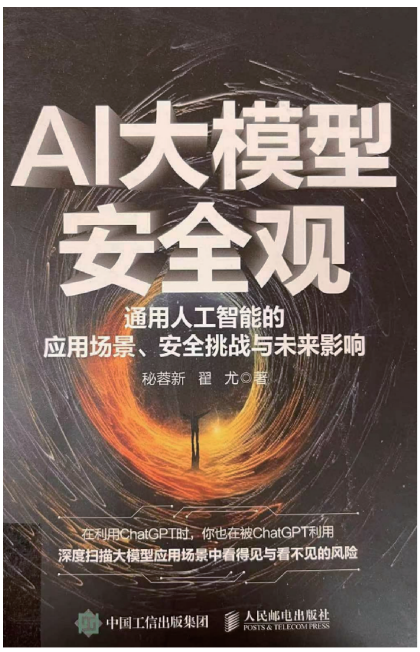
多模态 AI 大模型的安全风险与综合监管策略
——评《AI 大模型安全观》

当今时代正处于以数字技术为引领,以人工智能为趋势的全新发展范式中,2023 年以来,ChatGPT 和大模型技术引起全球的关注,人工智能成为每个人工作和生活中都可以使用的工具。在惊叹人工智能技术快速发展的同时,也需要更加关注人工智能带来的安全挑战。随着人工智能技术的快速发展,多模态 AI 大模型通过整合视觉、听觉、语言等多种模态的数据,在金融、医疗、教育等多个领域展现出巨大潜力。然而,大模型可能带来的安全风险也不容忽视,这些风险包括但不限于数据泄露、算法偏见、对抗性攻击等。因此,识别多模态 AI 大模型的安全风险,并提出相应的综合监管策略,对探索大模型未来发展趋势,促进 AI 技术的健康发展至关重要。



书名:AI 大模型安全观
作者:秘蓉新,翟尤
出版社:人民邮电出版社
ISBN:9787115622761
出版时间:2023 年 7 月
定价:69.8 元

笔者在开展南京工业大学高等教育发展规划课题(HED2024C-04)的研究过程中,认真阅读了《AI 大模型安全观》一书,该书主要分析了 ChatGPT 在内容安全、网络安全、隐私安全、版权合理和伦理道德等方面带来的新挑战、新风险,并从人工智能技术监管角度给出一些策略和建议,让读者能更加客观、全面地认识 ChatGPT 和大模型,以及它们带来的安全方面的机遇和挑战。全书共分为 8 章,第 1 章介绍了 ChatGPT 底层大模型技术的基本特征和技术创新点以及未来科技创新的风口,并分析了个人如何把握这次人工智能的发展浪潮。第 2 章介绍了多模态大模型 ChatGPT-4 给人类生活带来的机遇和挑战,并指出要客观看待它的价值。第 3 章通过梳理大模型的发展历程,同时结合当前业内关注的焦点,重点分析了大模型创作过程中大众普遍关心的问题。第 4 章分析了人工智能时代的安全挑战,指出数字技术是一把双刃剑,一方面提升了人们工作和生活的质量,另一方面,网络攻击带来的破坏将引发新的安全挑战。第 5 章重点分析了大模型带来的安全挑战,并指出大型模型有时可能会产生不准确或不真实的信息。第 6 章分析了 ChatGPT 在网络安全、个人隐私保护、版权保护、伦理风险等领域暴露的诸多问题和挑战,为后续安全人员和监管部门制定相应的政策措施提供了实践依据。第 7 章指出技术本身是中立的,不具备固有的道德属性,如何把 ChatGPT 应用到网络安全防护,有效帮助网络防护人员提升工作效率,发现潜在的安全风险,提升安全能力。第 8 章从整个行业的角度来分析 ChatGPT 对经济社会的影响,并帮助大家更加客观深入地理解这轮人工智能发展浪潮的关键点。该书为学者提供了一个全面的视角来审视多模态 AI 大模型的安全风险和监管策略。不仅分析了问题的本质,还提出了切实可行的解决方案,对于政策制定者、企业决策者以及研究人员都具有重要的参考价值。

作者指出,目前多模态 AI 大模型的应用场景主要包括:在金融领域,AI 大模型被广泛应用于风险评估、欺诈检测、智能投顾等。通过分析大量的交易数据和用户行为,大模型能够识别潜在的欺诈行为,提高交易安全性。同时,大模型还能够为投资者提供个性化的投资建议,优化投资组合。在医疗行业,AI 大模型的应用主要集中在疾病诊断、治疗方案制定、医疗影像分析等方面。通过大量的医疗记录和数据分析,大模型能够帮助医生进行更准确的疾病诊断,制定更有效的治疗方案。在教育领域,AI 大模型可以提供个性化的学习体验,通过分析学生的学习习惯和能力,为学生推荐适合的学习资源和路径。此外,大模型还能够辅助教师进行课程设计

和教学评估,提高教学质量。在法律行业,AI 大模型的应用主要体现在智能法律咨询、合同审核、案件分析等方面。通过分析法律文本和案例,大模型能够为律师提供法律建议,辅助法官审理案件,提高法律服务的效率和公正性。随着技术的进步,AI 大模型将拥有更强的泛化能力,应用于更广泛的领域和场景中。

在享受 AI 大模型带来工作生活便利的同时,其安全风险也随之而来。在网络安全方面,主要包括数据隐私与安全,多模态 AI 大模型通常需要处理大量的个人敏感数据,如面部图像、语音记录等。这些数据的泄露或不当使用可能导致严重的隐私侵犯问题。同时,模型训练过程中可能存在数据投毒、模型窃取等问题,也会威胁数据安全。模型偏见与歧视一直存在,由于训练数据的偏差或模型算法的不完善,多模态 AI 大模型可能在决策过程中表现出性别、种族、年龄等偏见,导致不公平的决策结果。模型滥用与恶意操作,多模态 AI 大模型可能被用于生成虚假信息、进行网络攻击等恶意操作,对抗性攻击可能导致模型在特定情境下失效,影响其可靠性和安全性。在评估多模态 AI 大模型的安全风险时,首先需要认识到,这些模型由于其复杂性和处理多类型数据的能力,面临着独特的安全挑战。如对抗性攻击可以通过精心设计的扰动误导模型的决策,而数据隐私泄露则可能暴露个人敏感信息。此外,模型窃取和后门攻击等威胁也不容忽视,它们可能导致模型在特定条件下执行恶意行为。在社会安全方面,金融、交通、医疗、城市管理等基础设施领域遭到攻击,会导致整个产业链的瘫痪,影响社会稳定。多模态大模型在特定应用场景中可能引发法律和伦理问题,如自动驾驶汽车发生交通事故,如何判断责任方是一个复杂的过程,涉及汽车制造商、软件研发人员、网络服务商、汽车所有者、使用者和乘客等多方。此外,企业的数据质量导致的计算结果的可控性差、用户权益和隐私屡遭侵犯也是当前数据安全面临的巨大挑战。

为了应对这些安全风险,需要采取一系列措施,包括但不限于加强数据加密、差分隐私保护、安全多方计算、模型水印和指纹技术等,以及建立完善的大模型安全评估框架和探索有效的防御机制。主要包括:建立数据审计和追溯机制,确保数据的合法合规使用,提升模型透明度与可解释性,研究和开发可解释的 AI 模型,提高模型决策的透明度。通过模型可视化、特征重要性分析等方法,帮助用户和监管者理解模型的工作原理和决策依据。最后推动伦理与合规性审查,在模型设计和部署过程中,引入伦理审查和合规性检查,确保模型的应用不会违反伦理原则和社会价值观。由于这些模型的决策过程往往是不透明的,因此,很难判断它们是否会产生不公平或有偏差的决策。提高模型的可解释性不仅可以帮助人们理解模型的行为,还可以及时发现和纠正潜在的安全问题。

笔者认为,可以建立多层次监管框架,构建包括法律法规、行业标准、企业自律等多层次的监管网格,确保多模态 AI 大模型的全生命周期受到有效监管。首先,在政府层面,需要出台相应的法律法规,规范 AI 大模型的开发和应用。包括数据收集和处理的透明度要求、评估模型输出的准确性和公正性、以及对可能的滥用行为设立惩罚措施等。通过这些措施,可以在一定程度上确保 AI 大模型的安全和可靠应用。其次,在企业层面,需要加强自我监管和行业自律。企业应当建立完善的数据安全管理体系,确保数据的合规使用和隐私保护。同时,企业还应当积极参与到行业标准的制定中,通过合作和共享最佳实践,共同提升整个行业的安全水平。除了政府和企业的共同努力,还需要技术层面的创新和研究,通过引入更加先进的算法和技术,提高模型的鲁棒性和抗攻击能力。重视加强国际间的合作,共同制定全球性的 AI 治理框架和标准,以应对 AI 大模型带来的跨国安全风险,参考和借鉴其他国家和区域的先进经验,增强公众对 AI 大模型安全风险的认识和理解。通过教育和宣传活动提高公众的 AI 素养,鼓励社会各界参与到 AI 大模型的监管中来,形成政府、企业、公众等多方共同参与的治理格局。

综上,多模态 AI 大模型的安全风险管理是一个复杂而紧迫的问题,需要从技术、法律、伦理等多个角度出发,促进多模态 AI 大模型的健康发展。人工智能的发展不是一蹴而就的,大型算法的不断升级,各种不确定性问题也在叠加。未来,更高级别的通用人工智能的实现,将带来人工智能产品和技术的巨大突破。

(钱婷/南京工业大学应急管理学院)