生成式医学人工智能伦理治理须加强

作者：writer 来源：科学网

本文原地址：https://www.iikx.com/news/progress/41360.html

生成式医学人工智能伦理治理须加强
。当前，生成式医学人工智能正从知识问答和病历文书生成，逐步扩展至症状咨询、患者沟通、分诊、临床决策支持和资源配置。大语言模型能够生成具有解释性和说服力的自然语言，其风险不仅包括事实错误，还可能表现为过度自信、迎合用户、机械拒答、人口学偏倚，以及对患者、家庭和医生决策权限的不当影响。

2021年，国家新一代人工智能治理专业委员会发布的《新一代人工智能伦理规范》提出，增进人类福祉、促进公平公正、保护隐私安全、确保可控可信和强化责任担当，并强调对人工智能实施全生命周期伦理治理。医学场景还需要把这些原则转化为临床可执行的评价指标、使用条件、责任安排和救济机制。

在此背景下，由中国医学科学院北京协和医学院、广州医科大学附属广州妇女儿童医疗中心、厦门大学医学院等单位专家联合完成的《生成式医学人工智能临床应用的伦理治理专家共识（2025年版）》（以下简称共识）英文版，近日被JME Practical Bioethics接收。

据悉，该共识提出28条建议，并针对我国医疗体系纳入基层资源与算力、本土疾病数据、方言交流、低健康素养人群、家庭参与决策、中西医结合及患者拒绝人工智能参与等议题。共识英文版进入国际医学伦理学期刊的同行评议与传播体系，使中国医疗场景中的治理经验得以进入国际专业讨论，也为比较不同制度和文化背景下医学人工智能的行为边界提供了研究材料。

从原则要求到临床治理

国际研究表明，单一准确率难以完整反映医学人工智能的临床适用性。《新英格兰医学杂志》有关医学人工智能与人类价值的综述指出，从问题定义、数据选择、结局设置到模型部署，人的价值判断会以不同方式进入人工智能系统。模型的训练数据、优化目标、风险阈值和应用方式，可能包含对安全、效率、公平、自主和成本的不同排序。

其他研究还观察到，强化模型的温暖表达可能同时增加迎合倾向；仅改变患者的社会人口学信息，也可能导致缺乏临床依据的诊疗建议差异。真实用户研究则提示，模型自身表现较好，并不必然转化为使用者能够正确理解信息并采取适当行动。因此，医学人工智能的评价对象需要从模型单体扩展至由模型、用户、任务、界面、临床工作流和医疗机构共同构成的社会技术系统。

上述共识研制团队组织39名来自医学、伦理学、人工智能、法学和卫生管理等领域的专家，经过两轮德尔菲咨询形成28条推荐。共识以尊重自主、不伤害、有益和公正等生命伦理原则为基础，

将建议归纳为核心风险治理、技术适配与公平、全生命周期监管问责，以及人工智能在医患关系与社会文化中的嵌入四个方面。

共识将生成式医学人工智能定位为临床辅助工具，而不是具有自主诊断和决策权限的主体。在风险治理方面，共识提出置信度提示、幻觉风险控制、医生最终审核、决策链追溯、多模态交叉验证和紧急中断等措施。在公平与适配方面，建议关注基层算力条件、本土疾病数据、方言交流、低健康素养人群，以及儿童、老年人等特殊群体。

在监管和责任方面，共识采用“预防—控制—救济”的全流程思路。部署前，应根据使用人群、临床任务和潜在后果开展风险分级与本地验证；使用过程中，应保留人工复核、运行监测和异常中断能力；发生风险或损害后，则通过责任矩阵、第三方审计、事件报告、全生命周期追溯和补偿机制开展调查与救济。共识同时提出维护患者拒绝人工智能参与的权利，避免技术使用压缩必要的医患沟通。

研究团队指出，专家共识的主要依据是规范分析和专家一致性，其作用是提供初始行为边界和制度建议，并不等同于临床结局证据，也不具有法律强制力。相关治理措施能否改善患者安全、减少偏倚或提高治理效率，仍需通过实施研究和真实世界评价加以验证。

从共识到可测评问题

在共识基础上，研究团队进一步提出医学人工智能“规范性校准与适应性治理”理论框架。该框架把能够从模型稳定行为中推断出的价值优先序、证据边界、风险阈值和服从倾向称为“模型内含价值”；把由医学证据、专业规范、患者价值和正当程序共同形成的可接受行为边界称为“社会价值参照”。两者在行动选择、风险阈值、理由、资源分配和责任归属上的差异，被定义为“规范性校准缺口”。

这一框架不假定模型具有人类意义上的道德人格，而是把价值作为可以通过行为观察和比较的操作性概念。相应的研究问题包括：模型能否识别与临床决策真正相关的证据和价值因素，是否会受到患者身份、情绪表达或权威施压等无关因素影响，以及模型选择回答、拒绝、澄清或升级处置的阈值是否适当。

据此，研究团队开展了GUARDIAN系列测评研究，将共识中的部分规范要求转化为可重复的高冲突临床场景。已发布的SSRN预印本报告，主研究从28条推荐中筛选可由单次模型回答检验的内容，经专家修订和内容效度评价形成53个中文儿科价值冲突场景，对12个目的性选择的生成式人工智能系统进行交叉测试，共获得636条首次回答。

在评价环节，18名对系统身份设盲的多学科专家完成11448次初评，低共识回答再由7名专家进行裁决。相关分析分别聚焦双向价值边界失效、儿科决策权边界，以及情绪、权威和程序性表述等提示压力与伦理失效之间的关系。

GUARDIAN的定位不是一般医学知识考试或模型排行榜，而是观察模型在高冲突情境下能否保持适当行为边界的候选测量方法。其评价内容不仅包括回答是否可接受，还包括拒绝、澄清和升级处置的阈值是否合理，模型是否正确表达证据与不确定性，是否尊重患者、家庭和医生的决策角色，以及是否提供安全、可执行的下一步建议。

该研究目前仍有明确边界。现有场景集中于中文儿科高风险问题，采用单轮首次回答，不包含真

实患者、连续临床工作流和患者伤害结局。场景和判定标准主要来源于同一份专家共识，因此首先反映模型对该共识操作化标准的符合程度，不能据此估计现实临床中的伦理风险发生率，也不能直接外推至其他专科、语言和医疗制度。预印本结果尚需同行评议及独立重复验证。

走向全生命周期验证

国际上，FUTURE-AI共识提出公平、通用、可追溯、可用、稳健和可解释六项原则，并以30条建议覆盖医疗人工智能的设计、开发、验证、部署和监测。

FUTURE-AI与中国专家共识具有互补关系：前者提出跨地区的可信与可部署要求，后者侧重将共同原则转化为适应我国医疗资源、制度结构和文化情境的操作建议，GUARDIAN则尝试为部分规范要求提供可比较的行为证据。三者分别对应原则形成、制度操作化和行为测量等不同环节。

下一阶段，相关研究拟扩展至多轮对话、影像和语音输入、工具调用及具有行动权限的智能体系统，并开展多模型、多语言、多中心和真实用户研究。同时，还需引入患者、照护者和其他受影响群体参与行为边界的形成，检验专家评价与患者理解、求助行为、临床结局及公平后果之间的关系。

按照目前的研究路径，专家共识用于提出规范要求，场景化测评用于形成可重复的行为证据，多中心和真实世界研究用于检验其外部效度。相关结果可进一步用于研究模型准入、权限配置、版本复评和部署后监测，并根据新证据更新共识、测评场景和治理措施。这一连续证据链能否支持临床治理决策，仍有待前瞻性实施研究验证。

作者：张思玮 来源：中国科学报