论生成式人工智能“致幻”的法律边界

杨智

贵州财经大学，贵州贵阳，550025；

摘要：生成式人工智能的致幻现象已对现行法律秩序提出了尖锐的挑战。本文试图跳出将“致幻”简单归类为产品瑕疵或技术故障的传统思维，转而探讨其引发的深层法律问题与伦理张力。面对此种“非人类中心主义”的信息生产模式，既有的归责框架逐渐显露出窘迫。与其急切地寻觅一个全知全能的监管主体，不如回归至具体法律关系之中，审慎地厘定开发者、部署者与用户之间因“致幻”信息流动而产生的权利义务关系。在言论自由的保障与虚假信息的规制之间，在技术创新激励与公共安全维护之间，法律需要一场耐心且富有远见的规范对话与司法续造。

关键词：生成式人工智能；法律责任；算法

DOI：10.69979/3029-2700.26.04.083

1 生成式人工智能致幻现象的规范内涵

设想一位年轻律师面对信托纠纷，求助于法律大模型，模型给出援引详实的回答。而直至查阅原文，他才发现案例与阐释皆为算法编织的幻觉。这便是生成式人工智能“致幻”的典型图景。技术社群倾向于将其描述为模型在概率空间中的一次“游走失误”[1]。但模型并不“理解”法律，它只是在海量的语料训练后学会了人类关于法律的言说方式。当它被问及一个信息盲区时，便会忠实地模仿权威的论述口吻，用虚构来弥合知识上的裂隙。

将此种现象仅仅视为有待技术迭代去修正的“噪声”，恐怕是一种法律上的天真。技术噪声可以随算法精进而减弱，但它所引发的社会关系变动与信任危机，却已实实在在地构成了法律必须回应的规范性问题。当一台机器能够以极高效率生产出披着权威外衣的虚假陈述时，我们面对的就不再是一个单纯的产品质量问题，而是一种全新的信息风险样态[2]。既不同于传统媒体的不实报道，也不同于个人的主观诽谤，它是一种植根于技术架构本身的“系统性失真”。

这种系统性失真触碰到了法律的敏感神经。在这个依赖信息准确性的社会里，算法“致幻”产生的破坏不容小觑。如果这种“致幻”发生在医疗诊断中，其连锁反应更难以估量。而损害发生时，该向谁追责？是信任机器的当事人，是开发者，还是算法本身[3]？

2 生成式人工智能致害的责任分配重构

人工智能的制造者很轻易就能把因果关系作为突破口，免于风险责任的承担。而私法与公法中的行政法摆脱这一困境的方法是绕过因果关系，选择使用“过错推定”或者“无过错责任”为归责原则。但即便转向适用无过错责任框架下的产品责任制度，依然会有新障碍。面对生成式人工智能致幻带来的损害，法律传统往往向既有的归责体系寻求答案，即产品责任。但当我们真的试图将生成式人工智能装入“产品”的模具时，一种格格不入感油然而生[4]。正是如此，对传统路径进行系统反思，进而重构更为合理的归责制度体系确有必要。

传统的产品责任锚定“缺陷”，但生成式人工智能作为一种前沿技术形态，本质是算法自主性发展到高级阶段的集中体现。其归责需要从很多方面考量。其一，幻觉的产生具有概率性和情境依赖性。同一个问题在不同的提问方式下，可能会得到截然相反的答案[5]。这就导致人类很难为模型的“真实度”设定一个刚性的技术标准。其二，生成式人工智能并非一个完成时态的“产品”，它更像一个持续演进且更新的“过程”。用户输入的提示词和模型后续的微调，都在共同塑造着最终的输出结果。在这样的动态链条中，要界定一个“初始的”缺陷，并将其归责于开发者，难度可想而知。

在人机对话场景中，模型所输出的内容—是否应当被界定为一种“数字建言”[6]？循此逻辑，审慎核实义务是否亦应更多地配置于处在信息接收终端、并最终基于该信息采取行动的用户身上？沿着这个思路走下去，法律责任的重心便从开发者一端，部分地滑向了部署者和用户。部署者负有更为直接的警示与说明义务。一个向公众开放的法律咨询人工智能，其界面上是否以清晰、醒目的方式提示了产品存在的风险？若部署者未能尽

到此项义务，甚至有意无意地夸大模型能力、营造其“无所不知”的权威感，那在因“致幻”引发信赖损害的个案中，让其承担相应的过错责任，便有了坚实的法理基础。

至于开发者，直接让他们为每一次具体的“致幻”输出承担侵权责任，恐怕是一条死胡同。可行的路径是转向公法层面的治理义务[7]。开发者掌握着模型训练的核心数据，是防范和减轻系统性风险的第一道防线。法律可为设定注意义务的标准，让人工智能学会说“我不知道”。

3 算法输出的言论定性及其规制边界

一个更为根本的难题浮出水面：生成式人工智能生成的内容算不算一种“言论”？如果算，那么它属于谁的言论？其致幻性的输出，能否被纳入法律对“虚假言论”的规制框架[8]？

将算法生成的内容等同于人的言论，直觉上会遭遇巨大的阻力。言论自由作为基本人权，根基在于对人的理性、尊严和自治的尊重[9]。一个由概率模型驱动的人工智能产品，它没有作为权利主体的资格，无法成为言论自由的享有者。

然而人工智能的“发声”过程深刻地嵌入了人的表达之中。用户输入提示词是一个设定议题和方向的过程；模型生成回答，是一个依据算法填充内容的过程；最终用户对回答的选择传播，则是一个具有独立法律意义的行为。从这个角度看，人工智能的“言论”更像是一面扭曲的镜子，它折射的是开发者通过代码和数据注入的价值观，是部署者通过产品设计所引导的应用场景，也是用户通过提问所激发的信息需求。

如此对于“致幻”性内容的规制不能套用“事前审查”或“事后追惩”模式。如果法律要求一个提供通用对话功能的模型，必须为它所生成的每一个可能产生误导的回答承担责任，那将无异于宣判了这项技术的死刑。开发者为了避免风险，唯一理性选择就是极大地限制回答范围，使之无害。这对于言论自由所珍视的思想市场而言同样是伤害——因为一个活跃的、有时甚至会出错的人工智能，也可能扮演着“苏格拉底式助产士”角色。

挑战在于如何区分两种不同性质的“幻觉”。一种是“无害的幻想”，法律对此应保持宽容与克制。另一种则是“有害的虚假陈述”，尤其是在法律、医疗、金融等高度专业且后果严重的领域，人工智能一本正经地提供错误的法律条款、杜撰的医疗方案或虚假的财务数据[10]。这种“幻觉”直接侵蚀社会信任的基石，并对用户的财产、健康乃至自由构成现实威胁，法律必须予以明确规制。

区分的标准不在于内容本身真实性，而在于其传播场景和语境。带有批判性审视的学术运用和失察失责的职业过失理应有所区别。法律规制的着力点不应是那个作为工具本身，而应是负有专业注意义务和诚实信用义务的“人”。他们借助人工智能提高效率，但不能将自身的责任外包给一个概率模型。其伦理规范需要增加一条新的注脚：审慎核实人工智能生成信息的真实性，并对其最终的职业行为承担全部责任[11]。

4 生成式人工智能的分层治理与义务构造

致幻现象所触及的是现代社会运行所依赖的信任机制本身。信任的建立与维系需要以制度性担保为前提。患者信赖医方，系以医学教育体系与临床执业规范为基础[12]；当事人接受裁判，系以程序正义与救济渠道为保障。此种信赖关系，是相关职业共同体通过长期制度积累所形成的公共产品。

当生成式人工智能以拟人化交互方式、确定性表述形式介入上述专业领域时，其所引发的是一种信任关系的错置。该等技术产品并未取得人类专业共同体所持有的制度性授权，却在交互效果上分享了与之近似的认知权威。用户界面中“内容可能不准确”的风险提示，在格式上虽已满足告知义务之形式要求，然其警示效果受制于认知心理学所称之“晕轮效应”——模型流畅、自信的输出形式，往往削弱用户对风险提示的实际注意程度[13]。而法律所应承担的功能在于为此种新型人机交互关系设定清晰的权利义务配置规则。

就模型开发者而言，法律介入应以谦抑为原则。生成式人工智能技术仍处快速演进阶段，其技术架构与应用形态尚未定型[14]。刚性过强的强制性规范既难以与技术进步保持同步，亦可能对创新活动形成不当抑制。较为可行的治理路径应是推动行业自治与标准建设。核心要求在于，开发者应以可视化方式向部署者及公众披露模型的性能特征——包括但不限于不同应用场景下的准确率区间、幻觉发生率等关键指标。此种以透明度为核心的治理模式，具有双重功能：为下游部署者及终端用户提供决策参考依据；为因致幻输出所引发之纠纷，提供判断开发者是否已尽合理注意义务的事实基准[15]。

生成式人工智能的发展不可避免地带来了治理上的风险，就部署者一端而言。其义务构造可概括为三个

层面。首先是理应承担的是注意义务，部署者须以显著方式向用户明示该服务之生成式人工智能属性，及其内生的“致幻”风险。在法律、金融等专业应用场景中，此种警示不宜停留于首次使用前的格式性勾选同意，而应嵌入交互流程的相应节点，形成持续性的风险提示机制。这是因为格式化的单次警示难以有效对抗前述“认知晕轮效应”，用户在持续使用过程中对模型输出内容的信赖程度往往呈递增态势[16]。然后是审核义务,对于面向不特定公众开放的服务，尤其是允许用户将生成内容一键分享至公共传播渠道的应用形态，部署者应当建立与其服务类型、技术能力及风险等级相适应的内容审核机制。此项义务要求其对明显且可能引发重大法益侵害的致幻内容具备识别与处置的合理能力。审核义务的边界，应依服务场景的风险程度作差异化界定。最后一块拼图是信息可追溯性义务。部署者应确保其服务具备基础的日志记录与数据留存功能，使得因致幻输出所生之损害具备被追溯的技术可能。若因部署者未建立必要的日志机制，导致损害发生后的责任主体无法查明、因果关系链条因算法黑箱而中断，则该等情形本身即可构成对注意义务的违反。

至于用户一端，规制的重心在于认知能力的培育。在算法系统与人类判断力共存的交互格局中，算法素养的养成具有基础性意义。它指向的是对生成式人工智能能力边界的理解、对流畅性输出保持合理审慎的认知习惯，以及在涉及重大利益的决策节点，寻求人类专业判断作为最终确认的行动自觉。

5 结语

致幻现象集中呈现了吸纳颠覆性技术过程中所遭遇的法律与伦理命题。该现象并非技术过渡性瑕疵，而是技术与法律规范长期对话的起始。寄望于单行法典或专门监管机构，一劳永逸地消除致幻现象，忽视了技术演进的不确定性，难免陷于规范与现实的错位。

较为可取的法律回应路径应是正视算法的认识论局限，厘清多元主体的责任边界，并在此基础之上建构一种以责任分配为核心的协同关系框架。

在此框架内，开发者的任务在于提升模型输出的可靠性与运行机制的透明度；部署者则承担告知义务与审核职责；而用户始终保有判断权与核实义务——此项权责不可让渡。法律的功能定位将更接近于一种引导性治理：通过个案的积累逐步明晰规则，通过灵活的标准适应技术迭代，剪除那些可能侵蚀公共利益的风险，同时为创新与表达自由预留空间。

参考文献

[1]孟天广,李珍珍.治理算法：算法风险的伦理原则及其治理逻辑[J].学术论坛,2022,45(01):9-20.

[2]毕文轩.生成式人工智能的风险规制困境及其化解：以ChatGPT的规制为视角[J].比较法研究,2023,(03):155-172.

[3]张欣.生成式人工智能的算法治理挑战与治理型监管[J].现代法学,2023,45(03):108-123.

[4]吴悠然.生成式人工智能服务提供者的责任分配机制[J].学术研究,2025,(10):60-68.

[5]闫强,张倩语,魏娜.基于演化博弈的生成式人工智能幻觉应对分析[J].系统仿真学报,2026,38(02):399-415.

[6]王文玉,代金亮.生成式人工智能时代信息生态秩序的失衡风险与治理方案[J].网络安全与数据治理,2025,44(04):40-45.

[7]王旭,谢方,刘斌斌,等.生成式人工智能治理何以有效能?——基于30国政策法规的实证研究[J].图书与情报,2025,(05):47-60.

[8]罗敬蔚,李勇坚.人工智能大语言模型幻觉风险与制度因应[J].科学管理研究,2026,44(01):80-90.

[9]何雪锋.人的理性为法律立“法”——凯尔森的法律认识论及其现实意义[J].华东政法大学学报,2017,20(04):73-81.

[10]支振锋.生成式人工智能大模型的信息内容治理[J].政法论坛,2023,41(04):34-48.

[11]周学峰.生成式人工智能侵权责任探析[J].比较法研究,2023,(04):117-131.

[12]钱镇,刘俊荣.生成式人工智能辅助诊断下的知情同意挑战与应对[J].医学与哲学,2026,47(05):35-40.

[13]李智,赵雪蓉.生成式人工智能大模型的数据风险与治理机制[J/OL].上海政法学院学报(法治论丛),1-14[2026-04-12].

[14]金宇菲.生成式人工智能训练数据著作权侵权风险及应对[J].出版与印刷,2025,(03):38-49.

[15]陈嘉鑫,董紫来.信息生态理论视域下生成式人工智能虚假信息风险的防范化解[J/OL].情报理论与实践,1-11[2026-04-12].

[16]程啸.论生成式人工智能侵权责任中的因果关系[J].中国法律评论,2026,(01):73-87.