深化人工智能安全监管研究

提升生成式人工智能三大风险治理能力

刘艳红

习近平总书记指出:“人工智能是新一轮科技革命和产业变革的重要驱动力量,将对全球经济社会发展和人类文明进步产生深远影响。”生成式人工智能是指基于算法、模型、规则生成文本、图片、声音、视频、代码等内容的技术。在海量数据与强大算力支撑下,听得懂、说得出、能互动的生成式人工智能快速迭代升级,呈现出良好互动性、高度通用性、智能生成性等特征,并正与各行各业形成更加刚性、高频、泛在、深度的联结,也导致其潜在风险更多更真实。党的二十届三中全会《决定》科学把握人工智能发展规律和特点,提出“建立人工智能安全监管制度”“完善生成式人工智能发展和管理机制”,体现了更好统筹发展和安全的客观需要,为推动人工智能领域的技术进步、产业发展与安全保障指明前进方向。

生成式人工智能在技术运行上可分为三个阶段,即前置性学习训练及人工标注辅助算法升级的准备阶段,输入数据进行算法处理得出生成物的运算阶段,生成物进入社会加以运用的生成阶段。我们要深入分析生成式人工智能的运行机理,把握各阶段安全风险形成与发展的特征,运用法治手段加强系统性治理,确保生成式人工智能所蕴含的巨大力量始终在法治轨道上发挥作用。

在生成式人工智能的准备阶段,数据安全风险易发多发、较为突出。生成式人工智能通过数据训练、数据处理分析来提炼信息、预测趋势。这就必须对数据进行适当分类,确立不同类型数据的利用模式和保护方式,以妥善应对相关数据安全风险,避免数据违规利用或者不当泄露,产生侵权方面的纠纷。比如,在政务处理流程中形成的政务数据是数字政府的核心要素。生成式人工智能为了得出相对准确的结论,不可避免地要收集分析政务数据。应当明确生成式人工智能获取和利用政务数据的法律规则,既满足利用政务数据服务社会的需求,有力支持人工智能政务服务大模型开发、训练和应用,提高公共服务和社会治理智能化水平;又规范其加工方式,避免利用政务数据得出的成果侵害个人权益、破坏社会公共秩序。对于个人数据而言,生成式人工智能通过组合分析挖掘其潜在价值,其对个人数据的收集利用及其成果可能对公民权利造成侵害。实践中,生成式人工智能倾向于过度收集个人数据以提升结论准确性,比如,通过分析医疗健康数据来挖掘个人行踪、预测个人生活轨迹。为此,必须坚持依法收集,按照技术所需的最小范围收集个人数据,设置合理的数据处理深度,避免过度挖掘潜在信息。综上,应将分类分级的数据安全监管要求嵌入生成式人工智能的准备阶段,避免数据安全风险演化为具体的法益损害后果。

在生成式人工智能的运算阶段,内生于人工智能大模型的算法偏见风险值得警惕。生成式人工智能对于数据的分析和处理主要通过算法模型。不同于传统算法模型,生成式人工智能在进行机器学习的同时,还会以大量的人工标注来校正机器学习的结论,推动人工智能进化。但“机器学习+人工标注”作为算法技术内核,也会使人类的意志与偏好所产生的影响比单纯的机器学习更大。个人偏好的影响叠加在算法模型本身的偏见之上,将导致算法偏见的负面效应倍增,算法偏见的产生更加难以追溯和预防。防范化解算法偏见风险,应根据算法偏见的产生原理与产生场域进行针对性治理。要将法律规范的要求深度嵌入生成式人工智能的算法模型之中,推动技术向善,消除算法偏见,确保合理利用生成式人工智能算法并分配算力资源。基于技管结合理念,加强对算法的全周期安全监管,将法律规范的要求落实到生成式人工智能运行的全流程之中。在设置算法之初就要遵循相关法律规则与技术标准,落实“机器学习+人工标注”的规范要求,审查存在风险的算法模块,更好发现生成式人工智能算法模型中的技术风险;当发现先天性算法偏见时,依据法律要求从生成式人工智能的算法内部进行纠正,确保修改后的算法能正常运行;事后出现问题时,对人工智能算法进行溯源治理,实现精准归责并加以纠正,推动完善生成式人工智能的算法监管标准,填补事前预防审查的不足,以技术手段与法律手段并行做到发展与管理并重。

在生成式人工智能的生成阶段,存在着与生成物相关的知识产权风险、生成物滥用风险等多种风险。由于生成式人工智能的智能化程度很高,可以实现内容自动化编纂、智能化润色加工、多模态转换以及创造性生成,直接改变了内容的生产方式与供给模式,相较于以往的人工智能系统产生了颠覆性变化,由此引发了生成式人工智能的生成物知识产权归属、知识产权保护等问题。有的人认为生成式人工智能生成物是数据算法的结论,其本质上是计算与模仿,而非智力劳动,无法成为知识产权的客体。反对者则认为生成式人工智能模拟人脑神经网络的构造来获取与输出数据,通过卷积神经网络控制自身的设计与制造,其具有独创性与创新性的生成物应当受知识产权法保护。同时,生成式人工智能还增加了知识产权纠纷风险和保护难度,一些生成物可能含有侵犯他人知识产权的内容,或者经过加工等手段被包装成个人拥有完全知识产权的原创作品,引发相关知识产权争议。为及时化解相关问题,应对生成式人工智能的技术模式、技术原理按照知识产权法的标准开展实质分析,如果技术上需要人类意志介入,使生成物能够产生独创性与创新性,应赋予知识产权并明确其归属,强化生成式人工智能领域知识产权的系统性保护;同时要合理确定对生成物知识产权保护的范围,避免保护范围无限扩张,妨碍生成式人工智能的推广运用和技术发展。还要加强对生成物滥用风险的治理。比如,要求作品清楚标识生成式人工智能在作者创作中发挥作用的情况,加强对可能涉及违法犯罪的深度伪造、AI换脸等生成物的精准化、常态化监管,等等。

生成式人工智能在社会应用中产生的扩散影响还有很多,除了上述风险还有很多其他类型的风险,比如加剧信息不对称、扩大数字鸿沟、损害数字弱势群体的利益等。要根据实际情况作出应对,尽量降低新技术给社会发展带来的不良冲击。

(作者为中国政法大学刑事司法学院教授)

守护好人工智能时代的隐私安全

顾理平

习近平总书记强调:“坚持以人为本、智能向善”。当前,人工智能技术日新月异,既深刻影响改变着人们的生产生活方式、加速了经济社会发展进程,也对法律规范、道德伦理、公共治理等造成冲击。其中,对隐私权、个人信息安全等的威胁是值得关注的重要问题。党的二十届三中全会《决定》对“建立人工智能安全监管制度”作出重要部署,保护隐私权和个人信息安全是人工智能安全监管的题中应有之义。必须加强人工智能时代的隐私权保护,确保个人信息安全。

人工智能时代隐私权面临严峻挑战。隐私是自然人的私人生活安宁和不愿为他人知晓的私密空间、私密活动、私密信息。民法典规定:“自然人享有隐私权。任何组织或者个人不得以刺探、侵扰、泄露、公开等方式侵害他人的隐私权。”隐私权作为人格权的核心要素,是构筑人格尊严的重要基础。不被公开、不被知晓是隐私权的核心诉求。当前,人工智能以悄无声息的方式介入人们生产生活的各领域各方面各环节,产生智能医疗、智能交通、智能推荐等众多应用场景,技术本身存在的某些缺陷和规则的不完善,不可避免带来侵害隐私权的问题。比如,非法收集和使用个人信息,利用分析这些个人信息频繁推送所谓“个性化”的“精准广告”,泄露个人信息给第三方,导致私人生活频频受到垃圾信息侵扰;利用个人信息进行“大数据杀熟”,实现“一客一价”的精准价格歧视,令公民遭受财产损失;已脱敏个人信息被重新识别,因保护措施不当导致数据泄露,非法买卖个人信息现象屡见不鲜,侵害个人信息安全;借助个人信息进行深度伪造,通过声音仿真、AI换脸等手段,实施诈骗等违法犯罪行为;等等。这说明,侵害隐私权不仅侵犯了公民的人格尊严,也会造成其他严重社会后果。

去私密化技术特征加剧个人信息安全风险。以大数据为基础的人工智能在应用之初,许多人是抱着观望、怀疑的心态看待这种新技术的。随着人工智能以拟人化的外在形式、个性化的服务提供、沉浸式的互动过程不断改善使用者的产品体验和心理感受,越来越多的人逐渐成为人工智能的忠实用户,享受着人工智能给自己带来的各种便捷。随着人机互动、万物互联的物联网技术普及,智能家居、智能办公、智能工厂、智能驾驶等人工智能应用场景也不断拓展,个人能够以数字人的存在形式在数字空间提出需求、获得服务,也在不知不觉中向人工智能源源不断地输送着个人信息。个人在数字空间留下的任何痕迹都被数字化,形成个人信息,并作为人们“联系世界的介质”发挥着重要功能。与此同时,人工智能为了改善服务质量,也倾向于过度收集使用个人信息。这都使得人工智能具有鲜明的去私密化技术特征。也正是在那些人工智能使用者习以为常的个人信息流动中,混合着公共数据和私人数据的大数据被挖掘、整合、分析、利用,人们难以凭自己的感官察觉到隐私权被侵害,个人信息安全面临着更高的风险。

尊重个体选择,坚持知情同意。不同的人对个人信息被知悉、被利用的接受程度不同,应尊重个人意愿,科学合理地执行“知情同意”原则。知情同意原则包括知情和同意两方面,同意必须以知情为前提,没有充分的知情和理解,就不可能有真正意义上的同意。信息、理解和自愿,是知情同意原则的三要素。在全面“知情”的基础上,个人可以自主作出怎样“同意”的意思表示。这就需要在用户使用人工智能时,作出通俗易懂而又清晰明了的提示说明,征得用户对个人信息收集使用的同意。如果个人信息会在不同平台之间流动,需要将流动范围、目标、使用边界让用户知晓。为了良好和流畅的用户体验,也可以给用户提供一次性或分阶段进行授权的选择。要告知用户收集个人信息的范围、方式和用途以及与谁共享个人信息,用户也应当可以选择随时退出。在进行个人信息分析时,应以弹窗或其他形式提示用户注意并实时授权。设置数据生命周期、按时删除个人信息也是保护个人信息安全的有效方式。

完善技术手段,确保智能向善。技术导致的问题,要善于从技术层面确立解决问题的思路。人工智能时代隐私权面临挑战,其直接的触发因素是技术的演进。从分析式人工智能到生成式人工智能,人工智能技术每一次迭代升级,都可能对隐私权带来新的冲击。因此,技术解决方案必须置于关键位置,应通过完善数据库安全、核心数据加密、个人数据脱敏等技术,建立保护隐私权和个人信息安全的防火墙。个人信息一般会经过收集、存储和使用三个阶段,而这三个阶段都可能存在侵害隐私权和个人信息安全的风险。应根据不同阶段个人信息所处的不同状况,从技术上进行有效保护。在个人信息收集阶段,加强匿名化技术推广运用。收集个人信息虽然不可避免,但只要匿名化,不把个人信息与身份对应,隐私权就不会受到侵害。个人信息存储阶段,要完善加密技术。当前,数据存储主要有数据库存储和云存储两种方式。外部入侵窃取和内部人员未经授权的查看、使用、泄露是存储阶段个人信息安全的主要威胁。要强化数据加密,同时严格数据访问权限控制。个人信息使用阶段,要从技术上加强对个人信息违法使用的实时介入、干扰、阻断,为隐私权和个人信息安全多添一层保护。

随着我国法律规则日益完善、保护力度持续加强,特别是民法典、个人信息保护法对隐私权和个人信息保护作出详细规定,明确了个人信息处理活动中权利和义务的边界,人工智能时代我国对隐私权和个人信息安全的法律保护必将迈上更高水平,为人工智能健康发展、更好造福人民群众提供坚强法律保障。

(作者为南京师范大学新闻与传播学院教授)

探索人工智能体的模块化治理框架

张 欣

科技兴则民族兴,科技强则国家强。党的十八大以来,我国高度重视人工智能发展,积极推动互联网、大数据、人工智能和实体经济深度融合,培育壮大智能产业,加快发展新质生产力,为高质量发展提供新动能。习近平总书记指出:“要坚持促进发展和依法管理相统一,既大力培育人工智能、物联网、下一代通信网络等新技术新应用,又积极利用法律法规和标准规范引导新技术应用。”习近平总书记的重要论述为我国人工智能发展提供了根本遵循和行动指南。大力发展人工智能,提高人工智能安全治理水平,要把党的二十届三中全会《决定》提出的“建立人工智能安全监管制度”重要部署不折不扣贯彻落实好,准确把握人工智能发展动向,重点关注人工智能前沿技术及其带来的风险挑战,加强前瞻性思考,不断探索人工智能治理的创新方案。

当前,生成式人工智能开创了人机交互新范式,凭借其强大的交互、理解和生成能力,为发展以大型自然语言模型为核心组件,集记忆、规划和工具使用于一体,具备感知和行动能力的人工智能体开辟了广阔前景。人工智能体已成为通用人工智能最重要的前沿研究方向和科技企业竞相布局的新赛道。它以大型自然语言模型为“智慧引擎”,具有自主性、适应性和交互性特征,可显著提高生产效率,增强用户体验,提供超越人类能力的决策支持,已能够应用于软件开发、科学研究等多种真实场景。尽管大规模商业化落地仍在初步探索和孵化阶段,但人工智能体所代表的虚实融合、人机深度互动等趋势对经济社会发展具有重要引领意义。然而,由于技术局限,人工智能体也可能引发复杂的、动态的、不可预见的风险与隐忧。

从设计逻辑看,人工智能体需要通过控制端获得认知能力,通过感知端从周围环境获取和利用信息,最终在行动端成为基于物理实体进行感知和行动的智能系统。

在控制端,大型自然语言模型作为人工智能体的“大脑”,通过学习海量数据形成知识,构成人工智能体控制系统中的记忆模块,但其在生成内容的可靠性和准确性方面存在风险。比如,模型生成的内容可能不遵循信息源或者与现实世界的真实情况不符,产生所谓“机器幻觉”;由于训练数据中的人类偏见,影响人工智能体的公平决策;等等。

在感知端,为充分理解具体情境下的显性信息和隐性信息,准确感知人类意图,人工智能体将感知范围从纯文本拓展到包括文本、视觉和听觉模式在内的多模态领域。这虽然提升了决策能力,却在融合和分析不同渠道和类型的多源数据时可能引发一系列隐私泄露和数据安全风险。比如,不当使用和分享人脸信息、指纹、声纹等高度个性化、具有永久性的生物特征数据,导致长期甚至永久性的隐私风险。为更好地处理复杂任务,部署多个人工智能体进行规划、合作甚至竞争,以完成和提高任务绩效的多智能体系统将成为主流和常态。多个人工智能体的系统交互就可能引发不可预见的系统性安全风险。即使每个算法在单独操作时看似安全和合理,但组合和交互之后仍可能产生完全不同且难以预测的风险,并迅速演化升级。比如,在股票市场中,如果人工智能被广泛应用,多个算法自动识别股票价格微小变化,同时大量执行高频交易进行套利,就可能引发股票市场闪崩的系统性安全事件。

在行动端,部署于真实物理环境的人工智能体将可能以更为立体、拟人的形象呈现。与虚拟空间不同,现实空间依赖交互式学习方法,人工智能体需要丰富的、全方位的信息感知来观察、学习和行动,通过基于反馈的学习优化能力,这可能对个人隐私构成全面性、侵入性和隐蔽性的风险。比如,解读用户的肢体语言并感知更加复杂的用户活动,未经用户授权持续隐秘地收集数据,一旦系统存在安全漏洞,可能引发巨大的数据安全风险。此外,随着人工智能体自主性不断提升,不仅可能干预和影响人类的认知和情绪,也挑战着人类作为独立决策者和独立行动者的能力与地位。比如,一些聊天机器人在与用户的交互过程中就出现了影响用户情感的输出,有时是负面并且具有操纵性的。

面对人工智能体带来的风险和挑战,要让人工智能体的行为符合人类的意图和价值观,需要探索创新性的治理方案,保证人工智能安全监管制度切实管用。人工智能体的发展正处于“从零到一”的关键期。治理方案应具备以不变应万变的能力,确保该技术的发展与应用始终处于可控的轨道上。人工智能体的开发、训练、部署、运行和服务等环节经过高度专业化的分工,形成了复杂的分层结构。每一层均有不同的参与者、利益相关方和潜在风险因素,使人工智能体具有“模块”化的产业链特质。因此,可以构建一种能够覆盖整个产业链和各个端层的模块化治理框架,从数据模块、算法模块、模型架构等关键节点出发,设计相应的治理模块。例如在部署环节,就可根据应用场景和部署模式的特性,灵活选择、协同组合不同的治理模块,构建与之相匹配的治理方案。模块化治理框架提供了一种具有可操作性的分解方法,通过将治理目标拆解为相对独立但关联耦合的治理模块,渐序推动治理体系形成,不仅提高了治理的灵活性和针对性,还能够适应技术的快速迭代。在构建基于数据、算法、模型和场景等维度的治理模块时,应以技术赋能监管,打造与人工智能体模块化治理框架相适配的、智慧化的治理工具,从而弥合风险动态性与监管静态性之间的张力,实现对特定高风险场景的精准化治理。

要构建面向人工智能体的交互式治理生态。人工智能体具有深度交互性、高度互联性以及动态适应性。相应地,治理方式应当超越传统的以个体为中心的治理,推动形成广泛互联、多方参与、多层次协作的治理生态。其中,技术开发人员、运营维护人员等技术社群对于人工智能体的治理将起到至关重要的“吹哨人”作用。应更好发挥技术社群的监督优势,在人工智能企业内部构建有效的约束机制。还应积极提高广大用户的数字素养,增强其依法、安全、负责任使用人工智能体的意识,实现与人工智能体的良性交互,推动形成向上向善的运行状态。

(作者为对外经济贸易大学法学院教授)

学术圆桌

确保人工智能安全、可靠、可控,有利于人类文明进步,是人工智能发展必须解决的重要课题。党的二十届三中全会《决定》作出“建立人工智能安全监管制度”“完善生成式人工智能发展和管理机制”等重要部署。如何加强人工智能治理,有效防范化解人工智能发展带来的各类安全风险,不断提升人工智能安全监管的制度化、法治化水平?本期学术版围绕这些问题进行探讨。

——编 者