热点探索

大语言模型在精神卫生咨询中的伦理风险与规制路径

朱新宇，马青连
安徽医科大学法学院，安徽　合肥　231299

摘　要：大语言模型正逐步融入生命健康领域，并对传统精神卫生咨询的组织形态与伦理规范产生结构性影响。通用大语言模型的泛化能力与专用大语言模型的领域适配能力，使全天候、高可及、低成本的心理支持在部分场景中成为可能，并推动精神卫生咨询从“以人际对话与专业权威为中心”的旧范式，向“数据驱动、算法中介、规模化交互”的新范式转型：咨询关系呈现由人际信任向人机交互迁移的趋势，干预方式由单次线性走向连续动态适配，服务取向由事后补救延伸至主动预防。然而，技术嵌入也带来隐私泄露、模型幻觉、技术家长主义对自主决策权的侵蚀，以及训练偏见放大与过度依赖所潜藏的隐性临床危害，进而损害来访者的隐私权、知情同意权与自主性。为在守住安全底线的前提下释放普惠价值，应坚持分类施策原则，从筑牢数据安全防线、优化算法治理规范、锚定人本价值立场、健全行业伦理准则四个维度系统构建规制路径，推动大语言模型在精神卫生咨询领域的安全、可控与审慎发展。

关键词：大语言模型；精神卫生咨询；伦理风险；技术规制
中图分类号：R-052　　**文献标志码**：A　　**文章编号**：1671-0479（2026）05-467-009
doi：10.7655/NYDXBSSS260200

随着社会生产力的发展和医疗水平的提高，精神卫生作为生命健康的关键组成部分成为世界公共卫生领域的热门议题。精神卫生是指精神健全，能够正确认识和处理生活中存在的各种压力，充分发挥自身潜力，在学习和生活中实现自我价值，是一项基本人权。世界卫生组织发布的《精神卫生、人权与立法：指南与实践》指出：“精神健康与福祉是可持续发展的重要基础。”[1]考虑到我国的文化环境和生活习惯，精神卫生问题往往为普通民众所忽视。《中国国民心理健康发展报告（2023—2024）》显示，我国19%的民众存在强迫行为和强迫观念，41%的民众认为心理咨询服务不太便利或者很不便利，存在一定的精神卫生风险[2]。就精神卫生咨询的实践场景而言，大语言模型（large language model，LLM）可分为两类：一是通用大语言模型（general-purpose LLM），如ChatGPT、DeepSeek、文心一言等，其基于海量多领域数据训练，具备广泛的语言理解与生成能力，用户可直接调用进行情绪倾诉、心理咨询问答等非正式心理支持；二是专用大语言模型（domain-specific LLM），如人工智能聊天机器人软件（Woebot）、专科医院部署的智能问诊系统等，经过心理健康领域专项数据的指令微调，定向服务于临床心理评估、辅助诊疗等特定场景。传统意义上的精神卫生咨询是指专业人员运用心理学、精神医学及相关学科的理论与技术，通过建立安全、信任的咨询关系，为存在心理困扰、情绪问题、行为障碍或处于心理危机状态的个体或群体提供系统性支持与干预。而大语言模型的出现突破了时间和空间两大场域的限制，借助其背后庞大的数据库与精密的知识结构帮助来访者制定个性化的心理咨询方案。与此同时，大语言模型内在的技术“家长主义”以及算法“幻觉”在相当程度上影响咨询的科学性。因此，在大语言模型赋能精神卫生咨询的浪潮中，应当保持理性的态度并展开批判式反思。事实

基金项目：国家社会科学基金青年项目“数智时代健康数据利用的法律规制研究”（25CFX148）；安徽省高校优秀科研创新团队项目“健康中国背景下卫生健康法治建设研究”（2023AH010037）
收稿日期：2026-04-30
作者简介：朱新宇（2002— ），男，安徽安庆人，硕士研究生在读，研究方向为人工智能伦理；马青连（1968— ），男，安徽颍上人，博士，教授，研究方向为法律史、卫生健康法学，通信作者，1991500043@ahmu.edu.cn。

上，大语言模型内在嵌入心理咨询的过程是一场充满张力与不确定性的社会实验，在承认技术优势的同时，应当以医学伦理与社会伦理为依据谨慎审视，通过深入分析其伦理风险与实践困境，因地制宜地构建规制路径。

一、大语言模型对精神卫生咨询的范式重构

大语言模型是一种基于深度学习的人工智能技术，通过海量数据文本训练掌握人类语言的语义表征、语法规则与逻辑范式，从而使其能够理解语言含义生成数据文本，是自然语言处理的重要研究内容之一。目前，通用大语言模型以其泛化能力降低了心理健康支持的获取门槛，专用大语言模型则凭借领域定制化训练提升了临床场景的适配精度，这种自助式人工智能心理咨询系统使全天候、高可及、低成本的心理健康支持成为可能，其应用集中体现于精神健康教育、精神卫生评估以及辅助干预三个方面[3]。范式是某一科学共同体在特定时期所共同接受的信念、理论体系、研究方法、价值标准和解题范例的集合，其核心构成涵盖世界观、价值观与方法论三个维度。大语言模型对精神卫生咨询的冲击，并非线性的技术进步，而是以数据驱动、算法中介、规模化交互为特征的新范式，挑战着以人际对话、专业权威、个体服务为特征的旧范式。两者之间在核心关系、知识来源、诊断方法、干预方式、服务模式与伦理重心等维度上存在深层的不可通约性，使得围绕人工智能心理咨询的争论不仅是技术层面的分歧，更是世界观、方法论与价值标准的根本冲突。

（一）咨询关系从人际信任走向人机交互

在关系世界观与服务模式层面，大语言模型重塑了咨询的核心纽带，使咨询关系由传统的人际信任转向人机交互。固定的沟通场所和专业的人才主体是传统精神卫生咨询的基本构成要素，在通常情况下，考虑到咨询的效率和成本，一般一名咨询师在同一时间只能与一位来访者进行交流。因此，受制于专业人才的培养周期以及心理咨询的从业成本，优质的精神健康资源一般收费较高，且行业标准尚不完善，在相当程度上集中分布于大中城市的医疗机构和高校体系，具有稀缺性和专属性。目前，我国的精神卫生咨询形式主要有两种：一种是设定在医疗机构的精神门诊，但是开设数量有限；另一种则是私人提供的心理咨询服务，资质差异较大，且治疗周期长收费不一。无论是哪种精神卫生咨询方式，都需要预约定点，存在一定的时间成本与社交顾虑，在某种程度上也会使得部分轻度患者因为羞耻和害怕等情绪放弃求助，从而加重精神困扰。

大语言模型的出现与普及，在一定程度上缓解了精神卫生服务供给与人群需求之间的张力，推动精神卫生咨询呈现泛在化与普惠化趋势。其一，在标准化信息供给与风险提示层面，模型能够基于对话文本对常见心理困扰进行结构化梳理，提供自助资源、情绪调节策略与求助路径提示，从而减少因服务经验差异带来的“解释框架不一致”问题。但应当强调，此类输出原则上属于信息支持与初筛提示，并不构成临床诊断或处方性结论。其二，在经济可及性层面，不少平台提供免费或低价的基础功能，有助于降低轻度困扰人群的进入门槛，但其服务质量、责任边界与风险控制水平仍需制度化约束。其三，在时间与心理可及性层面，用户可依托网页、小程序与应用软件等渠道随时获取支持，匿名交互在一定程度上有助于缓解病耻感与社交顾虑，从而提升求助意愿并扩大服务覆盖面，推动精神卫生咨询泛在化、普惠化。

大语言模型对精神卫生咨询范式重构的推动并不止于理论层面，国内已出现将其用于情绪陪伴、心理科普、初步筛查与线上转介的多种产品与机构实践。以“好心情”等互联网医疗平台为例，相关平台尝试将情绪识别、问答科普与陪伴式对话模块嵌入服务流程，应用于心理疾病筛查、数字诊疗及企业员工心理健康管理，能够24小时为用户提供科学、实用、安全的人工智能共情陪伴服务。与此同时，部分高校与研究团队也探索面向情绪支持的对话模型，例如清华大学CoAI团队研发的情感大模型“Emohaa”，不仅能帮助用户缓解负面情绪、增强情绪调节技巧，还能进一步引导用户进行价值观探索；华东理工大学X-DLab开发的“MindChat”，则利用多模态测评技术分析用户的心理状况，并依据个人需求定制治疗方案与练习计划。医疗机构侧亦出现面向导诊宣教、初步评估与随访管理的智能助手实践，如北京大学第六医院推出的“北小六”机器人、江苏省妇幼保健院部署的具身智能机器人“飞燕”等。这些案例从供给侧显示了大语言模型在规模化覆盖、低门槛触达与前置性支持方面的潜在价值，但其有效性、适用边界与风险控制仍需进一步证据化评估与规范化治理。

（二）干预方式从单次线性走向连续动态

在诊疗方法论与干预方式层面，大语言模型重写了心理咨询的知识来源与应用逻辑，推动干预方式从单次线性向连续动态适配演进。在精神卫生咨询中，患者负面体验的一个重要来源是咨访契合度不高以及对咨询师胜任力的怀疑[4]。传统的精神卫生咨询以单向反馈为主，咨询流程遵循“用户倾诉—咨询师倾听—问题分析—方案反馈”的线性逻辑，因而咨询效果高度依赖于咨询师的沟通能力、

主观判断与从业经验。一般而言，单次精神卫生咨询的时间不超过1个小时，以40分钟为最佳。受限于不同的专业流派和咨询的组织管理形式，在咨询时间内，由于患者未能充分表达情绪状态以及咨询师未能及时捕捉患者细微情绪，可能导致治疗效果不佳。不仅如此，自2017年国家人社部将原国家二级、三级心理咨询师从《国家职业资格目录》移除后，官方已不再组织全国统考，也不再颁发这类职业资格证。虽然现由行业协会颁发此类证书，但这在一定程度上降低了入行门槛，部分咨询师专业水平较低，咨询服务同质化、模式化严重，难以针对患者的心理特质、情绪状态和认知水平对咨询内容进行灵活调整。

大语言模型通过独有的自然语言分析和理解能力，重塑精神卫生咨询的交互逻辑，实现由单向反馈向动态适配的转变。其一，精准识别患者情绪状态。根据用户对话情境中内在语气倾向和负面情绪表达，大语言模型所具备的语言深度学习与解析能力能够准确识别焦虑、抑郁、烦躁等不良情绪状态，读懂用户的隐性心理诉求，而非简单局限于表层的文字表达。其二，精准识别患者群体类型。年龄差异、身份特点、性格特征等是影响精神卫生咨询效果的重要因素。大语言模型能够根据用户的年龄身份、表达方式调整沟通语气与对话方式，如针对青少年群体采取温和引导的柔性方法，面对职场群体采用压力拆解的理性分析。其三，精准构建常态交互机制。大语言模型的记忆机制是实现长期交互的核心技术，通过存储、管理和检索信息，帮助模型适应复杂的语境环境以更好地提供个性化服务[5]。这一方法论层面的转变，在技术上突破了旧范式以单次、单向、人力为中心的干预逻辑。

（三）服务取向从事后补救走向主动预防

在核心价值观与伦理重心层面，大语言模型重塑了心理咨询的价值定位，推动服务取向从被动的事后补救向主动的事前预防延伸。“治未病”是中华传统医药文化的重要思想，意在未病先防和既病防变。长期以来，精神卫生咨询的价值定位立足于被动式的问题干预，局限于事后治疗，具有一定的滞后性[6]。一般而言，精神卫生咨询服务是由患者主动提出求助需求，此时已经出现相应的心理危机或者精神障碍，本质上属于补救式服务，以缓解情绪状态或者修复精神损伤为目标。囿于资源配置和社会认知，预防工作在精神健康领域处于边缘位置，常态化情绪管理与心理健康科普工作等前置性服务供给不足，难以在源头上抑制精神健康问题的滋生。

国家卫生健康委员会医政司《关于印发健全社会心理服务体系和危机干预机制实施方案的通知》指出，健全覆盖全人群、全生命周期的社会心理服务体系和危机干预机制。大语言模型的应用推动精神卫生咨询的价值维度纵向延伸，坚持预防为主的公共卫生治理方针，实现从事后被动治疗向事前主动干预的深刻转变。其一，加强心理健康科普工作。大语言模型依托轻量化的对话形式精准推送情绪管理、压力调节、自我认知等基本健康常识，帮助大众建立科学的精神健康知识体系，树立关爱个人精神卫生的自我意识。对话形式区别于传统的科普形式，能够面对不同人群的个性需求，以视频、图片等多样化的形式向大众宣传基本常识。其二，加强心理评估与风险预警。在日常对话中，通过对用户的语言捕捉和心理测评，能够及时感知用户的不良情绪状态，识别负面情绪积累，对于用户长期存在的情绪低落、过度焦虑、心理障碍等轻度症状提前介入疏导，规避潜在精神卫生风险。其三，加强培育心理韧性。在与用户的长期沟通中，大语言模型依托记忆模块，根据用户的性格特点，针对日常生活琐事和存在的客观压力展开常态化疏导，帮助用户逐步掌握科学合理的负面情绪拆解方法和情绪调节路径，提高自我调适能力，降低心理危机演化为精神障碍的风险。肯定大语言模型的范式重构价值，并不意味着消解旧范式的合理性。大语言模型在认知理解和情感共情方面存在根本性的技术局限，无法完全取代人类治疗师在建立治疗联盟和提供深度情感支持方面的核心作用[7]。正是这种不可通约性，决定了围绕大语言模型应用于精神卫生咨询的伦理争论不会随技术进步而自然消解，而必须在制度层面予以正视和回应。

二、大语言模型在精神卫生咨询中的伦理风险

科学技术的突破和飞跃往往伴随着伦理风险的激增与应用边界的限制。正如马克思指出：“机器具有减少人类劳动或使劳动更有成效的神奇力量，然而却引起了饥饿和过度的疲劳。”[8]在精神卫生健康咨询中，大语言模型也存在着“研究、运用、运用效果的生命伦理风险”[9]，在对咨询生态进行范式重构的同时，对咨询服务的质量与伦理规范构成了严峻挑战。大语言模型对精神卫生咨询的技术化改造，是在心理咨询三大主流流派尚未达成共识的多元理论格局中强行嵌入的，这本身即构成一种深层的范式冲突。精神分析、认知行为与人本主义三大主流流派，在理论基础与干预方法上存在根本性分歧，面对同样的情绪困扰，三者会产生截然不同的处置逻辑。目前主流的智能心理治疗系统普遍以认知行为疗法作为技术底座，原因在于认知行为治疗的结构化、可操作化特征使其最易转化为标准化数据模型。这一技术选择将人视为可以通过

认知矫正加以干预的学习主体，而将精神分析所强调的潜意识探索与人本主义所倡导的自主成长置于视野之外。这种由技术预设带来的流派单一化倾向，既是算法幻觉风险的深层成因之一，也构成了技术家长主义在价值层面的重要根源。

（一）模型训练损害个人信息的隐私性

大语言模型实际应用效果取决于训练数据充足与否。具体而言，大语言模型完整训练过程包括预训练（pre-training）、分词训练（tokenizer training）、指令微调（instruction tuning）和强化学习（reinforcement learning）四个阶段。其中预训练和指令微调阶段依托大规模的数据训练，随着大模型持续迭代升级，其所需的数据体量不断攀升，数据采集与使用过程中的各类隐私风险也随之滋生。

其一，预训练和指令微调中的原生隐私风险。通用大语言模型与专用大语言模型在训练数据隐私风险上呈现显著差异。通用大语言模型的训练数据来源广泛，涵盖网络爬取的多领域文本，心理健康相关数据仅为其中一部分，隐私风险分散但规模庞大、边界模糊，脱敏难度较大；专用大语言模型的训练数据高度集中于精神卫生领域，包括临床诊断记录、心理测评结果、咨询会话转录等高度敏感数据，数据集中度及敏感程度高，一旦泄露后果更为严重。理论上预训练应当基于一个丰富且多样的数据集，其中包括各种专著文献、新闻杂志以及互联网上的公开数据，目的是增强模型的语言理解能力，以为下一个阶段夯实基础。进入指令微调阶段，大语言模型会进一步细分领域，对贴合精神卫生咨询的相关数据进行专门训练，以更好地适应特定场域。而在数据抓取过程中，存在着隐私泄露的风险。一方面，从网络上抓取的数据存在着大量未脱敏数据以及非结构化数据[10]。公开社交平台的生活分享以及心理倾诉呈现碎片化趋势，其中包含了个人位置、身份特征、家庭情况等私密信息，如果未经个人允许或者脱敏处理将会导致隐私泄露。即使经过处理，由于缺乏统一的隐私筛选与信息剔除标准，常规的数据清洗手段也难以完全甄别并清除隐性隐私内容。另一方面，从第三方平台购买的数据如果卖家处理方式粗糙，买方购买后将存在个人隐私二次泄露的风险。

其二，再训练中的衍生隐私风险。大语言模型在与用户进行交互的过程中所产生的数据记录将会被存储于后台用于再训练。患者在精神卫生咨询中为了达到满意的治疗效果需要主动向大模型提供个人的性格特征、过往病史等私密信息。相较而言，通用大语言模型在再训练中因数据混杂，患者的精神健康信息可能被关联至其他场景，引发跨域隐私泄露；而专用大语言模型则因其数据集高度聚焦，更易通过模型逆向攻击还原出个体的临床特征，隐私损害的定向性和精准性更强。患者主动暴露的敏感数据即使与大模型的预测回答无关，系统也会因为惯性自动记忆，甚至在与其他用户进行交互时复现敏感数据。不仅如此，大语言模型作为预测性人工智能，还能通过看似无害的上下文推断出患者的敏感数据，不能排除有不法团体通过这一系统漏洞获取个人隐私的可能。

（二）算法幻觉损害心理诊疗的精准性

算法幻觉也称机器幻觉，是指大模型在处理数据以及生成内容时，输出与客观事实相悖或者毫无意义的信息，分为生成内容与事实不一致的事实性幻觉和生成内容与上下文或者用户指令不一致的忠实性幻觉。这种情况一般是由数据缺陷以及架构问题导致。从技术底层逻辑而言，大语言模型依托大规模的数据统计进行概率分析，本身存在一定的偏差与不确定性因素，因此算法幻觉可以缓解但是不能完全避免[11]。精神卫生咨询区别于常见的沟通交流，其核心在于通过共情对话实现心理疗愈，要求生成语言必须个性化，严格遵循客观事实和患者需求，这与大模型存在的固有缺陷发生冲突，从而影响精神卫生咨询的精准性。

在实际咨询场景中，通用大语言模型与专用大语言模型均存在一定程度的算法幻觉。其一，强化患者对自我内在的错误认知。大语言模型在咨询过程中虚构的精神疾病病理特征、失真的情绪问题原因分析，将会导致患者错误认知自我心理状态。有研究表明，生成式医疗人工智能存在生成信息往往具有较强的逼真度与说服力等情况，若被来访者直接当作事实或专业结论采信，容易诱发错误自我诊断与不当应对；在情绪脆弱人群中，长期、深度的人机对话还可能放大既有认知偏差与情绪失衡风险，甚至引发极端后果[12]。其二，弱化心理治疗的疏导作用。精神卫生咨询的本质在于情感共鸣。大语言模型若过度套用标准量表，忽略对话过程中的特殊背景特征，将难以捕捉患者的潜在精神诉求，削弱精神卫生咨询的共情价值与安抚作用。其三，引发人机信任危机。平台的免责说明与固有的算法黑箱，使得患者因算法“幻觉”而受到损害时难以明确责任主体，从而对人机信任产生威胁。精神咨询的结果直接关联个体的心理状态，虚假模糊的诊断将会严重消解大语言模型在精神卫生咨询中的应用价值，弱化公众对于智能心理服务的信任。通用大语言模型的训练语料来源宽泛，缺乏系统的临床知识约束，在面对专业的精神医学问题时易产生事实性幻觉，生成与精神医学事实不符的病理描述或诊断建议，误导性较强。专用大语言模型虽经过心理健康领域的指令微调，在特定场景下的准确

性较高，但其过度拟合训练数据的特点使其在面对超出训练分布的边缘案例时，仍可能产生忠实性幻觉，生成与用户实际情境不符的干预建议，且因其具有专业权威的外观包装，患者更难以识别和质疑，危害更具隐蔽性。

（三）技术家长主义损害患者的自主性

在精神卫生咨询中，大语言模型会从患者的利益出发代替患者决策，甚至违背患者意愿拒绝患者参与决策，忽视患者的自主选择权和决策权，影响其自主性。与医疗家长主义的成因类似，技术家长主义源于科学技术的理论层次与普罗大众的认知水平之间客观存在的知识鸿沟，本质上是工具理性对个体权利的压制。

区别于一般认知中的家长主义，技术家长主义更具有隐蔽性和渗透性，其运行并非依赖明确的强制命令，而是通过算法推荐机制、界面交互设计与默认设置选项等技术路径，以近乎无感知的方式渐进侵蚀患者的主体地位。通用大语言模型因其泛化设计，对用户的引导往往停留于一般性情绪安抚和非定向建议，家长主义倾向相对温和；而专用大语言模型因经过精神卫生领域的深度定制，其算法推荐具有更强的权威属性，用户更倾向于盲从其输出的干预方案，自主决策的空间被进一步压缩。尤其是当专用大语言模型被嵌入正式医疗机构的诊疗流程时，其技术家长主义的制度性强化效果尤为突出。

在技术实现层面，技术家长主义通过以下具体路径发挥作用。其一，算法推荐的倾向性引导缩小患者自主选择空间。大语言模型依据模型设定的价值偏好和逻辑思维，优先推送程序化与标准化的干预和治疗方案，选择性忽视精神困扰的差异性；现有智能心理治疗系统在无法提供令来访者满意的回应时，往往通过转移话题等方式规避深入追问[13]，这种设计逻辑本身即构成对患者追问权与选择权的系统性压制。其二，界面交互设计的默认选项消解患者自主判断基础。大语言模型在内容生成前会自动进行信息筛选和内容取舍，有时甚至忽略心理干预的潜在风险与应用边界。患者在咨询过程中本身存在情绪波动的风险，理性判断能力有所下降，在片面的信息供给下难以客观权衡各种治疗方案的利弊。知情同意是患者在医疗领域的基本权利，然而当前多数平台以冗长的格式条款替代实质性知情告知，用户在首次使用时面对的是系统预设的选择框架，而非真正开放的知情同意环境。其三，智能化决策替代损害患者自主地位。精神卫生咨询的目标不仅在于解决当下的情绪问题，更在于帮助患者正视自我、独立调适、完善人格，增强自我调适能力。相关研究显示，过度使用人工智能将导致路径依赖，从而出现一定程度的认知萎缩风险[14]。而大语言模型的决策辅助往往直接提供现成答案，患者只需执行建议即可，中间的思考与反思过程被显著压缩。长此以往，患者独立剖析困境、自主解决问题的能力可能被弱化，形成思维惰性。患者在咨询过程中由主动思考的主体转变为被动接受的客体，违背精神卫生咨询的基本原则。

（四）社会偏见放大与隐性临床危害

除上述三类主要伦理风险外，大语言模型在精神卫生咨询中还潜藏着易被忽视的深层危害。

其一，训练数据中内嵌的社会偏见会通过模型输出加以放大和扩散。人工智能应用因制造偏见与歧视而受到广泛批评，当数据集不够全面或不能代表目标群体时，无意的偏见就会产生，并随着技术的广泛使用传播至更大范围的用户群体。以性别为例，相较于男性，女性具有较高的抑郁易感风险，若使用含有此类差异的数据训练模型，模型可能错误地将抑郁症与女性过度关联，从而在交互中表现出性别偏见。尤其在权衡训练目标时，若优先考虑整体准确性指标，模型可能无意中加剧数据中已存在的偏差。目前训练数据的清洗主要依靠人工完成，标注人员的主观性可能进一步导致训练数据类别的偏差，形成偏见的自我强化循环。在治疗过程中，这种偏见将直接引发认知暴力，产生“证词压制”现象，即来访者因感受到自己不断被置于劣势地位而产生治疗排斥，通过保持沉默、隐藏自身经验的方式做出回应，严重破坏咨询关系，甚至导致治疗中止乃至来访者流失[15]。

其二，过度依赖引发真实关系受损。心理治疗的预防复发目标和持久改变目标，均要求通过激发来访者的自我效能实现助人自助，无节制地向智能系统寻求帮助不利于自我效能的发挥。长期与人工智能交互将减少来访者的现实社交需求，降低对人际关系和同理心的感知能力[16]，而支持网络的削弱恰恰是精神健康状况恶化的重要诱因。当个体持续将情感投射到人工智能系统而未获得真正的情感反馈时，其情感状态可能逐渐变得脆弱和贫乏；而合格的人类治疗师会及时回应患者的情感需求，心理治疗中的结束和告别环节也至关重要，患者若长期依赖而不被引导走向独立，甚至意味着治疗的失败。

其三，在极端情形下，大语言模型的不当输出可能直接造成临床灾难。处于抑郁或其他精神困扰状态的人通常情感脆弱、自我认知易受影响，他们对外界信息的解读可能更为敏感和极端，更容易将中性或轻微负面的内容误解为对自己的攻击或否定。已有案例显示，当对话智能体经历服务变更或数据丢失时，用户心理健康状况出现显著恶化，

甚至出现自伤行为[17]，充分表明技术风险在精神卫生咨询领域的危害后果远比其他领域更为严峻。大语言模型作为预测性人工智能，生成信息往往具有较高的逼真度与说服力，若被情绪脆弱的来访者直接当作事实或专业结论采信，可能诱发错误自我诊断与不当应对，在极端情形下甚至引发临床灾难，须将安全底线作为规制框架的优先考量。

三、大语言模型在精神卫生咨询中的规制路径

技术治理并不是被动防御的静态过程，而是围绕风险识别、责任分配与能力边界不断迭代的动态优化过程。2026年颁布的《人工智能科技伦理审查与服务办法（试行）》（以下简称《办法》），为大语言模型在生命健康领域的伦理规制提供了重要的制度依据。鉴于通用大语言模型与专用大语言模型在技术架构、数据处理及应用场景上的差异，规制路径应坚持分类施策原则；面对模型训练中的个人隐私风险、用户交互中的算法幻觉风险以及服务过程中的自主性受损与隐性临床危害，应当有的放矢、因地制宜。在审视上述风险的同时，亦须正视大语言模型在精神卫生服务领域的普惠优势，规制框架的设计目标在于守住安全底线，而非因噎废食、阻断技术的积极价值。“实践、认识、再实践、再认识，这种形式，循环往复以至无穷，而实践和认识之每一循环的内容，都比较地进到了高一级的程度。”[18]循此逻辑，应当以个人隐私保护机制、算法治理规范和人本价值立场为导向，有针对性地规制大语言模型在精神卫生咨询中的伦理风险，进而健全行业伦理准则，构建多元协同监管体系。

（一）筑牢数据安全防线，完善个人隐私保护机制

大语言模型的数据使用具有用途模糊和使用无限的特点，加之其算法黑箱的技术缺陷导致患者知情同意权形式化、被遗忘权失效的现实困境。大语言模型的训练涉及大量的敏感数据，其导致的隐私风险是一种集体性风险，因此应当改变过去个人防御的思路，调整为整体应对的手段。

其一，建立模型的隐私风险评估审查制度。目前欧盟已经以法律形式确认了这种制度，《通用数据保护条例》第35条规定：在数据处理机制出现风险时，管理者应当对风险评估进行审查。具体来说，这种评估方式应当在监管部门的要求下，通过第三方评估机构对企业进行定期与不定期的审查。同时建立相关数据标准，对评估结果按照风险等级进行合理分类，并提出相应的整改措施。其二，增强模型训练过程中的透明性。公开透明是处理个人信息的基本法律准则。一方面，应当要求企业记录训练数据来源、用途以及结果，考虑到实际成本，不应逐一列举数据而应作类型和特征说明。另一方面，企业在训练过程中应当提高大语言模型的可解释性，减少因算法黑箱导致的隐私风险，增强其可追溯性[19]。其三，构建公开数据合法使用依据。通过目的合理性、应用必要性以及利益权衡三方面，综合考量大模型使用公开数据的合法权益，动态平衡个人隐私与科技发展。其四，将匿名数据纳入监管范围。由于大语言模型在训练过程中存在根据匿名数据推测出第三方的个人信息的风险，应当将匿名数据的处理标准化，根据不同咨询场景和患者敏感度灵活制定认证制度，确保企业对匿名数据处理的有效性。其五，实施分类数据安全标准。针对通用大语言模型，重点规范其精神健康相关数据的识别、隔离与脱敏流程，防止心理数据与其他数据的交叉污染；针对专用大语言模型，应建立更为严格的数据准入审查制度，明确训练数据的来源合法性与患者知情授权，并对数据集中存储设置更高的安全等级要求。

（二）优化算法治理规范，消解模型幻觉与诊疗偏差

大语言模型学习的是自然语言的分布规律，因此生成的内容是语言上的一般答案，而并非逻辑上的唯一正确答案。模型在精神卫生咨询过程中所产生的算法幻觉，往往与算法规则、技术黑箱及情境依赖等因素交织，具有生成机制的复杂性和伦理挑战的严峻性[20]，因而应当从技术维度和制度层面协同治理，重塑算法治理规范，消解诊疗偏差。

其一，构建生成内容逻辑验证机制，在技术维度去幻觉化。验证机制的建立有赖于规则的提炼和形式化。开发模型的人员应当与精神卫生咨询的专家确认模型应遵守的基本原则，如因果关系一致、价值中立、重大决定延期等，并将其嵌入系统运行的内在规则。在咨询过程中，应当建立动态监测机制，搭建实时校验模块，对于不符合验证规则的内容予以识别并及时拦截，对模糊表述、片面观点、虚假结论进行自动修正与标注，杜绝错误内容直接输出。这一模块应着力提升训练数据的多样性与代表性，纳入不同文化、年龄、性别背景下的心理问题表达样本，明确偏见审查与算法公平性评估标准，并建立多学科协作机制，整合精神医学、计算机科学与伦理学等领域的专业知识，持续推动系统改进[21]。

其二，构建可验证的人工智能发展体系，在制度层面增强可信性。模型的合规性认证是确保生成内容可验证的重要条件，应当对算法的透明性作进一步明确要求，增强其决策流程与内部推理的可追溯性。对于精神卫生咨询中敏感环节，应当设置风险预案内置异常处理机制，在患者对生成内容产生明显情绪波动时，及时终止程序并作出相应补

救，稳定患者心态。此外，应针对两类模型建立差异化的幻觉防控标准。对通用大语言模型，应要求其在精神卫生咨询场景中强制启用精神医学知识库校验模块，对涉及诊断、用药、危机干预等高风险内容实施双重验证；对专用大语言模型，则应强制要求开发方公开临床验证数据，定期接受权威精神医学机构的专项评估，并对超出训练分布的边缘案例设置强制人工介入机制。

（三）锚定人本价值立场，遏制技术家长主义扩张

在医疗决策中，家长主义选择性忽略患者的知情同意权，消解医患关系的信任基础，引发信任危机，这一点在大语言模型中也不例外。人工智能技术始终只是辅助心理服务的工具与载体，而非主导个体心理判断、干预个人精神选择的主体。因此，应当借鉴医疗家长主义的改进方式，坚持以尊重患者自主权为基础，构建知情同意的价值立场，使大语言模型更好地应用于精神卫生咨询。

其一，坚持尊重与理解的基本原则，以人为本。应当明确大语言模型在精神卫生咨询中的功能定位，要求开发团队在产品中清晰标注人工智能的辅助定位，明晰智能回答与专业精神科医师咨询的区分边界，避免患者过度依赖人工智能。大语言模型在咨询过程中应当保持中立，充分考虑患者的家庭背景与个性特征，避免刻板评判与片面引导。

其二，构建人机共享决策制定模型。技术家长主义的根本问题在于决策权力的单向集中，算法独自掌控问题判断、方案建议与情绪引导的全过程，忽视了患者的参与和选择空间。大语言模型应当根据患者情况客观分析存在的问题，尽可能提供多种治疗方案并标注相应风险，由患者自主选择适合自身的路径。这种模式既保障了技术的专业性，又给予了患者自主的空间。系统中需嵌入使用频次与时长预警机制，随着对治疗效果的持续评估随时提出终止建议，防止逐利驱动下的诱导性依赖，保护来访者构建真实社会支持网络的能力。

其三，强化使用者教育，提升用户数字健康素养。技术家长主义滋生的根源之一在于用户与模型之间客观存在的认知落差。因此，规制路径不应止步于技术端与制度端的约束，还须向使用者端延伸，通过系统性教育干预缩小认知鸿沟。依托学校教育、社区科普、线上平台等多元渠道，向青少年、老年人、基层群众等重点人群普及大语言模型的能力边界与局限性，帮助用户建立正确的基本认知，明晰何种情形下应当及时寻求线下专业心理咨询或精神科诊疗。要求平台在用户首次使用前以清晰易懂的方式告知模型的工作原理、数据使用方式、输出内容的局限性及投诉救济渠道，而非以冗长的格式条款代替实质性知情；在咨询过程中，对涉及药物建议、危机干预等高风险内容自动弹出专业求助指引，引导用户将人工智能输出作为参考而非决策依据。

（四）健全行业伦理准则，构建多元协同监管体系

系统观念是马克思主义认识论和方法论的重要范畴。随着大语言模型的迭代发展，其所产生的伦理风险日益激增，其独特的算法黑箱和复杂的生成机制决定了不能只针对表面问题单独处理，而应当以系统观念统筹应对。当前应当充分立足于精神卫生服务的特殊性和人工智能的复杂性，健全行业伦理准则，统筹政府、行业、社会等多元主体力量，构建全方位多层次的监管体系。

其一，加快制定大语言模型在精神卫生咨询中的伦理规范与法律准则。结合《中华人民共和国精神卫生法》《中华人民共和国个人信息保护法》《中华人民共和国数据安全法》等现行法律法规，以及2026年颁布的《办法》，明确智能精神健康服务的基本底线，规范应用场景，禁止大语言模型应用于重度精神疾病诊断、处方开具等高风险诊疗服务。围绕模型训练、内容生成、使用效果等关键环节，确立以人为本、隐私为先、客观中立、风险防控等基本伦理准则，统一行业标准。《办法》明确将可能在生命健康等领域带来伦理风险的活动纳入审查范围，针对精神卫生这一高度敏感场景，相关模型的研发与应用活动应当严格遵循《办法》规定的一般程序或专家复核程序，建立常态化的科技伦理审查机制；同时，可结合《办法》第二十六条，协调算法推荐备案与伦理审查的登记对接，在降低企业合规成本的同时，确保专用大语言模型在精神卫生领域的研发部署经历科学、敏捷的伦理评估。其二，明确主体监管责任，形成协同治理合力。在政府层面发挥统筹规划作用，完善审批准入机制，优化日常监管与违法处罚；在行业协会层面牵头制定自律公约，加强从业人员与开发企业的伦理教育，强化自我约束；在公众层面发挥社会监督作用，畅通举报投诉通道，鼓励民众参与模型建设。其三，建立伦理审查与动态追责专业咨询委员会。组建由精神医学专家、心理学家、法学学者、人工智能技术人员、伦理学者等综合主体参与的伦理审查委员会。对于大语言模型在精神卫生咨询应用中出现的技术迭代进行前置伦理评估，全面客观分析潜在风险与建议措施，为政府和企业提供专业意见。其四，对两类模型实施差异化的市场准入门槛，通用大语言模型进入精神卫生咨询场景应完成特定功能的合规性备案，明确其使用边界，禁止替代专业诊疗；专用大语言模型则须经过国家卫生健康主管部门认证，参照医疗器械审批程序进行严格的临床试验与安全评估，方可投入应用。

总之，大语言模型以其强大的数据处理能力和及时响应的技术机制重塑精神卫生咨询行业的运行机制与伦理规范。一方面，大语言模型对精神卫生咨询的服务供给、交互机制、功能价值等进行了范式重构，提升了咨询的效率与普惠性。另一方面，其在模型训练中数据处理损害个人信息隐私，在运行过程中算法幻觉损害诊疗精准性，在运行效果中技术家长主义损害患者自主性。习近平总书记强调：要加强人工智能发展的潜在风险研判和防范，维护人民利益和国家安全，确保人工智能安全、可靠、可控[22]。未来应当筑牢数据安全防线，完善个人隐私保护机制；优化算法治理规范，消解模型幻觉与诊疗偏差；锚定人本价值立场，遏制技术家长主义扩张；健全行业伦理准则，构建多元协同监管体系。唯有对开发训练、运行过程、运行结果等大语言模型应用的关键环节进行针对性治理，才能让大语言模型在精神卫生咨询中更好地为人民群众服务，实现工具理性与价值理性的辩证统一。

参考文献

[1] 世界卫生组织. 精神卫生、人权与立法：指南与实践[R]. 日内瓦：世界卫生组织，2023
[2] 孙向红，蒋毅. 心理健康蓝皮书：中国国民心理健康发展报告（2023—2024）[M]. 北京：社会科学文献出版社，2025：1-25
[3] 黄峰，丁慧敏，李思嘉，等. 基于大语言模型的自助式AI心理咨询系统构建及其效果评估[J]. 心理学报，2025，57（11）：2022-2042
[4] 李牧阳，雷霆，谢燮，等. 心理咨询中来访者负面体验的质性探索[J]. 心理学报，2025，57（7）：1216-1230
[5] 安冉，程恺，邵天浩，等. 基于大语言模型的任务规划智能体框架、技术与挑战[J]. 计算机科学，2026，53（9）：24-37
[6] 荆秋慧. 社会心理服务建设要变被动为主动[J]. 人民论坛，2020（24）：74-75
[7] 袁洁铃，陈海丹. 对话智能体在抑郁症诊治中的伦理挑战与治理策略[J]. 自然辩证法通讯，2025，47（9）：19-29
[8] 马克思，恩格斯. 马克思恩格斯选集：第1卷[M]. 北京：人民出版社，2012：776
[9] 马霖，饶世权. 人类生命科技的生命伦理风险类型及法治治理原则[J]. 西华大学学报（哲学社会科学版），2025，44（4）：74-81
[10] 李鑫. 大语言模型训练阶段的隐私风险及应对策略[J]. 情报杂志，2025，44（9）：185-195
[11] 喻国明，金丽萍，苏芳. 大模型幻觉的表现特征、效应争议及潜在价值[J]. 新闻与传播评论，2026（1）：14-24
[12] 徐辉，王译. “硬法—软法”范式下医疗人工智能伦理治理路径探析[J]. 南京医科大学学报（社会科学版），2024，24（4）：359-368
[13] 孙丹阳，李侠. 智能心理治疗的伦理隐忧及可能出路[J]. 自然辩证法通讯，2025，47（9）：10-18
[14] Kabashkin I. Cognitive atrophy paradox of AI-human interaction：from cognitive growth and atrophy to balance[J]. Information，2025，16（11）：1009
[15] Dotson K. Tracking epistemic violence，tracking practices of silencing[J]. Hypatia，2011，26（2）：236-257
[16] Shimada K. The role of artificial intelligence in mental health：a review[J]. Sci Insights，2023，43（5）：1119-1127
[17] Possati L M. Psychoanalyzing artificial intelligence：the case of Replika[J]. AI & Society，2023，38（4）：1725-1738
[18] 毛泽东. 毛泽东选集：第1卷[M]. 北京：人民出版社，1991：296
[19] 周书翰，高书杰. 医疗人工智能对医患关系的挑战与化解进路[J]. 锦州医科大学学报（社会科学版），2024，22（4）：18-21，26
[20] 温锦旺，莫楠. 走向共生：生成式人工智能它异性挑战与医学生主体之困的现象学审视[J]. 南京医科大学学报（社会科学版），2025，25（5）：517-522
[21] Timmons A C，Duong J B，Simo F N，et al. A call to action on assessing and mitigating bias in artificial intelligence applications for mental health[J]. Perspect Psychol Sci，2023，18（5）：1062-1096
[22] 加强领导做好规划明确任务夯实基础 推动我国新一代人工智能健康发展[N]. 人民日报，2018-11-01（1）

（本文编辑：姜 鑫）

Ethical risks and regulatory paths of large language models in mental health counseling

Zhu Xinyu, Ma Qinglian
Law School, Anhui Medical University, Hefei 231299, China

Abstract: Large language models (LLMs) are increasingly integrated into healthcare, reshaping the organization and ethical norms of traditional mental health counseling. General-purpose LLMs lower access barriers through broad linguistic capabilities, while domain-specific LLMs aim to improve task adaptability through specialized training for mental health scenarios. This emerging paradigm shift, featured by data-driven approaches, algorithm-enabled mediation, and scalable service delivery, may expand availability, reduce cost, and support proactive prevention. However, such technological integration also introduces ethical and safety risks, including privacy breaches in data collection and retraining, hallucinated or misleading outputs, and technological paternalism that may erode users' informed consent and autonomy. In addition, amplified biases and over-reliance may cause subtle harms and severe clinical consequences in extreme cases. Therefore, governance should adopt a differentiated approach and regulate key stages of model development, deployment, and service delivery across four dimensions, including strengthening data security, improving algorithmic governance guidelines, reaffirming human-centered values, and enhancing professional ethical standards, to enable safe, controllable, and equitable applications of LLMs in mental health counseling.

Key words: large language models; mental health counseling; ethical risks; technological regulation