分解重组机制完全掌握技术系统的因果链条。这种认识论预设支撑着现代法律对“合理注意义务”的界定，要求行为人在自身认知疆域内预见并防范风险。但人工智能的涌现特性使工具首次获得了超越人类理解能力的认知维度：AlphaGo 在围棋博弈中创造的不遵循传统定式的新策略，既非对棋谱的机械复现，亦非对人类棋手的经验模仿，而是深度学习模型在蒙特卡洛树搜索中形成的策略空间涌现。[⑧]当工具开始具备创造性的认知能力时，使用者便陷入双重异化：既是技术红利的受益者，又是系统风险的最终承担者。这种困境在医疗人工智能领域尤为显著，Watson 肿瘤系统的诊疗建议常包含超出临床指南的激进方案，但医生既无法追溯决策逻辑，亦难以用既有医学知识验证其合理性，[⑨]传统医疗过失责任中的“理性医生标准”在此沦为无本之木。

法律系统对技术变革的回应迟滞，折射出规范科学与社会事实之间的深刻张力。莱布尼茨在《单子论》中构想的预定和谐宇宙观，在法律领域具象化为“全知立法者”的神话——相信通过精密的概念演绎便可预见并规制所有社会关系。但生成式人工智能的或然性本质，使得任何试图将概率空间离散化为规范条文的努力都注定沦为西西弗斯式的徒劳。适应于确定性工具时代的静态规范体系已无法有效解决或然性工具时代的问题，目前已经初现端倪的新模式是动态合规体系，通过算法透明度分级与“人在回路”的机制设计，在技术赋能与控制效能之间寻求平衡。[⑩]动态合规体系作为应对人工智能技术不确定性的制度响应，其核心在于通过立法授权建立持续更新的适应性规则框架，要求高风险人工智能系统开发者实施全生命周期风险管理流程，包括部署前基于标准化测试场景的合规验证、运行中实时监控及异常行为自动上报机制，并配套设立跨部门监管平台对系统迭代进行动态评估。该体系允许地方政府在中央统一安全底线约束下开展差异化场景试点，将企业实践反馈转化为规则优化依据，形成“测试—反馈—规则迭代”的闭环机制，从而在保障技术创新的同时控制系统性风险。

生成式人工智能作为或然性工具的技术特性

类比是法律思维中最重要的一种，法律人习惯于借助类比将过往情境中的规范性因素迁

移至当下和未来的事实情境，哪怕社会事实已经发生了翻天覆地的改变。这种思维惯性在人工智能时代面临严峻挑战。以北京互联网法院审理的“AI 文生图案”[11]为例，法官将 Stable Diffusion 等生成式人工智能类比于照相机、画笔等传统工具，认为用户通过输入提示词和参数设定“进行了智力投入”，进而认定人工智能生成内容构成著作权法意义上的作品。这一判决延续了“创作工具论”的法律逻辑，却忽视了生成式人工智能与传统工具的本质差异——前者并非确定性工具，而是一种全新的或然性工具。因此，尽管法官承认“技术的发展过程，就是把人的工作逐渐外包给机器的过程”，“技术越发展，工具越智能，人的投入就越少”，但仍然认为“人工智能生成图片，只要能体现出人的独创性智力投入，就应当被认定为作品，受到著作权法保护”。

确定性工具的特征在于输入与输出的线性因果关系完全可控。以照相机为例，摄影师对拍摄对象、角度、光线等要素的选择直接决定了影像内容，不同设备在相同操作下产出高度一致的结果。而生成式人工智能的本质是基于概率论框架构建的非线性技术系统，其核心特征表现为三重或然性。其一，输入输出间存在概率性关联。用户输入的提示词仅能划定内容的大致方向（如“中国风少女”），却无法决定具体的表达性要素（如人物姿态、构图细节）。正如“重现测试”所示，相同提示词、参数和模型在不同硬件配置下生成截然不同的图像，证明输出结果主要取决于人工智能内部的黑箱运算而非用户指令。其二，操作过程具有涌现性与部分不可溯源性。人工智能通过海量数据训练形成万亿级参数矩阵，其决策逻辑是分布式权重计算的结果，而非预设规则的机械执行。这种涌现性使得系统行为难以预测，且错误输出（如“幻觉”）无法通过传统审计路径溯源修正。其三，结果生成呈现非必然性与多元可能性。同一提示词可能生成无限变体，用户无法像控制相机快门般精准实现“所见即所得”，只能被动接受系统概率性输出的诸多可能之一。

生成式人工智能的或然性特质不仅解构了工具行为的因果链条，更在技术输入端制造了前所未有的规范性裂隙。新加坡国立大学的西蒙·柴斯特曼教授在《好模型借，伟大模型偷》一文中指出，当前人工智能训练数据的获取实质上是“系统性知识劫掠”（systematic knowledge plunder），其核心矛盾在于创造性破坏的双重性——既通过数据融合催生技术革新，又侵蚀了知识产权制度激励人类创新的伦理根基。这种劫掠并非偶然失范，而是内生于概率驱动技术范式的必然：大语言模型必须吞噬海量文本（如 Books3 数据集盗用 7 万部作品）才能构建统计关联网络，而传统“先授权后使用”的版权规则在此遭遇规模性失灵。更严峻的是，技术红利与法律保障的张力呈现马太效应：当 Adobe Firefly 等“合规模型”因采购授权数据而成本激增时，黑箱模型却通过盗用数据获得竞争优势，迫使创作者陷入“反公地悲剧”——其个体维权收益远低于诉讼成本。这种困境揭示了或然性工具时代的根本悖论：法律若僵守机械论时代的授权逻辑，将扼杀人工智能创新；但放任数据掠夺，则会瓦解知识经济的正义基石。破局之道在于构建双轨制数据治理框架：对事实性数据适用宽泛的文本与数据挖掘例外（如欧盟 DSM 指令第 3 条），而对创造性表达则建立延伸性集体许可机制，通过著作权集体管理组织向人工智能开发者征收法定许可费（参照音乐产业的 Content ID 模式）。唯有通过此类制度创新，方能在技术野性与创作者尊严之间维系动态平衡。[12]

这种或然性彻底颠覆了传统法律规制的底层逻辑。确定性工具时代建立的权责分配机制，预设了工具完全可控且责任可归因于使用者单一主体的前提。然而当工具本身具备非确定性、涌现

性和部分自主性时，沿用“画笔—画家”的类比不仅牵强，更可能掩盖技术引发的系统性风险。法律若继续将人工智能简单归类为“新型创作工具”，无异于用马车时代的交通法规管理自动驾驶汽车——既无法解释为何相同“创作行为”（输入相同提示词）产生截然不同的“作品”，也难以回应算法偏见、训练数据侵权等衍生问题。当工具从被动执行者转变为主动参与者时，法律必须重新审视“人—工具—结果”三元关系的根本性重构，这正是或然性工具对现代法律制度提出的深层挑战。

（一）系统性幻觉的内生机理：自回归架构与概率抽样本质

生成式人工智能的系统性幻觉根植于其技术架构的机理，这种内生性特征对传统法学归责范式构成了根本性挑战。自笛卡尔机械论世界观确立以来，工具始终被视为因果链条的确定性中介，霍布斯将心脏喻为发条、将神经视作游丝的机械隐喻，正映射着确定性工具观对法律归责体系的塑造：设计者意图与使用者操作通过线性逻辑传导至结果，责任归属如同钟表齿轮啮合般严丝合缝。然而，大语言模型的自回归架构彻底颠覆了这一认知框架，其本质是通过注意力机制捕捉海量语料库中的统计关联，在概率抽样过程中重组知识要素。凯恩斯在《论概率》中早已预见，世界的本质或然性被启蒙理性强行塑造为确定性表象。他从主客观两个维度以及两者之间的互动关系来定义概率（本文所称的“或然性”）：“在最根本的意义上，我认为它（概率）指的是两组命题之间的逻辑关系……从这种意义上衍生出另一种意义……当‘或然’一词被用来指代因知晓某些关于基本逻辑意义上概率关系存在的二级命题而产生的理性信念的程度。此外……我们还可以将‘或然’一词用来指称作为可能理性信念对象的命题，该命题与构成证据的命题之间具有相应的概率关系。”[13]生成式人工智能以技术的方式验证了这种洞见——它打破了休谟“太阳明天将升起”的经验预期，使每个推论都成为基于证据权重的概率判断。问题在于，当工具系统本身嵌入或然性基因，“一本正经地胡说八道”便不再是可修复的漏洞，而是技术本质的必然显现。我们习惯于将人工智能幻觉归咎于训练数据质量或算法优化不足，却不愿承认系统幻觉（system hallucination）正是或然性工具的认知范式特征。

面对系统性幻觉带来的归责真空，[14]法学理论亟需超越机械论范式的路径依赖。凯恩斯在重构布尔逻辑体系时提出的非数值概率理论，为理解生成式人工智能的或然性特征提供了认识论启示：法律不应强求算法输出具备传统工具意义上的逻辑必然性，而应通过动态合规框架容纳概率空间的认知不确定性。这要求我们重新审视“制造者知识”背景下的机械论，在神经网络时代构建基于过程导向的新型归责模型——将算法透明度、人类中心验证机制与职业伦理再造相结合，在技术创新与权利保护之间建立动态平衡。当技术本体突破确定性边界，法律唯有在谦抑中重构认知框架，方能在数字时代延续其规范生命力。

就生成式人工智能在法律工作中的应用而言，可以采用两个要素来定义幻觉。其一是准确性（correctness），指生成内容在事实层面的正确性与相关性，要求回答不仅符合法律事实（如正确描述判例中的判决理由或法律条文内容），而且与用户查询内容直接相关。如果回答包括错误陈述，例如虚构判决理由或法律条文，则被判定为不准确，此种生成内容被称为“事实错误型幻觉”。其二是有依据性（groundedness），这对于法律职业有特殊的重要性，强调生成内容必须严格基于权威法律渊源并正确引用，要求提供的法律主张需有明确且匹配的引证支持（如准确援引判例、法规或文献）。若回答虽正确但所引用的来源无关、错误或虚构（例如引用不存在的判例内容），则被视为“依据错误型幻觉”。不过，有学者认为还需要引入第三个要素来评估生成式

人工智能工具对法律工作的有用性，即完整回应性（complete responsiveness），指回答需全面覆盖用户提出的核心诉求，避免遗漏关键信息或提供无关内容。若系统拒绝回答、回避问题或仅提供部分有效信息，则被归类为“不完整回应”，这不属于幻觉，但同样影响生成式人工智能在法律工作中的有用性。[15]

认识到幻觉的成因和类型，便可以有针对性地建立分层治理机制。在操作层面，可采用“事实—依据”二元校验框架：前者要求生成内容符合客观事实（如法律条文内容、判例核心要旨），后者强调结论必须严格锚定权威法律渊源。例如，当人工智能分析合同效力时，不仅需正确表述《民法典》第 144 条关于无效民事法律行为的规定，更需准确援引最高人民法院相关司法解释的具体条款。这种双重校验机制虽无法根除幻觉，但能显著降低“虚构判例”或“错误归因”的风险。在技术手段上，事实吻合和依据锚定都可以通过训练阶段约束机制（包括知识蒸馏和基于人类反馈的强化学习 RLHF）、推理过程控制技术（包括检索增强生成 RAG 和置信度阈值控制）以及输出层验证系统（包括元数据溯源机制和对抗性压力测试）来实现。更深层的解决方案在于重构人机协作模式——将人工智能定位为“初步法律意见生成器”，而专业人士则承担“终局验证者”角色。法律场景中人工智能生成的类案分析、条款解读等，必须经过法律人的批判性审查。这种“生成—验证”循环不仅符合“人在回路”的伦理要求，更能通过反馈机制训练人工智能提升输出精度。例如，法官利用人工智能辅助撰写裁判文书时，可通过标记幻觉内容、补充正确法律依据等反馈行为，逐步优化系统在特定法律领域的表现。

（二）设计者意图的穿透困境：参数空间的黑箱效应

法律体系对工具责任的认定，长久以来建立在一个看似稳固的基石之上：设计者意图的可穿透性与使用者意志的可追溯性。无论是罗马法中对“形式因”的强调，还是现代产品责任法对“设计缺陷”与“制造瑕疵”的二分，其底层逻辑均预设了工具行为与人类意志之间存在一条清晰、可逆的因果链条。设计者的理性构想——无论是钟表匠对机芯的精密设计，还是汽车工程师对转向系统的安全预设——被认为能够通过物理结构或预设规则完整地投射至工具的功能边界，并最终决定其行为后果。使用者则被视为这一意图链条末端的执行者或偏差制造者，其责任源于对可控工具的误操作或对清晰指令的偏离。这种“设计者可控—使用者可责”的归责框架，构成了确定性工具时代法律规制的核心范式。

然而，生成式人工智能的崛起，以其独特的技术本体论彻底动摇了这一范式的根基。其核心运作机制并非基于笛卡尔式机械论世界观下的线性因果推演，而是依托深度神经网络在高维参数空间中的概率采样与分布式表征。当开发者将初始算法架构投入数据的洪流进行训练时，数万亿参数在反向传播过程中形成的复杂拓扑网络，已然超越了传统法律认知中的“设计意图”所能涵盖的范畴。设计者的伦理预设或功能目标——例如“不得虚构法律条文”或“确保算法公平”——如同投向深潭的石子，在参数空间的湍流中迅速消弭其初始的指向性。这种消弭并非源于设计者的疏忽或恶意，而是内生于技术本身的特性：神经网络的隐层空间阻断了意图的完整渗透。

现代法律体系在此遭遇的根本性挑战，在于神经网络的隐式知识表征与传统因果追溯路径的断裂。以 Transformer 架构中的多头注意力机制为例，每个决策节点的激活状态都受数十亿参数的动态加权影响，其决策逻辑是海量数据统计关联的涌现结果，而非预设符号规则的机械执行。即便开发者设定了明确的初始约束，模型仍可能因训练语料中的统计偏差、潜在空间中的语义纠

缠或实时交互中的随机变量，在特定语境下生成符合形式理性但实质错误或有害的输出。这种现象在技术层面被称为“分布偏移下的涌现特性”，在法学维度则表现为设计者意图与系统行为的实质性脱嵌。我国《个人信息保护法》第 24 条对算法自动化决策公平性的规制困境，恰是这种断裂的典型例证：当算法通过潜在空间的非线性组合生成信用评分时，即便开发者遵循“公平性”的伦理承诺并进行了初始的偏见检测，隐藏在 Embedding 层中的文化偏见或社会经济刻板印象，仍可能通过余弦相似度的数学运算或上下文关联的概率抽样，无声无息地渗透至输出结果，导致系统性歧视。法律试图追溯这种歧视的源头至“设计缺陷”，却发现其无法精确定位于某个可识别的代码片段或预设规则，而是弥散于整个参数矩阵的动态交互之中。

更深层的法理悖论源自技术本体论层面的认知革命。传统工具的责任归属遵循亚里士多德“形式因主导质料因”的范式，设计者的形式构想被认为完全支配着工具的物质形态及其功能边界。但生成式人工智能的创造性输出，本质上是海量训练数据在降维映射过程中形成的统计重构，设计者的初始代码仅构成可能性空间的边界条件（如模型架构、损失函数、初始参数范围），而非确定性法则。正如强化学习中的策略梯度优化会不断突破初始奖励函数的约束范围，大语言模型在人类反馈微调（RLHF）过程中形成的价值取向，已然混杂着用户行为数据、平台运营目标与社会反馈信号的复杂博弈。这种动态演化特性使得“设计者责任”的概念变得支离破碎且难以锚定。当系统通过持续学习（如在线微调或模型版本迭代）获得超越初始训练集的推理能力时，法律既无法将责任固定于某个特定版本模型的开发者（因为后续演化已非其可控），也难以在时间维度上划定可归责行为的清晰起点。例如，一个初始设计用于法律检索的模型，可能在用户交互中逐渐习得并强化了某种特定的司法偏见，而这种偏见的形成路径在参数空间中呈弥散式分布，难以逆向追踪至某个具体的训练批次或微调指令。

这种“穿透困境”在技术架构层面表现得尤为尖锐。以稀疏激活的专家混合架构（Sparse Mixture of Experts, SMoE）为例，其通过路由机制动态选择部分专家子网络处理输入，仅激活总参数的一小部分（如总参数 70B 中仅激活约 13B）。这种设计虽能显著提升计算效率，却进一步加剧了决策过程的“黑箱”效应：最终输出的内容依赖于路由机制对专家组合的瞬时选择，而该选择本身又是高维参数空间中概率采样的结果。设计者意图不仅无法穿透神经网络的隐层空间，更在动态路由的随机性中被进一步稀释。使用者输入的指令（提示词）则如同试图在湍流中投下航标，虽能划定大致的航向（如生成“中国风少女”图像），却难以规训模型内部复杂的概率采样过程，无法精准控制具体的表达性要素（如人物姿态、服饰细节、构图风格），更无法阻止系统性幻觉（如虚构的判例或法律条文）的随机涌现。

面对参数空间的黑箱效应及其引发的归责困境，法律体系亟需摆脱对机械论范式的路径依赖，重构适应或然性工具本质的责任框架。这并非要彻底抛弃“设计过错追责”的传统机制，而是需要为其注入动态化、过程化的监管维度，构建一种复合责任框架。

1. 参数轨迹存证与连续审计机制。法律可强制要求高风险人工智能系统的开发者与部署者建立完整的参数演化轨迹记录系统。这不仅包括初始训练日志和最终模型参数，更应涵盖关键训练节点（如预训练完成、微调阶段、重要版本更新）的权重快照、训练数据批次信息及使用的超参数。通过将神经网络的权重更新纳入可审计范围，为事后归责提供可追溯的数据基础。同时，引入独立的第三方机构进行连续审计，评估训练过程中的数据合规性、算法偏差变化及潜在风险累积。[16]

2. 涌现行为的概率风险评估体系。借鉴复杂系统理论中的涌现预测模型，构建针对算法行为的动态风险评估体系。这包括开发基于对抗样本的压力测试工具，模拟极端或罕见输入条件下模型的输出稳定性；利用不同随机种子下的输出方差分析，识别模型的系统性偏差或不确定性阈值；建立算法行为的“热力图”，可视化其决策敏感区域与潜在脆弱点。此类评估应贯穿模型的全生命周期，从设计、训练、部署到持续运营阶段。[17]

3. 司法证明环节的对抗性解释技术。在诉讼或监管调查中，突破对“算法可解释性”的绝对化追求，转而采用务实的技术手段辅助司法证明。例如，通过“对比解释法”（如 LIME、SHAP 等局部解释技术），展示在相同输入下，不同决策路径（对应参数空间的不同区域）如何导致差异化的输出结果，以此揭示模型决策的随机性或系统性偏差；利用“反事实生成技术”，模拟若输入数据的某些特征（如性别、种族）发生变化，输出结果将如何改变，为歧视性决策提供间接证据；探索“算法行为克隆”技术，训练简化但可解释的代理模型（如决策树）来近似模拟黑箱模型在特定任务上的行为模式，为法官和陪审团提供直观的认知接口。

4. 结果导向向过程控制的范式转变。法律规制的焦点需从静态的“设计者意志”转向动态的“参数演化路径”。这意味着监管标准不应仅关注最终输出的合规性（如是否产生歧视性结果），更应关注训练数据治理的规范性（如数据代表性、偏见清洗）、算法开发流程的透明度（如文档完整性、版本控制）、持续监控的有效性（如漂移检测、实时干预机制）以及反馈回路的健全性（如用户申诉渠道、错误修正机制）。通过设定过程性合规义务，引导开发者与运营者在技术全链条中嵌入风险防控措施，而非仅寄望于对最终输出的结果审查。

唯有将法律的目光从对“黑箱”内部的徒劳凝视，转向对其输入输出动态、演化路径及社会效应的系统性监控与调节，方能在技术自主性与人类价值观之间维系脆弱的规范性平衡。这要求法律科学自身完成从机械决定论向概率主义范式的认知跃迁，承认不确定性并非需被消除的技术缺陷，而是或然性工具的内在属性需被容纳与管理。在此过程中，传统的“设计者—使用者”二元归责框架将被重塑为涵盖设计、开发、训练、部署、监控、反馈等多主体的动态责任网络，其核心在于确保技术野性在制度理性的引导下服务于人类福祉的创造性释放而非系统性失控。

面向或然性工具的法律范式创新

传统法律归责体系根植于机械论世界观预设的确定性框架，将工具视为线性因果链中可完全解析与掌控的客体。这种“可控即可责”的归责逻辑在确定性工具时代具有自洽性，例如《民法典》第 1165 条确立的过错责任原则，预设行为后果可追溯至特定主体的可预见性控制范围。然而，生成式人工智能的或然性特质彻底颠覆了这一认知基础。仍以北京互联网法院审理的“AI 文生图案”为例，原告通过输入包含 24 项正向指令及 129 项反向约束的提示词生成图片，法院虽认定该图片构成著作权法意义上的作品并归属于用户，但其论证逻辑暴露了传统归责范式的结构性危机：相同的提示词在不同硬件环境下生成构图差异显著的图像，印证了 Transformer 架构的多头注意力机制将提示词解构为高维向量后，输出结果由参数空间中的概率分布主导，而非用户指令的线性映射。这种技术特性使得传统法教义学中的“可预见性”要件失去实践支点——正如自动驾驶系统因传感器融合误差导致事故的情形所示，设计者的安全冗余设计与使用者的合规操作均

无法完全规避算法黑箱中的随机性风险。

（一）技术本体论变革对法律归责体系的冲击

在或然性工具的冲击下，传统法律体系面临的根本性挑战在于其认识论基础的动摇。确定性工具时代预设的“设计者可控—使用者可责”的线性模型，建立在笛卡尔机械论世界观与霍布斯“制造者知识”理论之上。这种模型要求工具行为完全遵循设计者意图，且使用者能够通过分解重组机制理解技术系统的因果链条。然而，生成式人工智能的自回归架构与概率抽样本质彻底颠覆了这一认知框架——神经网络的万亿级参数空间阻断了设计者意图的完整渗透，强化学习的动态调优消解了使用指令的确定性约束，其结果生成呈现非必然性与多元可能性。当 AlphaGo 在围棋博弈中创造出超越人类棋谱定式的策略，当 Watson 肿瘤系统提出突破临床指南的诊疗方案时，工具首次展现出超越人类理解能力的认知维度，传统法律归责体系赖以为生的“意志—行为—责任”因果链遭遇本体论层面的消解。

这种技术本体论变革呼唤法律思维从机械决定论向概率主义范式的认知跃迁。罗马法中的追加性质的诉讼（actiones adiecticiae qualitatis）制度为此提供了历史参照。该制度通过法律拟制解决奴隶经商的责任分配问题：奴隶虽无独立法律人格，但其商业行为产生的债务由主人承担补充责任。这种设计不依赖主体拟制，而是构建“行为—风险—责任”的动态映射关系，恰与或然性工具的责任分配困境形成跨时空呼应。体现在我国《生成式人工智能服务管理暂行办法》等规范中的“人在回路”原则，正是这一法理传统的现代转译：要求高风险人工智能系统必须为人类监管预留空间，确保人类监督者具备“必要能力、培训与权力”来理解系统运作机制并保持决策自主性。这种设计不追求穿透算法黑箱，而是通过架构性约束重建人类在概率空间中的价值锚点。

人机协同的法律范式转换需重构三大核心概念。其一，法律主体性应从意志载体转向控制节点。传统法律主体理论预设单一意志中枢，但自动驾驶系统的避险决策是环境感知传感器、路径规划算法与执行器协同运作的分布式决策结果。借鉴区块链分布式账本技术理念，应将法律主体性定位为动态控制能力的制度化表征——正如《民法典》第 1249 条突破“控制力递减则责任递减”逻辑，规定遗弃动物致害时原饲养人责任不因丧失直接占有而中断，揭示了责任主体从“意志控制者”向“风险承担者”的嬗变。其二，因果关系应从逻辑必然性转向概率相关性。医疗人工智能基于贝叶斯网络给出的诊疗方案，其决策路径无法通过传统涵摄技术追溯，却可能显著提升患者生存率。这要求法律接纳“统计显著性”作为因果关系的新型认知框架，如同最高人民法院在区块链存证案中承认时间戳的盖然性证明力。其三，归责基准应从主观过错转向过程控制。深度伪造内容在社交平台的指数级传播，使 Web1.0 时代的“通知—删除”机制形同虚设。应当建立算法透明度分级制度，通过参数轨迹存证、实时风险监控与对抗性测试构成的过程合规体系，将责任焦点从静态的“设计过错”转向动态的“演化路径监督”。

更深层的困境源自或然性工具对主客体关系的认知颠覆。在医疗人工智能诊断场景中，当预训练模型与实时文献数据概率耦合生成诊疗建议时，临床医生因专业壁垒无法有效验证算法逻辑，传统医疗事故中的过错认定标准便沦为空洞的形式要件。例如某三甲医院引入的人工智能辅助诊断系统将良性肿瘤误判为恶性，但追溯责任时发现，错误既非源于初始训练数据偏误，亦非操作失误，而是模型在线学习过程中形成的策略漂移所致。这种责任断层印证了神经网络隐层激活路径的不可溯源性，其本质是设计者意图在梯度流变中被重新编码，使用者指令在概率采样中失效，

平台监管能力在数据权属模糊中虚化。现行法律体系在此遭遇三重悖论：若严格适用产品责任法追究开发者责任，将抑制技术创新必需的风险容错空间；若沿用使用者自负原则，无异于将普通个体置于无法承受的认知负荷之下；若强化平台责任，则需直面算法透明性与商业秘密保护的冲突。

（二）法律范式转型的核心路径

法律回应或然性工具挑战的核心在于重构“可控—可责”框架的认知论基础——从追求因果必然性的意志论转向容纳不确定性的控制论。具体而言，需建立动态合规体系：其一，通过算法透明度分级制度强制要求高风险场景提供可验证决策路径，[18]如《生成式人工智能服务管理暂行办法》第 19 条要求提供者披露训练数据来源及算法原理概要；其二，构建“人在回路”机制保障人类终极决策权，[19]如自动驾驶系统须设置实时人工接管接口；其三，推行对抗性测试常态化，例如法律人工智能系统需内置模拟恶意提示词的压力测试模块，通过定向攻击检测系统抗干扰能力。这种过程导向的治理模式，本质是将法律规范转化为算法空间的约束性参数，使法理逻辑通过人机协同迭代内化为技术架构的组成部分。[20]

更深层的制度创新在于责任配置的梯度化重构。传统侵权法以牛顿力学式的线性因果关系为归责基础，但生成式人工智能的决策过程遵循统计力学式的概率云分布。以金融 DeFi 协议因预言机数据偏差引发崩盘为例，开发者、矿工与用户行为均符合行业惯例，但损害仍不可避免。此类案例要求法律从结果回溯转向风险前瞻，建立弹性责任框架：一方面，借鉴《个人信息保护法》第 69 条的过错推定规则，对算法幻觉导致的损害实行举证责任倒置；另一方面，构建“行为贡献度图谱”，根据主体对风险的控制力分配责任。[21]例如，在自动驾驶事故中，若系统自动化程度达 L4 级且制造商已披露隐层激活路径的置信区间，可适度减轻人类驾驶员的注意义务；反之，若算法黑箱未通过国家网信部门的安全评估，制造商应承担严格责任。这种分层归责机制既避免“全有或全无”的裁判困境，又能激励技术开发者提升系统的可解释性。

（三）人机协同范式的制度实践与未来调适

我国智慧法院建设已展现人机协同范式的实践雏形。上海法院“206 系统”通过刑事证据标准库引导法官办案，其实质是将法律规范转化为机器可处理的特征向量；杭州互联网法院的异步审理模式，通过离散化时间节点重组诉讼流程，突破传统诉讼的共时性约束。这些实践揭示了法律演化的新路径：规则不再依赖先验的教义学建构，而是通过司法场景中的持续人机互动动态生成。正如强化学习中的策略梯度优化会不断突破初始奖励函数约束，大语言模型在人类反馈微调（RLHF）过程中形成的价值取向，已然混杂着用户行为数据与平台目标的复杂博弈。这种动态演化特性要求法律放弃追求终极正解的确定性幻想，转而构建容纳多元可能性的或然性框架。

未来法律体系的调适应聚焦三重维度。在规范层面，建立“简单规则—行为激励—多元协同”的适应性法治框架，参照《个人信息保护法》的“告知—同意”基础规则与场景化例外并行的弹性结构，对算法偏见等新型风险设置底线禁令，同时授权地方政府开展差异化监管沙盒试点。在技术层面，研发法律增强现实（Legal AR）系统，通过可解释性人工智能技术生成判决书的法律推理链，将隐层激活路径转化为可视化的论证图谱，使系统性幻觉在对抗性解释中显形。在伦理层面，推动法律人从“权威解释者”转型为“算法决策评估师”，培育概率敏感性思维。

或然性工具的时代悖论在于：技术越是逼近人类智能，法律越需通过“人类中心验证机制”

守护其规范性根基。当大语言模型生成具有表面合理性的法律论证，当扩散模型创造出无法追溯创作意图的视听内容时，唯有重构“生成—验证”的人机协作闭环，才能在技术创新与风险防控之间维系动态平衡。古罗马法谚曾言：“法律乃善良与公正之术。”在概率主义框架下，这一古老箴言的新生在于：公正是概率分布中的价值锚定，善良是随机变量间的人文关怀。

最终，法律范式的创新需回归认识论层面的革命。[22]传统法律训练强调形式逻辑的严谨性，却缺乏处理不确定性输出的认知框架。最高人民法院推行的“AI 辅助裁判能力认证”试点项目，其核心考核指标并非操作熟练度，而是法官对模型置信度参数的批判性解读能力。这要求法律共同体培育概率敏感性思维——既不能全盘接受陷入“自动化偏见”，亦不应简单拒斥技术赋能。

结论：在科技创新与制度弹性之间

在传统工具理性的疆域内，法律通过行为可预测性与结果可归因性的双重锁定构建起稳定的责任框架。一本关于侵权法的经典著作精准地表述了这种线性关系维度上的确定性思维：“支付损害赔偿的责任首先取决于与 x 和 y 相关的事实情况 F1 的存在，其次要求无免责事由（F2）适用，最终这种状态必须是损失的因果关系结果。”[23]哪怕是注意到风险这种带有不确定属性的因素，传统法律思维仍然希望通过个人为结果承担责任的方式将其纳入确定性范畴：“侵权法中的风险分配是具体的分配决策，涉及加害者与受害者之间损失的矫正，因而与个体结果责任（individueller Ergebnisverantwortung zusammenhängt）相关。”[24]然而生成式人工智能的或然性特质，使得技术行为的因果链不再是线性展开的必然性序列，而是概率云中动态漂移的粒子轨迹。工具理性批判的当代使命，在于超越机械论时代“全知设计者—被动执行者”的认知范式，直面算法黑箱中涌现的认知盲区与价值裂隙。法律亟需发展出适应量子化责任形态的弹性框架：既非简单移植产品责任中的设计缺陷标准，亦非套用高度自主系统的严格责任原则，而是通过动态风险分配机制将技术透明性义务、过程审计标准与结果矫正功能整合为连续链条。当系统性幻觉成为技术内生的创造性代价，归责逻辑应从结果回溯转向风险前瞻，在概率权重与价值排序的动态平衡中重塑问责路径。

人工智能深度嵌入人类决策系统，催生出人机混合行动者这一新型规范实体。传统法律主体理论建立在生物人与拟制人格的二元分立之上，难以解释提示词工程师与语言模型协同生成决策的复合行为结构。破解这种困局需要构建法律主体性的拓扑模型：通过行为贡献度图谱将人类意图输入、算法参数调校与系统自主响应解析为多维度的责任向量，在梯度化归责框架中实现控制力与预见性的动态匹配。[25]这要求突破既有法律中雇主责任、产品责任与平台责任的模块式划分，发展出穿透技术层级的穿透性义务体系。当自动驾驶系统在强化学习中形成超出训练集的应急策略时，责任分配不应固守“人类最后决策权”的形式教条，而应根据系统透明度水平与风险控制能力配置差异化的注意义务。[26]

概率主义时代的自由意志尊严，体现为人类在技术赋能与价值守护之间的审慎平衡。生成式人工智能既可能通过认知外包削弱人类的反思能力，也可能借助知识增强拓展理性的疆界。理解生成式人工智能的或然性特质并基于这种理解重塑规则的意义在于培育人机协同中的批判性自觉：通过算法可解释性强制规范维持必要的认知主权，借助人类反馈强化学习机制构筑价值校

准通道，运用对抗性训练模型防范系统性偏见固化。[27]这要求法律跳出“控制”与“放任”的二元对立，转而构建激励相容的治理生态——当大语言模型生成虚构内容时，制度回应不应止于错误纠正，而需推动法律知识库和法律人工智能的协同进化，使技术幻觉转化为法律范式更新的催化剂。唯有在接纳不确定性的认知革命中，法律才能完成从秩序维护者到文明塑造者的范式跃迁。

注释：

①⑦ 参见Marcus P. Adams, *Hobbes's Two Sciences: Politics, Geometry, and the Structure of Philosophy,* Oxford: Oxford University Press, 2025。

② 是指一组由裁判官创设的特殊诉讼程序，旨在解决奴隶或家子等处于他人权力支配下的人从事商业活动时的责任分配问题。其核心逻辑是：通过法律拟制，让主人或家父对奴隶或家子的商业债务承担补充责任，从而在奴隶/家子无独立法律人格的前提下实现交易安全。

③ Klaus Heine and Alberto Quintavalla, “Bridging the Accountability Gap of Artificial Intelligence — What Can Be Learned From Roman Law?” *Legal Studies*, vol. 44, no. 1, 2024, pp.65-80.

④ F. W. Maitland, “Moral Personality and Legal Personality,” in *State, Trust and Corporation*, edited by David Runciman and Magnus Ryan, Cambridge: Cambridge University Press, 2003, pp.62-74.

⑤ 笛卡尔：《谈谈方法》，王太庆译，北京：商务印书馆，2000年，第44页。

⑥ 托马斯·霍布斯：《利维坦》，黎思复、黎廷弼译，北京：商务印书馆，2009年，第1页。

⑧ David Silver, et al., “Mastering the Game of Go with Deep Neural Networks and Tree Search,” *Nature*, vol. 529, no. 7587, 2016, pp. 484-489.

⑨ Eric Topol, *Deep Medicine : How Artificial Intelligence Can Make Healthcare Human Again*, New York: Basic Books, 2019, chapter 3.

⑩ 郑戈：《人工智能伦理的机制设计》，《中国法律评论》2025年第1期。

⑪ 参见北京互联网法院（2023）京0491民初11279号民事判决书。

⑫ Simon Chesterman, “Good Models Borrow, Great Models Steal: Intellectual Property Rights and Generative AI,” *Policy and Society*, vol.44, no.1, 2025, pp.23-37.

⑬ John Maynard Keynes, *A Treatise on Probability*, Cambridge: Cambridge University Press, 2013, pp.11-12.

⑭ 参见邓建鹏、赵治松：《生成式人工智能的法律规制：以侵权法为视角》，《东北师大学报》（哲学社会科学版）2025年第3期；丁晓东：《论算法的法律规制》，《中国社会科学》2020年第12期。

⑮ Varun Magesh, et al., “Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools,” *Journal of Empirical Legal Studies*, 2025, pp.216-242.

⑯ 参见 Jonas Schuett, “Risk Management in the Artificial Intelligence Act,” *European Journal of Risk Regulation*, vol.15, Special Issue 2: Special Issue on the Future of Food Law, 2024, pp.367-385.

⑰ 张恩典：《算法影响评估制度的反思与建构》，《电子政务》2021年第11期。

⑱ 参见刘辉、雷崎山：《生成式人工智能的数据风险及其法律规制》，《重庆邮电大学学报》（社会科学版）2024年第4期。

⑲㉗ 参见郑戈：《人工智能与法律的未来再思考》，《数字法治》2023年第3期。

⑳ 参见张凌寒：《人工智能法律治理的路径拓展》，《中国社会科学》2025年第1期。

㉑ 梁远高：《论人工智能大模型训练数据风险的分层规制》，《郑州大学学报》（哲学社会科学版）2025年第3期。

㉒ 参见余成峰：《法律人工智能新范式：封闭与开放的二元兼容》，《中外法学》2024年第3期。

㉓㉔ Nils Jansen, *Die Struktur des Haftungsrechts: Geschichte, Theorie und Dogmatik außervertraglicher Ansprüche auf Schadensersatz*. Band 76, Jus Privatum, Tübingen: J.C.B. Mohr（Paul Siebeck）, 2003, s.66, s.107.

㉕ 参见陆小华、陆赛赛：《论生成式人工智能侵权的责任主体——以集体主义为视角》，《南昌大学学报》（人文社会科学版）2024年第1期；陈晨：《人工智能侵权中的责任主体及归责原则》，《中外法学》2025年第2期。

㉖ 参见唐林垚：《人工智能时代的算法规制：责任分层与义务合规》，《现代法学》2020年第1期。

编辑　**杜运泉**