【意识形态】

生成式人工智能意识形态安全风险探析

谢 波 曹亚男

【摘 要】当前生成式人工智能的快速发展在带来新机遇的同时，亦给国家安全带来现实挑战和潜在风险。尤其是随着生成式人工智能逐步嵌入我国社会生产生活各领域，其悄然无声地塑造着意识形态，对意识形态安全之影响日渐显现。在生成式人工智能深度学习、算法推荐、交互方式发展过程中，意识形态安全风险的生成有其独特的理论、技术和实践逻辑，并呈现四种典型样态：开辟信息空间制造输出虚假信息导致信息失真、单向传播编织受众“信息茧房”形成算法霸权、通过资本僭越预设立场影响受众政治态度、以“去中心化”压缩主流意识形态生存空间。为牢牢掌握人工智能时代意识形态斗争主动权，中国需在总体国家安全观引领下多方面综合着力，打好网络意识形态主动仗，包括重塑人工智能技术与价值平衡观、坚定主流意识形态一元中心指引、构建主流意识形态具象传播框架以及推建生成式人工智能范域监管机制等。

【关 键 词】生成式人工智能；意识形态安全；信息茧房；算法霸权

【作者简介】谢波，西南政法大学国家安全学院副院长、副教授，主要从事国家安全学基础理论、新型领域安全研究；曹亚男，西南政法大学总体国家安全观研究院助理研究员，主要从事国家安全治理研究。

【原文出处】《国家安全研究》(京)，2024.1.102～127

【基金项目】本文为2022年重庆市社会科学规划智库重点项目“人工智能对国家政治安全的影响及应对策略研究”(编号2022ZDZK23)、重庆市高等教育学会2023-2024年度高等教育科学研究课题“数字化时代加强大学生国家安全教育的实践路径研究”(编号cqgj23019C)的研究成果。

生成式人工智能(Generative Artificial Intelligence，GAI)，是一种专注于生成或创建新内容的人工智能，它利用现有的文本、音频或图像等大型数据集进行机器学习，然后生成全新内容。[①]自2014年现世起，生成式人工智能呈现出指数级爆发态势，已落地文本、图片、音频和视频等多项应用。与以往决策式人工智能不同，生成式人工智能聚焦于创作新内容，它可基于语言模型(Language Model)、图像模型(Image Model)，利用深度学习技术(Deep Learning)，分析并学习数据库，自动生成新的文本、图片、音频和视频等内容。随着大数据、深度学习、云计算等技术与智能算法深度融合，生成式人工智能开始广泛应用于市场各行业、各场景，催生出许多改变行业认知和现实的创新应用，对世界政治经济、文化传播和人类生活皆产生了极其深刻的影响。全球知名IT研究与顾问咨询公司高德纳(Gartner)认为，生成式人工智能是“最具影响力和快速发展的技术之一，可以带来生产力革命”；[②]基辛格等战略家指出，“没有哪个大国可以忽视人工智能的安全维度”。[③]对我国而言，发展新一代人工智能是关系我国核心竞争力的战略问题，也是必须紧紧抓住的战略制高点。[④]

伴随着互联网日益成为意识形态斗争的主阵地、主战场、最前沿，[⑤]显然“掌握网络意识形态主导权，就是守护国家的主权和政权”。[⑥]在当前网络空间意识形态斗争错综复杂的形势下，作为网络空间传播思想文化的重要载体，生成式人工智能的意识形态安全风险把控仍存在漏洞，诸如偏好性提取学习基础数据信息、通过算法推荐制造信息茧房(Information Cocoons)，对我国意识形态进行“去中心化”

(Decentralization),进而挤压我国主流意识形态的生存空间,形成并维护意识形态霸权。对此,本文直面数据挖掘、深度学习和算法推荐的技术进步,追寻算法中保持意识形态安全的适度张力,强调科学技术和价值观之平衡,并基于此种立场阐释生成式人工智能意识形态安全风险的生成逻辑及表现样态,进而在总体国家安全观引领下提出应对路径。

一、生成式人工智能意识形态安全风险的生成逻辑

正如有学者指出,新技术的涌现引发新一轮科技产业革命,这将深刻改变社会生产、生活方式和全球分工体系,并冲击国际政治议程。[⑦]生成式人工智能在技术属性上不单承载着生产力功能,亦承载着意识形态功能。它依托海量数据库,依赖深度学习技术,通过算法推荐技术多维广泛应用于社会生产生活的各方面,对用户意识形态发挥着"无意识"的调节与渗透作用,并通过理论、技术和实践三重逻辑催生意识形态安全风险。

(一)理论逻辑:生成式人工智能的科技属性天然影响意识形态

意识形态(Ideology)一词最早来自法国哲学家德斯蒂·德·特拉西,他试图建立一门名为"意识形态"的新兴学科,以"把握人性,尽可能地减少烦恼",并"根据人的需要和意愿来安排社会和政治秩序",[⑧]但这种"观念科学"(Idea-logy)[⑨]并未建立起来。现代意义上的意识形态是指"一组相对稳定的价值观念"(本质是关于事物真伪善恶且可以支配人们思想与行为的原则或信念),[⑩]以政治、法律、道德、哲学、艺术、宗教等具体社会意识形式表现出来,代表某一阶级或集团的根本利益并为其服务的系统化思想体系。马克思将意识形态理解为在一定历史时期内,个人或群体对世界及社会所持有的各种见解和观点的总和。申言之,意识形态是其所代表的阶级或集团的情感、表象和观念的总和,集中体现了特定阶级或集团的利益和政治要求。因此,它被视为阶级利益的理论化,影响并制约着人类行为和活动。马克思和恩格斯强调意识形态的上层建筑属性并认为经济基础决定上层建筑,即"物质生活的生产方式制约着整个社会生活、政治生活和精神生活的过程"。[⑪]恩格斯还指出:"每一时代的社会经济结构形成现实基础,每一个历史时期的由法的设施和政治设施以及宗教的、哲学的和其他的观念形式所构成的全部上层建筑,归根到底都应由这个基础来说明。"[⑫]

科学技术作为经济基础的重要组成部分,深刻影响并决定着意识形态体系。后者的产生和演进依托于人类生产生活的思维方式和行为模式,存在于人类生产生活的各层面,是对人类集体经验的总结与加工。从人类现代化历史经验来看,科技革命必然会带来生产、生活方式的巨大变革,推动社会治理方式的嬗变。20世纪以来,科学技术取得的重大成果迅速转化为科技产品,深入到社会生产生活方方面面,全面影响并改变了人类的思维方式和行为模式,深刻重塑了大众的价值观和社会意识形态体系。如同蒸汽机和电力一样,生成式人工智能正作为新时代的重要推动力,给当下世界生产内容的方式、人们汲取知识的方式和思维模式带来巨大变动。

德国哲学家尤尔根·哈贝马斯曾将科学技术看作一种隐形的意识形态,认为"技术统治的意识同以往的一切意识形态相比较,'意识形态性较少',因为它没有那种看不见的迷惑人的力量",因此"比之旧式的意识形态更加难以抗拒,范围更为广泛"。[⑬]生成式人工智能在一定程度上简化了信息技术应用流程,降低了信息技术使用门槛。而且,随着生成式人工智能向高度智能化方向发展,人类对其依赖性将不断加深,这会潜移默化地影响着人的思维观念、价值取向、伦理道德,悄然无声地改变着部分民众的价值观,进而冲击主流意识形态。生成式人工智能在意识形态层面的嵌入逻辑就是以生成内容为基础,经由大数据收集检索和对算法推送的系统整合,以内容生产者的角色在意识形态领域实现认知、思考、应用和观念等多方面价值观重构。[⑭]当生成式人工智能技术应用于生产生活时,隐藏在社会生活中的意识形态也从"无形"变得"具体""有形"。[⑮]海量数据的积累、庞大算力的提高和精确算法的迭代助推

生成式人工智能成为当下维护意识形态安全最具争议性的技术之一。实际上，ChatGPT成为人类历史上增长最快的消费者应用程序[16]足以说明其强大的功能和变革性意义。

(二)技术逻辑：生成式人工智能的技术特点易受价值偏好影响

生成式人工智能依托自动生成内容的方式，在生成式AI模型、训练数据(Training Data)基础上，自动生成文本、图片、视频、音频、代码等多样化内容。以ChatGPT为例，GPT译为生成式预训练模型(Generative Pretraining Transformer)，该模型被用于分析和预测语言，提供类似机器翻译、自动文摘和快速问答的自然语言处理服务。其惊人之处是可在文本中自动学习概念性内容，即根据上下文记忆并在短时间内自动预测、直接输出下一段相关内容。GPT无需人工设计特定的自然语言处理系统，可根据所“投喂”的文本信息自动生成语法正确、内容连贯、相关度高的文本。[17]在经历了GPT-1、GPT-2、GPT-3和GPT-3.5四个阶段发展后，OpenAI公司2023年3月正式发布了ChatGPT迭代产品GPT-4，并在近期升级到GPT-4 Turbo系列。4.0版的发布标志着生成式人工智能正式进入“强人工智能”时代，[18]目前的ChatGPT已经扩充到2023年的最新数据库，最高支持用户一次性上传10万字的指令信息，可根据用户指令快速生成文本、图片、视频、音频、代码等内容，甚至推出了语音交互功能(Voice User Interface)。如此多样化的内容与技术飞跃密切相关，它所依托的是具有海量文字、图片、视频等内容的数据库，为预训练模型提供足够丰富和无限可能的资料来源；所依赖的基础模型是一种新型机器学习技术(Machine Learning)，可助生成式人工智能分析大量自然语言数据(Natural Language)；所背靠的是大型神经网络(Artificial Neural Networks)，通过在已有文本或数据库中找到有关自然语言的规律来学习。

每个单独的生成式人工智能数据模型都需海量数据库提供生成内容的“原料”。[19]以ChatGPT为例，其数据模型有百万亿个参数，预训练数据集包含百亿词汇，规模庞大至45TB空间，数据类型来源于截止到2023年的最新网页、社交传媒、新闻报道等各类数据信息并在源源不断增加。尽管生成式人工智能背靠如此大规模的数据库，但它在学习时并非全盘吸收，而是根据所设定的学习模型偏好性地提取基础数据信息。即生成式人工智能要学习何种数据是由预训练模型设定者决定的，设定者的主观意志与价值选择很大程度决定了生成式人工智能的意识形态。在诞生之初，模型设定者“投喂”给生成式人工智能的语料库主要产生于西方意识形态和价值观体系，基于中华文化和世界其他国家文化的语料库内容甚少，具有严重的意识形态导向。当生成式人工智能对该数据库加以学习后，很容易生成含有意识形态偏见的文本。此外，不同于百度(Baidu)、谷歌(Google)、必应(Bing)等搜索引擎，与生成式人工智能对话是一个动态的交互过程，提问者的每一个问题都能推动机器学习。当它的作答获得第一位提问者正向反馈后，该答案就会被不断强化，反复出现在相关问题的回答中。反之，如果其根据数据库作答并未收到提问者的肯定，那么这一答案将不会被再次提起，取而代之的是提问者更想看到的其他答案。可见，语料库本身就具有显著的意识形态性。加之生成式人工智能不断学习使用这些具有明显意识形态问题的语料，而当提问者也带着本身意识形态和价值观与生成式人工智能交流时，后者不可避免地会受到语料库意识形态的影响。[20]

生成式人工智能输出符合提问者预期的回答与其所依托的学习模型密切相关。Transformer模型是可以让人工智能进行自主深度学习的核心架构，其核心技术是“利用人类反馈强化学习”(Reinforcement Learning from Human Feedback，RLHF)的训练方式。[21]深度学习技术源于人脑生物神经网络机制，能够模仿人脑自动对数据进行特征提取、识别、决策和生成。在生成式人工智能训练时，第一步通过人工标注的方式生成微调模型，标注团队对现有数据样本进行标注，形成“提示词—正确”的数据反馈；第二步训练用于评价答复满意度的奖励模型，对生成式

人工智能给出的多个答复进行排序，其中隐含了人类对模型输出效果的预期，为模型的答复提供评价标准；第三步是通过生成式对抗网络不断锻炼、提升生成式人工智能的回答质量。按照上述算法进行学习的生成式人工智能通过循环不断地强化学习，直至彻底自主掌握人类的偏好，最后变成提问者满意的模型。而在这一过程中，人类偏好答案的标注和创建以及评价标准，均由生成式人工智能的控制主体和设计者决定。可以说，这一算法设计生成人工智能吸引用户使用的最大亮点，同时又不可避免地蕴含了控制主体的政治倾向和偏见。

(三)实践逻辑：生成式人工智能应用实践中价值取向调节困难

面对OpenAI公司在市场投入ChatGPT后引起的巨大浪潮，国内外各大科技巨头纷纷投身生成式人工智能的蓝海。在国内，阿里巴巴达摩院正研发测试ChatGPT类似产品，阿里巴巴还发布了“通义千问”大型语言模型并展现出强大的市场野心；国内领先的AI科技公司百度发布“文心一言”，并逐步将其与百度搜索引擎结合起来；互联网大厂360也积极布局生成式人工智能。[22]国际上，互联网大厂均加速了在该领域的科技布局。近日，谷歌公司正式发布迄今功能最强大、最通用的多模态人工智能大模型——Gemini(中文称“双子座”)，并宣布自2023年12月7日起将使用GeminiPro的微调版本进行更高级的推理、规划和理解，以英语版本在170多个国家和地区提供服务(不包括中国大陆地区)。

目前，生成式人工智能已强势多维介入生产生活实践、全景式嵌入应用场景并建立起多元化综合输出渠道。得益于数字孪生技术(Digital Twin)的发展，生成式人工智能赛道日益广阔、细化，相关产品种类和应用场景愈发丰富。一方面，产品服务种类愈发多样。从最初的文本扩展到声音、图像、视频等领域，再到目前独立设计核心软件，生成式人工智能不但满足了消费型内容亟待扩充的需求，也快速产出了多样化的内容形态。2021年，高德纳公司就曾预测，至2023年，将有20%的内容由生成式人工智能创建；至2025年，生成式人工智能产生的数据将占所有数据的10%(2021年不足1%)。[23]另一方面，应用领域愈发多元。在实体领域，它可为制造业、建筑业等自动生成设计图纸；在数字媒体行业，可自动生成高端游戏、新闻资讯、宣传图片、智能客服等；在教育行业，可帮助论文写作、生成学习材料、提供虚拟语言导师、协助完成作业等；在金融行业，可自动预测股市波动，为股民、企业规划最好的投资方案选择……生成式人工智能的涉足领域还在不断扩展，已然在人类生产生活的多领域“攻城略地”，下沉为一种人皆可用的技术工具。

生成式人工智能对意识形态主体行为的影响通常不在于以理性逻辑的力量说服人，而是隐匿地介入实践活动，从而无意识调节主体的思想和行为。[24]不同于虚拟现实技术(Virtual Reality)“多对一”的批量交互模式，生成式人工智能并不利用泛化的技术对主客体进行感知包裹，而是以人的主观能动性发挥为基础，通过“一对一”输入输出的方式增强人的智能体验。[25]正是这种信息生成和分发模式决定了它与使用者之间建立起了“私人订制”的数字交流形态，即生成式人工智能根据使用者的历史问题和偏好，为使用者提供个性化产品和服务，进而提升使用者正向反馈度。[26]在对话过程中，生成式人工智能的单一对话框场景满足了使用者的私密性需求，大大提升了使用者对人工智能机器的信赖度；同时，其对使用者文字进行情感和价值观立场分析以迎合用户需求，令回复内容紧贴用户立场。因此，生成式人工智能最终输出的答案一般与使用者的预期相符，在即时正向反馈中将主客体拉入学习模型和思想行为的交流过程中，将经过处理和筛选的特定内容与人们的意识形态相连接，以其内在思维范式和调控原则介入人的行为实践，进而使人们不知不觉“无意识”接受生成式人工智能的潜在规训。[27]

二、生成式人工智能意识形态安全风险的典型样态

自GPT-3发布以来，生成式人工智能始终是数字经济生态中的加点，其“破坏性创新”[28]的属性深刻

影响着人工智能产业的发展方向。该产业创新者和深耕者一方面惊叹于大模型、大数据、大算力和大算法支撑下的生成式人工智能在自主深度学习、内容生成和算法推荐等方面呈现出的丰富功能和强大性能，不断加大投资规模，深入推进生成式人工智能技术演进研究；另一方面，也针对相关产品在可靠性、完备性、客观性等方面带来的安全风险和伦理挑战展开了深层次思考。[29]当前，生成式人工智能意识形态安全风险在以上风险生成逻辑下，呈现出四种典型样态。

(一)信息失真：开辟信息空间输出虚假信息

首先，易于生成恶意信息大兴“舆论战”。生成式人工智能输出内容完全取决于信息源和训练数据库。如果信息源和训练数据代表了社会上普遍存在的价值偏见，便会造成信息的深度伪造，致使生成式人工智能成为虚假信息的制造者和传播者。以新闻报道为例，新华社早在2017年就推出我国首个能够自主写稿的AI平台“媒体大脑”。在2019年全国“两会”报道中，“媒体大脑”通过收集、分析和对比近6年政府工作报告的异同后，推出名为《一杯茶的工夫读完6年政府工作报告，AI看出了啥奥妙》的文章，[30]可见生成式人工智能已然广泛地应用到新闻生成中，并产生了深刻影响。然而，利用生成式人工智能技术进行新闻生成必然会依赖大数据，但从网络中抓取数据本身的真实性、客观性就令人存疑。人工智能难以了解信息来源、相关人物、事件缘由等深层次问题，假新闻的出现就不可避免。例如，2023年2月，有网民利用ChatGPT编写传播《杭州2023年3月1日起取消限行》虚假新闻，引发舆论负面影响；[31]2023年5月，甘肃省平凉市公安局网安大队侦破全国首例利用人工智能技术炮制虚假不实信息的案件。该案犯罪嫌疑人利用ChatGPT对近年社会热点新闻要素进行修改编辑，编造虚假信息并非法牟利。[32]此外，由于生成式人工智能具有强大的信息储存和结构文本生成能力，能收集和储存用户就个人和公共问题与其交流时产生的个体意见，这就为某些国家、组织和个人窥探和监视我国公民隐私、社会热点、政治风向、群体心态和国家机密，进而以此为素材并通过断章取义、恶意裁剪、歪曲抹黑等手段大兴舆论战以攻陷我意识形态阵地提供了极大便利。[33]

其次，提取输出内容碎片化片面化。在经典传播时代，信息传播方式大多是自上而下的精英式传播，大众掌握的信息质量和数量逊于精英群体。而随着互联网和生成式人工智能的广泛应用，信息传播速度和广度呈爆炸式增长。一方面，信息传播走向碎片化，即脱离传统的完整叙事结构截取片段呈现出来，通常表现为碎片化制作和碎片化传播，然而如此呈现的信息易导致信息输出“去逻辑化”；另一方面，现今快节奏的生活方式塑成日渐浮躁、快速阅读的学习心态，用户更易接受碎片化、便利传输的信息，反而无法接收阅读时间长、内容逻辑完整的信息，因而也很难对一件完整的事情作出客观、整体、有准度的认知和评价，往往是凭借一时的情绪牵引形成激进、偏差的观点。生成式人工智能输出内容“短平快”的特点完美契合了当下年轻用户的需求和特点，向此类用户输出能够引发更多情感宣泄的内容，而忽视对复杂问题和意识形态观点的深入探讨、忽视对事实和逻辑的深入分析，主流意识形态的完整性便在某种程度上被肢解。

最后，提供传播极端言论的隐秘新渠道。当前，网络空间意识形态斗争异常错综复杂，西方敌对势力利用网络空间散布错误、虚假舆论制造思想混乱，千方百计破坏我国意识形态安全已成常态。生成式人工智能长于根据交互用户的偏好不断强化学习，直至使回答完全契合提问者的需求，这就为少数极端分子宣传造势提供了发挥空间，甚至可能沦为以美国为首的西方国家对我国进行网络意识形态渗透干涉的隐秘路径。此外，生成式人工智能在输出端所展现出的内容生产和人机互动能力均基于输入端数据投喂和学习训练，这使其在汲取海量网络语料过程中可能为极端主义者、宗教极端主义者和暴恐分子等制造极端信息或通过人为“数据投毒”(Data Poisoning)来训练AI模型留下可乘之机，由此，生成式人工智能最终可能演变为输出有害极端言论的新

渠道。

(二)算法霸权:单向传播编织受众信息茧房

其一,弱化行为主体独立思考能力。首先,在人工通用智能(Artificial General Intelligence,AGI)模式下,生成式人工智能在面对不同语种、语境提问时,可以模拟用户发言(模拟本身就是人工智能获取新知识的一种强有力工具),[34]针对性、选择性地生成用户满意的目标言论,且生成的答案简洁有力、易于理解而令用户习惯性地认为这是正确答案。有学者就此表示:“算法既不知道我们,也不了解我们,但画像技术可以为每个用户(读者、观众)提供与其兴趣和需求相关的有针对性的信息。”[35]其次,生成式人工智能通常采用“一问一答”的方式与使用者对话,不仅通过标注序号“一、二、三”并以总结性话语结尾的方式展现富有层次、通顺连贯的答案,还能够承认错误甚至纠正和拒绝不合理请求,拟人化的机器形象进一步增强了话语权威。最后,生成式人工智能只提供一种答案且不提供信息来源,阻断了其他观点影响行为主体感知的可能,以避免行为主体产生批判性观点。如果人类不具备相关知识和鉴别力,就会被恶意行为者所操控,降低对机器的警惕性和防御性,彻底沦为技术的傀儡。

其二,固化行为主体思维认知范畴。从传统媒体时代的媒介霸权、文化霸权到如今的算法霸权和人工智能霸权,愈发精准、智能的科技也在慢慢催生“数字利维坦”,[36]其表现之一便是“信息茧房”的形成。[37]在数据标注、深度学习和算法推荐技术的助力下,生成式人工智能可为用户设定偏好标签,持续、片面地向某一用户推送或者回复固定意识形态的内容。加之,机器意识依赖于目前我们仍未完全把握的一些变量:公众对人造意识的需求、对有知觉的机器安全性的担忧、基于人工智能的神经假体与神经增强是否成功,甚至还有人工智能设计者的一时兴起。[38]这不仅阻碍用户接收真实、客观、多样的信息,令其长期沉溺于所谓“用户主动选择”的海量信息中而缺乏与外界不同观点和意识形态的沟通,不仅加剧了互联网世界的用户圈层分化,导致极端、偏执观点的接收群体愈来愈深陷于这一观点,而且还带来自身意识形态的偏见和极化。这一现象持续发展将会导致不同群体形成各自偏好的不同意识形态和价值观,像“信息孤岛”一样环伺在主流意识形态周边并不断扩张,围猎和侵蚀主流意识形态,逐渐使其失去领导力,陷入孤岛困局。

其三,退化行为主体信息处理能力。尽管网络信息库不断丰富,但信息却越来越碎片化,人脑对信息收集、加工、吸收的成本和难度日益加大。区别于传统人工智能技术带来的信息影响力,生成式人工智能依托大模型、大数据、大算力和大算法仿照甚至超越人脑的信息处理,在短时间内,根据指令接收信息、回溯经验、诉诸理论、提出假设、调研情况、检验假设、推导结论,以此代替人类思考。人类的思考过程逐渐简化为“输入指令—得到结论—复制粘贴”,对生成式人工智能的依赖度大大提升。长此以往,人们会越来越依赖利用人工智能快速获得看似简单、清晰且权威的结论,而提问者思维惰性势必增大、信息处理能力将严重蜕化,掣肘我国掌握生成式人工智能使用主动权。

其四,加剧意识形态领域监管难度。作为一种用户无限制使用的大型对话性语言模型,生成式人工智能就像一部意识形态言论输出机器,源源不断生产真假难辨的信息。[39]这些信息或包含虚假信息,或包含偏见看法、错误思想、仇恨言论等,甚至某些国外生成式人工智能对我国社会主义制度和主流意识形态存在偏见、敌视和攻击倾向,带来主流意识形态的全域、动态、复杂风险,导致监管工作的难度大大提升。对于国外生成式人工智能,虽然我国尚未引入ChatGPT等应用,但大量出国留学人员和赴外旅游用户都可以使用ChatGPT,而通过备案审查和“翻墙”使用的人也不在少数。由于基础设施设在域外,我国监管法律和制度无法对其乃至其控制主体进行监管,政府的监管很难做到无死角。对于国内生成式人工智能,政府资源的有限性也决定了监管的失效性。生成式人工智能所生成的数以亿计且真假难辨的信息高度拟人化,无不给政府在信息监控、

审查和筛选方面带来极大压力。

(三)资本僭越:预设立场影响受众政治态度

第一,人工智能并非意识形态的制造者,而是偏见的继承者。意识形态在内容上具有“阶级性”,[39]国外学界的研究表明,以ChatGPT为代表的部分生成式人工智能带有鲜明的美西方资本主义思维和明显的左翼自由主义倾向。例如,学者大卫·罗萨多对ChatGPT进行了15种不同的政治倾向测试,结果一致表明其回答表现为“左倾”观点偏好(如反对死刑、自由市场,支持政府补贴、向富人征税、堕胎和为拒绝工作的人提供福利等),而非其所声称的立场中立客观。[41]学者约亨·哈特曼在对政治选举中的ChatGPT意识形态进行研究后,也认为必须注意ChatGPT的左翼自由主义政治倾向。[42]而ChatGPT具有鲜明的美西方资本主义思维特色,根本原因在于它是由美国研发、主要基于美西方国家大数据和人工训练生成的。在美西方价值观长期浸染下,ChatGPT生成内容必然具有鲜明的政治立场,对我国带有天然“国家偏见”。类似的美国研发的生成式人工智能产品都大同小异。[43]因此,以ChatGPT为代表的部分生成式人工智能存在左翼自由主义政治倾向和政治偏见,同社会主义主流价值观的政治立场和价值取向根本对立和冲突。在当前认知领域仍被美西方主导的情况下,部分生成式人工智能将会成为贯彻西方意识形态、直接影响用户的高效宣传工具,必须予以高度警惕。

第二,强化资本主义意识形态渗透力。大众对社会的了解和认识通常源于统治阶级的宣传和引导,进而形成长期潜移默化的思维与行为模式,这也是美西方通过生成式人工智能影响受众意识形态的最主要方式。在中美意识形态博弈加剧的背景下,美国依托科技霸权,放任部分生成式人工智能利用“算法黑箱”[44]将具有反华导向的内容产品投喂受众,毒化国际受众对华认知。而受资本主义意识形态浸染的生成式人工智能不仅可针对我国和国际社会受众心理弱点提供隐形渗透我国意识形态安全阵地的内容产品,助力美西方构建高维度、全方位、多视角对华意识形态渗透和攻击舆论战场,剥夺、涣散、冲垮党和政府在国内国际意识形态领域的主导权、主动权,影响我国乃至国际受众的政治判断与政治选择,加速对华和平演变进程,导致“颜色革命”风险隐患陡增。

第三,助推强化资本主义价值观联盟。在网络信息时代,资本和权力的介入严重干扰了网络空间的意识形态和价值观。国外大部分生成式人工智能的数据训练需巨额资金投入,仅靠企业自身大都无力承担,背后必然有国家资本支持。在人工智能时代,国家资本运作的目的转变为通过“技术封锁”[45]“数据垄断”[46]等手段散播和传递有利于自身阶层和利益的意识形态内容,强化美西方价值观的支配地位。生成式人工智能以技术工具的方式被有效嵌入政治体系、制度框架等上层建筑的子系统中,或干预社会政治文化体系,或引发社会问题与社会危机,成为资本主义国家资本谋取剩余价值、进而维持资本主义制度体系运转的技术工具,其中“算法歧视”[47]与“算法黑箱”就是最明显的表现形式。[48]进入数字时代,掌握这两种技术的主要是发达资本主义国家,其驱动生成式人工智能利用“算法歧视”和“算法黑箱”精准定位并投喂受众,利用技术霸权加深发展中国家对其技术依赖,散播包含大量偏见、歧视和虚假信息的西方政治倾向和价值观念,强化资本主义价值观联盟,从而在全球范围内建立起资本主义意识形态体系。

(四)去中心化:压缩主流意识形态生存空间

作为一种思想体系,意识形态的目标是社会成员的广泛认同。“主流”的涵义主要有二:一是在深广度上对社会公众产生强烈的影响;二是常依靠政治权威发挥作用、维持影响力。在此意义上,我国主流意识形态是指社会主义意识形态,主要包括执政党指导思想、政治信仰、民族精神和道德文化等内容,其必然要为上层建筑和经济基础服务。[49]而非主流意识形态是指与主流意识形态相对,存在于不同阶级、阶层和利益群体中不占主导地位的价值观念和社会思潮。在生成式人工智能飞速发展的今天,意

识形态领域日益失去固定的中心点，任何阶级、阶层或利益群体都可能成为某种意识形态的中心点，我国主流意识形态也日渐被非主流意识形态“去中心化”，生存及其影响空间受到严重挤压。

一是压低主流意识形态话语深度。意识形态话语表达具有相当的复杂性，其作用在于塑造意识形态，从而巩固其在精神领域的支配地位。[50]以往我国主流话语在形成、传播和发挥作用的过程中，均由政治立场鲜明、马克思主义理论知识储备扎实的理论工作者把关，马克思主义意识形态根基深厚。随着生成式人工智能嵌入社会并参与某些重大概念和事件的解释，凭借自身的普及性和便利性引导公众习惯于第一时间寻求人工智能解释路径。然而，其所展现的多元话语选择、个性话语生产和精准信息推荐功能，本质上就是对已有信息的重新处理，无法产出富有深度、高度和创造性的政治话语。此外，这种不筛即用的方式势必带来意识形态混乱风险，将严重削弱官方阐释的说服力、渗透力和思想深度。

二是削弱主流意识形态话语权威。意识形态“不是纯粹的理论说教，而有明确的价值导向，具有排他性”。[51]这一特征决定了主流意识形态对非主流意识形态的排斥和控制行为，从而为自身所代表的阶级或集团提供稳定的政治认同及社会控制力量。主流意识形态话语权代表主流意识形态引导社会舆论、宣扬社会思想的权威性，是国家软实力的重要组成部分，直接关乎主流意识形态建设。美西方国家借由生成式人工智能散播自由主义、保守主义、新自由主义、新保守主义、民族主义、民粹主义等资本主义意识形态，在生成言论、音频和视频过程中宣扬“中国威胁论”“中国崩溃论”“普世价值”等西方话语，意图通过削弱我国公民的爱国精神和家国情怀来从内部击溃中华民族的意识形态防线。

三是引发话语体系发展不平衡风险。在生成式人工智能空间，可以说谁制定了规则，谁就掌握了话语权。以美国为首的西方发达资本主义国家凭借强大的资金技术优势，占据着网络空间的霸主地位优势开发生成式人工智能，而包括我国在内的发展中国家由于起步晚、市场占有率不足和技术资金相对劣势，很难对国外生成式人工智能形成市场制约。同时，部分国外生成式人工智能的话语运行规则逻辑特色鲜明。譬如，在使用以ChatGPT为代表的国外生成式人工智能时，用中文提问得到的答案信息量比英文少得多。其原因在于ChatGPT等生成式人工智能训练是以西方知识体系为基础，生成文本的基本格式也以英文为主，这无疑强化了英文及内嵌的话语优势。而中文语料在国外生成式人工智能学习库中占比很低，无法成为回答的主要内容，这会严重影响中文话语的有效性和广泛性，使本就在世界话语体系中不占优的中文话语权持续减弱。[52]

三、生成式人工智能意识形态安全风险的应对路径

对意识形态之于国家安全稳定的关键作用，马克思早已指出：“如果从观念上来考察，那么一定的意识形态的解体足以使整个时代覆灭。”[53]习近平总书记也强调：“新形势下，意识形态领域斗争复杂尖锐。历史和现实都警示我们，思想舆论阵地一旦被突破，其他防线就很难守得住。”[54]当前，世界正在经历百年未有之大变局，我国正处在民族复兴关键期，必须在道路、方向、立场等重大原则性问题上守好底线红线，牢牢把握战略主动权。面对复杂严峻的生成式人工智能意识形态安全风险，更须坚持以总体国家安全观为统领，在价值引领、中心恢复、以疏代堵、全面治理等方面整体着力、系统推进，筑牢维护国家安全的防线。

（一）在价值引领中重塑人工智能技术与价值平衡观

著名技术哲学家卡尔·米切姆提出：“未来的技术哲学研究，将更注重对技术应用过程的意义与价值的考察”。[55]作为一种新兴智能内容生产技术，生成式人工智能遵循一定的科技价值观是必然的。欧盟委员会早在2018年即提出人工智能价值观，在2020年发布的《人工智能白皮书：通往卓越和信任的欧洲之路》中再次强调：“考虑到人工智能可能对我们社会产生的重大影响以及建立信任之需要，欧洲

的人工智能必须以欧洲价值观和人类尊严、隐私保护等基本权利为基础。”[56]生成式人工智能依赖强大功能对人的思维指向、价值观念、道德取向等均发挥着重要作用——制造信息茧房,利用使用者个人的价值标准固化社会公众价值取向,成为意识形态渗透攻击的新武器。因此,在该领域树立共识价值观、以价值引领技术向善,对实现生成式人工智能内容创造、技术进步和社会主义核心价值观二者间的平衡至关重要。

综合而言,需要确立生成式人工智能核心价值导向,树立以人为本的技术观,向生成式人工智能注入价值理性来驾驭工具理性;同时强化价值理性以规避意识形态撕裂风险,避免科技发展和人文价值的零和博弈。具体要坚持在“守正”基础上进行创新,重点把握好三方面工作。其一,政府、生成式人工智能制造者、应用者都需明确社会主义意识形态的核心价值所在,包括强调人民利益和社会公平,着重提高技术普惠性,确保生成式人工智能技术发展惠及社会大众。其二,强调人的主体地位和尊重个体权利,捍卫人的尊严和权利,确保技术发展不侵犯人权,树立“技术为人所用”的价值观念,防止技术取代或剥夺人的自主选择权和自由发展权。其三,将社会主义意识形态融入算法推荐核心技术中,在其中输入社会主义意识形态的“血液”,提高算法推荐中社会主义意识形态输出优先级;通过明确核心价值、树立生成式人工智能创新和应用价值导向,为社会主义意识形态多元化优质内容的创作和传播搭建合适的道德框架。

(二)以恢复中心为目标坚定主流意识形态一元中心指引

首先,坚持马克思主义的一元指导地位。在生成式人工智能领域,主流意识形态出现被“去中心化”困境的根本原因在于,马克思主义理论在实践中的应用缺失。去中心化的本质就是意识形态多元化和非固定化,这就导致生成式人工智能输出内容时形成马克思主义理论的“真空地带”。中国共产党成立伊始便把马克思主义作为意识形态建设的思想旗帜与政治灵魂,马克思主义是我们立党立国的根本指导思想。背离或放弃马克思主义,我们党就会失去灵魂、迷失方向。[57]而意识形态的地位最终取决于其自身的影响力。[58]由此,必须巩固马克思主义在生成式人工智能意识形态安全风险治理工作中的一元指导地位,创新马克思主义思想政治理论在该领域的话语表达,实现意识形态话语的通俗化、大众化转换,[59]增强马克思主义在智能科技领域的说服力、吸引力,减少人民群众对宏大政治叙事话语的疏离感、距离感,真正使马克思主义话语体系贴近社会公众生活,强化社会主义意识形态话语权,压倒“资本至上”的意识形态态势。

其次,坚定中国共产党的一元领导地位。党的二十大报告指出:“意识形态工作是为国家立心、为民族立魂的工作。牢牢掌握党对意识形态工作领导权,全面落实意识形态工作责任制,巩固壮大奋进新时代的主流思想舆论。”[60]历史经验表明,主流意识形态若要占据我国生成式人工智能市场主导地位,就必须坚持党在该领域意识形态安全风险治理工作中的绝对领导,这也是“坚持党对国家安全工作的绝对领导”这一根本原则的必然要求。[61]其一,坚持党性原则,合理利用生成式人工智能产品,将社会主义意识形态植入生成式人工智能数据训练库,使其成为党的“新喉舌”,保证产出内容充分体现党的意志、宣传党的主张。其二,将“党管媒体、党管数据”原则贯彻到治理工作全过程各方面,监管生成式人工智能模型设计者和提供者,构建党领导下社会各方力量协同参与的意识形态责任制,完善社会主义意识形态风险评估机制,保证我们党在智能革命中的主动、优势地位,掌握意识形态交锋主动权。

最后,提升民众人文与科技素养,筑牢意识形态安全防线。生成式人工智能带来的意识形态安全风险牵动着国家和个人的当下和未来,必须筑牢人民群众的意识形态安全防线。一是要提升人民群众网络与媒体素养教育。随着生成式人工智能下沉为社会公众均可使用的社交工具,部分“三低”(即低年龄、低收入、低文化程度)网民成为其输出意识形态

的重点对象。各级政府应加强网络意识形态工作，引导网民的社会主义核心价值观体系建设，通过讲堂、道旗、短视频等多种形式将网络媒介素养教育覆盖全民，拓宽其知识结构，打破“算法牢笼”。同时，教育民众不断学习拓展认知范畴、提升独立思考能力、增强明辨是非能力，坚守意识形态安全底线，不问、不答、不交流有损国家意识形态安全的问题。二是应提升意识形态工作者智能素养。当务之急是成立一支专门化、常态化、制度化的生成式人工智能意识形态安全风险治理工作队伍，完善智能素养培训机制，利用学术讲坛、讲座报告、课程教育和典型案例分析等途径，使意识形态工作者掌握生成式人工智能的传播特性、输出内容和应对策略，增强其技术认知和风险预判预防能力。

(三)通过以疏代堵构建主流意识形态具象传播框架

不可否认，生成式人工智能为意识形态领域带来了诸多边界性问题，但它的成功孕育于人工智能时代信息技术的发展，已成为顺应时代发展的必然趋势，一纸禁令无法阻挡其广泛应用。因此，面对生成式人工智能宜“疏”不宜“堵”。实际上，人工智能技术是一把“双刃剑”，[9]其发展进步在带来危机的同时也带来了机遇，恰当管理可更好地为我所用。主流意识形态的具象传播框架可以从传播方式具象化和传播内容具象化两方面构建，推动传播符号体系向抽象化、纵深化方向发展。这就要求我们加速科技创新实现“弯道超车”，构建“安全、可靠、可控”的生成式人工智能，[10]同时结合国情创新主流意识形态内容表达机制，突破意识形态国际传播壁垒，推动社会主义国家话语通过先进技术面向世界传播全球。

一是构建自主可控的生成式人工智能。其一，必须坚持党的领导。历史经验表明，党的领导是国家“集中力量办大事”的坚强保证。统筹谋划和整体布局自主可控的生成式人工智能建设，必须围绕党中央面向世界科技前沿、经济主战场、国家重大需求、人民权利作出的科技创新部署，以《新一代人工智能发展规划》《中华人民共和国国民经济和社会发展第十四个五年规划和2035年远景目标纲要》《生成式人工智能服务管理暂行办法》等文件为政策遵循和制度保障。其二，实现科技自立自强。若要突破目前发达国家打造的“小院高墙”，就必须跳出“算力困局”，在大模型、数据库和算法上设计出独立于美西方的生成式人工智能系统。现今，百度“文心一言”系统和复旦大学的MOSS系统正在填补此方面空白。而构建自主可控生成式人工智能的关键是赋予国内生成式人工智能丰富的数据和语料库系统，打造中国式应用场景，守牢意识形态阵地，实现用自己的话语、语料库和应用场景来抵御国外生成式人工智能带来的意识形态冲击，最终形成中国式人工智能新生态，广泛吸引世界人民参与其中，共同推动构建人类命运共同体。

二是加速科技创新，赶超发展赛道。其一，加大政府投入，释放政策红利。设立人工智能研究基地，打造世界级人工智能研究中心；借助政策红利激励人工智能研究机构通力合作、共享成果、协同发展，构建全球人工智能研究和创新的前沿阵地。发挥政府的引导作用，推动生成式人工智能研发、部署、应用产业一条龙的合理布局，避免同质化重复建设，实现资源利用最大化、最优化。其二，重视人才培育，实现算法优化。跨越技术鸿沟最关键的是培养生成式人工智能领域顶尖科技人才。这需要政府与头部企业全方位合作，组建相关高新技术攻坚科技团队，延长和健全人才培养链，推动生成式人工智能人才接受一体化、贯通式基础教育、高等教育和继续教育；同时完善人才审查制度，确保培养一支高政治性、高素质、高水平的人才团队。研发团队应研究算法关键技术的运行模式，加强算法的可解释性、可溯源性、可追责性，将生成式人工智能嵌入意识形态安全建设各环节，研发意识形态安全风险应对软件。其三，企业筑牢安全高墙，破除算法黑箱。企业是生成式人工智能产品的主要输出者，突破算法黑箱、算法歧视等技术必须企业助力。企业在构建生成式人工智能应用时，需要为其提供安全和有价值的数据输入，借助智能识别、算法推荐、智能引擎充分挖掘

社会主义意识形态的正向内容，从数据源头进行治理；在训练过程中监控大模型安全运行，避免算法黑箱、算法歧视、诱导沉迷等技术植入，以技术手段应对潜藏危机，以技术力量化解技术危机。

三是创新主流意识形态内容表达机制。习近平总书记在全国宣传思想工作会议上强调："宣传思想阵地，我们不去占领，人家就会去占领。"[64]首先，要守好主流意识形态阵地，发挥主流媒体舆论引导作用。党的新闻宣传战线要主动作为，率先占领生成式人工智能领域话语传播空间，将社会主义意识形态融入数据库和算法推荐；根据国情和社会主义特色表现形式构建数据网，确保生成式人工智能的应用场景符合主流意识形态建设方向。其次，要提高意识形态工作者素质，推动内容供给侧改革。组建官方优质创作团队，鼓励民间优秀创作者，确保具备不同背景和经验的创作人才积极参与主流意识形态内容创作表达工作。依托中华优秀传统文化、民族精神、红色文化，挖掘教育资源，利用其他人工智能数字技术讲好中国故事，推动社会主义核心价值观掌握意识形态交锋主动权，为公众防范生成式人工智能价值观陷阱提供正确的价值观指引。同时，丰富主流意识形态元素的生活连接场景，增强生成式人工智能产品的地域性表达；在人民群众熟悉的场景中应用生成式人工智能，构建多元主题的主流意识形态传播场景，以此获得民众的拥护与认可，拓展主流意识形态的生存空间。最后，要及时回应民众的舆论关切。人民群众对社会热点事件的态度直接影响对社会主义意识形态的信任度，其态度不免投射至民众与生成式人工智能的交流互动中。对于网络上民众集中反映的热点问题，官方媒体要及时回应、消除误解、避免扩大；对确需改进完善的地方要第一时间调查、处理，及时公布调查结果，回应群众舆论关切。切忌通过网信部门向企业或平台施压，意欲以行政手段抑制舆论生成或扩大，这样不但不会收获正向结果，反易激发民众逆反心理，导致群众对政府不信任，生成两者间沟通壁垒，给境外势力侵入以可乘之机。

(四)推动全面治理，丰富生成式人工智能范域监管机制

第一，要落实大模型、数据库和算法备案机制。其一，构建安全评估和备案管理制度。可将需要进行安全评估和备案管理范围限定为"具有舆论属性或社会动员能力"的生成式人工智能产品，细化相关产品的准入、退出机制，并推行龙头企业在网信部门备案制度。此外，对来自境外的生成式人工智能产品实施严格的安全评估程序、准入标准、备案管理制度和审查标准。其二，强调行业主体自律，重视约谈工作。[65]引导企业签订行业协议，构建行业组织自律体系；以生成式人工智能产品提供者为治理枢纽，强化服务提供者内部训练数据输入端和生成内容输出端的主体责任。服务提供者发现违法内容应当及时采取相应措施整改并向主管部门报告，助推生成式人工智能规范落地运用和长久发展。政府要加强与产品提供者的联系，重视对企业的约谈工作；通过行政监督强化政府对生成式人工智能意识形态安全风险的重视，督促企业向善向好发展生成式人工智能，不做不利于民族团结、政治稳定和国家安全之事。其三，赏罚并行，树立典型。加大对生成式人工智能意识形态有害产出的处罚力度，对有违《生成式人工智能服务管理暂行办法》的行为进行明确处罚；选取生成影响国家安全和统一、民族团结以及其他扰乱经济社会秩序的违法性、歧视性的内容产品进行通报批评、关停整改或依法取缔，维护清朗的生成式人工智能使用环境。同时，积极引导树立正面典型，厚奖能够有效抵御意识形态有害内容产出、重视生产社会主义意识形态内容的企业。

第二，要加强意识形态安全风险预判能力。习近平总书记指出："我们必须把防风险摆在突出位置"，"力争不出现重大风险或在出现重大风险时扛得住、过得去"，[66]同时要求在网络意识形态斗争上"坚持管用防并举"。[67]具体而言，一方面，完善意识形态安全风险审查制度。生成式人工智能的意识形态输出具有很强的隐蔽性，要利用关键词(包括形近词、同音词等)扫描技术对其输出内容中的非主流意

识形态和错误内容进行屏蔽，阻绝使用者对一些政治敏感词汇进行提问、传播。对生成式人工智能输出的图片、音频、视频等非文字内容进行人工后台监测，识别其中的隐喻内容，屏蔽涉及政治、思想的不良信息。在这一过程中，应把握好审查尺度，切忌“一刀切”；避免侵害使用者正常言论自由，授敌把柄攻击我国言论氛围、社会制度和党的领导。另一方面，依托技术支持预判风险。经过多年部署推进，我国已经具备较为成熟的互联网治理技术。政府可借助国家防火墙(国外称GreatWall，又称“防火长城”)管理域名、检查代码、过滤内容来监控舆情，及时识别和管理生成式人工智能在国内外政治思想、经济文化、社会民生等领域造成的意识形态风险点，科学预测可能出现的意识形态危机；对错误信息和不良言论导向要早发现、早介入、早解决，防止危机爆发。

第三，要以系统思维倡导全球共同治理。生成式人工智能已在各行业展现出广阔的应用前景，既具有各国地域特色，也具有全球性特征，堪称在技术上塑造着人类命运共同体。一是加强国与国之间的人才、技术交流。当前，我国人工智能技术和人才资源尚显薄弱，迫切需要学习、吸收和借鉴发达国家和地区的先进技术；潜心打破发达国家技术壁垒，依托科技向善的文化理念，以文明交流互鉴消除技术隔阂，携手具有共同目标的国家合作防范技术霸权，为技术领域治理贡献中国智慧、中国方案。其二，推动国际主体协同共治生成式人工智能。在保证自身社会主义意识形态话语权的基础上，积极号召国际主体参与涉及人工智能、机器学习等议题的国际会议；邀请各国教育、科技、伦理、法律等领域顶尖专家学者共同论证，就监控和规避生成式人工智能领域的意识形态安全风险商讨国际解决方案；制定生成式人工智能道德准则，重塑生成式人工智能价值观，推动共商、共建、共享安全无患的生成式人工智能长久发展。

注释：

①本书编写组：《国家人工智能安全知识百问》，人民出版社，2023年，第20页。

②"Gartner Predicts the Future of AI Technologies," https://www.gartner.com/smarterwithgartner/gartner-predicts-the-future-of-ai-technologies.

③[美]亨利.基辛格、埃里克·施密特、丹尼尔·胡滕洛赫尔著，胡利平、风君译：《人工智能时代与人类未来》，中信出版社，2023年，第172页。

④中共中央宣传部、中央国家安全委员会办公室编：《总体国家安全观学习纲要》，学习出版社、人民出版社，2022年，第117页。

⑤中央网络安全和信息化委员会办公室编写：《习近平总书记关于网络强国的重要思想概论》，人民出版社，2023年，第44页。

⑥习近平：《论党的宣传思想工作》，中央文献出版社，2020年，第22页。

⑦中国现代国际关系研究院：《百年变局与国家安全》，时事出版社，2021年，第49—50页。

⑧Caverly and Matthew Mark "Ideology as Theology," Conference Papers, Southern Political Science Association, 2013, pp. 1-15.

⑨[英]大卫·麦克里兰著，孔兆政、蒋龙翔译：《意识形态》，吉林人民出版社，2005年，第7页。

⑩刘少杰：《当代中国意识形态变迁》，中央编译出版社，2012年，第5页。

⑪《马克思恩格斯选集》(第二卷)，人民出版社，2012年，第2页。

⑫《马克思恩格斯选集》(第三卷)，人民出版社，2012年，第796页。

⑬[德]尤尔根·哈贝马斯著，李黎、郭官义译：《作为“意识形态”的技术与科学》，学林出版社，1999年，第69页。

⑭代金平、覃杨杨：《ChatGPT等生成式人工智能的意识形态风险及其应对》，《重庆大学学报(社会科学版)》，2023年第5期，第101—110页。

⑮陈丽荣、吴家庆：《大数据时代党的意识形态话语权探析》，《思想理论教育导刊》，2018年第8期，第105—109页。

⑯邱恒明：《AI新世界与超级个体崛起》，《中华读书报》，2023年2月22日。

⑰Carey Swan, "Stats Incredible: ChatGPT," Acuity, Vol. 10, No. 3, 2023, pp. 57-58.

⑱“强人工智能”一词最初是由约翰·罗杰斯·希尔勒针对计算机和其他信息处理机器创造的，认为“计算机不仅是用来

研究人的思维的工具，相反，只要运行适当的程序，计算机本身就是有思维的”(See John R. Searle, "Minds, Brains, and Programs," Behavioral and Brain Sciences, Vol. 3, No. 3, 1980, pp. 417-424)。由此，强人工智能是具有独立意志、能在设计的程序范围外自主决策并采取行动的人工智能，这也是强人工智能与弱人工智能的最大不同。

⑲Joaquín Borrego-Díaz and Juan Galán-Páez, "Explainable Artificial Intelligence in Data Science: From Foundational Issues towards Socio-technical Considerations," Minds & Machines, Vol. 32, No. 3, 2022, pp. 485-531.

⑳蓝江：《生成式人工智能与人文社会科学的历史使命——从ChatGPT智能革命谈起》，《思想理论教育》，2023年第4期，第12—18页。

㉑该训练方式包括三个步骤，即初始模型的选取和初步优化、评价性反馈记录和跟进、迭代初始和评价模型的优化，其中在第二阶段“评价性反馈记录和跟进”中，深度学习技术对生成式人工智能作用最大。

㉒丁磊：《生成式人工智能：AIGC的逻辑与应用》，中信出版集团，2023年，第46—47页。

㉓丁磊：《生成式人工智能：AIGC的逻辑与应用》，第13页。

㉔Yuanyuan(Gina)Cui and Patrik van Esch, "Autonomy and Control: How Political Ideology Shapes the Use of Artificial Intelligence," Psychology & Marketing, Vol. 39, No. 6, 2022, pp. 1218-1229.

㉕王少：《ChatGPT介入思想政治教育的技术线路、安全风险及防范》，《深圳大学学报(人文社会科学版)》，2023第2期，第153—160页。

㉖方旭：《从ChatGPT看人工智能时代意识形态领域运作表征与风险防范》，《毛泽东邓小平理论研究》，2023年第5期，第35—44页。

㉗王少：《ChatGPT介入思想政治教育的技术线路、安全风险及防范》，《深圳大学学报(人文社会科学版)》，2023第2期，第153—160页。

㉘“破坏性创新”(Disruptive Innovation)亦称“破坏式创新”，启蒙于奥地利经济学家约瑟夫·熊彼特提出的“创造性毁灭”理论，旨在阐明技术革新引发的垄断和效仿行为对竞争与创新关系的影响。其后，美国学者克莱顿·克里斯坦森对相关理论进行了引申完善，提出了“破坏性创新”概念，进一步阐述了新旧技术的冲击、竞争和代替给市场规则、市场格局和社会生活带来的颠覆。

㉙钟力：《生成式人工智能带来的数据安全挑战及应对》，《中国信息安全》，2023年第7期，第83—85页。

㉚《一杯茶的工夫读完6年政府工作报告，AI看出了啥奥妙》，http://www.xinhuanet.com/politics/20191h/2019-03/05/c_1210073771.htm。

㉛《这条‘假新闻’是ChatGPT写的?！警方已介入调查》，https://www.thepaper.cn/newsDetail_forward_21958652。

㉜《涉嫌寻衅滋事罪！甘肃公安侦破首例利用AI人工智能技术炮制虚假信息案》，http://www.chinapeace.gov.cn/chinapeace/c100064/2023-05/16/content_12657510.shtml。

㉝蒲清平、向往：《生成式人工智能——ChatGPT的变革影响、风险挑战及应对策略》，《重庆大学学报(社会科学版)》，2023年第3期，第102—114页。

㉞[美]赫伯特·A.西蒙著，陈耿宣、陈桓亘译：《人工智能科学》，中国人民大学出版社，2023年，第21页。

㉟[意]埃琳娜·埃斯波西托著，翁壮壮译：《人工沟通与法：算法如何生产社会职能》，上海交通大学出版社，2023年，第61页。

㊱“利维坦”(Leviathan)原指《圣经》中上帝在第六天创造出来的海上巨兽，后经英国哲学家霍布斯引用与改造，被指代为某种由人类自己制造出来却又日益脱离人类控制的异化力量。随着大数据时代的到来，人们生活方式因技术变革而日益便捷化，但同时人们也越来越深切地体会到一种“受缚于数字”的无奈之感。这一现象背后隐藏着大数据技术滥用的新型危机，即“数字利维坦”(Digital Leviathan)。

㊲2006年，美国学者凯斯·R.桑坦斯在其著作《信息乌托邦：众人如何生产知识》(毕竞悦译，法律出版社，2008年)中首次提出“信息茧房”概念。他认为处于信息茧房中的人们只愿意接受自己选择和愉悦自己的信息。随着智能算法社会的到来，算法推荐技术愈发推动用户信息茧房的形成——个人或群体被包围在一个信息壁垒之内，自主或者不自主地把信息选择行为固定在特定种类信息范围内，进而在思想和情感方面产生对这类信息的亲近和对其他类型信息的排斥。

㊳[美]苏珊·施耐德著，方弦译：《人工的你：人工智能与心智的未来》，湖南科学技术出版社，2022年，第50—51页。

㊴Jasmine Clark, "Artificial Whiteness: Politics and Ideology in Artificial Intelligence," College & Research Libraries, Vol. 82, No. 5, 2021, pp. 775-776.

㊵陈锡喜：《意识形态：当代中国的理论和实践》，中国人民大学出版社，2018年，第10页。

㊶David Rozado, "The Political Biases of ChatGPT," Social Sciences, Vol. 12, No. 3, 2023, pp. 1-8.

㊷ Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte, "The Political Ideology of Conversational AI: Converging Evidence on ChatGPT's Pro-Environmental, Left-Libertarian Orientation," https://ssm.com/abstract=4316084.

㊸例如,美国人工智能初创公司Anthropic开发的聊天机器人Claude、CharacterAI推出构建用户自己的聊天机器人、Hugging Face打造的开源社区等。

㊹“算法黑箱”(Algorithm Black Box)是指算法设计者在设计、开发和使用生成式人工智能时,将某些特定意识形态概念或者价值观念预先嵌入编码设计中,赋予其某些特定的意识形态属性或价值观,进而使此类产品在不为人察觉的情况下影响社会意识形态体系。参见谭九生、范晓韵:《算法“黑箱”的成因、风险及其治理》,《湖南科技大学学报(社会科学版)》,2020年第6期,第92—99页。

㊺“技术封锁”(Technological Blockade)来源于贸易保护主义,是贸易中的技术优势方借助于比较优势维持超额利润的一种手段,集中体现于高科技领域。参见程仲鸣、陈宇航:《技术封锁对企业创新策略选择的影响——基于美对华“实体清单”的经验证据》,《中国科技论坛》,2023年第2期,第135—145页。

㊻“数据垄断”(Data Monopoly)是指经营者利用对数据这一关键生产要素的控制优势,实施产生排除、限制市场竞争效果的行为。参见姜博文:《数字经济领域中数据垄断行为的识别与规制——以滥用市场支配地位中的数据垄断行为为例》,《财会研究》,2023年第10期,第73—80页。

㊼“算法歧视”(Algorithmic Bias)是指人工智能算法在收集、分类、生成和解释数据时产生的与人类相同的偏见与歧视,主要表现为年龄、性别、消费、就业、种族、弱势群体歧视等现象。参见[瑞]大卫·萨普特著,易文波译:《被算法操控的生活:重新定义精准广告、大数据和AI》,湖南科学技术出版社,2020年,第15—17页。

㊽杨爱华:《人工智能中的意识形态风险与应对》,《求索》,2021年第1期,第66—72页。

㊾张玲娜:《新时代主流意识形态建设研究》,知识产权出版社,2019年,第58页。

㊿严君、郭明飞:《意识形态话语表达的历史形态与现实转向》,《贵州社会科学》,2023年第7期,第11—18页。

(51)孙乃龙:《社会意识形态危机与规避:当代中国社会思潮的本质及导引研究》,中国社会科学出版社,2013年,第43页。

(52)张生:《ChatGPT:褶子、词典、逻辑与意识形态功能》,《传媒观察》,2023年第3期,第42—47页。

(53)《马克思恩格斯文集》(第八卷),人民出版社,2009年,第170页。

(54)习近平:《论党的宣传思想工作》,中央文献出版社,2020年,第23页。

(55)Carl Mitcham, "Notes toward A Philosophy of Meta-Technology," Society for Philosophy & Technology Quarterly Electronic Journal, Vol. 1, No. 1/2, 1995, pp. 13-17.

(56)European Commission, "White Paper on Artificial Intelligence-A European Approach to Excellence and Trust," https://ec.europaeu/info/sites/info/files/commission-white-paper-artificial-intelligencefeb2020_en.pdf

(57)习近平:《在庆祝中国共产党成立95周年大会上的讲话》,人民出版社,2016年,第9页。

(58)邓卓明主编:《改革开放以来中国共产党引领社会思潮研究》,人民出版社,2017年,第113页。

(59)范树成:《国外意识形态新变化对中国的影响及其对策研究》,社会科学文献出版社,2017年,第328页。

(60)习近平:《高举中国特色社会主义伟大旗帜 为全面建设社会主义现代化国家而团结奋斗:在中国共产党第二十次全国代表大会上的报告》,人民出版社,2022年,第43页。

(61)《习近平主持中央政治局第二十六次集体学习并讲话》,https://www.gov.cn/xinwen/2020-12/12/content_5569074.htm。

(62)《国家人工智能安全知识百问》,第12页。

(63)《总体国家安全观学习纲要》,第116页。

(64)中共中央文献研究室编:《习近平关于社会主义文化建设论述摘编》,中央文献出版社,2017年,第30页。

(65)作为行政行为的“约谈”,一般是以行政监管机关(或上级部门)或其工作人员为一方(约谈方),以作为被监管对象的行政相对人(或下级部门)或其委派的代表为另一方(被约谈方),由约谈方就被约谈方在运营或管理工作等事宜中存在的违法违规或不当行为进行行政指导。约谈的主要内容包括但不限于情况说明、意见交换、警示劝诫、责令整改纠正或风险及后果提示等。

(66)中共中央党史和文献研究院编:《习近平关于总体国家安全观论述摘编》,中央文献出版社,2018年,第9页。

(67)习近平:《论党的宣传思想工作》,中央文献出版社,2020年,第22页。