生成式人工智能“幻觉”侵权中服务提供者的过错认定

王子逸，朱伯玉

（山东理工大学法学院，淄博 255049）

摘要：生成式人工智能的广泛应用推动社会变革，但其“幻觉”现象也引发了各种新型侵权纠纷。由于生成式人工智能技术的复杂性，《中华人民共和国民法典》侵权责任编的过错认定规则难以直接适用。对此，应在明确服务提供者主体范围的基础上，从事前、事中、事后 3 个阶段构建注意义务，并进一步以概括预见能力、现有技术水平、控制力、“成本-效益”为标准认定过错。四维度标准共同构成适配生成式人工智能“幻觉”侵权的过错认定框架，避免责任的过度扩张或真空，有助于人工智能产业持续、健康发展。

关键词：生成式人工智能；人工智能幻觉；过错认定；服务提供者

中图分类号：F424；TP18 **文献标志码**：A **文章编号**：1671－1807(2026)17－0010－09

生成式人工智能在推动信息生产方式变革的同时，也因“幻觉”输出引发人格权、财产权和知识产权等新型侵权风险。由于此类内容通常由模型自主生成，而生成式人工智能自身又不具备主体资格，所以无法承担侵权责任。现有研究多主张由生成式人工智能服务提供者（以下简称“服务提供者”）对生成内容造成的损害负责，并且考虑到人工智能产业未来的发展，对此应采取过错责任原则[1]。因此，如何实现鼓励创新与规范治理的统一是《中华人民共和国民法典》（以下简称《民法典》）侵权责任编面临的重大挑战。在此情况下，服务提供者的过错认定，既是其承担责任的法律依据，也决定着责任边界的合理划定。现尝试对生成式人工智能“幻觉”侵权中服务提供者过错认定面临的难题及其认定标准加以探讨，以期为实务中相关纠纷的解决提供指导。

既有研究已经注意到生成式人工智能侵权的过错责任、注意义务和技术标准问题，但仍有必要将国内规范解释与域外风险治理经验进行更充分的对照。欧盟《人工智能法案》以风险分级为核心，将不可接受风险、高风险、有限风险以及通用目的人工智能的透明度义务区别配置，提示中国在“幻觉”侵权中的过错认定也应避免单一化标准，而应根据应用场景和损害强度配置差异化义务[2]。美国算法黑箱讨论及相关判例则表明，技术不可解释性并不当然阻却责任，关键在于主体是否能够证明其采取了合理的验证、记录、解释和纠错措施[3]。同时，Mata v. Avianca 等案件突出用户在专业场景中的核实义务，Moffatt v. Air Canada 则表明经营者不能以聊天机器人为“独立工具”为由排除其对输出信息准确性的管理责任[4-5]。因此，在既有研究基础上，将比较法中的风险分级、技术性安全注意义务和案例中的事实核实义务转化为中国侵权法上的分层注意义务与过错判断因素。

1 生成式人工智能“幻觉”的风险与归责困境

生成式人工智能“幻觉”侵权的过错认定应以厘清其风险结构和传统规则适用限度为前提。其生成内容具有自主性、不确定性，传统过错认定规则难以直接适用。

1.1 “幻觉”的风险

“幻觉”（hallucination）原本是人工智能和计算机科学领域用于指代大型语言模型生成不准确、无事实依据或脱离上下文信息内容的术语[6]。在法律语境下，“幻觉”不再仅仅是一个描述模型输出不准确的技术现象，而是被构建成一种可归责的技术缺陷，其原因在于：①生成式人工智能的固有缺陷可能造成他人权益受损，而服务提供者作为模型的开发者和部署者从中获取了利益，理应对此风险有防范义务；②“幻觉”侵权的特殊性在于侵权内容并非

收稿日期：2026-04-22

基金项目：山东省社会科学规划研究项目（25CFZJ11）

作者简介：王子逸（2003—），男，江苏无锡人，硕士研究生，研究方向为民商法律实务；通信作者朱伯玉（1967—），男，山东青岛人，博士，教授，硕士研究生导师，研究方向为经济法学、数字法学。

由用户上传,而是模型在与用户交互过程中“无中生有”地生成新的虚假信息[7]。这一特性使其区别于传统的网络侵权。这突破了《民法典》第一千一百九十四条所规定的传统网络侵权模式,从而要求法律必须发展出新的过错认定标准,以应对新的技术挑战。

从技术角度看,生成式人工智能的“幻觉”主要来源于以下4个方面:①模型的训练需要大量的语料,而模型自身所提供的训练数据中也往往含有不少虚假信息和错误信息,在模型的训练过程中会被内化;②模型是基于统计相关性而非事实逻辑而产生“认知”,在模型的训练数据不完备或者有矛盾时,模型会按概率补齐相关信息,从而产生虚假的或者是表象合理但内容失真的内容;③模型通过自回归方式生成输出,每一步都依据前面的结果作出概率选择,会导致局部误差不断发生;④模型虽然以确定的表达风格输出答案,但真实性有待考察。在人类反馈强化学习(RLHF)、内容过滤等机制下,模型有可能在没有依据时提供更为完备、更加肯定的答复,从而掩盖不确定性。

“幻觉”侵权在实践中主要表现为3类:①人格权损害,如捏造特定主体违法事实、不实经历或者生成高度相似的肖像、声音、视频;②财产权益损害,如生成失实金融、投资、商业信息,诱导用户作出错误交易决策;③知识产权与竞争秩序损害,如错误使用、归属或贬损他人作品、商标及商业标识。上述风险的共同特点在于,生成内容具有较强的真实外观,用户容易基于信赖而进一步传播或使用,从而扩大损害后果。

1.2 传统过错认定规则的失灵

(1)一般注意义务的适用基础被削弱。《民法典》侵权责任编通常以行为结果具有相对稳定的可预见性为前提[8],但生成式人工智能的输出“因时、因人、因问”而异,服务提供者通常无法具体预见某一提示词会对应何种内容,也无法像传统平台那样事先锁定审查对象。

(2)内容审查义务面临困境。传统平台侵权以用户上传的既有内容为审查对象,平台可以通过人工审核、关键词过滤、删除和屏蔽等方式进行后置治理;而生成式人工智能的输出内容是在交互过程中即时生成的,既没有稳定的待审查对象,也不存在传统意义上的“先生成、后审核”时间差[9]。更关键的是,当前算法本身对虚假信息、歧义表达、隐含贬损等复杂语义的识别能力有限,服务提供者难以借助现有工具完成精细化的事实审查。

(3)可预见性标准与合理性标准均遭遇困境。生成式人工智能的运行过程具有显著的“算法黑箱”特征,即便开发者也难以解释模型为何会选择此表达;而损害结果又往往取决于用户理解、后续传播和具体应用场景,呈现多阶段、链条化、不确定的扩散特征。因此,若仍采用静态、一般化的合理人标准和结果可预见性标准,既不能体现服务提供者的真实风险控制能力,也容易造成过错认定的主观化与空洞化。

2 生成式人工智能过错认定中的主体界定与注意义务构建

过错认定不仅取决于注意义务内容,更取决于责任主体界定。生成式人工智能运行涉及模型开发、平台部署和场景应用等多元主体,不同主体的风险控制能力不同。只有区分服务提供者的法律性质及产业链位置,才能实现责任与能力相匹配。

2.1 服务提供者的主体界定

根据《生成式人工智能服务管理暂行办法》(以下简称《暂行办法》)第二十二条第(二)项的规定,生成式人工智能服务提供者,是指利用生成式人工智能技术提供生成式人工智能服务(包括通过提供可编程接口等方式提供生成式人工智能服务)的组织、个人。但是《暂行办法》规定的生成式人工智能服务提供者的概念过于抽象和宽泛,需要对其进行更精细的区分。

(1)生成式人工智能服务提供者的法律属性应界定为服务提供者而非产品生产者。根据《暂行办法》第九条“提供者应当依法承担网络信息内容生产者责任”的规定,虽然在条文的表述上与《中华人民共和国产品质量法》中的生产者类似,但这并不必然等同于传统意义上的产品生产者。一方面,从形式上看,生成式人工智能不是一次性交付的独立商品,而是通过平台接口或者在线系统以动态交互的方式向用户提供内容生成服务,其提供模式更接近服务供给而非产品销售。另一方面,若将服务提供者等同于产品生产者,则意味着其应当适用严格责任,这与生成式人工智能运行过程中不可避免的技术缺陷不相适应,服务提供者也将承担过重的法律责任,不利于人工智能产业的健康发展。

(2)生成式人工智能服务提供者更适宜被界定为新型网络服务提供者。有观点认为,生成式人工智能服务并非一定通过互联网提供,所以不应纳入网络服务提供者的范畴[10]。但是实际上绝大多数

生成式人工智能系统是通过互联网、平台、API接口向人们提供服务，其服务行为与传统网络服务提供者高度相似，均通过在线环境向用户持续输出信息。因此，从服务方式和技术路径来看，生成式人工智能服务更适合被视为一种网络服务。但是生成式人工智能与传统网络服务之间存在本质差别。传统网络服务的内容来源于人工编辑、组织或用户上传，而传统网络服务提供者仅提供网络接入、信息存储或转发等中立性服务。相比之下，生成式人工智能的内容并非源自平台主动提供的信息，也非用户直接上传的信息，而是在用户指令与算法推断共同作用下即时生成的新内容，所以生成式人工智能服务提供者更应被界定为一种兼具内容生成功能与技术支撑属性的新型网络服务提供者，这一性质也决定了其过错认定必须区别于既有的网络服务提供者模式。

在此基础上，生成式人工智能产业链中的服务提供者还可进一步区分为模型层主体、服务层主体和应用层主体。模型层主体主要包括模型研发者、算法设计者等，负责预训练、数据筛选、算法架构设计和参数调优，对模型安全、数据质量和“幻觉”源头治理具有较强控制力；服务层主体主要包括平台运营者、部署者等，负责将基础模型转化为面向公众的具体服务，其控制力集中体现在安全策略设置、内容过滤、使用规则制定和用户行为管理等环节；应用层主体主要包括API接入者、二次开发者等，其通过将生成式人工智能嵌入具体场景，影响最终内容的呈现方式、使用方式和损害后果，尤其在高风险领域应承担领域化、情景化的审慎义务。由此可见，对服务提供者的过错认定，既要以其“新型网络服务提供者”的法律性质为前提，也要结合其在产业链中的具体位置，按照其风险控制能力和注意义务范围进行分层判断。

2.2 注意义务的分层构建

三阶段注意义务并非单纯的学理抽象，而应当与现行规范形成对应关系。《暂行办法》第七条、第八条分别规定训练数据来源、质量提升和标注管理要求，可作为事前语料处理与算法对齐义务的直接规范依据；第十条关于公开适用人群、场合、用途并引导用户理性使用的要求，以及第十二条关于生成内容标识的要求，构成事中披露、提示和标识义务的规范基础；第十四条、第十五条关于违法内容处置、用户违法活动处置以及投诉举报机制的规定，则对应事后纠错、限制功能、停止传输和反馈机制。2025年《人工智能生成合成内容标识办法》进一步区分显式标识与隐式标识，使文本、图片、音频、视频等不同形态输出的提示义务具有更明确的操作路径。据此，三阶段义务应理解为行政监管规范向侵权法注意义务的转化，而不是脱离现行法的独立创设。

事前审查义务旨在从源头减少“幻觉”，包括语料处理义务和算法对齐调优义务。①语料处理义务是预防“幻觉”的基础。训练数据的合法性、质量与代表性直接决定了模型是否存在“内生幻觉”倾向，是生成式人工智能的根本风险来源[11]。模型层主体控制语料筛选与清洗，应承担源头治理义务，包括审查数据来源合法性、进行去偏去污处理、验证数据代表性并制定语料安全规则，防止模型吸收虚假、失真或偏颇信息[12]。服务层主体虽然无法直接支配训练数据，但是应当对介入的基础模型进行合规性审查，确保不使用来源不明或者存在明显缺陷的第三方模型。应用层主体通常不参与训练数据处理，因此其义务相对较弱。②算法对齐调优能防止“幻觉”在算法层面出现。模型层主体应该采用RLHF（reinforcement learning from human feedback，基于人类反馈的强化学习）、安全对齐、拒答等手段降低“内生性幻觉”；服务层主体需要结合具体场景进行“二次对齐”，如增加模型阈值、过滤敏感场景等；应用层主体应该在医疗、法律等高风险场景设置更严格的安全阈值，减少衍生风险。

事中披露义务针对内容生成与输出阶段，通过实时过滤、内容标识和风险提示降低“幻觉”引发侵权的可能。①内容审查义务是生成式人工智能最独特的义务。相较于传统网络平台，AI输出不存在固定的待审查对象，不再是对已生成内容的删除、过滤，而是通过技术方法对即时生成内容进行实时检测和安全过滤[13]。模型层主体应设置基础过滤机制，如拒答集合、敏感词库和高危识别逻辑；服务层主体作为输出端，应通过识别工具和过滤算法阻断明显违法或高危险内容；应用层主体在法律、医疗、金融等专业场景中，应加强场景化审查，并附加必要免责声明。②生成式人工智能的“幻觉”具有高度表达自信的特征，用户很容易将生成内容误认为是真实信息，因此服务提供者应履行明确的内容标识与风险提示义务。模型层主体应当在系统界面中强调生成内容的不确定性以及适用边界。服务层主体应当在文本、图片、音频、视频等输出内容上强调“AI生成”并且提示是否存在虚假或偏差或

者不适用专业判断等问题。应用层主体应当在具体场景中加强风险提示,如在医疗场景中加强“非专业诊断建议”的提示,避免用户对于生成内容过分信任。

事后处置义务旨在防止风险在侵权内容传播后进一步扩大,包括用户行为管理义务与纠错和反馈机制义务。①在用户行为管理方面,生成式人工智能的风险不仅来自模型自身,也来自用户的不当使用行为。其主要分为两种:一种是用户故意诱导AI生成侵权内容,如用户通过“越狱”绕过系统的安全防护机制,引导AI生成虚假信息并进行传播;另一种是用户自身存在过错,利用AI自主生成的侵权内容造成损害,如用户对AI生成的明显虚假信息不加甄别就直接使用、传播,最终造成他人权益受损[14]。模型层主体应在技术上支持越狱检测与敏感输入识别。服务层主体则负有最强的用户行为管理义务,应当建立输入检测机制、阻断恶意提示词、制定多次违规限制规则等。应用层主体应结合其使用领域,对用户身份、使用目的进行必要审查,防止生成内容在特定场景中被滥用。②在纠错与反馈机制方面,服务层主体应建立投诉处理机制,在接到通知后及时删除、屏蔽或限制相关内容,并对风险内容增加补充提示或使用限制。模型层主体应根据反馈持续修正模型的“幻觉”并更新参数,以减少未来“幻觉”出现的概率。应用层主体应对具体场景中已造成误导的内容进行补正,防止其进一步传播[15]。

2.3 多层主体义务边界及责任衔接

模型层、服务层和应用层主体的义务边界,应以“谁最接近风险源、谁最能够低成本控制风险、谁因商业化使用获得主要利益”为基本判断线索。模型层主体主要承担基础模型安全、训练数据合规、对齐调优、红队测试和重大缺陷修复义务;服务层主体主要承担模型选用审查、接口调用规则、实时过滤、输出标识、用户管理和投诉处置义务;应用层主体主要承担场景适配、专业限制、用户提示、结果复核和具体传播控制义务。

在义务移转与衔接上,应区分“基础能力提供”与“场景化风险再配置”。模型层主体向服务层主体提供API或者基础模型时,应通过技术文档、模型卡、风险说明、更新日志等方式披露模型能力边界和已知风险;服务层主体在部署时若改变模型参数、开放高风险功能或者重新设计交互界面,则应就其新增风险承担相应义务;应用层主体若将通用模型嵌入医疗、法律、金融等专业场景,则不得仅以基础模型已经通过平台审查为由免除其场景化核验义务。

在责任分配上,同一损害可能由多层主体共同造成。若模型层缺陷具有普遍性且服务层、应用层无法通过合理措施识别或阻断,应侧重评价模型层过错;若模型本身风险已被披露,而服务层未设置必要过滤、标识或投诉机制,则服务层过错更为突出;若损害主要来源于应用层对生成结果的再包装、专业背书或商业传播,则应用层应承担主要责任。该分层规则可以避免将所有风险简单上推至基础模型提供者,也可防止终端应用者以技术外包为由逃避注意义务。

3 “幻觉”情形下服务提供者的过错认定标准

生成式人工智能“幻觉”侵权中,过错认定不能以损害结果倒推责任,而应综合风险认知、技术可行性、控制能力和义务成本。相应地,可从概括预见能力、现有技术水平、控制力和“成本-效益”4个维度构建判断标准,以平衡责任扩张与责任真空。

3.1 概括预见能力标准

判断行为人是否存在过错通常以其是否尽到注意义务为核心标准。在生成式人工智能“幻觉”侵权场景下,服务提供者通常无法预见生成式人工智能何时、以何种方式生成致害内容,但这并不意味着注意义务无法适用。在《民法典》侵权责任编中,注意义务包括行为上的注意义务和结果上的注意义务。行为人是否尽到结果上的注意义务又可根据其预见程度分为具体预见能力和概括预见能力,前者要求行为人在行为时能够明确预见损害对象、损害类型和损害结果,后者仅要求行为人能够认识到其行为具有一般危险性[16]。为避免对行为自由过度约束,原则上不以行为人仅具备概括预见能力为由而直接要求其履行相应的注意义务,但生成式人工智能具有“幻觉”的固有缺陷,使服务提供者天然地不具备具体预见能力,却能够认识到其技术存在造成一般性损害的危险,这种一般化、类型化的风险认知即概括预见能力。

概括预见能力要真正能够为司法适用提供指引,就不能停留在抽象层面,而应当在法官能够操作的事实判断中落实,其核心在于服务提供者是否能够认识到自身技术具有一般危险性、在特定应用场景中风险可能被放大以及自身控制能力存在边界。首先,任何生成式人工智能模型都是基于概率

推断运行，其输出很可能包含虚构内容或者不准确的表达。正因如此，服务提供者应当能够认识到此类风险具有普遍性与不可避免性，“幻觉”是生成式人工智能的内生现象，即使无法预见哪些内容会构成侵权，其对一般风险的认知仍可合理期待。如果服务提供者无法认识到这一点，即表明其对技术基本属性缺乏必要理解，难以认为其已尽到一般注意义务。其次，生成式人工智能的侵权风险在不同使用场景中存在显著差异，尤其在医疗、法律、金融等专业领域，错误输出更可能导致重大损害，服务提供者无需预见具体损害结果即可认识到其在高风险场景中带来的风险扩张，因此有必要采取额外的限制、提示或过滤措施。如果明知服务可能用于高风险场景的情况下仍按一般场景处理，则不能认为其已尽到概括预见义务。最后，服务提供者对于自身控制能力的限度应具有合理认知，应当认识到模型自身的不可解释性，包括“幻觉”不可完全消除、过滤机制存在上限、用户可能通过“越狱”(Jailbreak)或输入特殊提示词规避安全措施等情况。如果服务提供者在明知技术存在不可控性的情况下仍缺乏必要的安全机制即投入运营，或者未能采取合理的一般性风险控制措施，则应认定其存在过错。

为防止概括预见能力标准被过度主观化，司法适用中应采取“先客观、后主观、再场景”的判断顺序：首先审查同类服务提供者在损害发生时是否普遍知悉该类“幻觉”风险；其次结合服务提供者的规模、专业能力、既往投诉、内部测试记录和公开风险提示，判断其是否应当认识到风险；最后再根据医疗、法律、金融、公共传播等具体场景调整注意义务强度。概括预见能力只解决“是否应当启动注意义务”的问题，现有技术水平则解决“能够要求其采取何种措施”的问题，二者不能相互替代。以 Mata v. Avianca 案为例，律师将 ChatGPT 生成的虚假判例提交法院并受到制裁。该案所体现的并非服务提供者当然免责，而是说明在法律等专业场景中，“幻觉”风险具有高度可预见性：服务提供者应当通过显著提示、引用来源限制或者检索核验功能降低误信概率；专业用户和应用层主体则负有独立核实义务，不能将模型输出直接替代法律检索与事实核验。

需要指出的是，如果生成式人工智能“幻觉”侵权是由用户过错导致的，如用户对生成内容未尽合理核实义务、对明显虚假信息传播与利用或者通过恶意提示、“越狱”行为故意诱导 AI 生成虚假信息，最终造成损害，则应认定用户存在过错。但是，若服务提供者疏于审核而导致损害结果的发生或扩大，则其应当承担相应的补充责任。服务提供者仅仅是为用户故意侵权提供了途径，但是用户基于主观故意实施的侵权行为性质更为恶劣，应受法律更严厉的责难[17]。在责任形态上，服务提供者与用户可能构成共同侵权、分别侵权或者补充责任。若服务提供者明知用户实施侵权仍提供技术便利，或与用户形成共同传播、共同获利关系，可考虑共同侵权责任；若用户独立传播造成主要损害，而服务提供者仅在投诉处理、风险提示等环节存在瑕疵，则更宜按照原因力和过错程度确定按份责任或者补充责任；若受害人自身对明显虚假内容继续扩大传播，也应纳入过失相抵评价。用户过错还应作类型化处理。①用户以“越狱”、规避安全提示或者虚构事实输入等方式故意诱导模型生成侵权内容，并进一步传播、使用的，应以用户故意侵权为中心评价，服务提供者仅在未设置基本反规避机制、明知高频滥用而放任或者接到通知后未及时处置时承担相应责任；②用户虽无故意，但在专业场景中将生成内容作为法律意见、诊疗建议、投资判断等直接使用，且未进行合理核实的，应适用过失相抵规则减轻服务提供者责任；③若服务提供者通过界面设计、权威化表达或者商业宣传强化用户信赖，使普通用户难以识别“幻觉”风险，则不宜将核实义务过度转嫁给用户。

3.2 现有技术水平标准

《民法典》侵权责任编通常将是否履行“注意义务”作为认定过错的标准，但在生成式人工智能场景下，单纯考虑行为人的主观心态显然不足以应对其给法律带来的冲击。为了避免过错认定的主观化，应引入“现有技术水平”这一客观标准[18]。“现有技术水平”是指在特定时期内，本领域内可普遍获取和运用的知识和技能。以“现有技术水平”认定服务提供者的过错，应当考虑两个层面：一方面，在生成式人工智能生成虚假内容从而损害他人权益的情况下，认定服务提供者对此承担责任的前提是其具有过错。也就是说，按照损害发生时的技术水平，服务提供者能否通过技术手段或其他必要措施防止损害发生。如果可以但服务提供者没有采取，则其具有过错，应当承担侵权责任；反之则其没有过错，不承担责任。另一方面，在损害扩大的情况下，按照损害发生时的技术水平，如果服务提供者无论如何也无法预见并避免损害发生，则其不具有过错，不需要承担侵权责任[19]。但是，如果服务

提供者在技术尚不成熟、难以有效控制风险的情况下,仍将生成式人工智能投入运营,从而导致损害发生,则应当认定其在主观上存在过错,应当承担侵权责任。在这种情况下,应当认定服务提供者对潜在风险具有可预见性,而且将不可控的技术投入市场本身就是一种过错[20]。反之,如果在现有技术水平下无法预见损害发生,并且其原因无法得到解释,服务提供者的程序设计、运营管理等技术符合损害发生时的技术与管理水平的平均水平,加之其已切实履行安全保障义务,那么即使发生了一定损害,也不宜据此认定其具有过错。在服务提供者已经尽到与“现有技术水平”相当程度的事前预防和事中管控义务的前提下,残余风险应被理解为意外风险,不能简单地因为损害发生而要求服务提供者承担侵权责任。

为了使“现有技术水平”在司法适用时具备实质内容和可操作性,现有技术水平可从时间、行业和地域3个维度判断。①时间维度应当以损害发生当时的技术水平为依据,审查服务提供者是否存在过错,或者是否对损害的扩大存在过错。如果损害在现有技术水平下能够避免而未能避免,原则上应当认定服务提供者存在过错。同时,由于服务提供者自身技术和信息优势,其预见和控制风险的义务也相应地高于普通用户,所以其也应当及时、明确地告知和引导用户,提醒用户如果根据生成内容作出决策可能会面临潜在风险,并提供合理的替代方案或者明确要求用户自行承担风险。②行业维度是指在评价技术水平时应当以同行业内相同专业背景的专家普遍具备的平均技能与认知水平作为基准。在医疗、法律、金融等高风险场景中,服务提供者必须及时更新技术,利用算法和数据处理能力提高生成结果的准确性、可靠性,确保其技术水平处于行业前沿,以适应行业发展需要。需要强调的是,如果某项技术尚未在行业内大规模推广应用时就要求服务提供者达到该技术水准实属苛刻。在某项技术不成熟的情况下,服务提供者应根据行业发展态势灵活调整技术要求,并通过合作或创新弥补不足。③评价“现有技术水平”还需要考虑地域维度,即考虑不同国家或地区不同时期生成式人工智能技术水平、可获得性,特别是技术可能面临的管制、转让障碍等因素。然而,这并不表示中国在推动生成式人工智能技术发展时可以忽视国际上的整体发展水平。《生成式人工智能服务管理暂行办法》第六条规定,鼓励生成式人工智能算法、框架、芯片及配套软件平台等基础技术的自主创新,平等互利开展国际交流与合作,参与生成式人工智能相关国际规则制定。因此,服务提供者应当随时关注和把握国际前沿的技术发展动向。此外,许多技术性问题超出了司法人员的专业范围,在面临重大争议时仅凭司法人员的专业水平很难作出准确的判断,必须借助专家的知识和经验才能确保裁判公正。

Mata案还可用于说明“现有技术水平”标准的边界。在现有生成式模型尚不能稳定保证判例、法规和事实来源真实性的条件下,法院不能要求服务提供者完全消除法律检索类“幻觉”;但若同业已经普遍采用来源链接、检索增强、结果置信度提示、拒答机制等防控手段,服务提供者仍以通用聊天模式提供专业法律信息,则可据此认定其未达到损害发生时可期待的技术水平。

在“概括预见能力”与“现有技术水平”的关系上,应以前者作为义务发生的门槛,后者作为义务内容的上限。法官不能仅因服务提供者知道“幻觉”普遍存在,就要求其承担所有损害结果;也不能因某项技术措施理论上可行,就当然认定未采用该措施构成过错。更可操作的路径是:先判断风险类型是否处于可预见范围,再判断当时同类主体是否已有成熟、可获得、可部署的防控手段,最后结合具体场景确定措施强度。

3.3 控制力标准

在生成式人工智能“幻觉”侵权中,过错认定的核心在于判断服务提供者是否负有可期待的注意义务,而这种“可期待性”必须以其对风险的实际控制能力为限[21]。控制力反映的是主体对风险形成与扩大的技术支配程度,也是过错认定的能力前提。注意义务并非对所有行为人均等适用,其适用范围限于行为人现实能够控制的风险范围。若某一主体在技术和管理层面无法对风险施加实质性影响,那么要求其承担超出能力范围的义务将导致过错认定脱离现实基础。因此,控制力标准的功能在于确保过错认定建立在能力与义务相匹配的逻辑上,避免过度归责或者不现实的义务设定。

模型层主体掌握训练语料、模型架构与算法的底层机制,控制力看似最强,但并不意味着其在所有层面上都能够有效干预风险。基础模型的规模化架构、海量训练语料的学习以及算法优化的复杂性,使服务提供者无法通过简单调整对模型进行全面、细致的修正,并且结构性问题的修复通常需要

极高投入与超长周期的开发，所以模型层主体的义务应集中于那些确实能够施加影响的关键环节，如发现明显偏误和缺陷、修正影响模型稳定性的机制等。至于难以通过现实条件有效控制的风险，不能仅因拥有最高层级的控制力就推定其负有最严格的义务。

服务层主体虽然无法触及模型内部，但其在部署、调用、过滤和输出管理等环节具有较强的调节能力，其控制力主要集中在阻断模型输出时的可见风险，故其注意义务自然侧重于识别和处理明显不当的内容，优化用户交互规则以及必要的风险提示等方面。这几方面的调整通常更及时、可操作性更强，且所需投入较为可控，所以服务层主体在这些环节未采取合理措施时，其不作为更容易被认定为存在过错。Moffatt v. Air Canada 案更能说明服务层主体的控制力。加拿大航空将聊天机器人嵌入官网并用于回答丧亲票价退款规则，裁判机构认为聊天机器人仍属于网站组成部分，航空公司不能以“聊天机器人独立”为由切断其对网站信息的控制责任。对应到生成式人工智能服务中，只要服务层主体能够决定对外展示、接口调用、答案过滤、纠错更新和消费者提示，即应就这些可控制环节承担相应注意义务。

应用层主体虽然不能影响生成内容本身，但其对使用场景具有直接控制力，可以通过限制用途、调整情境、提供解释或引导用户正确理解生成内容等方式降低风险。应用层主体的控制力虽属末端，但其所能控制的范围仍具有实际作用。若主体明知某特定场景具有高度误用风险，但仍未采取相应的限制、提示措施，则该不作为构成未履行注意义务，应当认定其存在过错。

个案中判断某一主体是否“无法通过现实条件有效控制”风险，应综合考量 6 项因素：①是否能够接触或修改训练数据、系统提示词、模型参数和安全策略；②是否能够决定服务开放范围、调用频率、输出阈值和高风险功能；③是否掌握日志、投诉、测试报告等风险信息；④是否具备暂停服务、限制用户、回滚版本和紧急修复的能力；⑤是否通过合同、接口规则或审核机制影响上下游主体行为；⑥采取措施是否会导致服务基本功能丧失或产生明显不成比例的技术负担。只有在上述因素显示主体事实上不能识别、阻断或减轻风险时，才可认定其控制力不足。

3.4 “成本-效益”标准

生成式人工智能属于技术工具，本身并无特殊危险性，但是对大数据和自主生成内容仍然会产生知识产权侵权、人格权益受损等新型风险。因此，关键不在技术本身的危险性大小，而是服务提供者是否在合理程度内采取了可以防止损害发生的措施。在现有技术水平的条件下，当一种措施的成本高于实际可减少的损害时，难以要求服务提供者承担该义务，则应认定服务提供者没有过错。而当要求服务提供者采取现有技术水平无法满足的措施或者成本上无法负担的措施时，会使成本与风险严重失衡，影响生成式人工智能的正常工作。在防范成本与降低效益之间失衡的情况下，扩大注意义务的范围并不合理。尤其在生成阶段，对底层模型进行修正往往需要极高的成本，而在风险显现后通过删除、屏蔽、提示或者限制使用等方式进行控制，成本明显降低，义务负担也随之减轻。

“成本-效益”标准的本质不是简单的比较成本，而是根据采取的措施的技术可行性、边际风险降低能力、措施替代性等因素对注意义务的必要性和合理范围作出更加具体的判断。①成本不单单由直接的经济投入构成，还包括算法迭代成本、算力消耗、数据扩充与清洗成本、对模型性能（如生成速度、准确率）的影响程度等因素。许多看似简单的技术要求，实际上往往意味着对模型底层架构、训练语料体系、对齐机制的系统性调整，成本并非呈线性增加，而可能呈指数型增长。因此，对某项措施是否符合合理成本需要从技术的整体性和复杂性出发，不能以表面操作难易作为判断依据。②效益不能简单地定义为降低风险的数值多少，而是衡量其在特定技术中的边际贡献。任何技术措施都不可能彻底消除“幻觉”，因此关键不是完全消除风险，而是在现有技术水平下减少高风险输出或者降低受害者误信概率。例如，对生成内容标注真实性不确定的提示，虽然无法阻止“幻觉”出现，但能够显著影响用户的信赖程度，从而降低发生现实损害的概率，其效益远高于成本相近的其他措施。③还需考虑措施替代性和系统外部性。一些高成本措施可能存在低成本替代方案，如对模型底层架构进行大规模改造成本极高，但可以通过服务层接口设置输出限制、敏感场景过滤或风险提示，以低成本实现类似的风险防控效果。在这种情况下，即便高成本措施具有更高效益，也不应纳入义务范围，因为低成本的替代方案已经达到了可接受的风险水平[22]。此外，若某项措施虽能降低风险，但会给其他环节带来系统外部性的影响，如显著降低模型性能、阻碍技术迭代或使服务无法正常提供，则不应

将该措施纳入义务范围,因为这会对整个技术生态造成不合理的负担。Moffatt 案同时体现“成本-效益”标准的具体运用。对于经营者而言,将聊天机器人答案与正式票价规则保持一致、在答复中提示应以政策页面为准、对高频争议问题设置人工复核或固定答案,通常属于成本较低而风险降低效益明显的措施;若经营者未采取这些简单措施,反而让消费者在同一网站的不同入口获得相互矛盾的信息,则不宜以技术复杂性抗辩其未尽合理注意义务。

在上述基础上,“成本-效益”标准形成了一个动态的义务强度结构:成本较低、技术成熟且边际效益明显的措施构成服务提供者必须承担的义务;成本适中但具有显著风险降低能力的措施构成较强义务,应在特定场景中使用;而成本高昂、效果有限或存在相应低成本替代方案的措施,则不宜纳入义务范围;至于技术上尚不可行或成本远远超出合理水平的措施,则应排除在注意义务之外。

为了增强“成本-效益”标准的裁判可操作性,证据规则上可要求服务提供者提交模型评测报告、红队测试记录、内容过滤命中率、投诉处理数据、版本更新日志、风险提示页面、第三方安全评估和同业技术实践材料;受害人则可提交错误输出截图、传播链条、损害后果以及服务提供者明知风险的公开声明或历史投诉。对于技术成本、边际风险降低效益和替代措施可行性争议较大的案件,应通过专家辅助人、司法鉴定或者行业专家意见加以说明。法院在采信时不应仅比较绝对成本,而应审查该成本与服务规模、盈利模式、风险严重性和可替代措施之间是否成比例。

4 结语

生成式人工智能“幻觉”侵权中的过错认定应从传统结果导向转向风险防范措施的实质审查。概括预见能力标准解决风险认知问题,现有技术水平标准限定技术可行性,控制力标准实现多主体责任分层,“成本-效益”标准约束注意义务边界。四维度标准共同构成服务提供者过错认定的判断框架,既可避免因技术黑箱造成责任真空,也可防止因损害结果倒推责任而抑制技术创新。未来司法适用中,应结合具体场景、主体角色和风险控制能力,对服务提供者是否尽到合理注意义务作出个案判断。只有以此为基准重塑过错认定框架,才能在法律层面为中国生成式人工智能的发展提供基本的治理规范,使人工智能产业的发展在规范与创新的双重推动下实现良性循环。

参考文献

[1] 王利明. 生成式人工智能侵权的法律应对[J]. 中国应用法学, 2023(5): 27-38.

[2] European Parliament, Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)[EB/OL]. Official Journal of the European Union, (2024-07-12)[2026-03-10]. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[3] Supreme Court of Wisconsin. State v. Loomis, 881 N. W. 2d 749 (Wis. 2016)[EB/OL]. (2016-07-13)[2026-03-10]. https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf? content=pdf&seqNo=171690.

[4] United States District Court for the Southern District of New York. Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S. D. N. Y. 2023)[EB/OL]. (2023-06-22)[2026-03-10]. https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/.

[5] British Columbia Civil Resolution Tribunal. Moffatt v. Air Canada, 2024 BCCRT149[EB/OL]. (2024-02-14)[2026-03-10]. https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html.

[6] 张梦晗, 沈文乾. 错位的升级与结构性失序: AI 幻觉主导的信息迷雾风险与分类治理可能[J]. 传媒观察, 2025(10): 53-63.

[7] 谢凡. 生成式人工智能侵权诉讼的类型化建构[J]. 数字法治, 2025(4): 127-142.

[8] 任安麒. 网络服务平台算法推荐的著作权侵权认定规则[J]. 北京航空航天大学学报(社会科学版), 2023, 36(3): 190-198.

[9] 王利明. 生成式人工智能侵权的归责原则与过错认定[J]. 中国法律评论, 2025(4): 15-30.

[10] 姚志伟. 生成式人工智能服务提供者在私法上的法律性质[J]. 上海交通大学学报(哲学社会科学版), 2024, 32(12): 127-142.

[11] 毕文轩. 论生成式人工智能服务提供者的注意义务[J]. 苏州大学学报(哲学社会科学版), 2025, 46(2): 90-103.

[12] 张新宝, 卞龙. 论生成式人工智能服务提供者的过错推定责任[J]. 北方法学, 2025, 19(5): 5-19.

[13] 朱嘉珺. 生成式人工智能虚假有害信息规制的挑战与应对: 以 ChatGPT 的应用为引[J]. 比较法研究, 2023(5): 34-54.

[14] 沈森宏. 论生成式人工智能服务提供者过错的认定[J]. 现代法学, 2024, 46(6): 133-147.

[15] 王良顺, 李想. 生成式人工智能服务提供者的数据安全保护义务研究[J]. 南昌大学学报(人文社会科学版), 2023, 54(6): 72-83.

[16] 李中原. 析二元结构模式下的侵权过失评价机制[J]. 法学, 2013(8): 122-129.

[17] 李丹. 人工智能特殊侵权责任的理论证成及规范储备[J]. 东方法学，2025(5)：48-63.
[18] 王若冰. 论生成式人工智能侵权中服务提供者过错的认定：以“现有技术水平”为标准[J]. 比较法研究，2023(5)：20-33.
[19] 杨利华. 生成式人工智能服务提供者注意义务研究[J]. 比较法研究，2025(3)：54-68.
[20] BATHAEE Y. The artificial intelligence black box and the failure of intent and causation[J]. Harvard Journal of Law & Technology，2018，31(2)：890-938.
[21] 汪倪杰. 论先行行为与安全保障义务的关系[J]. 法学，2022(9)：106-121.
[22] 张华韬. 动态体系论下生成式人工智能侵权的归责与构成[J]. 法学杂志，2025，46(6)：82-101.

Fault Determination of Service Providers in “Hallucination” Infringement of Generative Artificial Intelligence

WANG Ziyi，ZHU Boyu
(School of Law，Shandong University of Technology，Zibo 255049，China)

Abstract：While social transformation is promoted by the widespread application of generative artificial intelligence，various new types of infringement disputes are triggered by its “hallucination” phenomenon. Given the technical complexity of generative artificial intelligence，the traditional rules of fault determination under Book Seven Tort Liability of the *Civil Code of the People's Republic of China* are difficult to apply directly. Accordingly，after the scope of service providers was clarified，the duty of care was structured across the pre-service，in-service，and post-service stages. On this basis，fault was determined through four standards：general foreseeability，the existing level of technology，control capacity，and cost-benefit analysis. A fault determination framework adapted to infringement caused by generative AI hallucinations is jointly constituted by these four dimensions. Thus，both excessive expansion and gaps in liability are prevented and the sustainable and sound development of the artificial intelligence industry is promoted.

Keywords：generative artificial intelligence；artificial intelligence hallucination；determination of fault；service provider