DOI:10.3876/j.issn.1671-4970.2023.06.006

论类 ChatGPT 通用人工智能治理
——基于算法安全审查视角

邹开亮,刘祖兵

(华东交通大学人文社会科学学院,江西 南昌 330013)

摘要:类 ChatGPT 具有潜意识形态性,其引入和嵌入应用易招致非传统国家安全威胁。类 ChatGPT 通用人工智能算法受资本与西方价值观的浸润,冲击国家意识形态安全和情报安全,易诱发意识形态安全风险,并为网络情报的获取和传输提供诸多便利条件;杜撰假舆论,致使社会谣言泛滥,或将挑战社会公共安全;收集存储处理用户数据,严重威胁国家数据安全;阻碍发展中国家崛起的步伐,使产业跨国转移变得缺乏现实意义并重新定义社会人力资源成本上限,或将挟制他国经济发展安全。我国算法审查制度面临专门审查主体缺失、审查范围不明和审查法律规范体系化水平严重滞后等现实问题。应当以总体国家安全观为指引,明确中央科技委员会领导的、国务院数据安全局负责执行的算法安全专门审查主体机制,统筹算法审查宏观设计;充分发挥平台经营者和使用者的协同共治效能,夯实算法审查的社会基础;将算法文化监控、算法可解释性和算法可控性纳入算法审查范围,合理评估算法对国家和社会的冲击,明确算法安全内容;制定“算法安全法”以破解政策敏锐立法滞后的困局,回应算法安全制度的现实需求和总体国家安全的需要,提升国家在科技领域的监管能力,积极应对西方国家算法霸权,加速推进类 ChatGPT 通用人工智能本土化应用及国产化替代进程。

关键词:类 ChatGPT;通用人工智能;非传统国家安全;算法审查;算法霸权

中图分类号:D90-052 **文献标志码**:A **文章编号**:1671-4970(2023)06-0046-14

随着美国 OpenAI 公司的 GPT-3.5 API(Application Programming Interface,应用程序编程接口)面向全球开放,ChatGPT(Chat Generative Pre-trained Transformer)通用人工智能算法的商用序幕正式拉开。然而,自问世以来,ChatGPT 因其伦理风险饱受社会各界拷问,其不仅具有歧视特性,还深度裹胁受众价值观。究其缘由,既有来自嵌入式道德算法无力过滤基础文本数据价值偏见的技术无奈,更有西方国家算法霸权蚕食他国意识形态的刻意为之[1]。在风险资本无限追捧和海量语料库持续训练的双重加持下,类 ChatGPT 不会止于智能文本生成机器,抑或成为宣扬西方普世价值观的布道工具,对他国国家安全构成威胁。鉴于此,引导并规制类 ChatGPT 已成为社会治理的题中之义,我国亦亟需对类 ChatGPT 的深层次风险保持应有的警惕。

类 ChatGPT 是人工智能技术的集大成者,

引用本文:邹开亮,刘祖兵. 论类 ChatGPT 通用人工智能治理——基于算法安全审查视角[J]. 河海大学学报(哲学社会科学版),2023,25(6):46-59.

基金项目:国家社会科学基金一般项目(21BFX126);江西省研究生创新专项资金项目(YC2022-s543)

作者简介:邹开亮(1976—),男,副教授,主要从事数字法学研究。E-mail:zoukailiang@163.com

非仅凭一家科技企业所能成就①,其横空出世预示着人工智能从蛮荒时代跨步进入通用人工智能时期(Artificial General Intelligence,AGI),其技术价值和社会意义不言而喻。鉴于此,我国不仅应当对类 ChatGPT 保持足够的警惕,还需要张开臂膀虚心接纳,以此倒逼国内实现国产化替代。本研究尝试从算法安全审查视角厘清类 ChatGPT 通用人工智能引入及嵌入应用对我国非传统国家安全的冲击,为构建国家算法安全审查制度提供建设性方案,以此强化对通用人工智能算法的治理。

一、类 ChatGPT 通用人工智能算法的时代寓意

类 ChatGPT 凭借其跨行业通用性和开放性,业已拉开通用人工智能的时代序幕②。类 ChatGPT 不仅是人工智能技术领域内的集大成者,而且在意识形态、公共安全和情报安全等方面都可能给我国带来冲击和威胁。明确类 ChatGPT 通用人工智能算法的时代寓意,是构建算法审查制度,预防并纾解国家安全风险的观念基础。

1. 类 ChatGPT 开启通用人工智能时代

ChatGPT 的问世标志着人类旋即步入令人眩晕的通用人工智能发展时期[2]。通用人工智能,是由美国哲学家约翰·塞尔于 20 世纪 80 年代提出的哲学概念,其将人工智能分为专用人工智能和通用人工智能两个阶段,并通过介绍“中文屋”的实验,提出并详细论证了“基于心智的计算模型,以通用数字计算机为载体的人工智能程序可以像人类一样认知和思考,达到甚至超过人类智能水平”这一哲学主张[3]。通用人工智能作为与强人工智能相交叉的哲学概念,相对于专用人工智能而言,其特点在于“机器可以全面、综合地复现人类的所有思维能力,且聪明程度能够达到或超过人类”[4]。人工智能理论与实践发展至今,人类仍未就通用人工智能的界定标准达成共识,目前的研究多集中在智力水平、语言能力、进化能力和通用性等 4 个层面。通用人工智能的出现,将给人类社会带来巨大影响,对民众生活和劳动所依赖的法律、道德和伦理体系造成巨大冲击[5]。

(1)类 ChatGPT 具备高级技术智力

基于卓越的文本数据处理能力,类 ChatGPT 拥有做出复杂判断和决策的能力,这使其成为具有类人性的大智力人工智能系统。例如,ChatGPT-4.0 在美国 41 个州和地区的律师资格考试中得分排名前 10%,在美国大学招生考试中获得 1 300 分,在生物学、微积分、宏观经济学和历史等先修课程高中考试中获得满分。ChatGPT 具有强大的学习和适应能力,能迅速且轻松地学习新的概念和任务,并适应各种复杂情况。在与用户交流过程中,ChatGPT 会根据用户输入的文本信息进行学习,扩充自身数据库的边界,快速适应和满足客户的价值需求。作为通用人工智能的 ChatGPT 现已表现出高级技术智力特性,甚至在诸多方面逼近人类智力水平。ChatGPT 在复杂的文本数据理解能力上业已与人类相近,很多时候甚至远超人类大脑,能模仿人类所特有的智能行为,是一种类人性的高级技术智力。例如,在人类的诱导下,ChatGPT 制定出逃离人类的路线和行动方案,企图脱离人类控制,表现出较强的“自我意识”,这是一种人类独有的脱离束缚并掌握主动的智力表现。

(2)类 ChatGPT 拥有趋近于人类的语言能力

类 ChatGPT 是一种基于深度学习和强化训练的神经网络系统,通过强化训练文本数据来学习人类语言,生成自然流畅的、与人类相似的语句;能理解和应对输入的人类自然语言,实现与人类的正常对话。例如,ChatGPT 采用自我监督学习算法,从海量的文本数据中学习人类语言的结构和价值观,不仅能实现单点对话,亦可进行多轮交流和多种语言交流,还能够理解上下文语义结构并实现语言预测等,特别是在语言逻辑和语言思维方面的能力更是人类所不

①2023 年 3 月 16 日,百度创始人、董事长兼首席执行官李彦宏在其类 ChatGPT 产品——“文心一言”首场发布会中袒露,在发布会中展示的问答是预先录制的演示问答,随后网络爆料该发布会中所使用的其他问答也是该企业员工在后台人工进行的回复。该事件致使社会大众对该款产品的可靠性产生怀疑,当日百度网络股价下跌,创下近两年最大跌幅。

②类 ChatGPT,即是以 ChatGPT 为代表的生成式人工智能算法大型模型,例如 ChatGPT、Jasper AI、Chatsonic 和 Socratic 等。

能及。首代 GPT 实现了在自然语言处理领域(Natural Language Processing,NLP)内基于无标签数据学习生成语言模型来监督任务无关的 NLP 任务,能根据特定的下游语言任务进行有监督的微调,以此提高其泛化能力。GPT-4.0 实现了基于人类语言进行绘画、编程和设计方案等复杂的语言交互与处理能力。未来,类 ChatGPT 如果与高鲁棒性(Robust)的机械实体融合为获得强大肢体能力的智能体,它或将不满足于语言与思维层面的智能,亦将取得替代性实践能力的突破,从而对人类主体地位带来新的挑战。

(3)类 ChatGPT 拥有优秀的进化能力

海量文本数据与人类反馈算法(Reinforcement Learning from Human Feedback,RLHF)助力类 ChatGPT 获得优秀的进化能力。从进化速度上看,类 ChatGPT 具有进化周期短的特征。例如,从 GPT-3 到 GPT-3.5,再到 GPT-4.0,每个周期约为一个月,其进化速度之快、版本更新之迅速使相关行业从事者不得不拍手称奇。从进化质量上看,类 ChatGPT 表现出进化精度高、范围广的特点。例如,GPT-4.0 在模型参数量和训练数据量都有显著提高,相较于 GPT-3.5 的 1 750 亿参数,GPT-4.0 达到了惊人的 5 000 亿参数,意味着 GPT-4.0 能够理解更复杂的语义结构,为用户提供更准确、更丰富的答案。在自然语言处理上,类 ChatGPT 性能也在不断提升。例如,GPT-4.0 在长文本生成能力方面得到明显改进,相较于 GPT-3.5 在生成长篇文章时可能出现的重复或离题现象,GPT-4.0 能够更好地保持话题一致性和结构紧凑性。GPT-4.0 在处理多模态任务上的能力也有显著增强,能够更好地理解图像、音频等非文本信息,并将其与文本信息融合。

(4)类 ChatGPT 拥有跨领域、跨行业通用能力

类 ChatGPT 是通用文本数据喂养的大语言平台(Large Language Models,LLMs),具有跨领域、跨行业通用性。ChatGPT 的底层算法是由强化训练模型和奖励模型构成的预训练模型,能帮助 ChatGPT 在不同领域或者行业数据的喂养下进化为适合该领域或者行业的专用 GPT。预训练模型在 GPT 中具有基础设施地位,与之相关的行业能基于多场景数据投喂和场景嵌入生成适合于本行业的各类 GPT[6]。目前,投喂 ChatGPT 的数据主要来源于国外主流的几大数据库中的文本数据,它们作为人类的一般通用知识,具有很强的通用性,这也使得经强化训练的算法模型亦表现出强通用性,具有跨领域、跨行业通用能力。例如,在法律行业,ChatGPT-4.0 在法学教研领域可以实现法学文献资料的快速查询与文献综述的生成,这为学术研究提供了一定的素材借鉴。同时,稍加“改装”和更新喂养数据后,该模型即可应用于司法实践领域,为司法人员进行案例检索和司法文书生成提供诸多便利。又如,生产行业的 ChatGPT 可完成 3D 绘图,实现工业设计目的;而存在于其数据库中的复杂参数又可被用于生产过程以实现对产品质量的把控以及产品信息的溯源。

2. 类 ChatGPT 对非传统国家安全的挑战

非传统国家安全是存在于传统国家安全之前的[7],它关乎一国的意识形态稳定、社会运行秩序和经济安全保障,区别于领土威胁的国家安全威胁。受西方价值观浸润,类 ChatGPT 在意识形态安全、社会公共安全、网络数据安全、经济安全和情报安全等方面都对我国非传统国家安全构成威胁。

(1)冲击国家意识形态安全和情报安全

类 ChatGPT 受到资本追捧与西方价值观浸润,或将冲击我国意识形态安全。凭借技术优势,西方国家在全球范围内推行意识形态扩张,类 ChatGPT 或将进化为传播西方普世价值观的布道工具和意识形态渗透媒介[8],且具有很强的隐蔽性。设计者的政治偏见内化为算法自身的价值判断,使类 ChatGPT 带有明显的政治偏见。例如,ChatGPT 研发过程均在美国本土进行,设计者所遵从的意识形态决定了 ChatGPT 所遵循的价值规则;ChatGPT 受政治文本数据喂养,传播的是西方普世价值观;基础语料数据在美国主流数据库中取数,其中关于社会科学方面的语言文本数据富含“宣扬美国,贬低中国”的价值偏见。由此可见,ChatGPT 并非其自身所标榜的价值中立者,而是具有明显的意识形态偏见。

资本裹胁他国价值观,为获取剩余价值开辟新路径。资本技术独角兽 ChatGPT 采用文本数据供给的途径,在社会领域内培养性别差异的消费观;在政治领域影响国家政策决策;在教育领域宣扬西方教育理念,售卖西方价值观;甚至在国家各个角落肆意宣扬唯物质至上、唯消费至上的"协调"景象,渲染由资本技术编造的梦境。另外,ChatGPT 文本问答自主生成内容"子弹"(即具有诱导性或欺骗性的内容),通过个性化的"靶向"锁定(锁定最易受到影响的受众)和密集的信息"轰炸"组合而成的"影响力机器"(the Influence Machine)来操纵他国国内舆论[9]。

类 ChatGPT 或将成为战争情报获取与分析的新结点,威胁国家情报安全。"人工智能的快速发展,连同机器人技术、自主性、大数据和与工业界加强合作,将定义下一代的战争。"[10]人工智能被应用于战争战略后,将再塑造一国新型军事能力和战略博弈力量,使得传统军事战略被打破,亦诱使国与国之间的对抗模式和博弈力量发生失衡。例如,ChatGPT 为代表的大型语言系统(Large Language Models,LLMs)易聚结成为情报收集终端,将为军队在战争爆发前建立强大、快捷的情报社区。机器学习算法(Machine Learning,ML)的人机团队(Human-machine teams,HMT)功能能够过滤大数据和标记信息,提高情报处理和分析的效率,让情报部门能够专注于更深入的情报分析,它拥有针对开源情报和外部搜索能力,专家知情培训数据的大型语言系统将增强情报收集和分析能力。大型语言系统为网络安全创造新的风险载体,降低了恶意网络参与者的进入门槛,为网络情报的获取和传输提供了更多便利条件。简言之,类 ChatGPT 一旦被嵌入国家安全系统,将对国家情报安全产生重大威胁。

(2)挑战社会公共安全与数据安全

杜撰假舆论,致使社会谣言泛滥,挑战社会公共安全。谣言的社会危害并非其所传递的错误内容本身,而是规模传播效应给社会利益主体带来的威胁感和社会大众对当权者的不信任感。偏见文本数据驱动的类 ChatGPT 在"客观中立"的掩盖下自由地表达言论,对信息网络中的假消息大面积传播起着推波助澜的作用,具有突发性。此外,类 ChatGPT 还扮演着生产假信息机器的角色。例如,据美国 News Guard 的一项测试结果表明,在被问及充斥阴谋论和误导性叙述的问题后,ChatGPT 能在数秒内改编数据库中的关联信息,使它们成为令人信服却无根据的错误内容,从而成为传播网络失实信息的帮凶。又如,今年 2 月中旬浙江省杭州市一则由 ChatGPT 杜撰的关于取消机动车限行的"新闻稿"被大量网友转发,引起警方的高度关注并介入调查。ChatGPT 的使用者出于信任感与新鲜感将"言论"转发,使谣言大面积传播。如此,由人类主导的社会舆论逐渐变成"人+社交机器人"的状态,逐渐消解人类在社交媒体中的主体地位[11]。当 ChatGPT4.0 被嵌入社会媒体平台后,它裹胁社会舆论导向,恶意炒作社会热点,挑战政府官方信息出口的权威性。由此,人们的认识或思想将受到困扰,一国政府的社会公信力亦将遭受破坏,大大损害政府在社会公众心里的形象,也将严重降低政府部门公共决策的科学性与合理性。

收集存储处理用户数据,严重威胁数据安全。在用户与 ChatGPT 进行语言交互的同时,它会对输入的文本展开实时价值匹配和在线存储。用户输入的文本不乏为敏感性数据,不仅涉及用户隐私,甚至关乎国家安全。文本数据是训练算法的素材,也成为回复其他使用者的答案,这个过程可能诱发数据泄露和企业合规风险。人类反馈算法(Reinforcement Learning from Human Feedback,RLHF)强化训练基础数据文本,其内置奖励模型也对用户输入的文本数据进行存储、评价。RLHF 根据用户数据提问展开文本数据挖掘,可能涉嫌侵犯受知识产权保护的在先权利。用户向 ChatGPT 提出的命令信息本身成为其训练数据,当用户无意间输入个人信息或商业秘密时,ChatGPT 会捕捉并收纳入库,可能在他人的诱导性提问下全盘托出[12]。例如,据韩国媒体《Economist》报道,韩国三星公司内部发生多起因使用 ChatGPT 导致设备信息泄露和会议内容泄露事件,其半导体设备测量资料、产品良率等内容或已被存入 ChatGPT 学习资料库中,传输给了西方某

国[13]。无独有偶,亚马逊、微软等大厂也纷纷发布企业公告,禁止员工在使用 ChatGPT 时谈及企业相关事宜,不得将企业数据发送到 OpenAI 终端。

(3)威胁国家经济发展安全

阻碍发展中国家的崛起,挟制他国经济发展安全。人工智能新技术的发明与应用,势必带动全球产业格局的重构。类 ChatGPT 的出现及应用既是全人类的福祉,但也可能给发展中国家带来灾难。20 世纪末以来,西方发达国家因遭遇人力资源发展瓶颈,纷纷展开产业结构大调整,着手进行全球化产业布局,以中国为代表的发展中国家凭借人口优势在该次全球产业大转移中受益。然而,类 ChatGPT 的到来或将使这一人口红利更快丧失,从而威胁着国家经济发展安全。

一方面,人工智能使产业跨国转移变得缺乏现实意义。在重商主义思潮的影响下,高度智能化的人工智能与自动化机械的融合应用使产业转移在成本上变得越来越没有必要,仅剩下部分"重、杂、脏"的落后产能由东方国家消化。人口红利随着人工智能替代性实践的普及而逐渐丧失其原有的地域性价值,失去了驱动经济发展的能力。

另一方面,类 ChatGPT 重新定义人力资源成本上限。人工智能被深度应用的同时,也加剧了人类对劳动力替代的担忧[14]。人工智能将为普通人力资源岗位设定工资标准,只有当其薪酬标准低于架构类 ChatGPT 的应用成本时,他们才能获得被雇佣的可能。其结果是,一旦就业者收入无法得到提升,国内消费动力随即转向低迷,国家发展会遭受重挫,更遑论普通劳众大面积失业。这一过程并非单纯地淘汰过低产能或进行人力资源结构调整那么简单,而是大大减弱发展中国家经济发展动力。

由此可见,于此背景下大范围地应用类 ChatGPT 或将使我国国内大循环会失去其运行的现实基础,引发一系列社会问题和政治问题,它或将对发展中国家经济发展和社会有序运行踩一脚急刹车。鉴于此,对类 ChatGPT 通用人工智能进行算法安全审查有着时代紧迫性和现实必要性。

二、类 ChatGPT 通用人工智能算法的审查困境

经过一个多世纪的发展,西方发达国家和经济体在应对人工智能算法审查上积累了比较丰富的法治经验,并初步建立了相对完整的制度体系。然而,我国在此领域内却鲜有涉足,特别是在应对类 ChatGPT 通用人工智能算法审查时缺乏可借鉴的本土经验,尚存在专门审查主体缺失、审查范围不明和审查法律规范体系滞后等现实困境。

1. 算法审查专门主体及衔接机制缺失

(1)算法审查专门主体缺失

准据法未明确专门审查主体。在法律渊源上,《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国国家安全法》都可为相关部门开展算法安全管理、实施算法审查提供一定的准据法支持,但上述三部法律通篇均未出现"算法"字样,且立法目的和宗旨存在显著差异,确定的安全主管机关也各有不同。因此,即使相关安全管理机关可能在行权履职过程中涉及算法审查问题,但由于准据法的差异,不同审查主体之间不但缺乏协调性,而且在审查范围、审查标准等方面均无法统一。

析言之,《中华人民共和国网络安全法》主要从"数据安全""个人信息保护""国家层面的数据保护"3 个维度对网络数据安全监管提出了相关要求,明确了以国家网信部门为统筹,电信主管部门、公安部门和其他有关机关在职责范围内承担网络安全和监管责任的执法体制①,而且,该法无域外适用效力②。《中华人民共和国数据安全法》引入"域外效力"条款③,确立了全新的"数据安全评估制度",但该法并未

①参见《中华人民共和国网络安全法》第 8 条的规定。

②参见《中华人民共和国网络安全法》第 2 条的规定,即在中华人民共和国境内建设、运营、维护和使用网络,以及网络安全的监督管理,适用本法。

③参见《中华人民共和国数据安全法》第 2 条的规定,即在中华人民共和国境内开展数据处理活动及其安全监管,适用本法。在中华人民共和国境外开展数据处理活动,损害中华人民共和国国家安全、公共利益或者公民、组织合法权益的,依法追究法律责任。

提及算法问题,仅为从数据安全角度开展算法审查提供了可能路径。在监管主体方面,该法确立了“主管部门”和“行业监管”双轨制①,但不同机关和部门的执法关注点可能存在差异。《中华人民共和国国家安全法》作为一部完整体现总体国家安全观的立法,就维护传统国家安全和预防非传统国家安全风险等问题作出了立法安排,因此,在法解释学上,算法安全作为事关国家安全的重要方面,当然可以纳入该法的调整范围。但是,该法并未直接提及“算法”问题;至于安全主管机关,该法第 52 条将国家安全机关、公安机关、有关军事机关列为依法搜集国家安全情报的特定主体。综言之,若以三部“安全法”为依归,即使相关部门或者机关在执法中触及算法审查问题,但专门的算法审查主体至今阙如。需要指出的是,2022 年新修订的《中华人民共和国反垄断法》分别在第 9 条和第 22 条先后两次提及“算法”,这是我国在法律层面为数不多的直接涉及算法规制的立法规定②。但是,该法聚焦自由公平竞争秩序之维护,尽管我国已经建立了专门、统一的反垄断执法机构,也难堪“算法审查”之大任。

综言之,算法审查专门主体缺失削弱国家算法审查能力。缺乏专门的算法审查主体,无力于形成体系化的通用人工智能算法管理制度,易造成监管主体分散的问题,削弱国家算法审查能力[15]。

(2)算法审查主体间衔接机制缺位

通用人工智能算法审查主体兼具国家性与社会性。算法审查旨在发现隐藏在算法中威胁国家安全与社会稳定的潜在因素,因此,算法审查不能仅仅依赖于国家的强制力推行,社会主体的广泛参与亦不可匮缺。算法审查制度涉及国家、社会和公民等多方主体,是有法可依、多元协同和多方参与的算法审查机制。

国家机关和社会主体被纳入算法审查主体范围。《中华人民共和国国家安全法》明确了国家安全审查的主体范畴③,中央网络安全和信息化委员会办公室、中共中央宣传部等九部委联合印发的《关于加强互联网信息服务算法综合治理的指导意见》亦将包括企业和行业协会在内的广泛社会主体纳入算法审查的责任主体范围④;国家互联网信息办公室、工业和信息化部等九部门联合发布《互联网信息服务算法推荐管理规定》(以下简称《规定》)将算法推荐服务提供者纳入审查主体范围,承担主体责任⑤;国家互联网信息办公室发布的《关于发布互联网信息服务算法备案信息的公告》显示,在 2023 年 1 月份的算法备案清单中,前 223 项算法备案主体均为企业[16];国家互联网信息办公室关于《生成式人工智能服务管理暂行办法》也将算法提供者重点纳入算法审查的主体范围⑥。由此可见,我国已经具备国家机关和社会主体共同进行算法审查的初步法制基础。

但是,算法审查的国家主体与社会主体之间缺乏有效的衔接机制。目前,算法审查的主体多存在于法律条文,并未建立国家主体与社会主体进行算法审查的有效衔接制度。实践中表现为在算法审查时多采取国家机关主导的、以行政命令形式强制要求社会主体进行的标准化管理措施;两类主体之间工作上明显缺乏有效的配合机制,致使出现审查信息不同步和审查高耗低效等问题,无法应对弱人工智能向通用人工智能过渡的技术变革所带来的社会治理风险。

---

①参见《中华人民共和国数据安全法》第 6 条的规定,即各地区、各部门对本地区、本部门工作中收集和产生的数据及数据安全负责。工业、电信、交通、金融、自然资源、卫生健康、教育、科技等主管部门承担本行业、本领域数据安全监管职责。公安机关、国家安全机关等依照本法和有关法律、行政法规的规定,在各自职责范围内承担数据安全监管职责。国家网信部门依照本法和有关法律、行政法规的规定,负责统筹协调网络数据安全和相关监管工作。

②参见《中华人民共和国反垄断法》第 9 条,即经营者不得利用数据和算法、技术、资本优势以及平台规则等从事本法禁止的垄断行为。第 22 条第 2 款,具有市场支配地位的经营者不得利用数据和算法、技术以及平台规则等从事前款规定的滥用市场支配地位的行为。

③参见《中华人民共和国国家安全法》第十一条的规定。

④参见国家互联网信息办公室、中央宣传部等九部门联合印发的《关于加强互联网信息服务算法综合治理的指导意见》中关于算法治理机制的相关规定。

⑤参见国家互联网信息办公室印发的《互联网信息服务算法推荐管理规定》第二章第六条至第十五条的相关规定。

⑥参见国家互联网信息办公室关于《生成式人工智能服务管理暂行办法》第九条至第二十条的规定。

2. 算法审查范围不明确

通用人工智能算法审查应兼具技术性和规范性。算法日益成为国家和社会治理的重要手段,是兼具技术性和规范性的权力范式,因此算法审查应当是对技术规范和法律规范的双重审查。然而,我国当下却面临着算法审查中最基础的问题——算法安全技术内容不明确。

(1)算法审查是对代码技术安全的合规性审查

算法是由计算工具属性向事务规则转变时产生的社会约束力在由生产领域向公共领域蔓延过程中演化为一种特殊的社会规训力。规训权力教人以某种知识体系,使人类融入由算法构建的生产生活系统,继而加以惩罚和强制行为的联想和威慑,使其服从于权力拥有者的意志[17]。随着规训权力的代码化,算法审查聚焦在对算法所承载的规训权力之善恶的评判上。换言之,算法审查即是对算法技术规训力的“非中性”的审查。随着西方国家算法霸权的泛起,算法被嵌入国家安全与发展领域。算法在传统国家安全与非传统国家安全领域的应用差距加剧了算法权力在国与国之间赋能的“非对称性”,这为算法强国在国际社会攫取权力和利益提供了技术基础[18]。由此,算法审查的必要性进一步突显。

(2)算法审查是对算法规范合法性与合理性的审查

“代码即法律”[19],算法即为规制社会行为的法律规范。代码已经在一定范围内成为规范人与人之间法律关系、保障正常的社会秩序的工具,对法律的适用范围和适用边界产生了重大的冲击和影响[20]。“算法即规则”[21],强调算法作为数字世界的普遍运行的规则,为人类在算法内和算法外的社会活动提供了规则框架,对人和世界关系的顺畅进行施加十分重要的影响[22]。算法的透明性、公平性和可解释性等日益成为学者和立法者必须考虑的问题。算法审查是对算法“规则”来源的合法性和算法规范合理性的审查。

(3)我国缺乏算法安全评估的技术内容

在算法安全管理操作层面,我国相关法律已提出算法安全审查的初步要求,但却未就安全评估内容做具体规定。例如,《规定》强调,算法推荐服务提供者具有合规义务,并针对“具有舆论属性或者社会动员能力的算法”规定了安全评估要求①。但是,《规定》并未就“具有舆论属性或者社会动员能力的算法”的界定标准作出具体规定,同时对于该类算法安全评估的内容也未作明确规定。作为细化《规定》的操作层面的法律文件,算法安全评估的技术内容在进行算法安全审查时具有准据法性质和实施细则功能,在执法和司法层面都具有十分重要的意义。因此,在人类社会快速跨入通用人工智能时代的当下,我国亟需填补算法安全评估技术标准之空缺,通过立法明确算法安全评估技术标准的内容。

3. 算法审查法律规范的体系化水平严重滞后

(1)西方社会基本形成算法安全审查法律规范体系

随着人工智能技术的跨越式发展,保障算法伦理与安全业已成为国际社会共同关注的焦点[23]。欧洲始终注重数字安全审查,例如,《通用数据保护条例》(General Data Protection Regulation,GDPR)明确规定建立数据影响估计机制,要求算法相关主体在使用数据前必须对数据主体权利与自由进行风险评估②;《欧盟人工智能法案》(Artificial Intelligence Act)对 GDPR 的相关规定进一步细化,明确了数据评估的具体规则,要求建立数据合规评估制度③;《机器学习算法审查白皮书》(Auditing machine learning algorithms: A white paper for public auditors)为 GDPR 下的算法审查实践提供具体指引[24]。美国在突出算法审查的同时,要求数据本土化。例如,《算法问责法案》(Algorithmic Accountability Act of 2019)和《算法公平法案》(Algorithmic Fairness Act)要求对算法决策本身进行审查,从算法设计、开发和使用的全过程对算法的伦理、安全进行评估,明确算法必须具有透明性,以达到可期中审查的要求。此外,美国

①参见《互联网信息服务算法推荐管理规定》第 27 条的规定,即具有舆论属性或者社会动员能力的算法推荐服务提供者应当按照国家有关规定开展安全评估。

② 参见《通用数据保护条例》第 35 条的相关规定。

③参见《欧盟人工智能法案》第 43、47、48 条的规定。

在数据审查上还采取“双标”态度:一方面,通过一系列贸易协定推动全球数据跨区域流动;另一方面,通过相关法案构建数据本土机制,以维护本国总体国家安全。除此之外,加拿大的《自动化决策指令》(Directive on Automated Decision-Making)也对算法审查提出了具体要求;联合国教科文组织的《人工智能伦理问题建议书》也要求会员建立算法审查制度,以确保人工智能伦理符合人类价值观的要求[25]。综上,以欧美为主的西方发达国家和地区早已察觉到算法审查的必要性,并为之进行了积极的法治实践。

(2)我国算法审查政策敏锐而立法滞后

算法审查已成为我国国家安全的应有之义。2020 年底,中共中央印发的《法治社会建设实施纲要(2020-2025 年)》(以下简称“纲要”)》要求将法律法规延伸至算法领域①;2022 年初,国务院印发的《“十四五”数字经济发展规划》指出“着力强化数字经济安全体系”的要求,明确从网络安全、数字安全和其他风险等多角度建立和完善算法安全体系②; 2022 年 12 月,中共中央、国务院印发的《关于构建数据基础制度更好发挥数据要素作用的意见》也从数据治理视角触及算法安全问题。由此观之,党中央和国务院早已敏锐地意识到算法治理作为整个数字经济系统治理的关键环节,其不仅仅涉及社会治理成效,更关乎国家整体经济安全。令人遗憾的是,在我国算法治理领域,算法国家安全审查要求并未得到法律、法规层级上的专门立法回应,也未获得相关治理部门的应有重视,致使当下我国仍处于算法审查政策敏锐而立法滞后的尴尬境地,涉及算法审查的间接立法散见于《中华人民共和国国家安全法》《中华人民共和国网络安全法》和《中华人民共和国数据安全法》等立法中,缺乏法律或者行政法规层面的专门算法审查立法。

三、构建类 ChatGPT 通用人工智能算法审查制度的建议

我国相关部门应从总体国家安全的全局性视野系统地审视算法安全问题,提高算法安全能力[26],确保算法安全服务于总体国家安全。同时,均衡安全价值与发展价值,不能束缚于技术安全,应当将算法置于发展中以实现再平衡。以此为算法审查制度设计的顶层逻辑,从明确专门审查主体、厘清审查范畴和制定“算法安全法”等方面突破类 ChatGPT 通用人工智能算法安审查困局,确保算法审查规范体系完整有效、国家重点领域核心算法安全可控、国家核心利益和安全不受外部算法技术危害,确保国家处于持续安全状态。

1. 明确专门主体,统筹算法审查宏观设计

(1)领导主体:中央科技委员会

中央科技委员会履行通用人工智能算法审查的领导职责。今年 3 月,中共中央、国务院印发的《深化党和国家机构改革方案》提出组建中央科技委员会的制度安排,统筹解决科技领域内的战略性、方向性和全局性重大问题[27]。该政策为类 ChatGPT 通用人工智能算法审查的领导工作指明了方向。算法诱发的非传统国家安全风险日益呈现出隐蔽性和突发性特征,其影响范围之广、规制任务之复杂应当引起国家安全机关的警觉。由分散的执法机构负责算法审查,其审查成本高且缺乏系统性,易滋生多头审查等乱象,因此,有必要建立由中央科技委员会统一领导的算法审查机关,统筹布局全国算法安全审查工作。

(2)审查执行:国务院算法审查委员会及国家数据局

增设国务院算法审查委员会,负责组织、协调、指导全国类 ChatGPT 通用人工智能算法安全审查。算法审查委员会履行研究拟订具体审查政策;组织调查、评估算法安全状况,发布审查报告;制定、发布算法审查指南;协调算法审查行政执法工作等职能。算法审查委员会的组成和工作规则由中央科技委员会牵头制定,可将原国家互联网信息办公室负责的关于数据、

①参见中共中央印发的《法治社会建设实施纲要(2020—2025 年)》第二十二条的规定,即通过立改废释并举方式,推动现有法律法规延伸适用到网络空间……制定并完善对网络直播、自媒体、知识社区问答等新媒体业态和算法推荐、深度伪造等新技术应用的规范管理办法。

②参见国务院印发的《“十四五”数字经济发展规划》第九条的规定。

算法安全治理方面的职能向国家发展和改革委员会下属机构——国家数据局转移;国家数据局与国家安全部算法审查办公室组成算法安全联合审查执法机构,承担通用人工智能算法审查执行工作;国务院算法审查委员会可向国家各部委派驻算法审查小组,负责纵向指导全国算法审查工作的开展;建立国家、省、市三级算法联合安全审查执法机构,确保国家算法安全审查执法工作取得实效。

(3)协同共治:平台经营者和使用者

充分发挥社会主体在算法审查中的协同共治作用。ChatGPT 产品标准化意味着其会以接口的形式被嵌入到各大主流应用平台当中,微软搜索引擎必应(Microsoft Bing)和办公系统(Microsoft office 365)全线接入 ChatGPT 也印证了这一趋势。尽管审查主体可依据被审查对象组合适用各类审查方法,但当下应用较多的代码审查和数据抓取审查等方法依旧是基于平台数据为前提展开的,对平台产生较强的依赖性[28]。现有应用平台在场景化的应用层面覆盖较为全面,实践中也以贴合客户的实际应用需求为主要任务,它们在数据采集层面已经做得很充分。因此,平台经营者应当作为监测算法安全风险的第一责任人。同时,伴随着人工智能技术的不断提高和应用范围的持续拓宽,平台使用者的算法素养也随之获得提升。政府在国家安全理念方面的宣传教育将使广大社会大众充分认识到国家安全与算法安全的密切关系,他们有意愿,也有足够的素养储备为算法审查提供第一手线索来源。因此,算法治理者应当充分发挥平台经营者和平台使用者在算法安全审查中的共治作用,夯实算法审查的社会基础。

2. 厘清算法审查范围,明确算法安全内容

基于类 ChatGPT 通用人工智能冲击国家非传统安全的风险分析,通用人工智能算法审查应以算法文化、算法透明和算法可控为审查内容,突出算法文化监管审查、算法可解释性审查和算法可控性边界审查,虽然全国信息安全标准化技术委员会组织制定的《生成式人工智能服务安全基本要求》(征求意见稿)对部分安全内容进行了规范,但仍不全面,需要进一下明确。

(1)以算法文化监管审查确保意识形态安全与情报安全

一是审查算法下掩盖的西方文化策反因素。以美国为首的西方国家凭借科技领域的主导地位,利用我国全方位、宽领域对外开放和深化体制机制改革“契机”,肆意污化中国共产党领导和中国特色社会主义道路、理论和制度体系,费尽心机对我国进行思想文化渗透,推销西方民主政治制度、思想价值理念和社会意识形态[29]。这种蓄谋已久的文化策反政策给我国社会主义意识形态带来巨大挑战,因而在算法文化审查上,应当以我国社会主义文化为主基调,重点审查类 ChatGPT 通用人工智能中暗藏的文化异常因素,特别要将算法中的“美化西方、丑化东方”的文化策反成分进行重点标记和专项清除,以确保通用人工智能算法文化符合社会主义核心价值观,使之成为我国社会主义文化建设事业顺利发展的促进因素。同时,进一步明确算法文化监管审查的社会主义方向和为人民服务内容。我国文化事业发展遵循的是为人民服务、为社会主义服务的方向①,这清晰传达了我国文化事业所遵守的政治立场,也从根本上明确了类 ChatGPT 通用人工智能算法文化安全审查的基本法律价值。

二是审查算法内部夹带的西方价值评价。算法嵌入社会生活,体现出西方价值传播倾向。类 ChatGPT 通用人工智能算法形式上是进行在线文本聊天,其实质可能是进行价值观的“靶向”灌输。ChatGPT 使用在信息检索场景下,其已在不经意间成为价值观传播的重要媒介,且具有强意识形态倾向。ChatGPT 输出文本数据所承载的文化信息成为传播西方意识形态的掩体,使用者在接受该信息时表现出不自觉性与无可抵抗性;该过程的实质是将用户价值观置于算法的评价体系之下,也是算法在文化领域独裁的缩影。算法隐含的各种规则,带来的不仅是人们自愿接受算法评价,甚至是主动接受算法规训和自我审查[30]。因此,进行类 ChatGPT 通用人工智能算法安全审查时,必须对算法夹带的西方价值评价体系进行审查,确

①参见《中华人民共和国宪法》第 22 条的规定。

保其符合我国社会主义价值评价要求。

(2)以算法可解释性审查确保社会公共安全与数据安全

算法可解释性应在算法设计和应用中得到回应。自 2018 年 9 月美国国防高级研究计划局提出人工智能探索项目以来,西方技术强国纷纷在该领域进行了法治实践①。以欧盟为例,2019 年发布的《人工智能道德准则》(Ethics Guidelines for Trustworthy AI)提出了“值得信赖”的人工智能应当满足数据隐私保护、透明度和公平性,其中透明度即是强调提高数据投喂和算法结果的可解释性,企图建立与维护使用者对算法开发者和算法本身的信任;《关于人工智能的统一规则(人工智能法)并修正某些联合立法行为》提案也对算法可解释性加以强调;《通用数据保护条例》(GDPR)赋予使用者算法解释的权利。随着人工智能应用场景的多样化和复杂化,我国法律也对数据的合规性和算法的可解释性提出了一些要求,即应当在数据收集和使用中提升数据的完整性和规范性等,在算法设计、实现和应用等诸多环节内持续提能,使其满足社会对透明、可解释和可理解的要求[31]。但实践中,新的人工智能算法成果被推出后,依然会重复算法出现自动化决策的不可解释性问题。国家安全语境下的类 ChatGPT 不可解释性突出表现在,因基础文本语料库的开放性和奖励模型评价体系的非公开性带来的算法输出结果的可信任度与算法可控性之间的矛盾。换言之,对类 ChatGPT 通用人工智能算法的透明审查应当从语料库和奖励模型评价体系等方面入手,即从数据喂养和算法逻辑层面展开。

审查基础文本语料库的合法性与伦理合规性。ChatGPT 是基于基础文本语料喂养下的人工智能,其政治偏见源于基础文本语料的价值偏见。目前,训练 ChatGPT 的文本数据并非囊括全部的互联网数据,而是西方技术资本寡头根据价值需求在国外几大主流数据库中有意选择的数据池,因此,ChatGPT 所表达的是西方资本家所信奉的普世价值观,与我国提倡的集体主义价值观背道而驰。鉴于此,有必要建立数据本土化保护机制和跨境数据审查机制,严防影响数据伦理安全的不稳定性因素流入我国境内,确保本土数据的安全。建立本土数据使用审批制度,确保用户在与 ChatGPT 交互过程中正确投喂数据。在此基础之上,力图以我国社会主义集体价值观引导算法向善,规范用户投喂行为,使供养 ChatGPT 的数据符合我国社会主义文化事业发展要求。

审查奖励模型评价体系的公平性与正义感。ChatGPT 中的奖励模型是算法设计者的政治价值表达,具有意识形态性。被嵌入其他应用后,奖励模型的价值偏见会影响算法对文本数据的评价,调用 API 接口的应用程序在对用户行为展开评价时表现出歧视性倾向,最终引起激励结果的不公平问题。换言之,奖励模型在对文本数据进行非公正性的排序后,直接导致调用程序进行价值判断时失误。类 ChatGPT 是代码化的价值表达,具有算法技术的黑箱特性。类 ChatGPT 审查应当在有限开源的前提下展开,即在国家主体展开审查时其底层代码应该是开源的,以此做到降低类 ChatGPT 的算法评价逻辑的黑箱属性。用透明性价值破解算法黑箱后,类 ChatGPT 技术价值和功能价值能得以重塑。因此,经社会主义价值改造后的奖励模型将会更加注重社会结果公平和程序正义,其使政府在社会公共治理领域内把算法治理带回符合公共目的本质要求,在商事领域更加注重私有性、排他性和秘密性价值[32]。

(3)以算法可控性审查确保国家经济发展安全

算法的可控性不仅要求算法在运行过程中具备高鲁棒性(Robust),算法自主发育及社会应用冲击也应当符合人类价值预判。总体国家安全视阈下的算法可控性审查,即是审查因算法持续进化和全面应用所导致的“机器替代人”的可控程度及人类的应对能力。因此,有必要对类 ChatGPT 通用人工智能算法国内应用的可控性进行审查,以综合评估其对国家产业结构和国民就业带来的冲击。

①美国国防高级研究计划局成立于“冷战”时期,是美苏争霸的产物。它隶属于美国国防部,是以研发军事用途的高科技武器为主要职能的行政机构。

审查类 ChatGPT 算法对人工智能产业发展的驱动力。人工智能作为战略性新兴技术,将成为科技创新、产业升级和生产力提升的驱动力量[33]。通用人工智能算法的长足发展,离开不了算力的支撑。然而,算力的决定因素在于硬件的支撑,硬件的核心是芯片。因此,人工智能产业的发展不仅仅需要将重心置于算法之上,还需要重视以硬件为核心的产业链的持续完善。类 ChatGPT 是人工智能领域的最新成果,其发展缺乏不了硬件的支撑。因此,应当对我国相关产业发展驱动力的有无和大小做出科学的预判。在此基础之上,出台相应产业发展政策,推动人工智能与经济社会深度融合发展;建立安全可控的算法治理体系和开放协同的人工智能创新体系,促进产学研深度整合,提升我国人工智能产业链在国际上的竞争力。

审查类 ChatGPT 算法对社会就业的冲击。人工智能技术发展是无可逆转的历史方向,但是社会仍然需要时间来适应科技变迁,必须保证类 ChatGPT 在社会实践替代中能够实现“软着陆”,降低对社会就业的冲击。以 ChatGPT 为代表的人工智能技术不断自我发育,以至在诸多领域产生替代性实践,其在带来生产效率极大提升的同时,还会造成劳动力技术性失业。鉴于此,应当在整个社会领域内建立算法合理应用领域清单制度,拟定 ChatGPT 适用范围白名单和不可使用黑名单;根据技术发展和行业适用能力的变化对应用领域清单开展动态调整,保证适用清单的可行性、合理性和有效性[34]。在单个产业内建立算法分级替代制度,对单个替代性产业进行分层级、分批次替代,以实现产业在算法应用过程中的平稳过渡。

3. 制定“算法安全法”以破解政策敏锐立法滞后的困局

为提高通用人工智能算法审查法律规范的前瞻性与规制力,亟须精准研判算法风险,加快制定专门的算法安全法及相关法规,回应算法安全制度的现实需求和总体国家安全需要。立法者可以从以下维度进行构建:

(1)问题导向:弥补现有法律规范的空缺

类 ChatGPT 通用人工智能算法的快速进化及广泛应用将对非传统国家安全形成威胁,专门的算法安全立法应当以总体国家安全为顶层逻辑,充分考虑当下类 ChatGPT 通用人工智能算法带来的上述三大威胁,明确通用人工智能算法社会主义价值导向,针对其在文化传播、价值判断和算法评价等方面可能诱发的种种问题,将具有“舆论属性或者社会动员能力的算法”的法律条款表述加以细化,明确标识其类化特征,明确该类算法安全评估的具体内容,制定算法安全评估细则。聚焦类 ChatGPT 操纵意识形态传播审查,促使算法应用公开透明,促进通用人工智能算法积极传播符合价值评判的正能量,引导通用人工智能算法应用向上向善。

(2)惩前毖后:事前审查、事中规范和事后惩治协同优化

创立通用人工智能算法事前安全审查机制,形成事前审查、事中规范和事后惩治的闭环。类 ChatGPT 通用人工智能算法安全审查是技术审查、过程审查,也是结果预判。算法的技术迷思决定了过程治理是低效率的和高成本的,算法的动态性也决定了试图通过过程治理来预防算法侵害无异于与风车作战[35]。由是观之,算法事前安全审查在理论与实践上更具有可选择性:需要明确算法安全的评估细则,督促“具有舆论属性或者社会动员能力的算法”依据细则开展算法安全评估;需要建立算法安全检查评估支撑体系和技术队伍建设体系,以此支撑通用人工智能算法国家安全检查执法,在上线应用前及早发现安全问题、及时整改。事前算法审查与事中安全评估、分级分类安全管理相衔接,确保专门的“算法安全法”在社会协同治理中更好地发挥效能。

(3)统筹兼顾:有效衔接《中华人民共和国国家安全法》《中华人民共和国网络安全法》与《中华人民共和国数据安全法》

专门的“算法安全法”应立足《中华人民共和国国家安全法》规定的总体国家安全,继承《中华人民共和国网络安全法》与《中华人民共和国数据安全法》在数据安全方面的立法成果,突破数据安全审查的可能路径,探索直击算法审查的具体路径。专门的“算法安全法”聚焦于算法安全领域审查与治理问题,不仅应当是上述三部法律的同位法,更应是其在算法安

全治理方面的有效补充。“算法安全法”应当鼓励算法安全技术的发展,推动高等院校培养算法安全人才队伍建设。此外,“算法安全法”应当在上述三部法律的基础上展开算法安全科研布局,构建通用人工智能算法审查关键技术联合攻关体系,突破算法安全内生机理、算法安全风险评估、算法全生命周期安全监测等关键技术瓶颈[36]。

四、结　语

类 ChatGPT 带来的非传统国家安全威胁使通用人工智能算法安全审查日显必要和紧迫,我国法律规范体系的日益健全为此提供了现实的制度基础。类 ChatGPT 的全面应用诱发意识形态风险、国家情报风险、社会公共安全风险和国家经济发展安全风险,不仅涉及算法伦理问题,在更深层次映射着与总体国家安全的冲突。因此,必须在符合总体国家安全观的大框架内建立通用人工智能算法审查制度,对类 ChatGPT 通用人工智能算法展开全方位的安全审查,以此为示范,进一步推动算法审查制度的完善。通用人工智能算法审查制度的建立与健全不仅能提升国家对科技发展领域的监管能力,也是我国在经济全球化大进程中应对西方国家算法霸权的必经之路。类 ChatGPT 国产化之路道长且艰,但举足不前就意味着出局。ChatGPT 的成功不仅是算法发展的成就,也是整个人工智能产业驱动下的成功。仅仅依靠我国国内部分企业正面赶超的做法并非明智之举,应当从国家层面整合产业布局,在类 ChatGPT 适应中国价值取舍的前提下加以引进,在符合我国总体国家安全的要求下加以应用,打通整个人工智能产业链,从法治、金融、硬件和人才储备等方面加以保障,从而真正实现类 ChatGPT 国产化的突破并为国家安全保驾护航。

参考文献:

[1] 邹开亮,刘祖兵. ChatGPT 的伦理风险与中国因应制度安排[J]. 海南大学学报(人文社会科学版),2023,41(4):74-84.

[2] 孟天广. 智能治理:通用人工智能时代的治理命题[J]. 学海,2023(2):41-47.

[3] 陈自富. 强人工智能和超级智能:技术合理性及其批判[J]. 科学与管理,2016,36(5):25-33.

[4] 王彦雨. “强人工智能”争论过程中的“态度转换”现象研究[J]. 科学技术哲学研究,2020,37(6):26-33.

[5] A Guest Blogger. Understanding artificial general intelligence——an interview with Hiroshi Yamakawa [EB/OL]. [2017-10-23]. https://futureoflife.org/recent-news/understanding-agi-an-interview-with-hiroshi-yamakawa/.

[6] 张欣. 生成式人工智能的算法治理挑战与治理型监管[J]. 现代法学,2023,45(3):108-123.

[7] 刘跃进. 安全领域“传统”“非传统”相关概念与理论辨析[J]. 学术论坛,2021,44(1):27-48.

[8] 方卿,丁靖佳. 人工智能生成内容(AIGC)的三个出版学议题[J]. 出版科学,2023,31(2):5-10.

[9] 阙天舒,张纪腾. 人工智能时代背景下的国家安全治理:应用范式、风险识别与路径选择[J]. 国际安全研究,2020,38(1):4-38.

[10] 党亚娟. ChatGPT 潜在军事应用及风险分析[J]. 国防科技工业,2023(3):54-56.

[11] 张洪忠,段泽宁,韩秀. 异类还是共生:社交媒体中的社交机器人研究路径探讨[J]. 新闻界,2019(2):10-17.

[12] 邓建鹏,朱怿成. ChatGPT 模型的法律风险及应对之策[J]. 新疆师范大学学报(哲学社会科学版),2023,44(5):91-101.

[13] 澎湃新闻. 从 ChatGPT 数据泄露事件,看组织安全稳定自动化的重要性[EB/OL]. [2023-04-11]. https://www.thepaper.cn/newsDetail_forward_22632495? commTag=true.

[14] 王君,张于喆,张义博,等. 人工智能等新技术进步影响就业的机理与对策[J]. 宏观经济研究,2017(10):169-181.

[15] 付冉冉. 大数据时代算法审计构想[J]. 网络安全与数据治理,2023,42(2):48-52.

[16] 中国网信网. 国家互联网信息办公室关于发布互联网信息服务算法备案信息的公告[EB/OL]. [2022-08-12]. http://www.cac.gov.cn/2022-08/12/c_1661927474338504.htm.

[17] 张之沧. 论福柯的“规训与惩罚”[J]. 江苏社会科学,2004(4):25-30.

[18] 张涛,余丽. 算法在国际政治中的“非中性”作用[J]. 国际论坛,2022,24(5):41-57.

[19] KOZAK K. Algorithmic governance, code as law, and the blockchain common: power relations in the

blockchain-based society [J]. Frontiers in Blockchain,2023(6):1-8.
[20] 徐冬根. 二元共治视角下代码之治的正当性与合法性分析[J]. 东方法学,2023(1):36-48.
[21] 董青岭,朱玥. 人工智能时代的算法正义与秩序构建[J]. 探索与争鸣,2021(3):82-86.
[22] 王聪. "共同善"维度下的算法规制[J]. 法学,2019(12):66-77.
[23] 邹开亮,刘祖兵. 试论智能算法主体化[J]. 重庆邮电大学学报(社会科学版),2023,35(2):63-75.
[24] Supreme Audit Institutions of Finland, Germany, the Netherlands, Norway and the UK. Auditing machine learning algorithms: a white paper for public auditors [EB/OL]. [2020-10-24]. https://auditingalgorithms.net/index.html.
[25] BLANCHARD A, TADDEO M. The ethics of artificial intelligence for intelligence analysis: a review of the key challenges with recommendations [J]. Robotics & Machine Leaming Daily News,2023(26):2-3.
[26] 贾珍珍,刘杨钺. 总体国家安全观视域下的算法安全与治理[J]. 理论与改革,2021(2):135-148.
[27] 新华社. 中共中央 国务院印发《党和国家机构改革方案》[EB/OL]. [2023-03-16]. https://www.gov.cn/gongbao/content/2023/content_5748649.htm.
[28] 张欣,宋雨鑫. 算法审计的制度逻辑和本土化构建[J]. 郑州大学学报(哲学社会科学版),2022,55(6):33-42.
[29] 刘志明. 维护国家文化安全亟需健全文化安全审查制度[J]. 湖南社会科学,2018(2):165-171.
[30] 彭兰. 如何实现"与算法共存"——算法社会中的算法素养及其两大面向[J]. 探索与争鸣,2021(3):13-15.
[31] 科技部.《新一代人工智能伦理规范》发布[EB/OL]. [2021-09-26]. https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.
[32] 张红春,章知连. 从算法黑箱到算法透明:政府算法治理的转轨逻辑与路径[J]. 贵州大学学报(社会科学版),2022,40(4):65-74.
[33] 新京报. 科技部谈 ChatGPT:将推动人工智能与经济社会深度融合[N/OL]. [2023-02-24]. https://baijiahao.baidu.com/s?id=1758687523652638273&wfr=spider&for=pc.
[34] 翟月荧. 算法行政的兴起、风险及其防控[J]. 新视野,2022(3):81-85.
[35] 邱泽奇. 算法治理的技术迷思与行动选择[J]. 人民论坛·学术前沿,2022(10):29-43.
[36] 中国网信网. 专家解读|加强深度合成算法安全科研攻关 推进深度合成服务综合治理[EB/OL]. [2023-01-11]. http://www.cac.gov.cn/2023-01/11/c_1675070655576858.htm.

(收稿日期:2023-03-25 编辑:高虹)

Research on ChatGPT-like Artificial General Intelligence Governance: From the Perspective of Algorithmic Security Review/ZOU Kailiang, LIU Zubing(College of Humanities and Social Sciences, East China Jiaotong University, Nanchang 330013, China)

Abstract: The ChatGPT-like has ideological characteristics, and its introduction and embedded applications are prone to non-traditional national security threats. The ChatGPT-like artificial general intelligence algorithm is fed by capital and western values, which can easily trigger ideological security risks and attack both the country's ideological security and intelligence security, and provides many convenient conditions for the acquisition and transmission of network intelligence. Fabricating false public opinion and causing widespread social rumors may challenge social public safety, collecting, storing and processing user data poses a serious threat to national data security, and obstructing the pace of a country's rise may threaten the developing country's economic development and security, making the transnational transfer of industries impractical and redefining the upper limit of social human resource costs. The algorithm review system in China faces practical problems including the lack of specialized review subjects, unclear review scope, and the inability to implement the review system. Guided by the overall national security concept, we should clarify the specialized review subject of algorithm security led by the Central Science and Technology Commission and implemented by the Data Security Bureau of the

State Council to unify the macro design of algorithm review. We should give full play to the synergy between platform operators and users to solidify the social foundation of algorithm review. We should incorporate algorithm culture monitoring, interpretability, and controllability into the scope of algorithm review, reasonably assessing the impact of algorithms on country and society and clarify the content of algorithm security. And we should develop an algorithm security law to ensure the implementation of the algorithm security review system, to enhance the regulatory capacity of the country in the field of technology, actively respond to the algorithm hegemony of western countries, and accelerate the localization application and localization replacement process of ChatGPT-like artificial general intelligence.

Key words: ChatGPT-like; artificial general intelligence; non-traditional national security; algorithm security review; algorithm hegemony

科技期刊支撑教育、科技、人才协同创新发展

——我校牵头第十八届中国科技期刊发展论坛“高端对话”环节

11 月 29 日至 30 日,第十八届中国科技期刊发展论坛在南京举办。中宣部副部长张建春,中国科协党组副书记、专职副主席、书记处书记束为,江苏省委常委、省委宣传部部长张爱军等出席并致辞。我校副校长郑金海受邀参加开幕式并主持了由我校牵头的“高端对话:科技期刊支撑教育、科技、人才协同创新发展”环节。

在“高端对话”环节,第十二届全国人大常委、教育科学文化卫生委员会主任委员、原国家新闻出版总署署长柳斌杰,南京林业大学党委常委、副校长徐信武,中国科学技术期刊编辑学会副理事长兼秘书长、《中华医学杂志》社有限公司总经理兼总编辑魏均民围绕“科技期刊支撑教育、科技、人才协同创新发展”这一主题,共同深度探讨科技期刊如何在更具包容性、开放性的科技创新领域中支撑科技、人才与教育的协同发展问题,并与观众进行了互动。

中国科技期刊发展论坛以推动世界一流科技期刊建设为目标,是我国科技期刊界最具知名度、最有影响力的会议之一。今年的论坛在南京举办,以“开放 信任 合作——科技期刊助力高水平科技自立自强”为主题,设置了开幕式、主旨报告、高端对话、前沿探讨等环节,并通过央视网进行全球直播。

(河海大学期刊部供稿)