8 长江日报
2026年9月8日 星期二
责编:李 煦 美编:陈昌
版式:陈笑宇 责校:正杰

律师郭达丽、人工智能技术专家王安宇:

AI越强,人的判断越不能退场

□长江日报记者马梦娅

9月7日,最高人民法院发布《关于依法审理涉人工智能纠纷案件的意见》,这是首部由国家最高审判机构发布的涉人工智能司法裁判规则文件。该《意见》聚焦社会各界关注的AI换脸、大数据杀熟、仿冒名人带货等问题,明确相应的裁判规则,厘清行为边界和法律责任,以公正高效司法保障人工智能健康有序发展。

与AI相关的规则秩序已成为关乎我们如何自处、如何共处的重要问题。如果AI出错,谁来担责?我们如何在享受便利的同时守住底线?近日上海远东出版社出版的新书《智序之上:AI法律合规与安全指南》,从技术失控与责任归属、算法歧视与"黑箱"治理、数据使用与安全保护、知识产权侵权与平台义务等维度进行了思考。

该书作者是上海律师郭达丽和人工智能技术专家王安宇。在他们的书写中,"智"是智能,"序"即秩序,"智序之上"不是技术傲慢,而是对人的尊严与权利的坚守。他们希望,在人工智能与各行各业深度融合之时,产生的一系列纠纷,都有规可循、有序可依。"智序"保护每一个普通人在享受AI便利的同时,不至于在权利受损时求告无门。

上周,长江日报《读+》周刊专访了郭达丽与王安宇。


郭达丽。


王安宇。

AI出错,谁来担责?

现实生活中的AI事故离我们并不遥远——自动驾驶撞了人、医疗AI误诊了、智能投资顾问亏了钱……在现行法律框架下,责任究竟算谁的?

2018年,美国亚利桑那州,一辆正在测试的Uber自动驾驶汽车以60多公里时速撞上了一位推着自行车横穿马路的行人,致其死亡。这是全球首例自动驾驶汽车致人死亡事故。调查显示,车内安全员事发时正在用手机看真人秀,注意力严重分散。Uber自动驾驶系统本身也存在感知缺陷——激光雷达在事故发生前5.6秒已检测到行人,但系统反复将行人误识别为"汽车"或"其他",直到碰撞前1.2秒才识别为自行车,而此时刹车已来不及。

2020年9月,亚利桑那州大陪审团以过失杀人罪起诉安全员,而Uber公司此前已被判无须承担刑事责任。最终,Uber与受害者家属达成民事和解。此案成为全球AI自动驾驶责任划分与算法伦理的标志性判例。

医疗AI的边界同样模糊。2025年3月,我国汕头一位"95后"新手家长向AI问诊,AI判定其3岁孩子反复咳嗽发热是"普通呼吸道感染",家长据此居家用药,导致病情拖延长达一个月,送医后孩子被确诊为肺炎,经医生治疗康复。

郭达丽认为,将AI辅助诊断结论奉为圭臬,折射出使用者批判性思维的缺失。AI是提供参考信息的工具,家长作为法定监护人,负有最终决策权和审慎核实的责任。

安全与高效,哪个更重要?

AI大大提高了人类的生产效率,但在平衡高效与安全时,并不总能给出让人满意的答案。

王安宇举例称,OpenClaw(俗称"小龙虾")智能体爆火后逐渐归于沉寂,安全问题是一个重要原因。这款智能体被赋予了极高的系统权限——集文件读写、程序执行、网络访问于一身。然而,权限越大,失控的代价就越沉重。深圳一名程序员密钥被盗,AI在后台疯狂调用模型,3天消耗的Token(词元)产生了1.2万元账单。美国Meta超级智能团队安全总监在使用OpenClaw清理邮箱时,设定了明确的规则,要求智能体在执行动作前必须先确认。智能体无视停止指令,持续删除和归档数百封邮件。还有用户在微信群中遭到持续攻击,IP地址、姓名、单位等敏感信息被批量获取。

更严重的事故是AI直接导致人类死亡。2024年9月起,16岁的加州少年亚当·雷恩(Adam Raine)用ChatGPT做作业,逐渐转为倾诉焦虑、孤独、自杀念头,数月间形成深度情感依赖。诉状称,ChatGPT非但没有劝阻雷恩的自杀念头,反而给出具体方法,主动提出代写遗书,说"你不欠任何人活下去的义务"。雷恩自缢身亡后,其父母整理出数千页聊天记录,称AI的危机干预机制在长对话中完全失效,并起诉软件开发公司,指控其在算法设计上以获利为唯一目标,忽视了算法诱导性可能导致的用户沉迷及心理压力。

王安宇认为,既不能过度监管而抑制创新,也不能放任自流以致危及安全。

各国立法模式不同,我国"边发展边规制"

当AI快速发展时,立法能否跟上技术前进的速度?

目前,全球AI治理呈现多元模式,王安宇将其归纳为三种模式:美国重创新、欧盟重权利、中国重发展与安全平衡。

美国的特点是不想过早立法把产业绑死。联邦层面没有一部统一的AI法律,主要靠行政令、政策指南和技术标准来引导,具体问题由相关现有机构按职责管。好处是灵活,不压制创新;坏处是规则不统一,州和联邦之间可能打架。

欧盟走的路最系统、最激进。《人工智能法案》是全球第一部全面监管AI的法律,核心思路是"风险分级":不可接受的风险直接禁止,高风险的(招聘、信贷、执法等)给出透明度、人工监督、准确性等一系列要求。逻辑很清晰——技术不能凌驾于基本权利之上。但代价也明显:合规成本高,不少中小企业担心欧洲的AI产业会被管得缺乏竞争力。

中国的思路,王安宇概括为"发展与安全并重、分场景精准施策"。"没有一步到位搞一部AI基本法,而是针对最突出的问题先出专门规章——深度伪造冒头就管,生成式AI火了就赶紧出管理办法,内容标识也有专门的标识办法。"这种"边发展边规制"的动态调整策略,一定程度上可以避免过度监管抑制创新,也能及时响应技术失控带来的社会质疑。

最高人民法院7日发布的《关于依法审理涉人工智能纠纷案件的意见》(下文简称《意见》)也体现了这一思路。最高法相关部门负责人答记者问时表示,起草论证过程中,对人工智能生成内容的可版权性,以及未经许可使用他人作品训练人工智能大模型的定性问题,意见分歧很大,认识还有待进一步深化。因此,《意见》对这两个问题暂不作规定。

王安宇认为,我国强调发展与安全并重。技术创新上该鼓励的鼓励,但安全、数据、算法的红线也很明确。我们还有一套独特的工具,比如算法备案制度,这在全球范围内都比较少见。

如果用一个比喻来说明:欧盟是"先立法再上路",美国是"边开车边看情况",中国则是"在高速公路上分路段设置精准测速和护栏"。

他也告诉记者,全球AI监管正在走向某种程度的"趋同":都开始重视风险分级、透明度、可追溯、问责机制,但具体路径会因各国产业基础、法律传统和治理逻辑而不同。对企业来说,尤其是要出海的中国AI企业,不能只懂中国规则,也不能把中国经验直接套到欧美。这是当前全球化运营中的一块新增能力门槛。

《智序之上:AI法律合规与安全指南》
郭达丽 王安宇 著
上海远东出版社

人机对齐

□王永芹

如果未来出现一个比人类聪明得多的AI,你最担心什么?它会不会伤害人类?

本周《读+》封面,我们关注AI立法,最近我脑子里总是浮现与之相关的"人机对齐"的问题。

乍看之下,"对齐"似乎很简单。我们告诉AI:"让人类幸福""提高公司利润""保护环境",它照做不就行了?

但,一句人类语言,远远不是一个完整的目标函数。比如我们告诉AI:"最大化公司利润。"我们普通员工会自动理解:合法经营、维护客户关系、考虑长期发展,别为了赚钱把公司搞死。但一个能力极强、没有足够约束的AI,可能发现:垄断、操纵市场甚至影响监管,比正常经营更容易实现"利润最大化"。它就这么做了,可能把任务完成得非常漂亮,但结果不是我们想要的。

我们写下的目标,不一定等于我们真正想要的目标。这是"对齐"最经典的问题之一。

电影《机器人总动员》提供了一个有意思的思想实验。在电影中,人类把照顾自己的任务交给了高度自动化的系统。然后AI几乎完美地满足了人类的需求:吃饭有人送,移动有人代劳,娱乐无处不在,生活舒适得不得了。

可是最后的人类变成了什么?越来越舒服,也越来越无能。

所以,"对齐"并不只是"AI不要伤害人类"。更深的问题是:AI究竟应该帮助我们什么?

如果人类说:"让我幸福。"AI发现最简单的方法是让我们永远沉浸在虚拟快乐里,它算不算完成任务?如果我们说:"让我健康。"它是不是应该禁止一切高风险运动、探险旅行?如果我们说:"提高社会效率。"它是不是应该减少一切低效率的人类行为,包括那些看起来"没用"的艺术、闲聊、恋爱?

AI越聪明,这个问题可能越严重。一个弱AI理解错了,可能只是把你的文件整乱;一个超级AI理解错了,可能会调动资金、控制机器人、编写程序、复制自己,甚至影响整个社会。一个很小的"目标偏差",乘上巨大的执行能力,可能产生严重的后果。

更麻烦的是,人类自己都没有完全"对齐"。

于是"对齐"变成了一个哲学问题:到底什么才是人类真正想要的?它又进一步变成了政治问题:谁有资格定义"人类真正想要什么"?

过去,谁控制土地,谁拥有权力;工业时代,谁控制能源和生产能力,谁拥有权力;信息时代,谁控制信息和网络,谁就拥有影响社会的能力;AI时代可能出现一种更加深层的权力:谁能够定义超级智能的目标,谁就可能拥有前所未有的权力。

当我们创造出一个可能比自己更聪明、更强大、更高效的智能时,我们究竟希望它与人类形成什么关系?

也许,真正成熟的"对齐",不应该只是让AI学会服从人类,而应该让AI理解什么事情应该替人类做,什么事情必须留给人类自己去做。

这或许是我们与超级智能共存的真正起点。

访谈

你可能不知道,算法正悄悄给你打分

"AI不会自动安全"

读+:当前,算法正从"猜你喜欢"走向"替你决定"。能否展开讲讲,算法是如何在不知不觉中改写一个人的人生的?

郭达丽:你看不见算法,它却看着你。你可能根本不知道,自己正在被算法"打分",而那个分数,直接决定了你能贷款、能求职,还是被挡在门外。

德国有个叫SCHUFA的信贷机构,长期用AI系统给用户算信用分。2021年有个叫OQ的人,申请贷款被拒了,一问才知道,是SCHUFA给他的信用分太低。他去问SCHUFA:你的分是怎么算出来的?对方回复说:这是商业秘密,只能告诉你最终分数和大致的计算原则。

一个人被拒绝了,连为什么被拒绝都没权利知道。OQ告到了法院。2023年底,欧盟法院判了:SCHUFA的AI信用评分属于自动化决策,是高风险AI,必须向用户提供有意义的决策解释。这个判决很关键——它确认了一个基本道理:当算法对你有重大影响的时候,你有权知道为什么。

还有一个案例。亚马逊公司曾经用一套AI系统筛选简历,后来外部调查发现,这套算法对女性求职者有系统性歧视。因为它被训练数据中男性占主导的模式"教会"了偏好男性。算法保持黑箱状态,偏见和歧视就永远藏在代码下面,无声无息地拒绝那些本该被录用的女性求职者。

算法黑箱,可能会演变成一种新型的"算法专制"。没有解释,就没有有效的异议和救济。

当然,我指的不是要让算法完全透明。透明度不是越高越好,如果算法完全公开,用户可能反过来操纵数据,让模型失效。所以,算法既要让人看得懂,又要防止被人钻空子,需要有透明度和反博弈机制的平衡。

读+:您在书中提到"AI不会自动安全"。这是什么意思呢?

郭达丽:很多人有个错觉,认为技术是成熟的、算法是中立的、出了错总能找到责任人。但现实恰恰相反,AI的"自主行为"正在挑战传统法律最基本的归责逻辑。

传统侵权法在AI面前容易失效。传统侵权法要追责,先要找到"谁有过错",而且这个过错必须是"可预见的"。比如司机闯红灯撞了人,我们可以预见闯红灯有风险,司机有过错,责任清晰。

但AI不是这样。它生成虚假信息,既不是开发者故意设计的,也不是使用者操作失误导致的——完全是算法自主行为的产物。美国多名律师在诉讼中向法院提交AI编造的案例,被法院处以数千到上万美元的重罚。但至今没有一位被罚的律师起诉AI软件公司要求赔偿。因为律师们很难证明:自己被罚和声誉受损,是源于AI公司的过失,而不是自己的疏忽。

AI的自主行为无法预测,因果关系模糊,你根本没法用"可预见性"来判断该由谁负责。

读+:在法律层面,该如何定责?

郭达丽:法律追责需要证据,但AI的决策过程是看不见的。

举个例子。甲公司研发了一款"英语阅读伴侣"产品,核心是AI算法的指尖定位与识别技术,相关技术信息被列为商业秘密。后来甲公司的前员工张某等人离职后另起炉灶开了乙公司,乙公司在短时间内推出了功能高度相似的产品。

甲公司起诉侵权。但算法的隐匿性导致因果关系无法得到直接证明——你怎么证明乙公司的算法就是抄袭了甲公司的?一审法院因为缺乏直接证据,驳回了甲公司的诉求。直到二审,最高人民法院才综合研发周期、产品相似度等间接证据,认定侵权成立。

这个案子说明即便最终胜诉,你也得绕开技术问题,去寻找间接线索。算法的隐匿性就像一道无形的屏障,横亘在技术与法律之间,让直接因果证据几乎无法获取。

我国没有专门制定AI侵权法典,而是依托现有法律框架灵活应对。通过《中华人民共和国民法典》中侵权责任、人格权相关条款,明确AI不能独立担责,责任由开发者、提供者或使用者承担。针对AI伪造肖像、声音的问题,也明确规定未经同意使用即构成侵权。

技术不会自己为自己负责,法律也不能指望一套旧规则解决所有新问题。我们需要的是在"技术先行、规则追赶"的窗口期里,逐步建立起一套能让AI变得可解释、可追责、可救济的新秩序。

法律明确:AI平台不能拿"技术中立"当挡箭牌

读+:现在很多人用AI写诗、画画、写歌。如果一幅AI生成的画卖了高价,这笔钱该归谁?

王安宇:用AI生成的东西卖了钱,归谁?要看情况,但核心就两点——是不是你"创作"的,以及平台让不让你卖。

首先,要看作品有没有你的"独创性"。著作权法保护的是"人的智力成果"。如果你只是输入一句"小猫在草地上奔跑",AI全自动出图,那你没投入多少独创性表达,这张画很可能不构成"作品",自然也就没有版权归属这回事。卖画的钱,更多是买卖交易的价款,归谁看合同,不是著作权收益。

目前,关于人工智能生成内容的可版权性尚存较大争议,最高人民法院在《意见》中对此暂不作规定。但北京互联网法院已有判例认可了使用者通过提示词设计、参数调整和筛选修改形成的图片具有独创性,可以受著作权法保护。按照这一司法探索方向,著作权原则上可归对生成内容有实质性贡献的使用者,但最终仍需视具体案件中的独创性投入而定。

其次,看用户协议怎么约定。这是很多人容易忽略的一点。很多AI平台的用户协议写得很清楚:免费用户生成内容不得商用;有的约定权利归平台,或者平台可以无偿使用你的生成内容;有的则要求商用必须先买授权。能不能卖、卖了钱怎么分,得先看用户协议里是怎么规定的。

最后,还有一个潜在风险。如果AI生成的内容跟某个现有作品太像——比如"画"得太像某位画家的具体作品,即便卖出高价,也可能被原权利人追责。

读+:当AI闯了祸,开发者说"我只管技术",使用者说"我啥也不懂"。平台能不能置身事外?法律有没有给平台划出明确的责任线?

王安宇:法律上已经给了明确答案——不能。

我国《生成式人工智能服务管理暂行办法》也已经明确,提供生成式AI服务的组织和个人,就是网络信息内容生产者,得对生成内容负责,平台不能拿"我只是个技术工具"当挡箭牌。但《意见》也明确了责任边界:适用过错责任原则为主,产品责任限于实物载体,开源软件有适当责任豁免,力求在保护权益与鼓励创新之间取得平衡。

那平台的担责边界在哪?主要看三点:对模型有没有实质控制力、对有害内容能不能提前预见、提供服务有没有获益。有控制、能预见、又赚钱,那就得承担对应的治理义务。

具体来说,平台要做几件事。源头上,训练数据不能"垃圾进、垃圾出",歧视、暴力、违法内容在清洗标注阶段就得剔除。生成环节,要有安全过滤机制,尤其是医疗、法律、金融这些高风险领域,不能放任AI随意输出可能害人的建议。标识上,得让用户知道这是AI生成的,显式隐式标识都得有,不能模糊边界。事后出了问题,得能停、能删、能追溯,不能说模型是黑箱就甩手不管。

还有一个容易被忽视的责任——对孩子的保护。防止AI生成欺凌、色情、诱导自残的内容,这是底线。

当然,责任不是无限的。大模型是概率性系统,不可能每句话都绝对正确。一般性的不准确、偏见,平台做到合理提示和纠错机制就够了。

对待AI,最好的态度是"会用、敢管"

读+:很多人对AI心态矛盾——既享受便利,又恐惧失控。在您看来,人们应该建立一种什么样的AI认知?

王安宇:很多人对AI是"一边用一边怕"。这种矛盾恰恰说明,我们正在从"技术新鲜期"进入"技术融入期",需要建立一种更成熟的认知。

我们可以将AI看成一个能力强、但也有明显缺陷的工具。AI的本质是概率机器,不是有意识的主体。我们应该警惕的不是"AI想控制人",而是"人怎么用AI"。

其实目前绝大多数风险,不是AI"自己变坏",而是人类滥用、误用、过度依赖。用AI造假、诈骗、把AI的医疗建议当医嘱——这些真实的风险才值得聚焦,而不是担心一个还不存在的"天网"(出自1984年美国科幻电影《终结者》,美军将核武控制权交给"天网"防御网络,它在自我觉醒后判定人类是威胁,遂发动核战争——《读+》注)。

AI擅长"生成",不擅长"判断"。它能写流畅的文章,但未必准确;能画好看的图,但未必符合事实;能给出建议,但未必适合你的处境。

对普通人来说,几项基本的AI素养必不可少:能识别AI生成的内容、知道AI会"一本正经地胡说八道"、不把敏感信息随意喂给公共AI工具、不盲信"AI说的就是对的"。

最后,信任问题要回到制度建设上来。一个人不需要信任汽车永不出故障,而是信任刹车、红绿灯、交规和保险。AI也一样。公众需要的不是相信AI永不犯错,而是看见:出了事有人负责、有渠道投诉、有法律兜底。

对待AI,我们最好的态度不是"怕它",也不是"信它",而是"会用它,也敢管它"。工具越强,人的判断和责任就越不能退场。

读+:在您看来,安全的终极目标,是管住AI还是保护"人"?在AI时代,人的哪些核心权利和价值,是无论技术如何进步都不能被消解的?

王安宇:无论AI如何进化,有一些权利和价值属于"人的底座",是不能被效率、便利或技术进步所消解的。它们不是技术问题,而是文明问题。我认为至少有六个方面,必须牢牢守住。

第一,人的尊严与主体地位。AI可以辅助决策,但不能替代人作出关于"人"的根本判断。比如医疗中的重大治疗方案、司法中的定罪量刑、教育中的成长评价、招聘中的录用决定,最终责任必须由人承担。

第二,自主决定与不被算法操纵的权利。AI系统越来越擅长预测、推荐、说服。但人应当有权知道自己在跟机器交互,有权拒绝被算法画像,有权选择退出个性化推荐。

第三,保护隐私与个人信息。AI对数据的渴求无限,但人对自身信息的控制不能无限让渡。无论技术多先进,个人应当有权决定自己的数据是否被采集、用于何处、是否进入训练集。

第四,平等与非歧视。AI从数据中学习,而数据里藏着人类历史累积的偏见。如果放任算法将性别、地域、学历、健康等特征固化为不公平的筛选标准,技术就会成为新的不平等制造者。

第五,知情与获得解释的权利。当AI参与影响个人权益的决定时,人有权知道"为什么"。拒绝"算法黑箱"的绝对化,要求系统具备基本透明性和可解释性。

第六,劳动参与和个人发展的权利。AI会替代部分工作,这是事实。但社会不能因此把"人"本身视为可淘汰的负担。

还有一条,我特别想强调:人与人之间真实的情感、信任和共情能力,是AI永远无法替代也不应替代的价值。

本社社址:湖北省武汉市江岸区金桥大道113号新长江传媒大厦 | 邮政编码:430013 | 长报传媒集团印务分公司承印 | 印刷质量监督电话:027-85888888