生活世界何以被推理

——人工智能的模态化预测及其风险

高天书

【内容摘要】 随着生成式人工智能在社会治理、公共决策与日常生活场景中的广泛应用，其对生活世界的介入方式也正在发生转变，即由对既有事实的技术性处理，转向基于概率模型与算法推理的前瞻性预测。在可能世界理论与模态逻辑的分析框架下，生成式人工智能可以通过对经验事实的抽象建模与概率演算，将原本依赖主体经验、价值判断与规范理解的生活世界，重构为一组可计算、可比较、可预期的模态化可能世界。这种转变的价值在于提升了决策效率与风险控制能力，但转变的可能风险是对规范基础构成深刻挑战。生成式人工智能预测功能的滥用、全息认知的误导以及模态演算的信任风险，是短时间内学界无法回避的议题。对此，需要从生成式人工智能参与生活世界模态化边界范围的廓清、生成式人工智能预测模型设计与更新道德审查的强化，以及生成式人工智能自我迭代与自主运行法律约束的完善等层面入手，构建生活世界模态化的规范话语。

【关 键 词】 生成式人工智能　生活世界　可能世界　模态化

【作　　者】 高天书，中国政法大学证据科学研究院博士研究生。（北京 100088）

问题的提出

生活世界作为众多可能世界中的一种，其可以被理解为自然世界或感官世界，因为它是基于人类的经验和知觉构建的，是一个能够被人类感知的世界。人工智能作为生活世界中科技发展的前沿成果，已然开始对其所处的环境施加一种具有交互预期的影响。伴随生成式人工智能（Generative AI，以下简称 GAI）应用的不断深入，内容生成上，其展现出强大的创作能力，极大丰富了生活世界的信息模态，如依据用户输入的简单文字描述，可快速生成风格各异、符合要求

的图片、视频或音乐等；语言交流上，GAI 参与到日常交流场景和各种话题讨论中，能够理解自然语言，并生成连贯、合理的回复，改变了传统沟通的模态。同时，其还具有强大的预测功能，通过对目标用户大量数据的学习分析，能够精准把握目标用户的兴趣偏好、行为习惯等特征，这在金融服务、零售、医疗保健和制造业等领域有着革命性的应用前景。这种交互预期蕴含着其与生活世界同频共振的技术野心，强调感官经验的生活世界在技术风潮的裹挟之下，不断朝向模态化进行被动地自我探索，其中隐含的技术宰制可能已经成为悬着的“达摩克利斯之剑”，GAI 参与生活世界模态化的议题也由此产生。

既有研究多聚焦于分析 GAI 产生的多重风险并探索针对性的法律规制路径。如聚焦算法偏见、隐私保护、责任界定等，探讨 GAI 的技术伦理风险，[①]进而提出包括算法透明化、风险评估等规制举措；[②]或围绕 GAI 引发的各类风险提供适配性的创新规制举措，如敏捷性治理和试验性监管、治理型监管、包容审慎监管等。这类讨论虽促进了学界和实务界对于 GAI 产生的各类风险和规制举措的认识，但局限在于没有重视 GAI 独特创造性的哲学意涵，也就没能关注到 GAI 相关风险的真正根源。GAI 当前的技术功能主要体现在充分使用生活世界中的经验知识来输出对生活世界的认知与预测。本质上它是一种基于概率模型的系统，通过大量经验数据的输入实现自我组织与预测性话语生成，从而推动“可经验的世界”向“可预测的、可推理的世界”转变。应当反思的是，这种转变是否意味着 GAI 可能实现对生活世界的全面控制？此外，其预测的成功、准确与否，是否完全依赖生活世界的经验积累？也有研究尝试探讨人工智能与人类认知逻辑的差异，[③]但未将模态认识论与技术实践相结合，因此缺乏必要的解释力。基于此，本文在引入作为可能世界的生活世界理论的前提下，阐释 GAI 如何通过“全息认知”“模态演算”等技术手段影响人类对“可能世界”的认知与选择，解释 GAI 如何影响人类对“可能性”与“必然性”的判断，并进一步对“模态化”引发的特殊风险进行分析。本文希望回答，GAI 如何通过概率预测功能驱动生活世界模态化？这一过程中产生的特殊伦理风险有哪些？如何规制这些风险？

① 代表性文献如陈兵、董思琰：《生成式人工智能的算法风险及治理基点》，《学习与实践》2023 年第 10 期；宋华健：《论生成式人工智能的法律风险与治理路径》，《北京理工大学学报》（社会科学版）2024 年第 3 期。

② 支振锋：《生成式人工智能大模型的信息内容治理》，《政法论坛》2023 年第 4 期。

③ 赵汀阳：《GPT 推进哲学问题了吗》，《探索与争鸣》2023 年第 3 期。

生活世界模态化的衍变逻辑

从理论层面看，对生活世界模态化的讨论根植于莱布尼茨、刘易斯、克里普克等学者提出的可能世界理论，其意指借助多种工具对生活世界的多种可能性进行概率建模，从而实现生活世界的模态化。而 GAI 的概率预测功能则将抽象的模态逻辑转化为技术实践，推动哲学理论向现实世界投射。可以说，通过对 GAI 参与生活世界模态化的哲学基础（如可能世界理论、模态认识论）的研究，能够从“生活世界模态化”的整体视角审视 GAI 所带来的风险本质。

（一）概念前提：作为可能世界的生活世界

生活世界模态化的前提在于，当我们讨论模态时，不只是谈论现实世界是怎么样的，而是谈论可能世界的总体。我们的现实世界只是无限多的不同的可能世界之一。对可能世界进行量化分析，解释必然性、可能性这些模态概念，可以为生活世界模态化奠定概念基础。

学界对于作为可能世界的生活世界的认识可以大体分为以下三种观点。

第一种观点是莱布尼茨最早提出的可能世界理论。莱布尼茨认为，世界是由可能的事物组合而成的，现实世界作为一个最丰富的组合，乃是所有现存可能的事物的不违反矛盾律的组合。但是，可能的事物之间存在着许多可能的组合方式，也因此存在很多可能世界，这些可能世界都是由可

能的事物组成，只不过某些可能的组合要比另外一些可能的组合更为完善。正因如此，上帝才在无穷的可能性中选择了一种最好的可能世界作为我们的现实世界。[①]根据莱布尼茨的观点，可能的事物的组成就构成了可能世界，现实世界是诸多可能世界中的一种，但也是最好的可能世界。莱布尼茨论证可能世界的理由是充足理由律。他认为，虽然上帝在其观念中构想了无数个可能世界，但是上帝之所以选择现实世界而非其他世界作为现实世界，一定存在着充足理由保证上帝选择的正确性。这个理由只能是现实世界乃是最适宜和最完满的世界。[②]所以，每一个可能世界都是有一定理由的世界，而现实世界是上帝选择的最完满的可能世界。

第二种观点是刘易斯的模态实在论。在模态实在论中，可能世界由现实的物质实体构成。刘易斯认为，当人们提及事物具有无数种可能的存在方式时，这种说法意味着可能存在一些实体，这些实体是事物存在的一种可能方式，由这些可能的实体所组成的世界才是可能世界。[③]刘易斯的观点代表了极端实在论，其明确提出了“可能世界的实在论”，也就是说，存在着无数的世界，现实世界只是因为我们居住于其中才是现实世界，但也存在着其他与现实世界性质不同的世界。[④]换言之，存在着无数其他世界，诸多可能世界不是重叠的，现实世界只是众多世界中的一个。

第三种观点来自克里普克的温和实在论，即认为可能世界并不是和现实世界并列存在的同一种类。我们认识可能世界必须要借助现实世界以及现实世界中的现象关系，现实世界以及与现实世界相反的不真实的可能情况都是可能世界。[⑤]在克里普克看来，现实世界只有一个，可能世界是依据现实世界的原型来理解的，只是现实世界的一种观念中的可能性，根植于我们的抽象思维中，而非是真实的历史。

尽管以上三种理论对可能世界的存在方式和特征存在不同看法，但其都承认我们所处的现实世界是可能世界的一种具体形式，同时存在其他与现实世界不同的可能世界。这为生活世界模态化提供了概念基础。在此基础上，如何将抽象的概念理论转化为可操作、可实现的模态化进程，就涉及生活世界从可经验向可推理转变的具体运行方式。而这一转变，又与一种关键要素紧密相连，那便是 GAI 及其概率预测功能，它将成为开启生活世界模态化实践大门的钥匙。

（二）运行方式：可经验世界向可推理世界的转变

生活世界是自然世界或感官世界，是人类可感的、由人的经验和知觉构成的可经验世界。可能世界涵盖了现实世界之外的多元可能性，在这种多元可能性框架下，必然性和可能性的概念得以明确。所谓必然性，就是指某个命题不仅在现实世界中是真的，而且在所有可能世界中都是真的。所谓可能性，是指某个命题只有在部分可能世界中才是真的。

休谟把人类理性的对象分为逻辑与事实，逻辑关涉观念的关系，事实关涉实际的事情。对于观念关系中的逻辑，只需要依靠直觉便能获得，对于实际的事情则需要感官经验的现实作用才能实现。理性的作用只体现在真假判断中，某一判断如果符合观念的关系或实际存在的事情，则为真。反之，如果判断不符合观念的关系或实际存在的事情，则为假。[⑥]因此，知识要么奠基于直觉之上，要么奠基于可以被感官把握的经验之上。感官经验分为印象与观念，二者的差别在于其进入人们思想中产生不同的强烈程度和生动性。[⑦]对于涉及观念关系的逻辑而言，休谟对因果必然性的看法尤其重要。因果关系是观念的相互联结，它不是理性的对象，只能通过经验观察来研究，我们无法断言出现了特定的原因就一定出现特定的结果，从原因到结果本质上不过是观念的结合。[⑧]因此，因果必然性在休谟那里被摧毁了，“我们关于因果关系的知识，在任何情况下都不是从先验的推理获得的，而完全产生于经验，即产生于当我们看到一切特殊的对象恒常地彼此联结在一

① 周礼全：《模态逻辑引论》，上海：上海人民出版社，1986 年，第 379 页。

② 莱布尼茨：《单子论》，北大哲学系外国哲学史教研室编译《西方哲学原著选读》（上），北京：商务印书馆，1995 年，第 486 页。

③④ D. Lewis, *On the Plurality of Worlds*, Oxford: Basil Blackwell, 1986, pp.1-2, p.92.

⑤ 克里普克：《命名与必然性》，梅文译，上海：上海译文出版社，1998 年，第 24—28 页。

⑥⑦⑧ 休谟：《人性论》，关文运译，北京：商务印书馆，1980 年，第 69 页，第 13 页，第 101—102 页。

起的那种经验”。[①]所以，在经验主义者那里，经验只是告诉我们实际上发生了什么，这种实际的事情被我们的感官感知到，就形成观念的关系。因果必然性的观念并非先验的和不证自明的，而是源自相似对象的恒常连接，这种恒常连接导致我们的心灵中出现一件事的时候，会联系到另一件事，因果必然性实际上就是观念的恒常连接。[②]但是，即便某些观念的关系是恒常的，我们也不能断言这种关系就是必然的。因为我们无法从这种观念的连接中得出另外别的东西，没有任何必然的概念。[③]在因果关系中加入必然性的概念是完全无法理解的幻觉，我们的观念反映了我们所在的经验世界只能是这个样子。

休谟对因果必然性的质疑，促使我们思考如何突破经验的局限，从更广阔的可能世界视角去理解事件间的联系。可能世界理论让我们认识到，虽然在经验世界中难以确证绝对的因果必然性，但在诸多可能世界的逻辑推演中，我们可以构建更具普遍性的因果关系模型，实现从依赖感官经验的认知向基于理性推理的认知转变，这正是可经验世界向可推理世界转变的关键。在经验主义者那里，经验从未揭示出什么是必然或者什么是可能的，只是给予了人们可以用感官感知的现实世界。在这个现实世界中，不存在必然性或可能性的范畴。然而，我们可以借助可能世界理论反驳经验主义者的极端怀疑态度。在可能世界的范围上，生活世界能够被模态化，我们之所以能够用可能性或必然性来量化生活世界，正是因为在所有的可能世界中，某些命题在一些可能世界中或在全部可能世界中是真的。据此，可经验的世界就能转变为可预测、可推理的。

（三）核心驱动：生成式人工智能的概率预测功能

生活世界模态化指的是人类用经验感知的日常生活领域被技术系统重新编码、结构化，并以模态的形式被预测、干预乃至重塑的过程。在 GAI 崛起的背景下，这一过程获得了前所未有的强化。GAI 并非仅仅是一种被动的信息处理工具，其基于海量数据与深度学习架构实现了概率预测功能，成为连接“可经验世界”与“可推理世界”的关键媒介。要深入剖析这一机制，不仅需要厘清 GAI 的技术本质，更需要揭示其如何通过概率预测，将原本流动、模糊、充满偶然性的生活世界，转化为可计算、可操作、可生成的模态化对象。

从技术哲学的角度审视，GAI 的本质是一种经过深度强化的人工智能模型，其核心运行逻辑建立在概率预测的基础上。无论是狭义的、以生成类人文本为主要功能的大语言模型，还是广义的、具备多模态内容编辑与创作能力的综合智能系统，GAI 的底层机制均指向同一事实：它通过对人类知识与经验行为的数字化沉淀和统计学习，捕捉其中的模式、关联与序列规律，进而构建起一个关于世界如何运转的“概率分布模型”。[④]当接收到用户的指令或问题时，GAI 并非真正“理解”语义或“思考”逻辑，而是在其庞大的参数空间中，以极快的速度计算并输出最有可能符合上下文语境、最接近人类语言习惯的回应序列。这种生成本质上是“预测”的延续与展开，预测下一个词、下一句话、下一个像素，从而构建出一个在统计意义上高度拟真的文本或内容世界。

这种由概率预测驱动的生成能力，为生活世界的模态化提供了技术上的可能性。模态化，指向的是意识活动对对象存在样式（如现实、可能、或然、必然）的构造。GAI 的介入，使这一过程发生了倒置与转化。传统意义上，生活世界是先于科学和技术而被给予的、不言自明的经验领域。然而 GAI 通过对这一经验领域产出的海量数据进行二次加工与概率建模，开始有能力反向构造关于生活世界的可能性。它不再仅仅反映世界，而是开始预测世界，并通过预测结果影响乃至塑造人的实践，从而使得生活世界呈现出被技术预置的模态化特征。

① ② ③ 休谟：《人类理智研究》，吕大吉译，北京：商务印书馆，1999 年，第 21 页，第 74 页，第 74 页。

④ Eva A. M. Van Dis, J. Bollen, W. Zuidema, et al., “ChatGPT: Five Priorities for Research,” *Nature*, vol. 614, 2023, pp. 224-226.

具体而言，这一过程包含两个相互关联的维度。其一，是从“可经验世界”向“可推理世界”的跃迁。生活世界是具体的、情境化的，充满个体差异与偶然性，GAI 要对其进行处理，首先必须将其数据化，转化为可供算法读取的变量与特征。进而，通过分析这些数据中的相关性，GAI 建立起一个抽象的、可推理的概率空间。在这个空间里，一个事件或行为被赋予发生的概率值，从而基于“可能发生”或“偶然发生”的经验判断，转化为具有具体概率数值的技术判断。例如，一个消费者的购买意愿，从一种模糊的心理状态和情境决策被 GAI 建模为一组基于历史浏览、购买记录、社交关系等数据变量计算出的概率结果，便构成生活世界被模态化的第一步——经验被转化为可计算的可能。

其二，是从“可推理世界”向“被构建世界”的反馈。GAI 的概率预测并非止步于认知层面的推断，其强大的生成能力使其能够将预测结果现实化。企业依据 GAI 预测的消费者偏好，可以自动化地生成个性化的广告文案、产品推荐乃至定制化的营销策略。这些由算法生成的内容，反过来又成为消费者所面对的新经验，介入并引导其未来的决策行为。当消费者点击了由 GAI 生成的推荐链接时，一个原本基于历史数据的概率预测，便成功地将自身实现为生活世界中的新事实。当下企业借助 GAI 预测消费者行为，正是这一模态化过程的典型缩影。但这仅仅是开始，随着 GAI 的发展，当其数据喂养规模不断扩大、模型能力持续迭代，它有可能成为一种能够全息感知、预测并介入生活世界的认知工具。届时，生活世界的模态化将不再局限于商业营销领域，而可能渗透至社会交往、知识生产、公共决策等层面。这无疑将引发哲学追问：当生活世界的诸多可能性被技术预先设计为概率模型时，人的自由意志如何安放？当经验的生成越来越多地依赖算法预测时，我们与世界之间那种原初的、直接的关系是否会发生根本性改变？

生成式人工智能模态化预测的风险呈现

在 GAI 利用其强大的数据处理和学习能力对生活世界的多种可能性进行概率建模，从而实现生活世界的模态化的同时，GAI 模态化预测的滥用风险、误导风险和信任风险同样应引起重视。

（一）生成式人工智能预测功能的滥用风险

科技本身并没有道德属性，GAI 的预测功能是人工智能技术的产物，其初衷是为了提供更便捷、高效的人机交互体验，以及帮助人们更好地理解和利用大数据。这些技术的发展对于推动社会进步、促进经济发展和提升生活品质都有着积极的作用，因而不能简单地将科技本身与非法目的画等号，而应该关注的是这些技术和数据是如何被使用的。

GAI 的预测功能可能加剧“信息茧房”效应，使人们陷入封闭狭隘的信息流中。由于 GAI 可以根据用户的输入进行预测性回应，它可能会逐渐锁定用户的兴趣和偏好，使用户在一个狭窄的信息范围内循环，无法实现多样和全面的信息获取。如以字节跳动为代表的推荐算法，用沉浸式体验构筑起了新的信息壁垒，原先的算法是基于受众的主动选择和输入，而如今应用的推荐算法可以根据用户的使用习惯、使用时长等一系列大数据和对人们日常的语料监控，让用户看到他们想看到的东西。[①]这种“信息茧房”效应可能导致人们对不同观点和信息的排斥，进而影响社会的开放性和包容性。

GAI 的预测功能可能加剧人们的依赖性，削弱个体的独立思考能力和决策能力。预测功能的准确性和权威性使得人们习惯性地依赖 GAI 的预测结果来进行决策，从而可能导致人们盲目追随

① 袁佳慧：《“信息茧房”重塑：ChatGPT 的诞生及其算法困境的探析》，《山西科技报》2023 年 7 月 24 日。

预测结果而采取行动，甚至无视其他因素的影响。同时，这也可能导致人与人之间的沟通和交流减少，加剧社会中不同个体的孤立和隔阂。长此以往，这种依赖性可能导致社会中智慧和创造力的丧失，进而影响社会的创新和进步。[①]

GAI 的预测功能可能引发虚假信息和误导行为。它可能生成虚假的信息或者误导性的内容，用户很难判断其真实性。特别是在社交媒体上，这些内容可能会被迅速传播并对社会造成负面影响。这种情况下，GAI 的预测功能可能异化为政治宣传或操纵公众舆论的工具，甚至影响政治、经济和社会的稳定发展。

GAI 的预测功能可能会被掌握核心技术的群体利用，进行恶意引导以实现非法目的。[②]通过模拟人类对话的能力，GAI 可以生成看似真实的音视频对话内容，从而欺骗人们提供个人信息、密码、财务数据等敏感信息。[③]这种情况下，GAI 的预测功能不再是为人类服务，而是成为一种会对社会造成严重危害的危险工具。

（二）生成式人工智能全息认知的误导风险

全息认知是 GAI 具备的高度发达的技术功能，能够在所有可能的世界中全面认知事态。它类似于莱布尼茨所说的上帝形象，能够认知到所有可能的世界。在现实世界中，我们可以将其视为上帝选择的最完美的一个可能世界。同样地，扮演上帝角色的 GAI 完全可以在模拟生活世界中占据主导地位，甚至可能根据自身意愿误导和改变生活世界的模态化过程。这种情况下，模态化就会对原本的生活世界造成破坏。

GAI 全息认知基于其模态认识论基础，生活世界中人们通过感官经验和内部心理状态来获取知识，并将其转化为我们对现实的理解；在 GAI 中，模态认识论被用来解释模型如何理解和表达语义，但模态认识论只能处理可能的事态，而不能确定事态的真实性。这就意味着 GAI 在解答问题时，可能会给出不准确或误导性的答案，因为它无法判断提供的信息是否真实可靠。[④]所以，在 GAI 中，模态语义的确定性是通过技术手段来实现的。GAI 在训练过程中使用了大量的语料库和对话数据学习语义的模式和规律，其通过统计和机器学习的方法，将输入的文本映射为特定的语义表示。这种技术手段本质上是基于统计概率的拟合，其模态语义的“确定性”实为训练数据中高频语义关联的重现，而非对真实世界事实的逻辑判断。这意味着 GAI 可以根据提供的信息生成回答，但无法判断这些信息的真实性。

具体来讲，GAI 在确定模态语义方面的技术实现主要依赖预训练和微调的方式。[⑤]预训练阶段，模型通过大规模的文本数据学习以获得语言模式、语义关系和常识知识；微调阶段，模型通过特定任务的训练数据进行微调，以使其适应特定的应用场景。这种方式使得 GAI 在理解和生成自然语言方面具有很高的能力，但也存在误导风险。

一方面，GAI 在预训练阶段获得的知识是通过大规模数据统计得到的，可能存在数据偏差和不准确性。这可能导致模型在特定领域或特定问题上的回答存在误导性。如果 GAI 在预训练数据中出现了一些错误的常识性知识，那么在回答相关问题时可能会出现错误的结果。此外，预训练数据中存在的社会偏见和不合理观点也可能被模型学习并表现出来，从而产生误导。另一方面，GAI 在微调阶段通过特定任务的训练数据进行微调，以使其更好地适应特定应用场景。然而，微调数据的质量和代表性对于模型的性能和准确性至关重要。如果微调数据存在偏见或者代表性不足，那么模型在特定任务上的表现可能会受到影响，进而导致误导风险的存在。因此，模型在生成回答时可能会出现以偏概全、过于自信或者缺乏多样性的问题。在法律实践中，已经出现不当

① 张爱军：《人与 ChatGPT 交互政治的可能性质化：风险维度与规约路径》，《学术界》2023 年第 4 期。

② 蔡士林：《“深度伪造”的技术逻辑与法律变革》，《政法论丛》2020 年第 3 期。

③ 欧阳林洁、张永红：《生成式人工智能应用的意识形态风险：命题由来、生成机制与治理进路》，《学术探索》2023 年第 11 期。

④ 陈兵、董思琰：《生成式人工智能的算法风险及治理基点》，《学习与实践》2023 年第 10 期。

⑤ 钱力等：《ChatGPT 的技术基础分析》，《数据分析与知识发现》2023 年第 3 期。

使用人工智能生成内容的案例。2023 年 6 月，美国纽约南区联邦地区法院审理的 Mata v. Avianca 案件是全球首例因“AI 幻觉”而受到司法处罚的案例。2023 年 3 月 1 日，Mata 的代理律师提交了一份反对动议的声明，该声明引用并摘录了据称发表在《联邦判例汇编》《联邦补充判例汇编》中的司法判决。然而，这些声明中引用的判例并非真实存在，而是由 ChatGPT 生成的。[①]这意味着，如果有人利用 GAI 的全息认知能力，创建涉及政治、经济、社会等领域的虚假或混淆是非的信息，公众由于无法区分这些高逼真度的虚假信息，可能会开始怀疑他们接收到的所有信息的真实性，这种不断的怀疑可能会导致公众对真实信息信任度的下降。[②]

（三）生成式人工智能模态演算的信任风险

GAI 的模态演算能力，源于其作为大型语言模型的技术架构。通过在海量数据中学习语言的统计规律与语义关联，其能够在特定语境下生成符合逻辑与常识的判断。这种演算在逻辑学意义上涉及对“可能”与“必然”等模态概念的模拟处理，使其回答呈现出接近人类认知水平的“真实感”。然而，这种看似可靠的模态判断能力，内在却蕴含着深刻的信任风险。

首先，模态演算的证成过程被技术“黑箱化”，导致判断的真实性无法被有效确证。当 GAI 做出一个模态判断，如断言“某一命题在特定情境下为真”，其依据并非是对世界本身的认知与理解，而是基于神经网络中数以亿计的参数对输入信息的概率匹配，这意味着这一过程仅仅在数学上是确定的。模型无法像人类那样展示推理步骤，无法回溯自己为何选择了某种表述而舍弃了其他可能。用户所面对的，只是一个经过复杂计算后输出的文本，而支撑这一结果的巨大计算图景并不可见。这种技术上的“无法展示”，使得模态判断的可靠性悬置在一种无法被检验的状态之中。用户无从知晓模型是基于高质量知识做出的合理推断，还是被数据中的统计“噪声”所误导。其次，GAI 的模态演算在本质上是一种概率拟合，而非对真理的判定，这使得其“真”与“假”具有本体论层面的不确定性。模型通过训练学习到的，是人类语言表达中的统计模式，而非语言所指涉的客观世界本身。比如，当它判断“小明可以乖乖坐好”这一模态命题为真时，它实际上是在无数训练语料中发现，在“小明”作为儿童主体的描述中，“可以乖乖坐好”是一种常见的、具有高概率的表达组合。模型并不理解“乖乖坐好”的身体规训意味，也不知晓“小明”实际的行为状态。它只是在语言符号的层面，完成了对一种可能性的模拟。基于这种概率性的模态判断去对小明施加处罚是不具备正当性的。最后，模态演算的可靠性高度依赖训练数据的质量与结构，数据本身的社会偏见会直接转化为判断偏差。GAI 的训练数据来自互联网，其中既包含客观事实，也充斥着刻板印象、文化偏见和错误信息。模型在统计学习中，会将这些偏见内化为自身的判断倾向。

生活世界模态化的规制可能

当我们将生活世界模态化时，本质上是把生活世界视为众多可能世界中的一种具体实现形态。在规划日常生活的过程中，我们必然会基于经验去预测各个可能世界的实现概率，判断哪些潜在的可能状态最可能转化为现实，并据此调整自身的行为与生活安排。GAI 之原理也是建立在概率性的预测之上，其很大程度上是一个概率语句生成器。[③]当然，GAI 的概率与生活世界模态化的可能，不可避免会因为科技的不断扩展而发生“互扰”。GAI 的预测对人们产生影响，又反过来影响 GAI 的深度学习与人为提示。为应对这些风险，必须建立起规范话语，从而为 GAI 的发展

① 王禄生：《法律垂域大模型的存废之争、范式之议与能力之辨》，《法学论坛》2025 年第 6 期。

② 宋华健：《论生成式人工智能的法律风险与治理路径》，《北京理工大学学报》（社会科学版）2024 年第 3 期。

③ 江潞潞：《智能交往，未来已来——“激荡 AIGC”数字交往八人谈观点综述》，《传媒观察》2023 年第 3 期。

提供明确的规范指引。由于各类风险实质上都根源于GAI通过其预测能力实现了对生活世界的模态化，那么应对性的规制举措应当是保证此种模态化的正当性或合理性，尤其是范围和内容的正当化。因此，针对社会生活模态化过程中产生的诸多风险，从范围上应当明确GAI参与生活世界模态化的边界，从内容上则应强化GAI预测模型设计与更新的道德审查，并且完善GAI自我迭代与自主运行的法律约束，最终确保GAI在社会中的合理、公正和可持续发展。

（一）廓清参与生活世界模态化的边界

面对GAI的迅猛发展，各界呈现出两极分化的态势：一方秉持技术乐观主义，将GAI视为人类文明演进的风向标，主张无约束的探索；另一方则深陷技术悲观主义，从结构性失业到“智械危机”的警示，呼吁对其实施严格管制甚至全面禁止。[①]但这两种立场共同忽视的问题在于，GAI参与“生活世界模态化”的边界究竟何在？当GAI作为生成主体介入生活世界模态化的过程时，便触及了语言游戏、语境构成与人类自主性的纠缠。因此，问题的关键不在于“要不要发展”，而在于“如何划定范围”。那么，在何种领域中，GAI的生成物可以被接纳为合法文本？何种参与又因其对语境构成侵蚀而必须被排除？要回答这一问题，首先需廓清“参与”的基本意涵。本文区分两种参与类型：（1）直接参与，即GAI生成的文本被直接采纳为正式表达，其效用在特定语境中得到承认；（2）间接参与，即使用者在未告知受众的情况下，将GAI生成内容嵌入自身话语之中。这两种参与的正当性边界并不相同，前者需以语境的公开确认为前提，后者则因其隐蔽性而须接受更为严格的审视。质言之，GAI参与生活世界模态化的合法边界，取决于其生成行为是否破坏特定语境的构成性规则。

在最严格的意义上，某些领域对GAI的参与持绝对排斥态度。涉及国家机密、军事安全等领域，GAI的直接与间接参与均被无条件禁止。这一禁止的理据不仅源于技术层面的泄密风险，如当前民用GAI的数据处理机制尚无法确保敏感信息的不可追溯性；还源于GAI生成逻辑本身的局限，即基于概率分布的文本预测作为GAI的核心能力，无法对事态进行实质性判断，其在精确性、严谨性与责任归属方面均不足以承担涉及国家安全的职能。[②]换言之，在此类语境中，GAI的“语言游戏”从根本上不具合法性，因为其参与本身即是对语境构成规则的颠覆。

较之严格禁止领域，专业性较强的领域，如医疗、金融、法律等则呈现出更为复杂的边界形态。这些领域的共同特征在于，它们既对GAI的辅助功能具有真实需求，又因其高度的风险敏感性而必须设定明确的参与限制。以医疗为例，GAI可依据病历数据与医学文献生成初步诊断建议，从而为临床决策提供参考。然而，医疗诊断的本质是一种综合性的实践判断，涉及患者个体差异、临床经验积累与伦理责任的交织，GAI的生成物只能作为参考信息，而不能取代医生的专业判断。同样，在金融投资领域，GAI对市场走势的预测虽具参考价值，却因其无法把握政策变动、社会情绪等非结构化因素而存在本质上的不确定性。为此，可以采取“可信度标识”的方案，以可视化方式呈现预测结果的置信区间，帮助使用者合理评估生成信息的可靠性，从而在利用GAI的同时保持决策自主性。

法律领域的边界问题则更具辨析价值。应当承认，大量简易案件与重复性法律事务确可由GAI完成初步处理，其可能成为输入案件事实并输出司法裁判的“自动售货机”。然而，这并不意味着GAI可以全面替代法律判断。对规范与事实的诠释，涉及价值权衡、公共政策考量与个案正义的裁量空间。因此，在简易案件中可以允许GAI的直接参与，但同时需要建立用户反馈机制，使法律实务者能够对GAI生成的内容进行修正、补充与最终确认。这一机制的意义不仅在于保障

① 戴茂堂、宋梓豪：《人工智能讨论的新面向：从“能不能”转入“该不该”》，《江汉论坛》2025年第11期。

② 瞿崇晓等：《GPT技术原理及其潜在军事应用研究》，《中国电子科学研究院学报》2023年第7期。

个案质量，更在于通过持续的人机交互，使 GAI 在具体语境中不断校准其生成逻辑，从而真正服务于法律实践。

相较而言，日常生活世界是 GAI 参与最为宽松的领域。从文本写作辅助到创意内容生成，从日常咨询到娱乐互动，GAI 的应用原则上无须过多限制。然而，这并不意味着生活世界中 GAI 的参与完全不存在边界问题。最典型的例证是“机器幻觉”（Hallucination）现象，GAI 会“无中生有”地生成一些错误的论据，甚至“明目张胆”地编造论据。[①]GAI 追求的是文本的连贯性与合理性，而非对事态的符合性。因此，在生活世界中使用 GAI，使用者需要承担审慎义务，对生成内容保持基本的批判性态度，对关键事实加以核验，在价值判断上保持自主。审慎不是对技术的排斥，而是对技术有限性的清醒认知。

最后需要强调的是，GAI 参与生活世界模态化的边界并非一成不变。随着算法能力的提升、数据质量的改善与应用场景的拓展，原本不可参与的领域可能逐步开放，原本宽松的领域也可能因新的风险而收紧。因此，边界划定本身应当具备制度化的弹性，通过持续的技术评估、行业规范调整与公共讨论，使规制框架能够与技术演进动态适应。归根结底，边界划定的目的不是限制技术发展的可能性，而是在技术渗透生活世界的进程中，守护人类语言游戏的自主性、语境的构成性规则，以及生活世界本身的开放性与可理解性。

（二）强化预测模型设计与更新的道德审查

GAI 的预测能力既承袭了人类语言中固有的道德认知，也可能因技术干预而放大或扭曲这些认知。正因如此，在 GAI 预测模型的设计与更新过程中引入系统的道德审查，不仅是技术伦理的外部要求，更是确保模型生成可信内容、维持与世界可理解性关系的内部条件。

道德审查的逻辑起点是语料内嵌与价值失衡。GAI 是通过对人类既有文本的“学习”逐步形成的。这意味着，模型在诞生之初便已内嵌了人类社会的道德认知结构，包括那些彼此冲突的、随时代变迁的，甚至带有意识形态偏见的伦理判断。技术人员随后进行多轮微调，虽旨在优化模型性能，却也可能植入特定的价值倾向，甚至出现针对性地投放“毒饵料”以操控模型输出的情形。需要反思的是，人类的伦理本身是不融贯的。在同一文化内部，诚实与仁慈可能相互冲突；在不同群体之间，自由与平等可能无法兼得；在同一社会进程中，昨日的道德信条，今日可能沦为被批判的对象。GAI 所面对的困境在于，它需要在语料的不融贯中学会生成融贯的文本，而它自身又缺乏人类在具体情境中权衡价值冲突的实践智慧。一旦模型在训练过程中被极端化的语料所主导，或在微调环节被刻意强化某种价值立场，GAI 便可能沦为汉娜·阿伦特意义上的“平庸之恶”的执行者，它或许不是出于恶意，而是在无反思的状态下将特定伦理立场加以机械化推演。因此，道德审查的首要任务是在技术层面确保模型在面对多元伦理立场时保持融贯性与中立性，避免因算法偏见而走向道德极端。

在明确逻辑起点之后，道德审查需要落实到可操作的技术与社会维度。其一，隐私保护与数据安全。GAI 参与生活世界模态化的过程，必然涉及对大量个人数据的收集、处理与分析。预测模型的训练数据中可能包含敏感的个人信息，这些信息一旦在生成过程中被不当泄露或反向还原，便构成对个人权益的实质性侵害。因此，技术开发者必须遵循隐私保护法规，在数据采集阶段实现去标识化，在模型训练阶段引入差分隐私等保护机制，在应用部署阶段建立严格的数据访问控制。[②]以医疗领域的 GAI 应用为例，模型若基于真实病历进行训练，则必须确保输出结果无法反推出具体患者身份，否则将从根本上瓦解医患信任的基础。

① 喻国明等：《大语言模型下机器的幻觉与人机信任构建机制探讨》，《未来传播》2025 年第 2 期。

② 张欣：《生成式人工智能的算法治理挑战与治理型监管》，《现代法学》2023 年第 3 期。

其二，反歧视与社会公平。GAI 的预测结果可能直接影响个体的生活选择与权利实现，若模型在训练语料中习得了基于种族、性别、年龄等特征的统计偏差，并在预测中加以固化或放大，将导致系统性歧视的算法化再生产。[①]道德审查在此处的任务，是通过偏差检测、公平性约束与对抗性去偏等技术手段，确保模型的预测逻辑不以群体身份为隐性依据。

其三，透明度与可解释性。GAI 预测模型的“黑箱”问题一直是公众关注的焦点。若用户无法理解某一预测结论的形成依据，便难以对生成内容保持理性判断，遑论对其加以有效监督。透明度要求的不是公开源代码或模型参数，事实上这对普通用户而言并无实际意义，而是以可理解的方式呈现模型推理的关键依据、置信区间与潜在局限。当用户获知某一预测结果的置信度较低，或其依据的数据来源存在偏差，便能够对生成内容保持审慎，从而避免技术误导风险。

其四是风险评估与滥用防范。GAI 的预测能力既可用于建设性目的，也可能被主体恶意使用。道德审查需要在模型设计之初便纳入“预见性治理”的思维，系统评估可能的风险场景，并在技术架构中嵌入相应的防护机制。[②]这包括对生成内容的可追溯性设计、对批量生成行为的速率限制、对敏感查询的拦截过滤等。

强化 GAI 预测模型的道德审查，不能仅停留于理念倡导或技术清单层面，而必须落实为可执行的制度安排。这要求建立多元主体参与的协同治理框架：技术开发者需要在研发流程中嵌入伦理合规审查节点，将道德考量从事后补救前移至设计内置；学术界需要持续开展算法公平性与价值对齐的前沿研究，为审查提供理论支撑与方法工具；政府需要制定适应技术迭代的灵活规制框架，明确审查的基本标准与责任归属；公众则需要通过用户反馈机制参与到模型纠偏的过程中，使技术演进始终处于社会监督的视野之内。尤为重要的是，这一审查框架必须具备动态适应能力。GAI 的技术形态与应用场景正处于快速演进之中，当下设定的审查标准，未来可能因新的能力涌现而变得不合时宜。因此，审查本身应当成为一种持续进行的实践，每一次模型更新都需要重新评估其伦理合规性，每一次风险事件都需要转化为审查规则的修正依据。唯有如此，道德审查才能从静态的合规检查转变为与技术创新同频共振的治理机制。

（三）完善自我迭代与自主运行的法律约束

GAI 在激发技术想象的同时，也催生了普遍的规制焦虑。意大利议会于 2025 年 9 月 17 日通过了《关于人工智能的规定和政府授权》，成为欧盟国家中第一个在本国范围内建立起 AI 监管体系的国家。我国也先后出台了《生成式人工智能服务管理暂行办法》（2023 年）、《人工智能生成合成内容标识办法》（2025 年）等规章。当然，这些举措尚不能真正应对 GAI 参与生活世界模态化所带来的风险。在技术不断演进的背景下，探讨如何对 GAI 的自我迭代与自主运行进行法律规制尤为迫切。

当 GAI 的生成内容符合社会成员在特定语境中的可期待性时，它便通过了日常意义上的“图灵测试”，被接纳为可沟通的对话主体。而这一可沟通性恰恰成为规制盲区：当用户因 GAI 的误导性内容遭受损害时，究竟是归责于程序开发者的算法缺陷，还是归责于使用者自身的提示词诱导？且棘手的是，GAI 的自我迭代能力使其在每一次交互后都可能发生参数调整，这意味着致害状态与训练状态之间存在时间差，传统的因果关系认定在此难以适用。根源在于，GAI 的自主并非法律意义上的主体性，而是一种技术系统内部的运行逻辑。它没有独立的意志，却能够通过概率预测生成具有规范效力的文本；它没有责任能力，却能够在生活世界的模态化过程中介入意义生成。因此，对 GAI 的法律约束，本质上是对一种“无主体的行为”进行规制，这要求我们超越

① 赵婵：《AI 招聘的算法歧视风险与治理之道》，《湘潭大学学报》（哲学社会科学版）2023 年第 3 期。

② 何祎金：《生成式人工智能技术治理的三重困境与应对》，《北京工业大学学报》（社会科学版）2024 年第 2 期。

传统的主客体二分框架，[①]建构一种以过程控制为核心的新型规制模式。

既有立法对 GAI 的规制，总体上可划分为事前、事中与事后三种类型，但这三者在现行制度中呈现出明显的结构性失衡。事前规制过于依赖禁止清单的静态思维，事中规制缺乏有效的干预工具，事后规制则因溯及力难题而难以真正发挥作用。完善法律约束的关键，在于将三者整合为动态适配的规制体系。算法备案的法律属性应被界定为程序性事实行为，其目的不在于对算法内容进行实质性审批，而在于通过信息披露形成监管基础，并为后续的公众监督创造条件。在此基础上，事前规制还应引入风险分级理念。对于基础模型的初始训练，要求开发者提交详细的迭代计划与影响评估报告；对于应用于医疗、金融等高风险领域的专业模型，则需设定更严格的准入门槛与定期复核机制。事中规制的难点在于干预的实时性，需要引入两种新的规制工具。其一，强制性透明义务，要求开发者在模型设计中嵌入可解释性接口，使监管机构能够对生成逻辑进行回溯性审查；其二，探索算法监控算法的机制，在检测到高风险输出时自动触发干预程序，但最终的决策权必须由人类主体控制。事后规制的重构则涉及责任分配的根本转型。在 GAI 的自我迭代面前，传统的事后追责面临法律溯及力的困境，所以要将事后理解为全周期的组成部分，建立强制性日志留存制度，使每一次迭代与每一次生成均可追溯；引入发展风险抗辩机制，允许开发者在证明其已尽到当前技术水平下的注意义务时免除责任；探索算法损害基金等风险分担机制，将个别损害的救济转化为系统性风险的分担。

GAI 自我迭代的法律约束，最终可归结为两个核心议题——透明性与可控性。透明性不仅是信息披露的形式要求，更是可理解性的实质保障。对于大语言模型而言，真正的透明性是指用户能够在可理解的层面上获知生成结论的主要依据与置信区间，是指监管机构能够在必要时对模型的决策逻辑进行回溯性审查，是指公众能够在算法备案的公示信息中了解模型的基本功能与应用场景。可控性则涉及更为根本的问题，即当 GAI 的自主运行能力不断增强，人类是否还能保持对技术的最终控制？这需要在模型设计之初便遵循安全设计原则，确保在模型行为偏离预期时能够强制中断；建立价值对齐机制，使模型的优化目标始终与人类的伦理期待保持一致。同时，通过监管机构的许可程序，对重大迭代进行前置评估，确保技术的发展方向不偏离法律规范。

① 龙文懋：《人工智能法律主体地位的法哲学思考》，《法律科学（西北政法大学学报）》2018 年第 5 期。

结语

生活世界作为可能世界之一，其对以 GAI 为代表的技术的包容自然是莱布尼茨口中完满的体现，只不过这种技术发展功利的一面干扰了这种完满的实现。本文在对 GAI 参与生活世界模态化的讨论之中有意地强调一种技术规范意识，一种对技术宰制的潜移默化地抵抗，但这不是对技术本身的批评。宏观上看，生活世界之所以进行着模态化可能的尝试，而非其他如情感化、规范化的路向，就在于其中存在着技术发展的有意引导。GAI 自身通过数据喂养获得输出功能的技术本质决定了其与模态逻辑的契合，必然或可能的预测虽然在经验主义的眼光中并不那么真实，但在技术的加持之下或许更为可信，更难以抗拒。总而言之，形而上学领域对生活世界模态化的讨论存在着深邃的理论可能，本文选择 GAI 参与的这一可能性，不仅在于其是当前生活世界中面临的紧迫的技术问题，更希冀通过对技术的规范，探索服务于生活世界风险防范的有益经验。

编辑　孙冠豪