为人工智能立“东方魂”

筛选海量信息分析出用户的健康隐私、招聘系统由于数据偏差歧视特定群体、智驾系统判断失误酿成交通事故……近年来，通用人工智能赋能千行百业，在给大众生活带来切实便利的同时，也暴露出伦理风险。尤其是在西方技术逻辑下，通用人工智能的伦理风险有被进一步放大的趋势。当前亟待为通用人工智能装上伦理“导航仪”，注入“中国智慧”。

在北京石景山区中关村科幻产业创新中心，工作人员展示动作捕捉技术。（郑焕松摄）

伦理风险难回避

伦理问题的本质，是技术发展与社会规则之间的冲突。尽管通用人工智能的自主决策能力远超传统工具，但这并不意味着它可以随心所欲，仍需在制度框架内合理运行。

早在1942年，美国科幻作家艾萨克·阿西莫夫就提出了著名的“机器人三定律”，即机器人不得伤害人类，也不得因不作为而使人类受到伤害；机器人必须服从人类给予它的命令，除非这些命令与第一法则相冲突；机器人必须保护自己的存在，只要这种保护不与第一或第二定律相冲突。

尽管阿西莫夫畅想的机器人充满智慧的时代尚未到来，但人工智能时代面临相似境遇。缺少类似“机器人三定律”的框架约束，人工智能能否好用、可控，已经成为摆在眼前的现实问题。

人工智能如同一个不知疲倦的信息猎手，全方位收集个人信息：从日常出行的位置信息，到网络冲浪的浏览记录；从手机通讯录里的人脉关系，到每一笔支付背后的消费习惯。这些海量数据经过复杂的算法整合与深度分析，如同拼图一般勾勒出一个精准且详尽的“数字画像”，可能将百姓不愿公开的隐私信息暴露在外。

生成式人工智能在给社会提供便利的同时，也存在泄密风险。三星公司在引入ChatGPT后，短时间内曝出多起机密资料外泄事件。经调查，泄密事件是由三星员工直接将企业机密信息以提问的方式输入ChatGPT中所引起的，该做法导致三星公司半导体设备测量资料、源代码、产品良率等机密内容进入学习数据库导致泄密。

人工智能还颠覆了传统的“行为—责任”，即“谁犯错谁担责”的责任伦理观念。比如在智能辅助驾驶事故中，往往不易确切判断是哪个具体技术环节出现故障或过错，确定承担责任的主体存在困难。

人工智能编制的算法也存在不平衡，如在外卖、网约车等平台上，新就业群体在算法的指引下疲于奔波，是因为算法规划的路径没有考虑到劳动群体的权益保障。面对消费者，算法又容易形成“大数据杀熟”，忽视了消费的公平性与平台的社会责任。

在人类情感方面，人工智能创造的“虚拟伴侣”正重塑人际伦理关系，在部分年轻群体中甚至有取代真人的趋势。浙江大学传媒与国际文化学院副院长赵瑜佩认为，人工智能伴侣其实就是聊天机器人在情感陪伴方向的具体应用，本质上仍然是基于统计模型和模式匹配的“模拟”，也就是依赖于如今的“情感计算”技术所赋予的情感能力。

西方技术驱动模式存在局限

既然通用人工智能的自然发展存在伦理风险，那么究竟什么样的发展逻辑才是可行的呢？对此，北京通用人工智能研究院院长朱松纯表示，西方主导的“大数据+大算力+大模型”模式曾被认为是唯一的发展路径。不过，几十年的探索让我们清醒地认识到，数据与统计方法驱动的人工智能发展道路存在自身的局限性。

以当下爆火的生成式人工智能为例，在某款爆火的工具诞生之初，记者试用发现体验感并不好。例如，对“刘翔在哪一年夺得世乒赛冠军”这样一个显然错误的提问，它给出了“刘翔在2004年获得了世界乒乓球锦标赛冠军”的回答。重复一遍提问，又给出了2005年的答案。而对于“泰山是济南的著名景点吗”，它第一次的回答是“是的，泰山是济南市的著名景点。它位于山东省泰安市，是中国五岳之一，有着悠久的历史和丰富的文化。”相隔一段时间再次提出相同问题后，它才对答案进行纠正。时隔近一年再次提出类似问题，该工具如今已经能够很好应对，不会再出现类似低级错误。

中国艺术科技研究所数字艺术部主任张宜春认为，看到人工智能带给人类新纪元曙光的同时，也必须看到当前的大模型在涉及文化判断和价值取向的领域，输出结果仍不尽如人意，在一定程度上存在着“胡说”“胡写”“乱画”现象，干扰了人们对社会主流价值观念的认同与判断。

西方逻辑下的人工智能滥用还催生出一个新问题，那就是人类数据库的“污染”。这是一个渐进式的过程，隐蔽性强，不易被察觉。

技术越进步，其生成内容的真假就愈发难分辨，对传统世界形成“吞噬”。很难想象，未来人们检索的图片、数据、问答等有相当一部分是经过人工智能修饰过的，甚至包括动物的外貌、植物的外形、书画的内容。将这些“美化”过的内容与真实世界对比的时候，未来人们又会以怎样的心态来看待这个世界，做出怎样的判断？

根据最新研究，将由人工智能生成的内容反馈给同类模型训练，可能导致模型质量下降甚至崩溃。这种自我吞噬现象被科学家们称为模型自噬。研究人员指出，大模型算法在图像、文本等领域取得了巨大进展，但持续使用合成数据来训练模型会导致模型变得封闭，并最终失去多样性和准确性。

用“中国智慧”规划未来路径

防范化解人工智能伦理风险，需要主动预防及构建制度化措施。尤其是着眼未来，要立“东方魂”，坚持“以人为本”“智能向善”，促进人工智能算法的不断完善。

早在多年前，我国即围绕人工智能发展认真谋划，不断提升人工智能的安全性、可靠性、可控性、公平性。2023年10月，中国发布的《全球人工智能治理倡议》提出“以人为本、智能向善”的人工智能发展方向；2024年7月，第78届联合国大会一致通过由中国主提的加强人工智能能力建设国际合作决议，该决议进一步强调了我国《全球人工智能治理倡议》中提出的人工智能发展应坚持以人为本、智能向善、造福人类的原则，积极构建“以人为本、智能向善”的人工智能领域人类命运共同体。

基于治理优势，我国可以发挥有为政府的超前引领作用。专家建议，可以建立人工智能伦理风险敏捷治理机制。在制度设计、框架确立、过程管理等方面，建立敏捷的反馈系统，尤其要加强动态跟踪、科学研判、风险预警和伦理事件应急处置。健全人工智能伦理风险等级评估体系，细化分级分类评估标准，做到准确识别、精准评估。鼓励多元参与和协同共治，更快速、高效地协调各方资源，确保人工智能技术符合伦理规范。

作为人工智能领域的先行发展强国，“中国智慧”可以拓展为全球价值共识。在维护人类福祉、坚守公平正义方面，我国可主动作为，就全球普遍关切的人工智能伦理风险治理问题给出建设性建议，为人工智能伦理风险治理相关国际讨论和规则制定提供参考蓝本。努力推动各国在人工智能伦理治理中实现权利平等、机会平等、规则平等，尤其为发展中国家争取代表性和发言权。

除此之外，在人工智能领域不断丰富中文语料正逢其时。张宜春认为，如果把大模型的训练过程看成是一个嗷嗷待哺的婴孩茁壮成长的过程，那么抚育婴孩成长的乳汁就是高质量语料数据。如今，高质量语料数据已经成为推进大模型建设的核心生产要素。政府相关部门要加快公立文化艺术相关机构的语料库建设和开放工作，尽快将主流声音、主流意识注入互联网中，使中文原则在人工智能时代的全球网络中占据一席之地。

（据新华社电《瞭望》新闻周刊记者 毛振华）