智能社会研究

（双月刊）

2024 年
第 2 期
2024 年 3 月 10 日出版

2022 年 11 月 10 日创刊

总第 9 期

目　次

论文

研究报告

译文

书评

访谈

CONTENTS

THESIS

RESEARCH REPORTS

TRANSLATED TEXT

BOOK REVIEW

INTERVIEW

走向负责任和可问责的金融大模型伦理治理*

段伟文**

摘要:作为计算智能和数字思维发展的新阶段,近期兴起的基于大模型的生成式人工智能已初具通用人工智能的特征,且实现了从现实空间、数据空间到生成空间的突破性进展。然而,生成式人工智能一方面在安全和伦理上存在可能危及人类生存的风险,另一方面存在一定的社会伦理风险。生成式人工智能在金融领域的应用,是金融行业在数字化转型大潮推动下的必然趋势。在金融领域,生成式人工智能的创新应用包括优化客户体验和定制个性化推荐、金融欺诈检测和预防、风险评估和信用评分、交易和投资策略优化、借助自然语言处理提升合规效率等。与此同时,金融大模型的应用也进一步加剧了人们的担忧,数据隐私、偏见、可解释性等伦理风险在该领域备受关注,而上述风险的共性问题主要在于生成内容的"幻觉"、数据投毒以及合成数据的使用。面对金融大模型的伦理和法律风险,对其进行伦理治理将更有利于对大模型潜在伦理和法律风险的整体治理。对于政府和行业主管部门而言:首先,应深入理解不同层级的伦理和治理原则,明确其背后的价值观和优先考量,从而根据具体需求确立金融大模型伦理治理的基本理念;其次,通过明确我国科技伦理和人工智能伦理治理的指导思想,探索金融大模型伦理治理的工作思路;最后,积极开展科技伦理审查工作,走向负责任和可问责的金融大模型伦理治理。

关键词:生成式人工智能　金融领域应用　伦理治理　数据隐私　偏见和歧视　负责任和可问责

---

* 本文系国家社会科学基金重大项目"智能革命与人类深度科技化前景的哲学研究"(项目批准号:17ZDA028)、中国社会科学院"登峰战略"新兴学科和交叉学科项目(项目批准号:DF2023XXJC02)的阶段性研究成果。

** 段伟文,中国社会科学院哲学所、中国社会科学院大学哲学院。

近年来,数据驱动的人工智能及其社会应用取得迅猛发展,呈现出诸多具有颠覆性社会影响的创新前景,迫使人们不得不系统研究如何认识其可能带来的社会、伦理、安全风险,进而采取合理的应对之策。特别是继深度合成技术之后,基于大规模预训练语言模型的生成式与对话式人工智能展示出强大能力,使其在包括金融领域在内的诸多商业领域拥有巨大的应用前景,但其在社会、伦理和法律等领域的风险也不容小觑。而从风险的动态感知和敏捷治理的角度来看,实现大模型有序创新和向善发展的关键在于系统认识其潜在社会伦理风险,并使伦理治理等规制伴随其全生命周期。

毋庸置疑,生成式人工智能的巨大变革力量将对社会和经济发展产生深远影响。国际货币基金组织最近发布的《金融中的生成式人工智能:风险考量(2023)》指出:人工智能在塑造经济和金融部门发展方面发挥着越来越重要的作用,并被视为通过提高效率、改进决策流程以及创造新产品和产业来提高生产力和促进经济增长的引擎;与此同时,人工智能还通过重塑金融中介、风险管理、合规性和审慎监管的性质,迅速改变金融业格局。随着基于大模型的生成式人工智能在金融领域的广泛部署,政府和行业主管部门必须充分认识这些部署可能导致和加剧的风险,探寻通过伦理治理等规则措施有效地加以审慎应对的措施。

一、走向通用人工智能的大模型与生成空间的形成

从工程技术上讲,近期兴起的基于大模型的生成式人工智能开启了对类人类智能的探索,其最具革命性的突破在于初步展现出通用人工智能的特征。一般而言,认知科学意义上的通用人工智能是指机器人和智能软件及算法具备像人一样可以普遍泛化的智能,它们能够作为独立的智能体进行学习、认识和决策,甚至可能具有独立的自我意识。虽然当前兴起的

ChatGPT、文心一言等大语言模型还不是认知科学意义上的通用人工智能，但由于它们的智能已经可以在一定程度上实现泛化，这也就意味着“通用人工智能”不再是一个理论概念，而是已经走在工程技术实现的路上，因此在研发活动和产业政策层面已将生成式人工智能归类为通用人工智能。

（一）大模型与生成式人工智能

大模型横空出世之后，人们就其所带来的突破性创新形成了一些基本共识。

首先，大模型生产的类似人类的连贯和流畅的语言内容表明，它已突破了连贯和流畅的人机交互的临界点。大模型不仅可以针对任意话题与用户进行高质量的对话，而且能准确地按照用户意图完成分类、问答、摘要和创作等若干应用场景的自然语言理解与生成任务，进而在流畅的人机对话指引下完成各种工作。

其次，大模型获得成功的关键是通过“模仿学习+强化学习”等新的学习模式，包括预训练和人工标注等工程方法，在工程上对齐了人机目标。通过系统地实施人机价值对齐工程，大模型能够对生成内容中存在的价值观念冲突进行较为有效的调节和矫正，从而对由此带来的风险加以控制。人机对齐工程固然建立在一些数学和算法理论之上，但本质上是一种工程，其地位和重要性就像航天工程和建筑工程中的可靠性工程和质量工程一样。人机对齐工程对于大模型创新政策的重要启示在于，大模型生成内容在价值观念上的偏差原则上是可控的，但考虑到价值的模糊性和过度的价值对齐对成本和效率的影响，也应该看到阈值的设置应该具有一定的包容性。

最后，在生成式人工智能的相关讨论中，主张暂停大模型研究的人工智能研究者的基本理由是大模型相当于人工智能的“奥本海默时刻”，强调大模型所开启的通用人工智能可能会带来威胁人类生存这一长期风险。但反对这一说法的人工智能研究者认为，目前的人工智能并不会带来这种长期

风险或生存风险,真正值得关注的是大模型等人工智能创新应用所导致或加剧的偏见、歧视、虚假信息、幻觉以及过度的能源与资源消耗等现实风险。尽管人们对大模型的风险认知存在分歧,但基本的共识是,这一强大的人工智能应用将非常广泛,应该密切关注其导致的包括伦理风险在内的多重风险,并对由此带来的相关社会、伦理、法律问题加以规制和治理。

值得指出的是,当前建立在大模型基础上的生成式人工智能并不是一种独立存在的智能体,而是一个由人类和机器智能构成的巨型智能生态系统,这个生态系统的运转实质上是通过工程和系统方法在人类智能和机器智能的复杂组合之上实现的。因此,生成式人工智能不能被简单地理解为文字、图像、音视频内容的自动化生产工具,其所开启的通用人工智能将成为未来社会的基础设施:一方面,它是建立在全数据基础大模型之上的知识生产的引擎;另一方面,它正在从根本上改变人类文明的操作系统。

(二) 从现实空间、数据空间到生成空间

为了深化对基于大模型的生成式人工智能的实质的认知,应该进一步看到,生成式人工智能是计算智能和数字思维发展的新阶段,其所带来的突破性在于其在现实空间(物理空间与社会空间)、数据空间的基础上形成了全新的生成空间。

计算机和互联网应用普及之后,逐渐完成了对现实世界的数据化,由此开启了一个新的空间——数据空间。数据空间可以被视为现实空间的镜像,如消费者的数字画像。现实空间的数据化所获得的数据可以分为不同类型:(1)特征数据,如人脸、基因组等生物特征数据;(2)位置数据,如反映现实空间中人和事物的坐标参数;(3)行为数据,如人们的搜索点击数据反映的是人们的行为倾向;(4)内容数据,如语言、图像和音视频等传递了有意义的信息内容。而且,数据的内涵会随着技术的演进被重新定义,随着情感计算等技术的发展,人脸和步态等生物特征数据还会进一步转化为情感

数据。不难看出,人工智能的伦理问题在很大程度上与这些数据的使用相关。

近来,随着深度合成和生成式人工智能的发展,内容数据的合成和语言的合成逐渐达到了连贯、流畅、合理甚至以假乱真的程度,合成的文字、图像、音频和视频内容呈现出海量趋势,这就在数据空间的基础上形成了由自动合成内容构成的生成空间。显然,生成空间来自现实空间和数据空间。认识到这一点,就能认识到网络信息、文献和各种数据的数量和质量等对生成式人工智能发展的要求是基础性的,就能理解现实空间的歧视偏见、数据空间的数据毒性可能导致生成空间对相关问题的放大与强化。反过来,生成空间中的错误内容——从虚假信息、看似有理但违背事实的陈述、随意编造的文献来源到以肯定的方式表达的幻觉——存在对数据空间和现实空间的反向侵蚀,以及进一步导致生成空间内容退化的可能性。

二、生成式人工智能的长期风险和现实的社会伦理风险

这一拨基于大模型的生成式人工智能发展中呈现的一个显著特征是,从研发者到普通用户都意识到了其研究开发和应用部署所导致的潜在风险和现实风险。有的人尤其关注生成式人工智能在安全和伦理上可能危及人类生存的长期风险或生存风险,另一些人则将其正在引发的社会伦理风险作为优先关切的事项。不论持哪种立场,人们普遍认为生成式人工智能的研发应该采取审慎发展和负责任创新的态度。

(一) 生成式人工智能的长期风险

最近,作为研发者代表的开放人工智能研究中心(OpenAI)将其首要核心价值设定为“我们致力于建立安全、有益的通用人工智能(AGI),这将对

人类的未来产生巨大的积极影响”,并强调“任何对此没有帮助的事情都超出了我们关注的范围”。OpenAI 的人工智能安全和价值方面的基本理念是通过多次优化和迭代实现安全和有益于人类的通用人工智能,力图避免强大但风险不为人类所能控制的通用人工智能在人们没有意识到的情况下突然降临。但是提出暂停研究人工智能的学者的主要理由是,在伦理和安全跟不上人工智能大模型的加速创新时,应该对相关研究采取更加审慎的态度,甚至有必要暂停或暂缓相关研究。问题是这一呼吁显然不能改变各国在该领域展开的激烈竞争,而且它与 OpenAI 的理念主要基于长期风险和生存风险的视角,不论该视角是否恰当,对于目前大多数以应用为目标的大模型创新来说并非优先事项。

尽管如此,国际科技界和产业界在应对人工智能的长期风险和生存风险方面还是达成了一些基本共识,如强调大模型和生成式人工智能的研发应该更加公开、透明,不能隐藏其研究、应用和部署。为了共同应对人工智能可能导致的长期风险,在科技界和产业界进一步推动开源创新势在必行。为此,一方面,应构建合作研究和开源创新平台,科学家以个人身份共同参与开发,共同研究生成式人工智能的形成机制与导致伦理安全问题的技术原因;另一方面,应通过合作研究和开源创新,发展一系列测试工具,对其安全和伦理问题进行量化测试。

然而,欧美的科技脱钩和去风险等政策的推行,使得生成式人工智能和通用人工智能的竞争不再是单纯的科技竞争,同时也为中国参与包括应对长期风险在内的国际人工智能伦理安全治理带来了困难。

(二) 生成式人工智能的社会伦理风险

当前,大模型呈现出泛在应用的趋势,人们更关心的是基于大模型的大量生成式人工智能应用所导致的社会伦理风险。

首先,生成式人工智能可能导致知识生产方式的根本性变革,对传统的

知识产权产生颠覆性影响。目前,人工智能在艺术创作领域的应用受到了大量的侵权指控,但作为一种内容生成技术,其不能没有数据作为原料,而其所生成的内容与生产原料存在区别,因此,是否侵权较难界定,并且会随着技术的发展而越来越难以界定。这使按照现有法律对人类创作内容的知识产权保护变得越来越困难;同时,对于人工智能生成内容的版权是否应该得到保护也成为一个新问题。鉴于生成式人工智能是一种全新的内容和知识生产方式,其模型训练必须大量使用现有内容,但目前法律上对内容侵权的界定是在这种技术出现之前提出的,因此,如果要发展生成式人工智能并对其实现有效监管,须重新界定其内容使用边界。

其次,生成式人工智能将对工作方式和就业产生根本性冲击。生成式人工智能与人类高度类似的对话能力表明,它将使人机交互的便捷程度大大提升,各行各业现有的工作流程都可能因此得到简化。值得指出的是,大模型生成语言和知识的能力迫使人们将其拥有的类似知识和能力当作“水分”挤了出来,这必然带来对现有工作内容的改变和工作岗位的挤压。特别值得关注的是,这次被取代的可能是受教育程度比较高的专业技术人员,如文秘、咨询人员、翻译人员、文创人员、设计人员甚至医疗和教育等领域的人员都可能面临失业。这实际上会导致大量专业技术人员教育投入的加速折旧,加大并形成新的数字鸿沟,造成规模更大和更复杂的社会问题。

再次,在社会认知层面,人们有可能因过度依赖生成式人工智能而将其视为知识权威、道德权威乃至精神权威,形成人工智能无所不知、无所不能的认知幻觉。随着生成式人工智能进一步发展,其很可能成为普通人日常生活中的助手,帮助其解答知识、辨别是非乃至区分善恶。鉴于人工智能并不能真正理解其所生成的知识内容以及判断是非善恶,而且会产生错误或随意堆砌和编造的内容,故对人工智能的过度依赖难免放大其生成的不准确内容和知识上的错误,甚至对社会认知产生结构性的负面影响。在金融

理财的相关应用中,基于大模型的人工智能顾问可能会出现这方面的问题。另外,生成式人工智能存在过度拟人化趋势。随着人机对话的场景应用日益普遍,在商业服务等应用场景中可能会出现滥用人工智能的风险,如利用客户的情感偏好设计误导性的人机对话等。

最后,对基于大模型的生成式人工智能的伦理风险的基本认识是,生成式人工智能的广泛运用会强化目前已经显现的各种社会伦理问题。一是偏见和歧视。如果用于开发人工智能算法和模型的训练数据有偏见,那么算法和模型也会有偏见,从而导致生成式人工智能的回应和建议出现歧视性的结果。二是信息误导。人工智能语言模型所生成的对话可能会向用户提供不准确或误导性的回应,进而导致错误信息的传播。三是信息滥用。人工智能大模型及其预训练需要收集和处理大量的用户数据,其中必然涉及技术和商业保密数据以及国家安全数据,隐私数据和敏感个人信息可能会被滥用。四是虚假内容及恶意使用。尽管生成式人工智能目前还没有实现大规模商业化社会应用,但从信息网络媒体、虚拟现实和深度合成等技术的发展经验中不难看到,生成式人工智能可能被用于制造不易识别的虚假内容,甚至被恶意使用,从而影响、干预和操纵社会舆论和政治过程。五是对个人自主性的干预。例如,生成式人工智能在商业上可能被用来影响或操纵用户的行为和决策。

三、生成式人工智能在金融领域的应用及主要风险

金融大模型应用创新与伦理治理之所以备受关注,关键在于人工智能在金融领域的应用是金融行业数字化转型大潮推动下的必然趋势。众所周知,金融领域无疑是最具竞争压力、最需要风险管控和复合监管的行业。在这些压力和要求的驱使下,提高效率、节省成本、重塑客户界面、提高预测准确性以及改善风险管理和合规性成为金融领域自我变革的方向。

（一）生成式人工智能在金融领域的创新应用

当前，拥抱数字技术、实现数字化转型成为金融领域推动创新和保持竞争优势的首要战略选择，因而生成式人工智能在金融领域得到迅速应用。根据相关报道和研究报告，摩根大通、高盛等知名企业已开启生成式人工智能在金融领域的应用，相关用例包括构建自动化文档处理功能、虚拟聊天机器人助理、欺诈检测和预防、后台流程自动化、内部软件开发和信息分析等。

从拥抱生成式人工智能的潜力以推动创新并获得竞争优势的目的来看，金融领域出现了一些值得关注的功能强大的创新应用：

（1）优化客户体验和定制个性化推荐。生成式人工智能不仅可以进行自动化日常交互和提供个性化建议，而且具有类似人类表达能力的聊天机器人可以快速高效地响应与客户的互动，实时回应查询。人工智能算法还可以分析客户数据，包括交易历史和浏览行为，为金融产品和服务生成量身定制的推荐，以提高客户的参与度和忠诚度。

（2）金融欺诈检测和预防。生成式人工智能可以通过分析大量交易数据并发现欺诈活动的模式来检测欺诈，金融机构可以运用这些根据历史数据训练的人工智能模型开发出高度复杂的欺诈检测系统，以此实现对金融欺诈的实时监控、异常检测、提前预警和主动打击。据报道，Capital One 和 JMPC 利用生成式人工智能增强欺诈和可疑活动检测系统，显著减少了误报，提高了检测率，降低了成本并提高了客户满意度。

（3）风险评估和信用评分。生成式人工智能算法可以通过分析包含财务记录、信用记录和其他相关信息的广泛数据集来生成预测模型，使贷方更准确地了解借款人的信用度，从而就贷款审批做出更明智的决策，优化贷款策略，降低用户违约的可能性。

（4）交易和投资策略。基于市场历史数据训练生成模型并整合实时市场信息，人工智能算法可以生成个性化的投资建议，优化投资组合策略，乃

至预测市场趋势。值得注意的是:一方面,生成式人工智能具有强大的洞察力,可识别人类观察者可能忽视的有价值的见解;另一方面,也要看到,人类的专业知识和判断力在投资过程中仍然至关重要。因此,生成式人工智能洞察与人类决策的集成将可能创建一种结合人工智能算法和人类直觉优势的协同方法。

(5) 借助自然语言处理提升合规效率。金融科技等行业受到严格监管,在保护客户数据隐私、防止洗钱、反欺诈等方面有严格的合规要求。面对日益强化的监管要求,金融科技等行业可以将生成式人工智能与自然语言处理(NLP)相结合,通过自动分析法律和监管文件来简化合规流程。在此过程中,生成式人工智能可以更高效地识别相关信息,提取关键见解并标记潜在的合规风险,以此减少工作失误、确保合规性(Shabsigh & Boukherouaa, 2023)。

(二) 金融领域生成式人工智能应用的主要风险

近年来人工智能在金融领域特别是金融科技领域的应用,引发了诸多对大数据分析和深度学习等相伴随的固有风险的普遍担忧,当前生成式人工智能在金融领域的应用及其部署前景则进一步加剧了人们对相关风险的关切。其中,备受关注的风险包括数据隐私、算法与模型中嵌入的偏见和歧视、机器学习过程的不透明和不可解释、系统稳健性、网络安全以及对金融稳定的影响等。

在数据隐私方面,机器学习等人工智能在金融领域的应用引发了一些伦理和法律难题,主要包括训练数据集中的数据泄露、通过推理披露匿名数据的可能性、在使用和丢弃数据后对训练数据集中个人信息的记忆、输出结果直接或通过推理导致敏感信息泄露等。导致这些问题的底层原因是人工智能的应用不能不使用个人数据,而使用个人数据就存在侵犯个人隐私和导致数据泄露的可能。因此,对这些问题的应对与缓解,就成为数据驱动的

人工智能应用中伦理、法律和监管等治理与规制的主要关切点及改进方向。随着大模型和生成式人工智能越来越多地嵌入金融业务,数据隐私方面的挑战变得更加严峻。一方面,为了获得足够的欺诈检测、信用评估等方面的能力,大模型需要从互联网和社交媒体等平台抓取信息,必然涉及个人信息的收集使用以及知情同意等各种可能存在争议的问题;另一方面,生成式人工智能系统需要不断通过用户的输入对其大模型进行训练和微调,可能存在敏感财务数据与个人信息泄露的风险。

在偏见嵌入方面,人工智能应用可能会因为训练数据的不完整和缺乏代表性、算法设计受到人类偏见的影响等导致对某些个人或群体不公正的偏见和歧视,如金融决策可能会因为人工智能系统嵌入的偏见而导致金融排斥。大模型和生成式人工智能可能会加剧这一问题。其中值得关注的问题包括:(1)偏见矫正更加困难,大模型在接受大量的在线文本和其他内容数据训练时,难免受到其中嵌入的人类偏见等数据毒性的影响,而且鉴于数据的广泛性和多样性,通过数据选择等方法减少偏见变得更加困难;(2)新内容的产生可能强化人类偏见,在人类提示下产生的新内容可能带有提示者的偏见,而这些新内容本身会成为后续生成内容的数据原料,有可能使得嵌入其中的偏见进一步固化;(3)对生成内容的过度依赖导致其生成内容造成误导,必须强调的是,人工智能目前并没有意识和人格,大模型不能像人那样理解其生成内容的意义和价值取向,人工智能生成内容需要人类强化判断,以尽可能减少不准确、歧视性和错误的内容。

类似地,在可解释性、稳健性、网络安全和金融稳定等方面,大模型和生成式人工智能也加剧了人工智能金融应用的相关风险。其中,具有共性的重要问题包括生成内容的“幻觉”、数据投毒和投毒以及合成数据的使用。生成内容出现“幻觉”是与大模型生成新内容的能力相伴随的一种输出风险,在用于金融服务对话的场景中更加严重,在金融安全和消费者保护方面的风险更大,可能极大地影响金融业务的稳健性。这就要求企业级大模型

的开发应通过更有针对性、质量更好和更透明的训练数据集,最大限度地抑制这一现象。数据中毒是指在训练数据中出现一些特定输入,其可能对大模型的构建产生破坏性的影响,数据投毒是利用这一缺陷破坏大模型训练的准确性等恶意行为。合成数据是通过算法创建的、可模拟真实数据的统计分布的人工数据,可用于测试模型的稳健性,用合成数据替代真实数据有助于降低数据隐私和数据安全方面的挑战,用它模拟消费者的行为模式可以获得更具可行性的见解。合成数据还可以生成更多样化的可能的数据集,更好地捕捉现实世界事件的复杂性,从而减少真实数据集合的代表性不平衡和偏见等问题,因而有助于构建能力更强、更具可解释性和满足监管要求的大模型。但同时,对合成数据的过度使用也会导致数据退化等问题。

四、负责任和可问责的金融大模型伦理治理初探

近年来,面对人工智能大模型的巨大机遇和风险,大模型的伦理和法律风险及其治理成为全球关注的话题。如上文所述,由于大模型一方面涌现出类似人类的强大智能,另一方面又存在数据来源难追溯、易产生难以分辨的虚假信息和知识幻觉、知识产权难界定、模型难审计以及对技能和就业的巨大冲击等新问题,对建立在其上的生成式人工智能的伦理治理和法律规制带来了全新的挑战。

当前,大模型部署应用的伦理治理主要聚焦于几个关键问题。首先,如何实施负责任且可问责的人工智能审计,这涵盖了大模型的决策机制、构建方式以及使用途径。有研究提出了包含治理审计、模型审计和应用审计在内的三级审计框架,以应对这一挑战。其次,如何以透明的方式向消费者传达人工智能的伦理标准,明确揭示潜在风险,这是政府与行业主管部门的重要任务。这意味着它们需要为大模型的构建方式制定清晰、广泛的标准,并向消费者和员工明确这些标准。最后,政府和企业的人力资源部门须迅速

感知并应对人工智能对技能和就业市场的冲击。人力资源部门应迅速识别这一冲击,并从战略层面应对这一前所未有的压力。他们须为员工提供技能提升的机会,并根据预测所需的新技能和劳动力调整相关策略。

由于生成式人工智能功能强大,且已经呈现出走向通用人工智能的可能,这就使得通过伦理治理和法律规制发展负责任和可问责的人工智能显得更加迫切。由于大模型和生成式人工智能技术的发展日新月异,鉴于法律规制的相对滞后性及其主要功能在于对结果的监管,再考虑到对伦理风险的识别有助于尽早感知法律风险,从全生命周期治理和过程管理的角度来讲,对大模型的伦理治理无疑更有利于对大模型潜在伦理和法律风险的整体治理。因此,一方面,应将大模型的伦理治理作为包括法律规制在内的大模型治理的基础性工作,对数据隐私和安全、偏见和公平、可解释性和透明度、可靠性和稳健性、模型风险评估和管理、人类监督和把关等基本的伦理关切提出必要的要求;另一方面,为了保持大模型治理的整体性,伦理治理与法律规制等应该遵循相同的治理原则,法律规制也应该与相关的伦理原则相互协调。

(一) 确立金融领域大模型伦理治理的基本理念

为了更好地开展大模型的伦理治理,首先必须确立伦理原则和治理原则等基本理念。在具体的大模型研究应用中,相关研发者、部署者和管理者可以对其所进行的研究与应用制定出有针对性和简洁可行的伦理原则、治理原则。

这些伦理原则的制定不应闭门造车,其实质是对目前已经发布的各个层级的规范中的伦理原则和治理原则的具体落实。目前,这些规范大致包括以下三个层级:

一是科技伦理原则和科技伦理治理原则。《关于加强科技伦理治理的意见》明确提出了增进人类福祉、尊重生命权利、坚持公平公正、合理控制

风险、保持公开透明等科技伦理原则,同时提出了伦理先行、依法依规、敏捷治理、立足国情、开放合作等治理要求。

二是人工智能伦理规范和治理原则。《新一代人工智能治理原则:发展负责任的人工智能》明确提出了和谐友好、公平公正、包容共享、尊重隐私、安全可控、共担责任、开放协作、敏捷治理等八项原则;《新一代人工智能伦理规范》明确了增进人类福祉、促进公平公正、保护隐私安全、确保可控可信、强化责任担当、提升伦理素养等六项基本伦理要求。此外,我国参与制定的《人工智能伦理问题建议书》提出了相称性和不损害、安全和安保、公平和非歧视、可持续、隐私和数据保护、人类的监督和决定、透明度和可解释性、责任和问责、认识和素养、多利益攸关方与适应性治理和协作等原则(联合国教科文组织,2021)。

三是金融领域和生成式人工智能相关的伦理指引和管理规定。《金融领域科技伦理指引》提出,在金融领域开展科技活动需要遵循守正创新、数据安全、包容普惠、公开透明、公平竞争、风险防控、绿色低碳等七个方面的价值理念和行为规范;《生成式人工智能服务管理暂行办法》提出,“国家坚持发展和安全并重、促进创新和依法治理相结合的原则,采取有效措施鼓励生成式人工智能创新发展,对生成式人工智能服务实行包容审慎和分类分级监管”,从中可见我国对生成式人工智能的监管和治理的基本原则。

确立金融领域大模型伦理治理的基本理念的第一步就是全面了解这些原则,认真理解和思考它们背后的价值观和优先考量,并由此展开价值观反思和优先事项讨论,进而明确大模型的研发和应用中应遵循的价值观和优先考量。然后,在理解和讨论的基础上,对这些原则进行归纳,并列出其所针对的问题;在此基础上探寻这些原则的基本内涵,反思其现实有效性和可能的含混之处,探讨不同的规范中相同或相近的原则之间的互操作性。最后,根据需要解决的问题和可能面对的问题,拟定出内涵相对明晰、互操作性强、既全面系统又重点突出的伦理原则和治理原则。而确立大模型伦理

治理的基本理念的过程,无疑有助于研发和创新应用者将价值观等理念纳入其心智模式和行为指南。

(二) 探索金融大模型伦理治理的工作思路

金融大模型的伦理治理要求进一步明确我国科技伦理和人工智能伦理治理的指导思想,并据此确立主要的工作思路。

正如《科技伦理审查办法(试行)》所强调的,我国开展科技伦理治理的宗旨是“强化科技伦理风险防控,促进负责任创新”。而《新一代人工智能治理原则:发展负责任的人工智能》这一文件,再次体现了我国开展科技伦理治理与人工智能伦理治理的指导思想是坚持促进创新与防范风险相统一,推动负责任的研究和创新。要使这一宗旨得到贯彻,关键在于以下三个方面的认知。

其一,对大模型和生成式人工智能兼具高度创新性和不确定性风险的技术的伦理治理和法律规制本身是一种综合性的协同创新,目的是使其充分发挥出技术创新潜力的同时又为社会所接受和信任。

其二,伦理治理和法律规制所面对的综合性的协同创新实际上是非常复杂和困难的系统工程,需要多方协同共治。从系统和整体创新的角度来讲,研发创新主体和规制监管主体之间的协同尤为重要:一方面,研发创新主体要开展负责任的创新,认真预见并实时发现各种不能忽视的风险,积极采取应对措施;另一方面,规制监管主体要充分了解生成式人工智能给社会与生产生活带来的颠覆性变化,把握其中技术逻辑的变迁,寻求对促进创新发展更具弹性和更为有效的治理。

其三,金融领域大模型的伦理治理应基于相称性原则,根据其产生的具体风险确定治理力度。《生成式人工智能服务管理暂行办法》提出,“对生成式人工智能服务实行包容审慎和分类分级监管”,强调主管部门将“针对生成式人工智能技术特点及其在有关行业和领域的服务应用,完善与创新

发展相适应的科学监管方式,制定相应的分类分级监管规则或者指引”。这一规定无疑体现了促进负责任的创新与敏捷治理的治理原则相结合的精神。对此,有法律学者指出,生成式人工智能的治理应改变我国原有的“技术支持者—服务提供者—内容生产者”的监管体系,实施“基础模型—专业模型—服务应用”的分层规制,并针对不同的层次适配不同的规制思路与工具。具体而言:基础模型层的治理应以发展为导向;专业模型层的治理应以审慎包容为理念,关注重点领域与场景的分级分类,设置合理的法律责任水平;服务应用层的治理应沿用原有的治理理念与监管工具,保证我国人工智能监管的协调性与一贯性,同时还应建立敏捷治理的监管工具箱,细化合规免责制度,给新兴技术发展留下试错空间(张凌寒,2023)。

最近,英国政府发布的《促进创新的人工智能监管方法》白皮书提出,在监管观念上,应避免可能扼杀创新的高压立法,而应采取更具适应性的方法来监管人工智能。白皮书概述了监管机构应该考虑的五项原则,以更好地促进在其所监控的行业中安全和创新地使用人工智能。这些原则分别是:第一,安全性和稳健性,人工智能的应用应该以安全和稳健的方式运行,并仔细应对管理风险;第二,透明度和可解释性,开发和部署人工智能的组织应该能够沟通何时以及如何使用它,并以适当的细节解释系统的决策过程,以与使用人工智能带来的风险相匹配;第三,公平,人工智能的使用方式应符合现行法律,并且不得歧视个人或造成不公平的商业结果;第四,问责制和治理,需要采取措施对人工智能的使用方式进行适当的监督,并对结果进行明确的问责;第五,可竞争性和补救,人们需要有明确的途径来对人工智能产生的有害结果或决定提出异议。这些观念对于探索大模型的治理具有一定的启发和借鉴价值(Secretary of State for Science, 2023)。

(三)加强科技伦理审查,走向可问责的大模型治理

使金融大模型的研发和应用实现负责任的创新,关键在于可问责的风

险评估和监督审查。对此,联合国教科文组织《人工智能伦理问题建议书》指出,应建立适当的监督、影响评估、审计和尽职调查机制,包括保护举报者,确保在人工智能系统的整个生命周期内对人工智能系统及其影响问责。为此,技术和体制方面的设计都应确保人工智能系统运行的可审计和可追溯。

2023 年 10 月,为规范科学研究、技术开发等科技活动的科技伦理审查工作,科技部会同教育部、工信部等十部门联合印发了《科技伦理审查办法(试行)》。这一办法的颁布无疑是我国加强科技伦理治理的里程碑,大模型的治理可以以此为抓手强化问责机制。根据审查办法,从事生命科学、医学、人工智能等科技活动的单位,研究内容涉及科技伦理敏感领域的,应设立科技伦理(审查)委员会。

值得关注的是,《科技伦理审查办法(试行)》对科技伦理(审查)委员会的职责提出了很高的专业要求和能力要求。除了开展科技伦理审查工作之外,还要按要求跟踪监督相关科技活动全过程,为科技人员提供科技伦理及其风险评估方面的咨询和指导,组织相关的业务和知识培训。为了更好地履行科技伦理审查责任,有效推进可问责的大模型治理,应该在加强相关认识和提升科技伦理素养的基础上推进相应的人才建设。

具体的建议是,从事金融领域大模型的机构和企业要注重培养乃至设置内部的科技伦理分析师和科技伦理架构师或科技伦理专员。其中,科技伦理分析师通过把握相关事实和证据,澄清相关观点、概念、问题与选择,为科技伦理审查等工作提供支撑;科技伦理架构师负责机构和企业的科技伦理治理的框架设计与整体协同,他们要在全面了解和参与内外相关政策和决策过程的基础上,负责协调机构和企业的科技伦理风险防范,在科技伦理审查工作实施中保障审查主体职责的履行。

参考文献

联合国教科文组织,2021,《人工智能伦理问题建议书》,https://unesdoc. unesco. org/ark:/48223/pf0000380455_chi。

张凌寒,2023,《生成式人工智能的法律定位与分层治理》,《现代法学》第 4 期。

Secretary of State for Science 2023, "A Pro-Innovation Approach to AI Regulation." https://www. gov. uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper.

Shabsigh, G. & E. Boukherouaa 2023, "Generative Artificial Intelligence in Finance: Risk Considerations." https://www. imf. org/en/Publications/fintech-notes/Issues/2023/08/18/Generative-Artificial-Intelligence-in-Finance-Risk-Considerations-537570.