李伦

湖南师范大学人工智能伦理研究中心主任李伦：

用伦理工程学方法防范 AI 风险

过去我们谈及数字风险，多认为问题来自人为操作或制度漏洞；而当下人工智能（AI）技术的发展呈现出一种新的风险特征，即“技源性”。这意味着风险源于技术本身的内在属性，而非人类主观上的不当设计或使用。

例如，“算法黑箱”源于技术的内在缺陷或不成熟，属于“技源性”风险；而“算法黑幕”则是人根据特定目的进行的不当设计，比如“大数据杀熟”，技术本身不存在缺陷。

面对技术内生风险，我提出“伦理工程学”治理方案——一种“以内嵌伦理框架的技术方法防范技术伦理风险”的方案。我在 2017 年提出“给 AI 一颗良芯”的口号，追求的便是从技术内生属性出发进行治理。2024 年，我进一步明确提出“伦理工程学”这条“技术 - 伦理”互嵌式治理路径。

伦理工程学不是纯粹的工程学，也不是纯粹的伦理学，而是“技术 - 伦理”互嵌式的方法。它旨在破解伦理学与工程学长期二元割裂的困境：伦理学倾向于宏观原则建构，缺乏实现路径，工程学则关注技术能力迭代，导致“技术先行，伦理善后”。而伦理工程学强调“伦理先行”，事先既要有伦理框架，又要有技术落地路径。

伦理工程学的实施分为两步。第一步是伦理框架的设计。根据技术系统的“技术 - 伦理”目标（如自动驾驶与手术机器人的目标不同）提出适宜的框架。这需要伦理学家主导，并与技术专家等展开合作。第二步是伦理框架的工程化嵌入。通过技术方法将伦理框架嵌入技术系统，使技术系统成为道德行动者，使其行为后果具有道德意义，或成为道德调节者，规范人的行为。

伦理工程学的思想源远流长，可追溯至中国传统儒家的“藏礼于器”。器物是礼制的物化形态，例如宫殿、龙袍龙椅的设计都贯彻三纲五常。这些器物设计完成后，反过来又规范人的行为，使人遵守礼制。这其实是典型的伦理工程学方法。

AI 与伦理工程学具有极强的亲和性：首先，AI 风险多源于技术内生属性；其次，AI 具有“自为性”，规则预设性强，通过代码和算法设计很容易将伦理框架嵌入。此外，AI 算法迭代周期短，为纠正和预防伦理风险提供了及时性方案。目前，我国及欧盟都高度重视以伦理工程学方法防范 AI 风险。

Anthropic 公司在 2023 年引入了“宪法 AI（Constitutional AI）”方案，并在 2026 年 1 月将其升级为“Claude 宪法”。“宪法 AI”的过程分为两步：一是伦理框架的设计，即制定一套“宪法”——16 条自然语言规则（如“不鼓励非法活动”“尊重平等”）；二是伦理框架的工程化嵌入，通过监督学习（SL）和基于 AI 反馈的强化学习（RLAIF），让 AI 根据“宪法”自我演化。

Claude 宪法的升级主要表现在伦理框架的设计，新宪法包含具有优先级的四大原则、核心德性要求（如诚实、保护自主、透明等）以及硬性约束红线（如绝不能生成儿童性虐待材料等）。Claude 宪法的升级体现了从“机械执行”到“理解意图”、从“规则导向”到“推理导向”的转变。它不仅要求 AI 背诵规则，更强调通过详细的道理说明培养 AI 的价值观和独立判断能力，使其在面对新案例时具有泛化处理能力。不过，当前核心的难题在于“可验证性”的提高。Anthropic 公司坦言目前的训练方法难以保证模型百分百遵循宪法，需要配合其他措施共同防范 AI 风险。

伦理工程学的落地具有多样性，这主要体现在伦理框架设计和伦理框架工程化方案的多样性。前者与技术系统的“技术 - 伦理”目标的多样性相关，也与技术系统所处文化背景的多样性相关；后者与技术的发展水平以及技术路径的选择相关。这种多样性表明伦理工程学的落地具有复杂性，需要技术界和伦理学界等多学科多领域的深度合作。

张欣

中南大学湘雅医院党委书记张欣：

AI 赋能医疗决策必须坚持“以人为本”

医疗决策的本质，在于综合研判患者信息、循证医学证据、诊疗指南及患者个人价值观，进而制定科学合理的诊断与治疗方案。随着人工智能（AI）加速融入临床实践，传统“医生 - 患者”二元决策模式逐步转向“医生 - 患者 -AI”协同参与的新范式。在此过程中，必须明确 AI 的辅助工具定位，不能越位成为决策主体。

国家有关政策及医疗机构 AI 应用与治理专家共识均明确指出，医疗 AI 应坚持辅助定位，恪守“人机协同，以人为主”的原则。尽管 AI 有助于提升诊断效率、优化医疗资源配置、改善患者就医体验、提高医院管理效能，但最终诊疗决策权与医疗责任仍须由医生承担。

当前我国发展医疗 AI 具备现实必要性。一方面，医疗服务需求持续攀升，优质医疗资源总量不足且分布不均，医务人员工作负荷逐步加重；另一方面，AI 已在临床诊疗、患者服务、医院管理及科研创新等领域展现出显著的应用价值。

医疗 AI 带来效率提升的同时，也伴随着新的伦理风险。当前大模型高度依赖海量数据训练，这对医疗数据质量提出越来越高的要求。如果数据本身存在偏差，再通过模型训练和算法放大，可能形成新的诊疗偏差。与此同时，算法“黑箱”、模型可解释性不足、算法依赖、情感剥离与医患关系物化，以及医疗安全与医疗责任归属等问题，都成为医疗 AI 推广过程中必须面对的重要挑战。

随着患者越来越多地借助 AI 获取医疗信息，医生也必须主动学习和了解 AI 技术，在充分参考 AI 分析结果的基础上，结合自身临床经验和患者个体差异作出最终判断。未来的医疗决策，既不会是单纯依赖医生诊疗的传统模式，也不应演变为由 AI 主导的决策模式，而应构建一种医生与患者共同参与、AI 辅助支持的新型共享决策模式。

推动医疗 AI 健康发展，应重点遵循以下原则：一是进一步提升医疗数据的质量；二是提升模型透明度与可解释性，确保决策逻辑清晰可溯；三是保障公平可及与算法包容，防止技术加剧健康不平等；四是充分尊重患者自主权、知情同意权与决策参与权；五是坚守医学人文底色，让 AI 兼具技术效能与人文温度。

置身于 AI 时代，医生不仅要精进医学专业能力，更要提升 AI 素养、强化伦理意识，坚守职业道德底线，切实履行在医疗活动中的主体责任。唯有如此，AI 才能真正成为提升医疗质量与患者福祉的有力工具，而非取代医生价值的决策主体。