DOI:10.3876/j.issn.1671-4970.2024.05.010

论 Sora 的伦理风险与我国治理因应

——也谈我国参与人工智能算法伦理全球治理的基本路径

刘祖兵

(同济大学法学院,上海 200092)

摘要:以 Sora 为代表的文生视频模型触发多重伦理风险,亟须探索复归理路。Sora 替代应用诱发弱化人类主体性风险,“绝对”失业潮恶化贫富差距,可能引发“新卢德运动”,影响人类社会稳定;Sora 接受偏见视频数据“浸润”,刻画对特定群体或观点的刻板印象,其生成内容欠缺价值多样性与包容性,诱发高度隐蔽性的社会歧视风险,影响人类社会公平;Sora 生成的虚假视频具有高度违法性风险,致使诈骗视频信息大行其道,诈骗犯罪高发,影响社会和谐。我国生成式人工智能伦理监管规范体系阙如,存在具体应用场景中的算法监管细则语焉不详、使用行为监管规则欠缺和算法伦理安全规则缺位等问题;视频数据使用者行为监管失衡,既有过度重视数据保护而阻碍数据要素丰富的宏观治理问题,也存在数据生产者行为规范缺位、部分法条可适用性不强的微观治理问题。应建立基于场景化的分类分级审查制度,纵向细化平台数据使用行为规则,进行算法听证,确保多元主体参与算法设计;从制度和技术维度强化视频数据使用者的行为监管,推动训练 Sora 的数据实现透明化和价值多样性。此外,我国应把握全球视野,参与生成式人工智能算法底层模型技术标准的制定与全球共享,致力于推动科技创新人才培养和域外专业人才引进;积极推动我国参与人工智能算法伦理治理的南北对话和南南合作,为人工智能全球伦理治理贡献中国智慧。

关键词:Sora;伦理风险;分类分级审查;数据透明;算法民主;南北对话

中图分类号:D92.11 **文献标志码**:A **文章编号**:1671-4970(2024)05-0099-14

人类在享受技术带来的狂欢时,技术伦理失范风险也悄然而至。2024 年,OpenAI 正式对外发布首款文生视频模型——Sora,基于 Transformer 架构、可根据用户文本指令生成高逼真的高清视频,其一经问世便迅速引起社会各界热议。Sora 极大释放了生成式人工智能(generative artificial intelligence,GenAI)的影视制作活力,打开了塑造影视产业新业态的大门,或将颠覆现有影视产业发展格局,给视频、电影行业带来巨大冲击[1]。Sora 面世的次日,以图像处理、视频制作软件为主营业务的奥多比公司的股价应声下跌超过 7 个百分点[2],至此,大量影视从业者陷入失业的忧虑之中。欧盟已充分认识到 GenAI 对人类社会带来的强势冲击,在立法上予以了有力回应。经过长时间的酝酿和广泛地征求意见,欧盟议会于 2024 年高票通过首部《人工智能法案》(AI Act),希望以此维持在人工智能治理领域的“布鲁塞尔效应”。

引用本文:刘祖兵. 论 Sora 的伦理风险与我国治理因应——也谈我国参与人工智能算法伦理全球治理的基本路径[J]. 河海大学学报(哲学社会科学版), 2024,26(5):99-112.

基金项目:教育部人文社会科学规划基金项目(21YJA820032);上海市 2022 年度软科学研究项目(22692104400)

作者简介:刘祖兵(1988—),男,博士研究生,主要从事数据法学和知识产权法学研究。E-mail:ncliuzb@126.com

该法案突出强调对高风险人工智能的伦理规制,为 GenAI 的功能性应用划定了明确的禁区,不仅包括对敏感信息的采集与识别,还特别将人们关注的用于合成照片以及音频、视频内容的深度伪装技术纳入规制范畴。此外,该法案还要求应用机构对 GenAI 的安全性进行评估,将技术滥用的负面影响个案上报至各国监管部门[3],这预示着欧盟将 GenAI 伦理风险规制提升到了国家治理新高度。

我国政府也正在为 GenAI 伦理治理做出持续努力,积极参与人工智能全球治理。在 2024 世界人工智能大会暨人工智能全球治理高级别会议上,我国发表了《人工智能全球治理上海宣言》(以下简称《宣言》),《宣言》倡导从促进技术发展、技术安全、人类福祉和打击犯罪等多个角度加强人工智能技术及应用监管,打造可审核、可监督、可追溯和可信赖的人工智能技术;倡导构建多元化的参与机制,在人类决策与监管下,提高人工智能治理的技术能力,以人工智能技术防范人工智能风险[4]。在此之前,我国立法者也预判 GenAI 应用可能诱发的社会风险,为此发布了首部治理 GenAI 的法律文件——《生成式人工智能服务管理暂行办法》(以下简称《暂行办法》),以期防范 GenAI 应用风险,促进人工智能技术健康发展。理论界也普遍认为 Sora 具有伦理问题,有学者甚至认为 Sora 具有潜在的助长网络犯罪和社会信任的风险[5],还有可能会诱发意识形态去理性化风险[6]。在此背景下,本研究在剖析 Sora 诱发的典型伦理风险的基础上,探析 GenAI 伦理治理的我国因应理路,希冀通过 GenAI 伦理敏捷治理实现人工智能算法至臻至善。

一、风险表征:生成式人工智能算法伦理失范的典型表现

在“工具理性”的驱使下,Sora 代表的文生视频模型以追求结果最优为导向,崇尚效率优位。受此影响,Sora 或将诱发多重伦理风险:一方面,Sora 的“技术替代性”将弱化人类主体性,诱发“绝对”失业潮,加剧人类贫富差距,影响社会稳定;另一方面,因受数据偏见的“浸润”,Sora 生成的视频饱含高度隐蔽性的社会偏见,影响社会公平。此外,行为人滥用 Sora 生成大量虚假的违法视频,致使网络诈骗犯罪行为高发,挑战人类法律底线。

1. Sora 替代性应用诱发弱化人类主体性风险

GenAI 的普及使技术持续加深职业替代危机,产生机器剥夺人类劳动权利的社会认知乱象,诱发弱化人类主体性的社会风险[7]。主体性实现的是以现实的改造活动为依托,在实践过程中表现出来的作为发动者、筹划者和决策者的功能[8]。“技术理性”无可避免地产生“技术替代性”,“技术性替代”节约了社会必要劳动力,也为资本摄取更多剩余价值提供了新的用途。然而,GenAI 时代的“技术替代性”诱发的失业问题更加突出[9],预计到 2055 年,在全球现有的八百二十种有薪职业中,愈七百一十种或将被机器取代,中国和印度可能成为受此影响最大的国家[10]。但是,与以往“技术替代性”失业不同的是,GenAI 带来的失业潮具有明显的不可逆性,它并不能同时为人类衍生出更多替代性的劳动岗位,因此,GenAI 诱发的失业潮是“绝对”的,这种失业潮并非存在于人类发展史中的暂时“阵痛”,该影响会比人类以往任何一次技术革命都要深刻,甚至可能导致“新卢德运动”。

GenAI 的职业替代并不能改变人类劳动的性质,其带来的“绝对”失业不能帮助人类实现“从必然王国到自由王国”跨越的愿景。申言之,GenAI 并不能在短期内将人类劳动从必要的“物质性劳动”转换为实现个人价值的“自觉劳动”,人类的社会劳动将长期保持着维持“生存必须”的必要劳动性质,该性质或将在很长的一段时期内不会改变。如,自动驾驶汽车(full-self driving,FSD)强势入驻武汉,大量廉价、高效且全天候的“萝卜快跑”出租车或将逐渐替代人类司机,致使大量司机面临失业风险。Sora 的普及应用亦是如此,其将使大量影视从业者失去维持生计的劳动机会,就像自动驾驶出租车造成大批量的失业劳动力一样。Sora 的“技术性失业”违背了人类劳动力市场供给的一般规律,或将使劳动力市场的调解机制失灵。

一方面,Sora 将削减大量相关从业者的数

量,致使从业者数量绝对过剩。对影视从业者的就业冲击可能不亚于机器生产替代手工劳动,尽管此前他们的工作机会一直被认为是专业性强、替代性弱的,是难以被机器替代的。反观现实,Sora 的“技术性替代”将使影视从业者失去必要的谋生手段,更有甚者,将把人类社会带入动荡的漩涡当中[11]。在未来可预见的时间跨度内,Sora 无力为影视制作行业提供同等数量的工作岗位,更不能实现对现在劳动岗位的合理替代,而是犹如“圈地运动”那样将从业者驱逐于职场,给予人类劳动力市场以重创,可能使劳动力的绝对过剩成为社会“难以承受之重”。资本对技术的裹胁加剧“绝对”失业潮。人类与机器的矛盾和对抗并非源于机器本身,而是机器背后的力量,即操控机器的资本主义生产方式,资本俘获劳动岗位后,其绝不会轻易地将这些劳动机会返还给普通劳动者。在技术发展的特定时间跨度内,技术不具备产生更多替代性工作岗位的能力,即失业者会永久性地失去原先的就业机会,因此,这种由 GenAI 带来的失业风险是“绝对”的。GenAI 将部分抗风险能力差的劳动者推向贫困,其带来的“绝对”失业潮会加剧社会贫富差距,人类必须清醒地认识到机器替代后面的力量,即机器并非造成社会“绝对”失业的根源,而是凭借机器运行的资本主义生产逻辑[12]。

另一方面,Sora 的“技术替代性”应用剥夺人类主体的实践能力。受资本操控的 GenAI 在逐利性的驱使下将人类劳动者驱赶到两种极端:一种是一部分人成为算法精英阶层,他们利用 GenAI 迅速积累社会财富,变成富有的人;另一部分则相反,他们处于算法搭建的“超级全景监狱”当中,受制于 GenAI、听命于算法决策,致使自身变得越来越贫困。“工具理性”强化人类的无用感,算法的效率优势使“工具理性”获得合理性,其低成本优势又推动着人类接受技术合理性,“工具理性”肆无忌惮地朝着资本绘制的路线传播。普罗大众逐渐沦为被算法算计的数字公民,他们手中的财富于无形之中涌入算法精英阶层手中。通过分析用户属性、价值观和精神状态等特征,被嵌入社会决策平台的 GenAI 可以精准刻画用户画像,变成定点投放广告、定向推送特定商品的数字工具,而用户往往难以抵制被精准分析的诱惑,只能越来越接受算法提供的“千人千面”的“温柔关怀”,致使贫困程度加深。在该过程中,人类逐渐失去参与社会生产实践的技能,他们的无用感会持续加剧,久而久之,社会总体收支杠杆出现失衡,这也随即把人类带入公平性缺失的困境当中。

2. Sora 受偏见数据“浸润”诱发社会歧视风险

人类在享受 Sora 带来便利的同时,也应该正视由其诱发的社会歧视风险。用于训练 Sora 的视频数据集饱含社会偏见,经强化训练后,这些数据偏见被算法逻辑继承,Sora 生成内容会把这些偏见放大,使所有内容都或多或少地体现视频数据偏见。Sora 的价值偏见风险多表现为生成视频的歧视性与欠包容性。

第一,Sora 刻画特定群体或观点的刻板印象,诱发社会歧视风险。Sora 在训练时可能会将视频素材中含有偏见特性的物理矢量转化为描绘某些群体或观点的固定形象特征。如,Sora 对视频素材中人群肤色的使用或将成为刻画种族歧视的特征,这种由颜色要素带来的刻板印象不仅是对部分人群甚至是种族的区别对待,还可能加深对不特定国家和地区人种的误解与歧视,甚至挑起敌意,可能使某些群体遭受被边缘化、忽视甚至孤立的处境。Sora 还容易生成带有性别偏见的视频,表达对某一特定性别的歧视性倾向。这种源于技术和数据的偏见不仅违背了公平原则,也可能对个体发展和社会稳定造成巨大的负面影响,给人类社会和谐带来不安定因素。

第二,Sora 生成的内容欠缺包容性。大模型无法消弭的“黑箱”效应加剧了生成内容的不可解释性和不确定性,容易产生错误和误导性的推断。用于训练算法的视频数据是有限性的,Sora 无法充分兼顾公众文化和背景的多样性,因此,经由大模型生成的视频天生不具备包容性:一方面,Sora 不能满足多数用户对价值多样性的要求,且其生成内容过于单一,无法真实反映社会主体价值的多元性,阻碍人类多元文化的发展。因受到隐藏在算法逻辑和视频矢量

之间的偏见“浸润”,Sora 生成的视频可能具有性别歧视、种族歧视和区域歧视等多重风险表征,其生成的饱含偏见的视频可能会加剧社会不平等,威胁特定群体的生存权利,在社会层面大范围地施加危害。另一方面,Sora 对用户的数字素养有着特定的要求,对用户缺乏包容性。Sora 生成行为是在先作品数据与用户文本指令融合生成的过程,其对大模型使用者的技能和文本指令的选择能力具有特定的要求,因为数字素养的不同,用户使用 Sora 生成的结果可能会存在明显的差异,容易诱发技术鸿沟效应。这种存在于算法领域的隐性数字歧视会加剧社会层级的极化,该风险不仅仅存在于个体之间、群体之间,甚至还可能上升至不同国别和民族之间,滋生数字霸权,威胁一国数字主权独立。

综言之,Sora 偏见的表现形式多样,业已渗透至算法生成的各种内容,并且具有较强的隐蔽性。在大范围商用后,Sora 用户将散布于全球各地,受众面较广,同时这种因使用者数字素养的差异性产生的欠缺平衡性和不公平性将给社会治理带来负担。鉴于此,社会治理者亟须强化监管,以期实现 GenAI 决策公平,从而推动实现社会总体公平。

3. Sora 生成的虚假视频诱发高度违法性风险

Sora 模型能根据用户文本指令自主生成高清晰度、高逼真的视频,或将助长虚假视频泛滥,致使诈骗视频信息大行其道。恶意使用者利用 Sora 伪造新闻事件、人物行为欺骗公众,或将引发公众恐慌、扰乱正常的社会秩序。Sora 模型还可能被用于制作具有煽动性或误导性的视频,影响、引导社会舆情,破坏社会和谐。

Sora 生成虚假视频,使换脸视频成为难以遏制的恶。人脸信息是高度隐私的生物信息,不仅能被用于精准识别个体,亦具有明显的社会伦理特征。在人类社会生活的各领域,人脸识别技术已经获得广泛应用,无论是算法应用解锁还是其他算法安全认证,生物人脸识别算法都发挥着不可或缺的作用[13]。然而,经 Sora 合成的假视频或将助长恶意“刷脸”和“换脸”等违法行为,再一次将人脸治理难题带回了公众视野,成为人类社会难以医治的顽疾。基于海量视频素材的“浸润”,Sora 能理解视频中的物理矢量关系,并对变量——“块”(patches)展开多样化的复制、模拟和组合应用。“块”将二维帧①拓展至“三维空间”,并且记录着帧的物理矢量,即帧在时间和空间上的相对关系。换言之,Sora 应用“时空块”理解视频物理素材,在该过程中,Sora 能精确处理视频中物理要素之间的衔接,摆脱了存在于二维帧之间松散的关联关系,实现前后视频之间的精准衔接,让同一物理素材在不同场景中获得很好的“映景感”。基于此逻辑,Sora 根据用户文本指令生成换脸视频,使深度伪造技术之滥觞变得难以遏制,非法换脸视频会成为犯罪工具,恶意使用者利用“眼见为实”的判断习惯或认知偏见实施诈骗,或者将生成的不雅视频在信息网络上传播,严重损害他人声誉,扰乱社会秩序[14]。

Sora 参与编造谣言,制造、引导社会舆情事件。为了能高质量地反馈用户文本指令,Sora 在精准理解视频素材中的物理矢量关系的基础上生成大量虚假的视频,虚假视频带来的负面影响能渗透至社会生活多个领域,不仅助长了社会谣言的泛滥,还将影响人类政治稳定,给社会治理带来诸多障碍。同时,这些并非仅仅是 Sora 所特有的,其他 GenAI 应用也难辞其咎,因此饱受社会诟病。如,此前在信息网络中大面积传播的虚假视频,即利用深度伪造技术编造的网络谣言,发布者企图以虚假事件带动社会舆情、引导社会舆论走向,实现不可告人的秘密。

Sora 被滥用的情形不仅止于此,其甚至还可能被用于生成含有色情、暴力内容或仇恨言论的假视频。这种假视频具有高逼真、来源隐秘和治理难度大等特点,这类滥用行为不仅侵犯他人合法权益,亦给社会和谐带来风险,为各国法律所不能容忍,给社会监管带来挑战。

二、制度梗阻:我国生成式人工智能算法伦理治理障碍

Sora 带来的多重伦理风险肇始于算法应用

① 帧,即组成连续视频的一张张画面,视频是由有限的帧构成。

安全和数据使用安全,因此,应当聚焦算法监管和数据合规。然而,在算法监管和数据合规层面,我国却面临多重治理障碍,这既有来自具体应用场景内的 GenAI 大模型监管细则的阙如,也有来自视频数据使用行为监管的漏洞。面对种种现实弊因,有待立法者查漏补缺。

1. GenAI 算法监管规范体系阙如

当下,我国尚未建立完整的人工智能算法监管体系,面临 GenAI 伦理监管体系阙如的现实困境。我国算法监管制度散见于《互联网信息服务算法推荐管理规定》(以下简称《推荐管理规定》)《互联网信息服务深度合成管理规定》(以下简称《合成管理规定》)和《暂行办法》等多部法律规范,诚然它们对算法使用原则、算法服务提供者和使用者的主体义务、算法的可解释性、算法决策的公平性等提出了若干基本道德要求,但目前尚未形成完整的监管体系,遑论具体应用场景内的 GenAI 伦理监管细则。

(1)具体应用场景内的算法监管细则语焉不详

《暂行办法》不仅涵盖技术发展与治理、服务规范,还就 GenAI 应用的监督检查和 GenAI 服务提供者的法律责任等作出了一些规定。然而,该办法对 GenAI 相关主体在具体应用行业内的行为规范多为原则性规定,未能细化为具体的行为义务。例如,虽然《暂行办法》第七条对算法服务提供者规定以多项禁止性义务,但是对于主体行为规范的量化规定仅仅是直接援引《中华人民共和国网络安全法》(以下简称《网络安全法》)《中华人民共和国数据安全法》(以下简称《数据安全法》)《中华人民共和国个人信息保护法》(以下简称《个人信息保护法》)的相关条款。令人遗憾的是,这 3 部法律中的多项条款也多为原则性规定,缺乏可供援引的细则,致使适用性与可操作性较弱。此外,GenAI 使用者的行为义务规范也明显缺位,《合成管理规定》中的相关规定多为原则性的禁止性规定,缺乏可量化的适用细则。当下,个人使用者的行为规范多以平台服务提供者单方拟定、提供的行为公约为主,并以用户遵守该约定为使用算法应用的前置条件。该类格式化的“行为公约”未能吸纳其他平台参与者的意见,未能反映他们对算法规则的期待。因此,多数情况下用户并不能理解条款的具体要求,致使用户行为规范沦为一种准规范,即他们在算法使用规范的引导下错误地将日常行为规范地转化为使用算法的行为准则。特别是在关乎社会综合治理质量的数据上,用户行为反馈的真实性和完整性尚有待考究,很难真实反映用户对算法平台规则的期待。

(2)具体应用场景内的算法使用行为监管欠缺

《数据安全法》从用户个人数据、政务数据的处理规范视角出发,要求在使用数据时遵守法律规定和伦理规范①;《推荐管理规定》亦仅仅强调算法推荐者的义务,未提及包括平台用户在内的算法使用行为义务。在使用大模型过程中,用户做出的算法使用行为关系到算法进化,但是用户行为规范缺乏可供适用的具体法律依据,因此,存在监管规则上的缺失。特别是用户在使用文本指令时,基于目的后面的数据生产者行为,海量不精确、错误甚至与社会主流伦理观相悖的视频文本数据被用于“浸润”Sora,数据价值随即过渡给了算法,使得 GenAI 遵循着数据所代表的不同利益集团的意志。加之大模型的“黑箱”效应,GenAI 生成内容显现出歧视性倾向,影响社会公平,如此不加规制任由其发展、进化,最终将导致算法异化,由此带来的社会影响可想而知。因此,立法者亟须重新审视细分场景下相关主体的权利义务设计,将部分透明义务向算法使用者(含用户)转移,以确保训练 Sora 的视频数据和用户使用的文本数据具备可靠性与透明性。通过规范用户实际使用 Sora 的行为,确保大模型接受用户“投喂”的文本数据符合人类价值取舍,促进算法向善。

(3)具体应用场景内的算法伦理安全规则缺位

细分应用行业内的算法评估规则作用范围局限于数据应用的安全领域,而未直接涉及算法伦理安全。我国现行算法评估体系并不完

① 参见《中华人民共和国数据安全法》第八条的规定。

善,存在较大的改善空间。《个人信息保护法》《推荐管理规定》《暂行办法》构成了我国算法风险评估体系的基本框架,然而,该框架并未预判 GenAI 可能诱发的“绝对”失业风险。同时,该框架寄希望于实现对算法应用结果的规训,缺乏必要的算法使用过程的管控,难以真正实现算法决策公平、正义,因此,亟须对 GenAI 替代性使用风险进行事前评估。从功能上看,我国算法评估制度在作用范围和可操作性上存在障碍。

一方面,制度的作用面过窄,且局限于个人隐私保护。实质上,我国现行算法评估制度仅仅是进行算法对个人数据安全影响的事前预估①,即算法在使用个人数据时是否能避免通过数据绘制个人画像的角度展开的,并非对 GenAI 伦理规范层面的评价,而是局限于对数据来源和使用目的的预判。因此,该制度强调的仅仅是数据保护要求,而未能扩展到 GenAI 伦理安全层面。

另一方面,评估制度的可操作性不强,且未触及算法伦理安全,治理效果或将偏离立法者预期。《推荐管理规定》细化了《个人信息保护法》中对个人信息使用前的安全评估要求,明文禁止 GenAI 将用户个人信息进行标签化处理。然而,这也在一定程度上恶化了嵌入应用后的 GenAI 伦理失范风险,为大模型实施个人歧视提供了一条更加隐秘的路径,致使算法决策偏离公平、正义的伦理要求,使其在实践中很难达到立法者的预期效果。此外,该制度的可操作性也受限于 GenAI 本身决策逻辑的不可知性和决策结果的不确定性,仅仅对数据进行审查并不能全面评估算法伦理善恶。实务中,GenAI 标签化地使用视频数据涉及的环节颇多,然而现行算法评估制度发生作用的环节仅仅局限于数据收集阶段,很难触及存在于其他环节的算法行为。此外,用户举证难题依然突出,被侵权者获得权利救济的难度较大(此处不展开讨论)。可见,现行算法评估机制并不能有效应对 GenAI 侵权隐秘性强、形式多样的现实难题,有待进行必要的修葺。

2. 视频数据使用者行为监管失衡

以《个人信息保护法》和《数据安全法》为主的法律规范构建起了我国数据安全保护体系,在个人隐私保护和维护数据安全方面发挥着重要作用。然而,在使用在先视频数据训练 Sora 的场景中,我国视频数据使用监管制度仍然存在明显的不足,有待完善。

(1)宏观层面:过度重视数据保护阻碍数据要素丰富

我国面临过度重视数据保护而忽视数据要素丰富的难题。从立法目的上看,《个人信息保护法》和《数据安全法》是分别从个人隐私保护和国家数据安全角度进行的立法实践,是一种事后的强制性法律规范,主要调整的是民事主体与国家法益、社会公共法益之间的冲突,保护的是公民个人隐私和国家安全免受不当势力侵害、社会公共利益不受不当损害。从实务效果上看,《个人信息保护法》和《数据安全法》给数据处理者施以过重的注意义务,束缚了数据使用者的创造力,限制了他们在使用数据时的活力。《数据安全法》多以抽象的数据使用行为作为规制客体,且多采取禁止性规定,不以中立裁决立场调整主体行为,而是以命令式的发问要求“不为得”“不可为”,多存在与“坚持共享共用,释放价值红利”②原则相冲之虞,恐将难以有效带动数据要素丰富和数据市场高效活跃,特别是在 GenAI 呈现井喷式发展的数字经济时代,数据保护与行业发展处于明显的失衡状态。

(2)微观层面:数据生产者行为规范缺位与法律条文适用困难

数据生产者的行为规范缺位。两部数据监管法均未对数据生产者的行为作必要的安排,也并未从制度层面推动用于“投喂”GenAI 的价值观应当符合我国社会主义意识形态的要求。这一制度缺陷为用户在大模型中恶意“投喂”虚假的、错误的甚至反动数据的行为留下了法律“空子”。目前,用于规制该类行为的约束文件也仅限于平台使用规范,并未上升至公法可规制的范畴。在我国行政法规中,《中共

① 参见《中华人民共和国个人信息保护法》第五十五条第二款的规定。

② 参见《中共中央　国务院关于构建数据基础制度更好发挥数据要素作用的意见》工作原则第二条的规定。

中央　国务院关于构建数据基础制度更好发挥数据要素作用的意见》(以下简称《数据二十条》)明确了“安全可控、弹性包容”的数据治理原则,并要求建立算法审查制度以确保数据在流通中有据可查。但该文件并未就数据生产行为作出翔实的规定,甚至可以认为,在数据主体混同的当下(主要是生产者与使用者的身份混同),数据生产者行为规范的缺位是该制度的一项重大缺失。

部分法条在适用上存在困难。在数据使用规制方面,我国诸多法律规范用词模糊、宽泛,给司法适用造成困难,同时,细分行业内数据使用行为的适用细则明显阙如。《个人信息保护法》使用了诸多较为模糊的法律词汇,给予法官一定的自由裁量权,同时也造成司法和执法障碍,给实务工作带来诸多不便。例如,该法第十八条的“最短时间”、第十三条第五款的“合理的范围”和相关条文中多次出现的“必要”等规定,这些都有待出台具体的司法解释加以明确和规范,否则给予裁决者过于宽松的自由裁量权或将滋生乱象。《推荐管理办法》也多为原则性规定,至今仍然未就此出台正式可适用的细分行业应用准则和司法适应细则。《生成式人工智能服务安全基本要求》(TC260-003)虽然对 GenAI 使用在先作品数据提出了合法性要求,但也并非具体适用性的监管细则,多为一般禁止性的行为义务,事前的可适用性并不强。

在先权利人权利保护困难。虽然我国《个人信息保护法》明确赋予个人用户以知情权、同意权、删除权和拒绝权等数据权利,但是用户在使用算法的过程中往往以无条件同意预先拟定的格式条款为使用平台的代价,行使上述权利即意味着用户远离算法应用。机械地赋予用户上述权利,而缺乏必要的保障机制往往使权利被束之高阁,成为华而不实的愿景。大模型在进行文本数据挖掘和机器学习过程中必然使用海量的视频作品数据,GenAI 服务提供者和用户在违背“知情同意”原则下使用在先权利人的作品数据。囿于黑箱属性和 Sora 融合创作的隐蔽性,在先权利人往往无法察觉侵权行为和侵权结果。同时,举证责任分配也欠缺合理性。即使在先权利人发起诉讼程序捍卫其权利亦面临举证难题,当事人在诉讼中地位的不平等和 Sora 侵权的隐蔽性,使“过错原则”在适用上存在障碍,甚至常常发生举证不能。较重的举证义务将导致权利人丧失获得司法救济的机会,“无救济,则无权利”,这也将变相助长 Sora 使用者滥用他人在先作品的行为,破坏人类知识产权秩序

三、制度调试:我国生成式人工智能算法伦理治理理路

为应对 GenAI 伦理失范的多重风险,立法者亟须完善我国算法监管制度和数据使用合规制度,明确我国人工智能算法伦理治理的因应理路:一方面,强化算法监管和算法民主化,实现人工智能算法公平和算法伦理多元共治;另一方面,强化数据监管,推动 GenAI 训练数据具备透明性和多样性,确保 GenAI 能够接受不同观点和群体价值观的“浸润”。

1. 建立算法分类分级审查体系,强化算法监管

单纯寄希望于事后规制不足以有效应对 GenAI 伦理失范风险。GenAI 生成内容具有歧视性特征,因此,需要在其设计完成后、上线应用前进行伦理审查。我国现有算法审查规范多基于赋予私主体的“算法解释权”、信息处理者的“说明义务”、行政机关的“算法监管”展开的,发挥着一定的社会功能[15]。然而,可应用于通用场景的审查标准多依赖于技术的透明性,这对行政机关的数字素养要求较高而对 GenAI 本身的行业技术规范要求偏低。因此,现有算法监管制度中必然出现前述存在于监管和评估层面的治理障碍。鉴于此,针对我国 GenAI 算法监管和数据监管所面临的制度短板,建议建立基于场景化的分类分级审查制度。同时,通过算法设计听证,吸纳多元主体参与算法设计,推动算法逻辑符合多数人的价值观。

(1)分级分类算法审查,强化算法监管

根据生成内容的性质将 GenAI 应用分类,明确各类 GenAI 应用的伦理审查标准。在算法应用中,一旦使用场景发生变化,算法的性质就会迥然不同[16]。算法的黑箱属性使得同一

数据在不同应用场景下触发的伦理风险等级存在差异,因此,监管要求也应当应场景的变化而有所区分。因此,意欲消除我国 GenAI 伦理治理障碍,就应当在符合透明性的一般要求上,根据不同的使用场景特征明确 GenAI 风险审查标准与审查的相关细则,同时规定相关主体的审查义务、责任分配等具体要求,而不能仅仅停留在寄希望于抽象的一般原则性规定。

然而,场景化的算法审查制度并不能仅仅聚焦于单一抽象的场景下讨论审查细则,而是应当在具体的应用场景中展开纵向剖析,把审查标准落实在具体的应用环节和使用细节上。如,用户指令 Sora 生成风景类视频,一般情况下,其生成的内容应当是环境友好型的,能对引导人类保护自然生态发挥促进作用,对 Sora 生成的描绘人居环境的视频也应当是符合“和谐”价值观要求的。除非出于满足特殊的场景刻画需要或实现特别的反面教育要求,表现人性恶的 GenAI 生成内容才能是符合规则要求的。因此,应当分别量化“环境友好”和“和谐”等的基本要求。寄希望于此,以终为始,倒逼数据使用行为的合规性。

综合评估生成内容的风险,确定 GenAI 风险等级。算法黑箱不可避免地带来了算法偏见风险,因此需要“打开算法的黑箱,提升算法的透明度”[17]。在此之前,需要预判 GenAI 在具体应用中的使用风险,通过衡量不同应用环节的风险等级(高决策风险场景、中决策风险场景、普通决策风险场景),确定不同解释主体的解释义务和相应的问责制度。一般认为,应当将关键基础设施、教育、就业和刑事追溯等领域列为高风险决策场景,这些场景中的 GenAI 应用必须接受严格监管[18]。特别是对影响劳动者就业和容易诱发犯罪风险的算法应用而言,我国应当出台单独的监管细则,严格规范其应用场景和应用范围;对于易诱发社会歧视风险的算法应用,因为其受众面较广,也应当被纳入高决策风险场景(至少应当是中决策风险场景)加以规制。同时,在实践中还应当坚持比例原则,平衡审查与公开的界限,合理把握算法披露的程度和限度,不能以牺牲特定主体的正当权利换取 GenAI 行业的繁荣。避免侵犯在先权利人的知识产权,以此保障人类知识产权制度的顺畅运行。

在不同等级的决策风险场景中,解释程序亦应当有所差别。具言之,对于高决策风险场景,行政机关有权要求 GenAI 服务提供者进行系统性解释,即由国家行政机关发起审查程序,或由负责的行政机关应民间社会团体的申请发起,依靠强有力的公权力保障和推动程序运行。如,根据 Sora 生成视频的使用目的,确定其可能产生的影响范围、程度,对于影响范围较广、后果较严重且作出歧视性决策结果的 GenAI 而言,可以由行政机关主动发起披露程序,从劳动平等、性别平等出发,追查其技术根源,要求 GenAI 服务提供者提交合理性解释与说明,必要时责令限期整改,以期符合算法安全要求。普通决策风险场景可以使用个案解释程序,即就单个用户的单次使用结果进行的有针对性的解释程序,也是应用户的申请对存在于 GenAI 中的决策疑虑发起的披露程序。个案解释程序不具有普遍解释效力,仅对该案的决策结果负责。鉴于提升算法服务效率和控制服务成本的考量,算法服务提供者在进行公众披露时可以只遵守个案披露程序,但是需要特别注意,进行行政披露时必须启动系统解释程序。

(2)算法听证,多元主体参与算法设计

尝试构建算法设计听证制度,确保公众参与 GenAI 设计,吸纳多元主体参与 GenAI 伦理共治。广大民众是算法规则的随受者,应当确保他们能参与 GenAI 规则的制定过程。与其他算法一样,GenAI 技术使用一整套看似规范和贴心的专属方案为用户提供个性化的定制服务,然而,透过这层虚伪的面纱可以发现,接受 GenAI 服务的用户越多,被嵌入应用的 GenAI 服务方案就愈精准,用户个体价值就愈容易被算法裹挟。GenAI 使用单一的代码为受众划定行为范围和价值禁区,歧视那些实施偏离算法预设价值规则行为的用户。在社会治理领域亦是如此,治理者借助各种代码逻辑实现其政治价值,圈定权力场域,温柔地迫使困于其中的人们在“安全范围”内活动。由此可见,算法制度的设计需要保证多元化主体的参与,公众参与 GenAI 设计的核心任务在于保障算法参与者的

利益,通过制度设计来满足最广泛主体的偏好差异[19]。通过算法设计听证,受众能清晰地了解 GenAI 的技术逻辑、价值取向和应用影响,在算法设计和运行过程中,持有不同价值偏好的受众能表达其价值诉求,以平等的身份与 GenAI 设计者、服务提供者交换价值取舍,保障他们能参与算法设计的过程,力图使 GenAI 能反馈广大公众的多元化价值诉求。

2. 强化算法训练数据监管,推动数据透明多样

我国对 GenAI 数据使用行为的监管力度明显不足,无力有效应对 Sora 诱发的多重伦理失范风险,建议从制度维度和技术维度加以应对。从制度上强化数据监管,明确数据生产者、使用者的行为义务,强化责任担当,确保数据透明、可责。推动全球数据共享机制的建设,助力实现 GenAI 训练数据的具备多样性,尽量避免 GenAI 生成内容中的价值偏见。

(1)数据监管机制:明确数据透明义务,强化责任担当

完善我国现有数据监管制度,要求用于训练文生视频模型的数据具备透明性。在动态发展中平衡数据要素丰富与数据安全,出台司法解释细化上位法的模糊性规定,明确相关主体的责任分配,以此便利执法和司法。以分类分级算法审查制度为基础,强化数据使用中的科技伦理审查。以我国《关于加强科技伦理治理的意见》《新一代人工智能治理原则——发展负责任的人工智能》《新一代人工智能伦理规范》等治理方案为底基,尝试区分应用场景及应用环节明确数据生产、使用中的伦理要求,制定细分行业的伦理审查标准,使之成为数据生产者、使用者的强制性行为规范。在此之上,对 GenAI 服务提供者施加数据使用解释义务,明确 GenAI 在具体应用场景下的数据透明性要求。另外,GenAI 数据监管制度还应当将数据安全风险的事前评估制度和数据安全事件的应对预案纳入其中,以此强化 GenAI 设计者和运营者具备数据安全底线思维,使数据生产者、使用者在符合数据安全的前置要求下训练 GenAI,约束伦理偏见,在数据来源、数据质量和数据安全上考虑用于训练大模型的数据伦理风险敞口[20]。

一方面,建立数据透明备案制度,确保训练 GenAI 的数据具备透明性。适时向公众公示用于训练 GenAI 的数据,包括数据来源、种类及传输加密机制,确保数据来源可靠、数据真实、不被篡改。对经公示的数据使用进行备案,确保训练 GenAI 的数据具备可解释性。训练 GenAI 数据的透明性要求也应当以合目的性为前提,即 GenAI 使用数据必须符合算法目的,是实现 GenAI 算法目的所必需的,不符合算法训练目的的数据不得被用于训练算法。以 Sora 为例,文生视频模型在各细分场景下训练所需要的数据是不同的,因此,应当就不同场景进行数据备案,而不能进行整体性备案。虽然整体备案在一定程度上能缓解公众对数据多样化的期待,但并不符合场景化治理理念,其终将导致 GenAI 生成过程的不可知,伦理失范或将是必然。

另一方面,出台司法解释,提高现行法律的准确性与可操作性。鉴于存在于我国数据治理领域的“部分法条的适用仍然存在困难”和“法定权利的可操作性不强,个人权利保护困难”的问题,需在立法层面加以应对。出台具体的司法解释,就存在于《个人信息保护法》中模糊性的法律条文进行限缩性解释。同时,就《推荐管理办法》的原则性规定制定司法适用标准,以期提高司法可适用性。建议从行业自律层面着手,通过科技企业自律的方式,增强保护用户权利的意识,以此解决部分法条在实践中可操作性弱的短板。例如,强化 GenAI 提供者的行为义务,增加用户投诉渠道,提升用户问题反馈的便捷性等。除此之外,为降低在先权利人在诉讼中的举证负担,建议尝试探索“举证缓和”制度。在尊重“过错原则”的基础上,适度限缩在先权利人的举证义务。当在先权利人已经穷尽必要的举证手段和最大限度地收集、提出证据后,赋予法官自由裁量权,减轻诉讼弱势方——在先权利人在案件中的举证负担。基于此种制度设计,希冀降低在先权利人维权难度。

此外,为明确数据生产者的行为义务,需在

立法层面使数据生产者对其生产的数据安全性负责。在数据生产者与使用者主体混同的情形下,数据生产者的数据安全义务不仅包括使用 GenAI 生成的视频数据是合法、安全和可靠的,而且还应当要求用户使用文本数据驱使 Sora 生成时,其数据是合法的(一般是私法),也是符合人类伦理期待的。

(2)数据共享机制:丰富数据多样性,杜绝数据滥用

为保障用于训练 Sora 的视频数据具备多样性,建议建立覆盖全球范围的数据共享制度。一般认为,数据偏见隐藏于数据的逻辑关系当中,没有不存在偏见的数据。数据偏见是记录数据源的人类情感反馈,零散的数据偏见常常未达到歧视的程度。但是,海量的数据“浸润”GenAI 后,基于数据偏见喂养的 GenAI 即具有歧视性。数据不仅普适记录生产者的意志,还在特定语境下代表特定民众的价值取向,隐含他们的价值偏见。非中立性的数据在参与 GenAI 生成时,受其价值“浸润”导致算法偏见,数据质量会使算法权力生成量级偏差,致使出现歧视性结果。鉴于此,应从数据的透明性和多样性入手,进行制度构建。数据多样性是提升 GenAI 泛化能力的前提条件,也是确保 GenAI 生成内容符合公平性要求的基础。因此,建议在全球范围内建立数据共享协议,明确数据的开放条件和开放范围,促进数据要素实现跨域流动和跨行业融通。在完善数据开放平台的基础之上,提供统一、全面、准确、便捷的数据共享服务,鼓励各方主体共同参与数据开放和共享。

为确保数据共享机制能顺利运行,建议建立 GenAI 数据多样性监测与影响评估配套机制,明确具体应用场景下用于衡量 GenAI 数据多样性的监测指标,防止数据滥用。监测指标不仅应该涵括数据来源、数据源的分布、数据类别的覆盖率,还需要将使用数据的 GenAI 逻辑进化情况纳入其中。此外,还应当定期展开数据多样性的事中评估,根据评估结果适时调整数据收集、存储、管理和应用策略,确保用于训练 GenAI 的数据持续丰富、透明与可控,力图使其符合人类多样性的价值观要求。

四、全球视野:推动我国参与人工智能算法伦理全球治理

一般认为,全球治理已成为人工智能伦理之治的人类共识。GenAI 伦理失范风险已成为亟须人类共同应对的全球性问题,当灾难来临时,没有任何一个国家能独善其身。当下,我国人工智能产业发展呈现出“重应用,轻底层”的显著特征,这既是产业政策失灵的无奈,也是资本逐利短视之必然,如若人类不能有效协调技术发展与社会治理的关系,或将陷入“科林格里奇困境”。我国人工智能算法伦理治理理论研究和法治建设的基础较薄弱,不仅需要完善本国科技伦理治理体系,还应当在保证国家安全的基础之上积极参与全球性治理合作,加强区域性的治理技术交流。为确保我国能积极参与人工智能算法伦理全球治理,还需要从技术创新和制度建设两个方面持续加大投入。

1. 以人工智能技术创新规制技术伦理失范

我国在大力发展大模型上层应用的同时,需要更加重视 GenAI 底层模型的研究,希冀通过人工智能技术创新规制技术理性越位。为了实现 GenAI 伦理全球治理,强化我国协同创新能力,建议从底层算法技术标准共享和全球人才创新机制建设着手,实现以人工智能技术创新规制 GenAI 伦理失范之目的。

(1)积极参与 GenAI 底层模型技术标准的全球共享

一方面,借鉴域外人工智能技术发达国家的成功经验,重视 GenAI 底层大模型的技术创新,推动底层技术标准的全球共享。底层模型的搭建是支撑人工智能技术应用的基础,也是促进技术伦理转向的关键步骤。当前形势下,我国需要十分重视底层模型的研究,认真审视 GenAI 伦理风险的技术根源。以美国为首的北方技术强国在 GenAI 底层模型的研究上投入较多的人力、物力和政策支持,在技术成果上也取得了显著的成绩,从 GPT3 到 GPT3.5,再从 ChatGPT4.0 到 Sora、ChatGPT turbo 和 GPT-o1,GenAI 正在且已经发生指数级迭代。这些技术强国不仅成为行业标准的制定者,引导其他国

家纷纷跟随,也是 GenAI 技术发展的既得利益者。鉴于此,我国应当在政策层面与北方国家接轨,在实现技术创新的同时积极参与全球人工智能标准的制定,跻身于推动 GenAI 底层技术标准的全球共享。

另一方面,强化我国 GenAI 技术发展的独立性。我国需要在大模型和算力等底层技术领域实现创新,在数据处理技术、隐私计算和道德模型等前沿技术领域攻克难关,突破技术发展瓶颈,特别是要避免对受西方意识形态喂养的 GenAI 算法模型产生过度依赖,致使诱发不可控的伦理失范风险。应当清醒地看到,西方 GenAI 技术强国走的是“重视底层模型、以底层倒逼应用”的路径,这是一条强科技发展的道路。反观我国,目前正在与域外截然不同的技术发展路径上举步维艰。寄希望于“拿来主义”的“重应用、轻底层”的做法暂时让国内一些科技企业得到了微弱的利益回报。但从长期来看,这种短视的做法无异于作茧自缚,终将受制于人,必将逐渐丧失技术发展的自主性。血淋淋的例子就发生在眼前,美国 OpenAI 于 2024 年 6 月 29 日突然单方面邮件通知所有中国用户,自 7 月 9 日起终止提供 ChatGPT 的 API 接口服务,致使我国大部分科技企业面临终止运营、切换平台等不利局面,也把我国国内 GenAI 平台使用者带入低质量平台甚至无平台可用的囧境[21]。

在参与国际标准的制定过程中,治理者务必根据我国国情制定人工智能战略政策和法律法规,为此需要充分发挥上层政策的引导作用,科学把握 GenAI 技术发展方向,着力于解决关键领域“卡脖子”的底层技术创新瓶颈,坚持自主设计、自主研发,牢牢掌握算法、算力技术发展的主动权。在展开 GenAI 算法设计时,明确道德算法阈值的设定应符合我国社会主义核心价值观的基本要求,以社会主义核心价值观引导 GenAI,使其符合我国社会基本价值取舍,也应当保证其与整个人类的基本价值观对齐。在提升算力时,需要制定适宜的产业发展政策,逐步攻克芯片等关键技术的发展难题,应当着眼于整个人工智能产业链的长足发展,以上游产业技术带动下游产业集群的持续发展。

(2)致力于科技创新人才培养和域外专业人才引进

以建设创新人才机制为底基,为 GenAI 技术创新提供持续的人才储备。为加快我国 GenAI 技术人才储备的步伐,需要在人才创新机制上下狠功夫,强化《中华人民共和国科学技术进步法》对我国人才创新制度的指导价值,在教育和制度上保障人工智能技术实现持续创新。

在教育上,专注人工智能专业人才培养:一方面,应当把握全球视野,加强教育培训与国际人才交流合作,在全球范围内提高人工智能从业者素养与技能水平。具言之,应当加快人工智能专项教育立法,建立健全人工智能教育法律体系,以此实现建立多部门协同、多主体参与的人工智能应用教育体系目标,细化国家人工智能教育发展战略和行动计划,使创新人才培养项目有章可循。另一方面,创新全球人工智能教育应用路径,促进数字教育公平[22]。如,将优质、前沿的人工智能教育资源投入人类教育实践,丰富教育内容,变革教育方式;将 GenAI 赋能教育方式和教育内容,实现全球教育均衡,促进全球教育公平;为产学研融合发展体系的建立提供“软基础”,实现我国教育体制改革的“软着陆”;提高广大师生的数字素养,强化教育从业者的人工智能伦理教育。

在制度上,健全人工智能国际产业转化机制:建立健全人工智能教育的国际合作制度体系和多元化的人才培养制度,从政策层面引领政府、学校、企业和科研院所投身国际人工智能人才培养事业,建立国际人工智能教育应用合作平台,探索可信赖人工智能教育应用实践。与此同时,充分利用人类现有知识产权制度体系,从产权保护视角和立场把握、促进 GenAI 大模型技术创新建设和产业转化,希冀有效应用技术治理人工智能伦理失范风险,引导 GenAI 伦理符合我国社会主义伦理的发展要求。

此外,适时启动国际人工智能专项人才引进计划,即在符合总体国家安全的前提下推进国际人工智能专项人才跨国别流动,不仅需要将人才送出去,还要从北方 GenAI 技术强国引

进优秀人才。结合人才强国计划,打造复合型人工智能人才梯队,依据城市发展和产业布局需要,建立国际人工智能人才引进计划。以东部沿海城市为引导,以 GenAI 产业发展为驱动,从国家政策、配套制度和人才发展等方面确保人才政策实现持续优化[23]。

2. 积极参与人工智能算法伦理国际治理与区域协作

当下,全球人工智能伦理治理进程正在加速演进,形成了复合型的机制建构路径[24]。于我国而言,应当积极跻身于构建人工智能全球治理体系,从国际治理和区域协作两个维度开展治理网络建设,各相关国际组织、国家或地区应当出台人工智能相关伦理监管政策,以期实现 GenAI 伦理治理与技术创新的平衡[25]。

一方面,积极推动国际组织展开全球算法伦理治理布局,发挥国际组织在全球治理上的积极效能。倡导建立全球范围内的人工智能治理机制,支持联合国发挥主导作用。GenAI 伦理治理议题业已进入国际治理议程,从起初的非国家主体倡议、政府和国际性机构参与,演变为重要国际组织介入,形成了多主体、多角度、多层级的全球治理机制复合体[26]。鉴于国际组织具有尊重文化多样性的号召力和协调力,应当在尊重国别差异的基础上强化全球治理的伦理标准,发挥国际组织在制定国际性行业规范和标准上的引领作用。同时,人工智能诱发的伦理风险并非简单的一国国别内的新兴技术风险,技术发展带来的是地缘政治的复杂化。因此,人工智能伦理风险也伴随着国别竞争的成分[27],在此情境下,积极参与人工智能国际治理具有现实迫切性和挑战性。因此,我国应当积极参与国际人工智能伦理治理研究和治理实践,强化南北对话,既要为制定全球性的伦理规范和治理制度建设提供中国智慧,又要力争扩大我国在国际人工智能伦理治理浪潮中的话语权,更要在治理中切实维护我国国家利益。

另一方面,强化区域间算法伦理治理协作。在底层算法伦理治理和数据伦理及其相关方面强化区域性合作,促进区域及不同国家之间的政策互认、互信,加强南南合作,提升南方国家在全球人工智能伦理治理中的代表性、发言权。积极鼓励区域组织、企业、研究机构、社会组织和个人等多元主体参与人工智能治理体系的构建和实施。不仅需要就区域内的数据安全、算法伦理安全达成基本合作意向,还应当就建立 GenAI 关联违法犯罪专项治理进行多边磋商,特别是就治理 GenAI 诈骗犯罪建立长效治理机制。需要在南南合作中防范恐怖主义、极端势力和跨国组织犯罪集团利用人工智能技术从事非法活动,共同打击窃取、篡改、泄露和非法收集利用个人信息的行为,应对 GenAI 诱发的高度违法风险。这在 2023 年 8 月公安部网络安全保卫局发起的“严打行业内鬼泄露个人信息 紧盯 ChatGPT 等新应用”专项行动中已取得初步成效[28],因此,该机制具有现实的可行性。

当下,我国在人工智能全球治理领域已经具备了必要的法治基础,如我国提出《全球人工智能治理倡议(2023)》,在人工智能发展、安全、治理三方面系统性阐述了中国治方案,形成了具有广泛共识的全球性人工智能治理框架和标准规范。倡导坚持伦理先行,建立并完善人工智能伦理准则、规范及问责机制,形成人工智能伦理指南;建立科技伦理审查和监管制度,明确人工智能相关主体的责任和权力边界,充分尊重并保障各群体的合法权益,及时回应国内和国际伦理关切[29]。《全球人工智能治理倡议(2024)》再次强调打击利用 GenAI 实施违法犯罪行为的国际性倡议。但是,我们还需要为人工智能伦理治理的南南合作做出更多的努力,需要就该倡议文件与南方国家达成共识,就此建立 GenAI 细分应用领域的长效合作机制。希冀在 GenAI 伦理治理领域内加强与发展中国家的合作,弥合存在于区域内的智能鸿沟和治理差距,借此与其他国家一道践行人类命运共同体理念,共同构建“开放、公正、有效”的人工智能算法伦理治理长效机制。

五、余　论

以 Sora 为代表的文生视频技术正在发生飞速迭代,其伦理风险彰明较著。Sora 重新定义了视频生产,同时,经由其生成的视频不仅饱含价值偏见,亦威胁人类就业市场,甚至挑战人类法律底线,应当从算法透明与数据多样路径

展开治理机制构建。然而,算法与数据分庭治理业已成为当下人类社会治理的基本态势,弊端突出,不仅徒增社会治理成本,还造成治理责任模糊甚至部分责任主体真空的后果。为此,需要尝试构建中央网信办和数据安全局为主体的协同治理模式,强化不同治理部门之间的协调性和工作的联动性,联合多元社会主体,形成分工明确、层次清晰、责任明确的协同治理体系[30]。坚持审慎态度平衡 GenAI 发展与伦理风险规制,审视技术发展与制度设计之间的矛盾。“不发展就是最大的风险”,赋予 GenAI 在发展过程中以必要的试错机会。GenAI 伦理治理是一个动态的过程,而非一蹴而就的结果,需要提高大众的算法素养和数据素养,使公众了解 GenAI 规则及其运行逻辑。欧盟《人工智能法案》强调 GenAI 算法应当与人类价值观对齐,并就此建立的诸如“监管沙盒”治理制度的做法是值得我国借鉴的,我国应当积极参与人工智能伦理治理的南北对话和南南合作,为人工智能伦理治理贡献中国智慧;参与国际标准制定,提高国际治理话语权。希冀通过以上努力,在自由使用 GenAI 的同时能有效预防并化解伦理失范风险,促进 GenAI 伦理符合人类预期,有力推动人工智能产业实现良性发展。

参考文献:

[1] 吴晓凌. 视频生成新模型 Sora 的突破与风险[N]. 新华每日电讯,2024-02-21(007).

[2] 人民网. 视频生成新模型 Sora 的突破与风险[OB/EL]. (2024-02-27)[2024-09-12]. http://world.people.com.cn/n1/2024/0227/c1002-40183853.html.

[3] 梁桐. 欧盟人工智能法案注重风险防控[N]. 经济日报,2024-03-25(004).

[4] 人工智能全球治理上海宣言[N]. 人民日报,2024-07-05(04).

[5] 邓建鹏,赵治松. 文生视频类人工智能的风险与三维规制:以 Sora 为视角[J]. 新疆师范大学学报(哲学社会科学版),2024,45(6):92-100.

[6] 曹克亮. Sora 的意识形态效应及其治理[J]. 统一战线学研究,2024(3):166-178.

[7] 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.

[8] 叶志华. 关于主体、主体结构和主体功能的一些认识[J]. 现代哲学,1991(3):18-20.

[9] JOHN S. Economic possibilities for our grandchildren:progress and prospects after 75 years[J]. Essays in Persuasion,2005:358-373.

[10] 韩东屏. AI 失业潮[J]. 21 世纪商业评论,2017(10):32-33.

[11] 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.

[12] 李韬,周瑞春. 生成式人工智能的社会伦理风险及其治理——基于行动者网络理论的探讨[J]. 中国特色社会主义研究,2023(6):58-66.

[13] 王毓莹. 人脸识别中个人信息保护的思考[J]. 法律适用,2023(2):15-24.

[14] 谢新水. 基于脸的治理及其转向:从刺脸、刷脸到非法 AI 换脸——兼论数智时代人的数字化问题[J]. 探索,2023(5):26-38.

[15] RAMPRASAATH R, MICHAEL C, ABHISHEK D, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization[C]//2017 IEEE International Conference on Computer Vision, Venice, 2017:618-626.

[16] 丁晓东. 论算法的法律规制[J]. 中国社会科学,2020(12):138-159.

[17] 李成. 人工智能歧视的法律治理[J]. 中国法学,2021(2):127-147.

[18] 沃尔夫冈·多伊普勒.《欧盟人工智能法案》的背景、主要内容与评价——兼论该法案对劳动法的影响[J]. 环球法律评论,2024(3):5-27.

[19] 陈家刚. 协商民主引论[J]. 马克思主义与现实,2004(3):26-34.

[20] 肖红军,张丽丽. 大模型伦理失范的理论解构与治理创新[J]. 财经问题研究,2024(5):15-32.

[21] 新京报. 突发! 国内开发者收 OpenAI 警告称 7 月 9 日起“断供”API[N/EL]. https://www.bjnews.com.cn/detail/1719303136169459.html.

[22] 人民网. 2024 世界数字教育大会发布“人工智能赋能教育发展”倡议[OB/EL]. http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/2024/2024_zt02/pxhy/pxhy_znll/pxhy_znll_mtbd/202402/t20240201_1113848.html.

[23] 孙鲲鹏,罗婷,肖星. 人才政策、研发人员招聘与企业创新[J]. 经济研究,2021(8):143-159.

[24] 鲁传颖. 全球人工智能治理的目标、挑战与中国方案[J]. 当代世界,2024(5):25-31.

[25] FRANCESCA F, TERESA S, ATIA C. Investing in

AI for social good:an analysis of European national strategies[J]. AI & Society:the Journal of Human-centered Systems and Machine Intelligence, 2023(38):479-500.

[26] PETER C, MATTHIJS M. M, LUKE K. Fragmentation and the future: investigating architectures for international AI governance[J]. Global Policy,2020,11(5):545-556.

[27] 余南平. 新一代通用人工智能对国际关系的影响探究[J]. 国际问题研究,2023(4):79-96.

[28] 中国日报网. 公安部:严打行业内鬼泄露个人信息紧盯 ChatGPT 等新应用[N/OL]. (2023-08-10)[2024-07-10]. http://cn. chinadaily. com. cn/a/202308/10/WS64d44990a3109d7585e48a7a. html.

[29] 中国南南合作网. 全球人工智能治理倡议[OB/EL]. (2023-10-19)[2024-07-10]. https://www.ecdc. net. cn/new/homenews/4339. html.

[30] 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.

On Sora's Ethical Risks and China's Governance Response: Also on the Basic Path of China's Participation in the Global Governance of Artificial Intelligence Algorithm Ethics/LIU Zubing (Law School, Tongji University, Shanghai 200092, China)

Abstract: The cultural video model represented by Sora triggers multiple ethical risks and urgently needs to explore the path of rationalization. The substitution of Sora application triggers the risk of weakening human subjectivity, worsening the wealth gap due to the "absolute" unemployment wave, triggering a possible "New Ludendorff Movement", and affecting the stability of human society. Sora accepts biased video data infiltration, depicting stereotypes about specific groups or viewpoints. The generated content lacks value diversity and inclusiveness, triggering highly covert social discrimination risks and affecting human social equity. The fake videos generated by Sora have a high risk of illegality, leading to the proliferation of fraudulent video information, high incidence of fraud crimes, and affecting social harmony. There is a lack of ethical regulatory framework for generative artificial intelligence in China, with unclear rules for algorithm regulation in specific application scenarios, insufficient rules for regulating usage behavior, and a lack of ethical safety rules for algorithms. There is an imbalance in the regulation of video data users' behavior, which includes macro governance issues such as excessive emphasis on data protection and hindering the enrichment of data elements, as well as micro governance issues such as the lack of behavioral norms for data producers and weak applicability of some legal provisions. We should establish a scenario based classification and grading review system, vertically refine platform data usage behavior rules, conduct algorithm hearings, and ensure the participation of multiple stakeholders in algorithm design. We should strengthen the supervision of video data messenger behavior from institutional and technological dimensions, and promote the transparency and value diversity of Sora training data. We should grasp a global perspective, participate in the formulation and global sharing of technical standards for generative artificial intelligence algorithm underlying models, and strive to promote the cultivation of scientific and technological innovation talents and the introduction of professional talents from outside the field. And we should actively promote China's participation in the North-South dialogue and South-South cooperation on ethical governance of artificial intelligence algorithms, and provide Chinese wisdom for global ethical governance of artificial intelligence.

Key words: Sora; ethical risk; classification and grading review; data transparency; algorithmic democracy; north-south dialogue