

习近平总书记强调，扎实推动科技创新和产业创新深度融合，要瞄准未来科技和产业发展制高点，加快新一代信息技术、人工智能、量子科技、生物科技、新能源、新材料等领域科技创新，培育发展新兴产业和未来产业。作为人工智能发展的重要方向和驱动力，生成式人工智能已展现出对经济社会发展的强劲推动力，我们需积极推动其在各领域的落地应用，并对其带来的一系列问题和现实挑战加以审慎应对，推动负责任的生成式人工智能发展。

推动生成式人工智能健康有序安全发展

激发生成式人工智能的创新活力与应用潜力

□ 曾大军

治理之道

习近平总书记强调，“科技创新是发展新质生产力的核心要素”，“要以科技创新推动产业创新，特别是以颠覆性技术和前沿技术催生新产业、新模式、新动能，发展新质生产力”。利用前沿科技培育和推动新质生产力发展，是建设现代化产业体系、推进高质量发展、提升国际竞争力的重要内容和必然要求。

生成式人工智能是发展新质生产力的重要驱动力。以大模型为代表的生成式人工智能技术，通过在技术上集成多年积累的深度学习模型与训练算法，依靠大算力，充分利用海量多模态数据，实现了从特定任务处理到跨领域通用的人工智能系统能力。通过自然语言理解、图像识别、语音识别、内容创作等多个领域展现了强大的能力和广阔的应用前景。大模型具有丰富的常识，拥有强大的通用任务解决能力、较好的工具使用能力、卓越的人机交互能力以及出色的指令遵循能力，这些优势使大模型能够显著提升应用效率，辅助产业实现数智化转型，从而带来巨大的经济效益。

近两年我国生成式人工智能取得了显著进展，技术路径基本探明，应用路径逐渐明确，产业化应用落地广泛。目前国产大模型数量已达数百个，涵盖金融、医疗、教育等多个行业，成为赋能千行百业的新抓手和支撑产业创新的新支柱。生成式人工智能的发展与应用趋势日益明朗。

生成式人工智能不仅是一种新技术，更代表着生产应用的新范式，能够催生新产业和新业态。大模型的学习模式类似人类大脑的处理机制，基于一个通用底座，通过学习不同的数据，能够形成多样化的能力，而无须针对特定任务开发定制算法。通过自然的人机交互，用户可直接通过对话生成应用程序，用户就是程序员，数据和场景本身就是软件应用。这一突破改变了软件开发范式，使软件开发由代码驱动向语言描述驱动转变，在需求分析、代码编写、测试、部署等软件开发的各个流程，都显著降低了技术门槛、提高了开发效率。这种新范式同时也给软件系统的

建设和应用带来了新模式，使各行业的体系变革和生态发展都有了新逻辑和新路径，促使新的产品、服务和商业模式不断涌现。新兴产业的形成和发展，将引领产业变革，成为新质生产力培育和发展的主阵地。

生成式人工智能能够推动传统产业深度转型升级，为实体经济注入新动能。通过将大模型应用于各个领域，可实现各行业的降本增效，提升产业自动化水平，助力企业的智能化升级，促使实体经济再上新台阶。在工业制造领域，生成式人工智能可辅助实现研发设计的自动化，为生产制造提供智能化的解决方案，通过预测需求、制定生产计划助力经营管理，并提供故障预测、健康管理等运维服务。在教育领域，可为师生提供智慧教育产品，实现个性化的因材施教。在医疗领域，可辅助诊断，成为医生的诊疗助理和患者的健康助手。在金融领域，可为金融机构提供智能客服、智能风控、智慧营销等服务，为客户提升服务体验。据有关预测，2035年生成式人工智能有望为全球贡献近90万亿元的经济价值，其中我国将突破30万亿元，且对我国经济的赋能效应显著优于美国等其他经济体。生成式人工智能与实体经济的深度融合，将打造“人工智能+”千行百业的产业新格局，形成现代化、智能化的产业体系，促进由传统生产力向新质生产力的转型。

生成式人工智能带来的安全风险和治理挑战不容忽视。由于大模型自身结构和生成机制存在漏洞，有被恶意攻击的风险，同时大模型自身的知识表达和学习模型还存在缺陷，经常出现常识性错误、杜撰内容等“幻觉”。

为此，应建立针对大模型全生命周期的安全监管与治理体系，从数据安全、模型安全、内容安全、伦理安全等多个角度，保障大模型研发和应用安全可靠。一方面，需制定技术标准，强化过程控制。建立严格的数据采集、存储和处理规范，确保数据来源合法可靠，同时保障数据隐私。对大模型进行安全攻击检测，提升其对恶意攻击的抵抗能力，并及时发现和修复安全漏洞。另一方面，系统推进大模型监管体系建设，制定和完善相关法律法规。强化内容审核和过滤机制，识别并过

滤生成的虚假信息、恶意内容和违法信息。明确伦理规范，建立伦理审查机制，确保人工智能发展遵循人类道德和价值观。通过建立全流程的安全保障体系，降低生成式人工智能的安全风险，推动技术的安全、可持续发展。

同时，为充分发挥大模型的带动作用，必须确保大模型的健康合理应用。然而，当前在大模型应用上仍存在一定误区。一是将生成式人工智能“工具化”，仅将其视为生成文案的办公工具。这种应用方式大大限制了生成式人工智能的潜力。大模型不仅可以用于文案生成，还能够在数据分析、决策支持、辅助管理等多个重要环节发挥作用，其应用价值远超办公自动化。二是将生成式人工智能“神化”，误将其视作万能助手。虽然大模型具备强大的通用能力，但在专业领域的赋能往往依赖专业知识，通用大模型难以满足需求，只有与根据专业领域的知识和需求定制专用模型相结合，才能实现更为精准和有效的支持。

树立立体思维，保障应用健康发展。着力促进各类大模型之间、人与机器之间的灵活配合与协同进化，充分激发大模型的创新活力与应用潜力，最大化释放大模型的应用价值。

一是形成各类大模型各司其职、共同发展的创新生态。大模型可分为三类：底座大模型、领域大模型和任务大模型。底座大模型通过在大规模无标注数据上进行训练，学习人类常识，能够形成举一反三的强大泛化能力，具有基础的通用功能，相当于完成了人类的通识教育。底座大模型可看作智能时代的基础设施，方便更多开发者在其基础上做进一步开发，从而促进大模型的普及和应用，加速构建更加活跃多元的技术生态系统。相比通用的底座大模型，领域大模型和任务大模型则更为专用。领域大模型专注于特定行业或领域的知识，通常基于领域相关专业数据进行预训练或微调，相当于完成了专业教育，成为领域“专家”。任务大模型针对具体任务进行微调和优化，以提升其在具体任务上的效果。应引导各类模型共同发展，培育多元化的创新生态。在实际应用中，应根据场景选择相应的模型，充分发挥各类模型的优势，从而更好地满足多样化需

求，实现最佳的应用效果。

二是形成大模型通专结合、高效协同的创新体系。通用大模型与专业大模型各有优势，它们之间的关系应辩证看待。通用大模型作为跨领域的通用平台，能够提供广泛适用的解决方案，满足不同用户的基本需求，具有适用范围广、灵活性强等优势。在依赖常识的通用场景，如主要面向公众的消费型业务和虚拟互联网应用等，能够发挥不可替代的作用。专用大模型专注于特定领域，通过深耕行业的专业知识，发挥领域纵深优势，提供针对特定业务场景的高质量解决方案，具有专业精度高、定制性强等特点。在实体经济、工业应用等更依赖专业知识的场景中，能够提供更加精准有效的支持。因此，应鼓励两类模型协同发展，支持互联网头部企业研发通用大模型，增强大模型的综合能力，同时支持各行业发挥行业优势，在通用大模型的基础上深耕专用模型，形成通专协同的创新体系。在实际应用中，通过实现两类模型的协同合作，使其充分发挥各自的优势，提高整体应用效果。

三是形成人机融合、混合增强的创新机制。大模型自然流畅的人机交互能力，颠覆了传统的人机交互范式，大幅提高了人机协同的效率和灵活性。为实现基于大模型的人机高效协同，需从业务体系层面进行系统性的重构。首先应明确界定大模型、人类和专家各自的职责范围，统筹发挥各自的优势。大模型在系统设计中应承担自动化任务者的核心角色，充分发挥其能够快速处理大量信息、高效执行重复性任务的优势，显著提升系统的整体运行效率；人类应发挥创造力强、决策能力强的优势，对系统起到整体设计和引导的作用；对于需要深入理解和判断的任务，例如复杂情境中的任务或非标准问题，需由经验丰富的专家团队来主导，确保决策的科学性和有效性。其次应面向实际需求，设计人机协同 workflow，建立有效的人机沟通机制和协作策略，确保人机能够高效合作。通过形成人机协同的创新机制，不断优化资源配置，提升系统的灵活性，实现更高效的系统性能，用人机融合新机制促进新质生产力的发展。

（作者系中国科学院科技基础能力局副局长、自动化研究所副所长）

观而思

当前，生成式人工智能作为人工智能技术发展的最新进展，依托大量丰富的数据和预训练的深度学习算法，比以往的人工智能具有更加强大、泛化的自然语言理解能力和内容生成能力，能够应对不同场景的使用并为多个行业提供应用支撑。

由于生成式人工智能的强大性能和通用性，针对不同领域、场景的定制化生成式人工智能将具有广阔的市场需求，相关产业发展前景广阔，因而有望成为智能化世界的基础设施，推动智能产业的发展。通过与相关产业融合，生成式人工智能将以智能赋能产业，深挖行业潜力、拓宽发展空间，推动产业智能化发展、引导传统产业转型升级，从而作为底层技术赋能千行百业。同时，生成式人工智能还将改变甚至塑造出新的生活习惯和消费方式。生成式人工智能使信息检索更加高效便捷，使商品信息、商家差异更加透明，这将导致企业更加注重产品质量与行业规范，倒逼产业升级。生成式人工智能在编辑、编程、翻译等方面也将减少繁杂的基础工作，给人们带来更多的休闲时间，间接创造新的消费需求和市场。另外，生成式人工智能将推动科研范式的变革。生成式人工智能在文献数据获取、结果分析等方面的应用将助推科学研究的知识模型快速生成和优化迭代，其在结构预测上的卓越性能可能会在药物研发、新材料研究上引发根本性变革。

但是，科学技术是一把“双刃剑”，生成式人工智能在带来诸多机遇的同时也会带来许多隐患。由于生成过程的复杂性，生成式人工智能的算法超越了工程师和用户的理解范畴，产生“算法黑箱”。而且人工智能呈现出的对文本、图像的掌握只是基于要素之间的相关性，并不真正了解对象的本质和因果联系，其生成的内容可能是在“一本正经地胡说八道”。若生成式人工智能产生的虚假信息被滥用，会给网络安全和国家安全带来严重后果。同时，生成式人工智能所依赖的算法设计和数据输入都蕴含了设计者的价值偏好，其输出内容也附着了价值观念甚至歧视偏见，也会存在被诱导输出虚假内容和有害信息的风险，引发算法歧视、学术不端等伦理风险。此外，数据来源和生成物可能引发知识产权纠纷。生成式人工智能的研发需要极大体量的数据进行训练，这让其数据来源的版权合法性难以得到保证，其生成物的版权分配也难以界定。大数据是生成式人工智能实现突破的重要基础，但是过度挖掘数据又将侵犯用户个人隐私，给用户带来困扰。

对于生成式人工智能这一新生事物，我们应对其持鼓励与包容的态度，但不能任其无序发展，必须推动负责任的人工智能发展。一是以技术规范技术。通过强化数据安全和隐私保护等技术的研发，加强对人工智能算法与模型的可溯源研究和审查评估研究，运用技术手段反制和鉴别人工智能所生成的内容。同时，积极探索生成式人工智能的问责机制与行为规范，努力打造安全可靠的人工智能。二是以政策法规保障技术健康有序发展。及时规范相关技术与产品的合理准入范围和应用边界，营造良好的发展环境。2023年，我国发布了《生成式人工智能服务管理暂行办法》，提出了包容审慎和分级分类监管等办法，为生成式人工智能在各领域的应用指明了健康规范的发展方向。2024年，我国又发布了《生成式人工智能服务安全基本要求》，进一步规定了生成式人工智能服务在安全方面的基本要求。这些政策法规有助于保障人工智能发展与安全并重、促进创新与依法治理相结合，必须切实推进实施。三是以伦理规范贯穿技术研发全过程。科技伦理是开展科学研究、技术开发等科技活动需要遵循的价值理念和行为规范，是促进科技事业健康发展的重要保障。生成式人工智能技术的发展需要遵循伦理先行、敏捷治理的要求，将科技伦理贯穿其科研开发应用全过程，加强伦理风险预警与跟踪研判，及时动态调整治理方式和伦理规范，以实现负责任的创新。

推动负责任的人工智能发展

□ 胡明艳

生成式人工智能的技术局限和伦理风险

□ 吴文峻

自ChatGPT面世以来，生成式人工智能成为人工智能领域发展的热点。生成式人工智能技术的基本形态，是采用海量数据来训练人工智能大模型，使其学会人类的语言文本和图像视频的统计模式后，能够在用户给出提示指令后，自动地生成所需的数字内容。与传统人工智能局限于分类、识别等特定功能相比，生成式人工智能更关注创造性内容的生成，展现出前所未有的文字对话、文档撰写、代码编写、图片创意、语音合成和视频生成等多样性的能力。

生成式人工智能技术综合了60多年来人工智能的研究成果，特别是最近10年来深度学习的技术突破。首先，能够从海量的语料数据中学习人类知识，熟记文本、语音、图像、视频等内在规律与统计模式，自动生成用户所需的新文本、图片、声音和视频。其次，实现了大模型的基于大量数据的训练学习通用特征与在特定任务上进行微调以优化性能的学习模式，并能够接受人类反馈进一步强化学习。不但可以记忆通用的人类知识，还可以持续学习专业的知识与技能，并在人类的奖励指导下，输出符合人类道德和法律约束的内容。最后，具备多模态数据的融合机制，动态融合文本、语音、图像、视频等数据，不仅能够生成更为丰富多样的数字内容，而且通过语音、手势、面部表情等的识别，支撑实现AI像人一样与人类交流互动。总之，生成式人工智能已经展现出了多模态、通用化的认知和交互智能，表现

出“由专到通”的强大能力，在自动化文档生成、自动化编程、智能客户服务、供应链管理、产品研发、智慧教育和智慧医疗等多个领域展现出极大优势，因而成为诸多领域的热门应用。

虽然这两三年来生成式人工智能展现出了前所未有的认知和多模态交互能力，但是其内在局限日益凸显，如何继续深入推进生成式人工智能技术发展，使之能够真正广泛落地应用，成为人们关注的焦点。首先，生成式人工智能所依赖的大模型在精准的认知理解和逻辑推理方面存在明显局限。一方面，由于大模型的幻觉现象问题，容易输出事实性错误的内容。另一方面，大模型的内容生成本质上属于概率统计性的词汇预测，所以它无法像人类一样完成长链条的动态逻辑推理。这两个方面的问题意味着很难将生成式人工智能直接嵌入实际业务。因此，如何面向落地场景的需要，消除生成式人工智能的认知幻觉，提升或弥补其逻辑推理能力就成为其进一步发展的关键问题。

其次，生成式人工智能面临着规模提升效益瓶颈。在大模型研发中，模型的性能提升和模型参数规模间满足幂律关系，即模型规模越大，输入的训练数据越多，模型预测能力就越强。因此，不少企业认为只要继续扩大模型规模，就能在不远的将来实现通用人工智能。但是，一方面，近5年来大模型的参数规模呈现指数级增强趋势，对智能算力的需求与日俱

增，最大模型GPT-4的参数规模已经到了万亿级别，需要上万张GPU卡的集群进行数月的模型训练和参数调优。而一些企业还在扩大GPU集群规模，正在投资建设10万张GPU卡，乃至百万张GPU卡的规模。构建和运行如此大规模的智算集群，需要克服集群供电和散热等能耗难题。另一方面，高质量和高密度数据语料库也将成为制约模型规模继续扩大的因素。众所周知，大模型的性能提升需要大量的高质量训练语料，目前语料库主要来自互联网公共领域数据的汇集。有报告指出，预计到2028年大模型训练将耗尽所有互联网公共数据资源，大模型增长不可避免地遭遇数据危机。为此，需要面向垂直领域，深度挖掘私域数据，扩大高质量数据共享，以支撑大模型更好适应垂直领域的需求。

由此可见，如果生成式人工智能企图单纯以扩大模型规模来实现绝对通用的智能模型，无论是在技术上还是在经济成本上，都不是可持续的技术路线。需要面向垂直领域的业务逻辑和落地场景，有针对性地业务流程中发挥大模型的长处，弥补其不足，以通专结合的技术路线，实现大模型在产业上的广泛落地与价值赋能。具体地说，就是瞄准领域的重点和难点需求，构造各种各样的专业小模型，与底座大模型相结合，打造领域适配的专业智能体，实现生成式人工智能对现有系统的赋能与升级。

生成式人工智能的发展使得相关的伦

理安全风险日益突出，大模型认知能力的涌现和内在缺陷给人工智能的社会治理带来了更多挑战。首先，基于大模型的生成式人工智能系统缺乏可靠的安全护栏，很容易因受到攻击而输出敏感信息或是价值观错误的内容。由于大模型的“算法黑箱”特征，其功能与行为的可解释性和透明性都是需要深入研究的难点问题。其次，生成式人工智能广泛应用带来的各类衍生风险层出不穷，最突出的是深度合成内容的治理问题。大模型生成的各类数字内容与人类创造的内容相比，很多时候已经到了真假难辨的程度，给网络诈骗和虚假信息传播等提供了新的技术手段。为此，急需系统地推行内容标识、水印验证等，建立可行的AI内容溯源管理平台，以确保健康的数字生态环境。最后，生成式人工智能研发与应用的敏捷治理将走向系统化和法治化。目前国内外都在积极制定生成式人工智能的法律法规制度，希望既有效地防范这一新兴技术带来的各类潜在和现实风险，又能够促进和规范这一技术的健康发展。为此，分级化的敏捷治理成为当前研究的热点，通过设置训练算力阈值，把治理的重点放在超大规模神经网络模型的研发和服务上，而对海量中小规模模型的研发活动采取简便的敏捷治理模式，从而实现新技术发展进步和有效治理的平衡。

（作者系北京航空航天大学人工智能学院教授）