

人工智能如何发展更安全

今年7月，中央政治局召开会议，分析研究当前经济形势，部署下半年经济工作。会议强调，要推动数字经济与先进制造业、现代服务业深度融合，促进人工智能安全发展。围绕人工智能安全发展的一系列问题，学习时报高端智库版组织了此次三方笔谈。

推进大模型赋能数智化应用健康发展

中国科学院自动化研究所副所长 曾大军

近年来，大模型技术飞速发展，极大提升了全球人工智能技术发展和应用迭代速度，被认为是“迈向通用人工智能的里程碑技术”。面对新一轮大模型引领的人工智能变革，加快推进大模型赋能数智化应用健康发展具有重要意义。

作为最成功的一类大模型，大语言模型是一种包含大规模参数、使用大量文本数据训练出来的深度学习模型，展现了大模型的能力和 basic 技术框架。目前，这一技术思路和框架已被用来处理图像、视频等多模态数据，并被推广应用于机器人控制与智能决策等场景。各类大模型正处于迅猛发展之中，众多高科技企业纷纷投身大模型建设，围绕大模型已经形成相对成熟的技术框架，但产品和生态尚在发展形成之中。总体而言，大模型的技术发展历程相比以往任何人工智能技术都更迅猛，其影响力也是史无前例的。

大模型的“大”主要体现在3个方面，即模型参数规模大、训练数据规模大和算力消耗需求大。模型参数方面，当前主流大模型的参数规模通常在千亿甚至万亿级别，远远大于以往各类深度学习模型。训练数据方面，大模型在前期训练阶段和后期调优阶段都需要海量的数据支撑，例如引爆大模型范式的聊天机器人 ChatGPT，预训练数据量达到了 45TB，其研发公司还在拉丁

美洲等地区招募了约 1000 名员工，专门从事与大模型相关的训练数据标注工作。算力需求方面，根据有关市场调查机构估计，ChatGPT 的运行需要上万片最先进的图形处理芯片同时工作，耗电量巨大。

大模型已经展现出惊人的能力，基于大语言模型的聊天机器人是以人机对话为接口的高效信息获取、智能处理和内容生成工具，能够实现高质量的信息整合、翻译和简单的问题求解与规划。其受到全社会广泛关注，主要是它已经初步具备通用人工智能的部分特性，包括通顺的自然语言生成、全领域的知识体系覆盖、跨任务场景的通用处理模型、通畅的人机交互接口。当然，聊天机器人只是大模型的起点，大模型未来的发展趋势也已初见端倪。

第一，大模型有望引领诸多行业的数智化创新发展。大模型的学习模式类似人类大脑，通过学习不同的数据，能够形成多样化的能力，不需要按照任务开发定制化的算法。通过自然的人机交互，智能化地解决复杂问题和实现辅助决策，推动各行业体系变革和生态发展。例如，未来的信息系统可以由领域用户通过与大模型的交互直接搭建和维护，“用户即程序员”，颠覆信息系统研发和应用范式。

第二，领域专用和轻量化大模型是当前技术研究的重点。大模型

虽然已经展现出诸多优势，但依然存在可靠性差、训练数据依赖、因果推理能力弱、搭建成本高等短板，在部分领域深度应用上的表现弱于专用小模型。为了更充分地发挥大模型优势，金融、教育、医疗等领域专用大模型已取得了阶段性成果，有望促进各项专业能力的涌现。

第三，大模型有望发展成为更加通用的人工智能。大模型起步于文本信息处理，现已涵盖图片、音视频等多模态数据。在不久的将来，大模型将超越信息域，结合硬件设施，发展成为与物理和人类世界互动的具象智能，逐步缩小与真正的“通用人工智能”的差距。

虽然大模型展现的能力已经让全社会看到了通用人工智能的曙光，但当前大模型也有一定的局限性。一方面，由于大模型自身结构和机制漏洞，有被恶意攻击的风险；另一方面，大模型自身的知识表达和学习模式还存在缺陷，导致其回答中经常出现“幻觉”，如常识性错误、杜撰内容等。

推进大模型赋能数智化应用健康发展，应坚持规划引领、需求带动、应用导向、安全为基，引导、扶持和监督大模型在更广泛的领域应用中中长计远虑、扎实落地、稳步推进。

首先，掌握技术主动权，加快自主可控的国产大模型全链条建

设。依托我国现有的人工智能人才基础，构建面向大模型与通用人工智能技术的平台和队伍。加大芯片、大数据、人工智能等产业的投入与整合力度，保障“海量数据”“顶尖人才”“算力支撑”的落地，推进我国全阶段自主可控大模型建设，掌握大模型技术发展的主动权和话语权。大力支持国产大模型的下游技术攻关和产品推广，促进国产大模型产品的应用落地和迭代升级。

其次，深化大模型应用生态建设，构建“通用大模型+专用小模型”的成熟应用生态。早在在大模型产生之前，基于领域经验归纳和业务知识构建的“小模型”已经广泛应用于各行业。应充分认识并应用好大模型和传统小模型各自的特性，在推广大模型产品的同时进一步推进小模型技术的升级和应用，发挥好小模型轻量级、高效率、善于解决特定问题的优势，逐步推进大小模型融合的应用生态。

最后，建立健全大模型安全审查制度与法律体系，从源头规避大模型发展风险。推动建立行之有效的数据审查机制和接入许可规范，从源头把控大模型内容安全性，并主动参与全球人工智能模型使用规范的标准制定。同时，对于大模型可能产生的风险进行合理评估与审核，如歧视、仇恨言论、私人信息泄露、虚假信息、协助犯罪等。

锻长补短 打造人工智能产业新优势

安徽省合肥市人大常委会副主任,合肥高新区党工委书记、管委会主任 宋道军

件等领域，推动核心算法、GPU 芯片、云计算等关键技术突破，构建自主可控大模型技术体系。二是推动类脑智能技术突破，研究神经网络学习、知识推理等创新计算模型和方法，构建新的通用人工智能发展路径。三是推动量子计算技术研发，突破量子比特操控、量子模拟加速、量子编译等核心技术，推动量子计算商用进程。

打造要素供给高地，强化产业支撑。我国智能算力基础薄弱，GPU 芯片面临“卡脖子”难题，可用于人工智能计算的智能算力严重不足，分布分散且异构严重，统一调度难度大，缺乏大规模公益性开放算力平台；训练数据供给不足，中文语音、词汇和语法多样复杂，数据清洗和数据标注难度较高，需在算力和数据支撑上开展统筹建设和有效积累。一是支持相关龙头企业联合攻关，构建千 P 智能算力集群，实现国产算力训练性能完全对标国际领先企业，突破大模型训练算力底座的“卡脖子”难题。二是推动大规模高质量多渠道数据资源获取、无监督预训练数据清洗、人机协同的多模态众包数据标注平台研发，构建 100TB 以上超大规模多模态多语种数据资源，争创国家级

数据训练基地。三是推动国产 GPU、自然语言处理、数据挖掘、深度学习等领域企业集聚，构建自主可控产业链体系。

打造场景应用高地，提升产业能级。庞大人口基数产生的海量数据和背后的巨大市场需求，是我国人工智能产业发展得天独厚的优势，需加速全时全域场景探索应用。一是推动“科研探索+”应用示范，充分发挥人工智能技术在文献数据获取、实验预测、结果分析等方面的作用，支持域内高校院所与人工智能企业联合设立科研合作专项，研发科学智能计算模型，充分挖掘分析实验数据，学习、模拟、预测和发现各种现象与科学规律，争取在量子信息、可控核聚变、深空探测等领域实现创新突破。二是推动“千行百业+”应用示范，推动国产大模型赋能地域聚力发展的产业地标，在相关重点产业开展大模型应用示范，推出新能源和智能网联汽车、工业互联网、智能家居等行业大模型，并形成一批可复制推广的细分领域应用解决方案和标杆型示范应用案例。

打造安全治理高地，优化产业环境。当前，我国人工智能安全监管面临一些挑战。数据保护方面，用户信息泄露和滥用涉及的网络和

数据安全技术有待突破，数据保护相关法律法规不完善；算法安全方面，对算法偏见、算法设计公正性和普适性的监管仍存在不足，需坚持统筹发展和安全，加强风险防控。一是高水平、高标准建设“中国（合肥）安全谷”，争取在 AI 大模型网络、数据安全、算法安全、云安全等关键技术领域取得重大突破，形成一批具有全国竞争力的网络与数据安全核心产品和解决方案，全面提升网络和数据安全防护能力。二是完善风险防控和处置机制，遵守人工智能发展有关法律、法规、伦理规范和技术标准，实行设计问责和应用监督并重的双层监管结构，防范和打击数据滥用、侵犯用户隐私等行为。

总之，打造人工智能产业新优势，需在技术、要素、场景和安全等多方面充分整合调度资源，各方合力推进。国家层面应重点支持大模型领域重大研发攻关项目，适度超前布局绿色智能的多层次算力设施体系，加强数据标准体系建设和完善数据交易机制，加快制定人工智能安全领域标准规范，鼓励地方和企业积极探索创新通用人工智能场景应用，推动我国人工智能产业发展形成新优势、取得新进步。

做好人工智能发展的风险防范

科技部新一代人工智能发展研究中心主任 赵志耘

当前，ChatGPT 类大模型加速了数字世界和物理世界的深度融合，引领人工智能进入新的发展阶段。与此同时，人工智能固有的不确定性、不透明性、难以解释、事实性错误等能力局限，对经济发展、国际政治和全球治理等方面带来深远影响，引发了全球对大模型广泛应用安全风险的热议。如何应对人工智能风险并把握人工智能发展的战略主动，有效保障国家安全，已成为我国国家安

全治理的重要议题。大模型技术引发新的安全风险和治理挑战。随着技术的更新迭代，人工智能的应用门槛持续降低，伴随着误用滥用情况的增加，经济社会领域的多重安全风险也在交织叠加。

一是放大隐私信息泄露风险。当前，人工智能开发者和服务提供者可能利用用户数据进行优化训练，但相关服务条款却并未对用户使用做出解释说明，可能涉及在用户不知情情况下收集个人信息、商业秘密等，安全风险较为突出。为保护个人信息安全，部分欧洲国家针对大模型进行了严格监管和审查，甚至考虑禁止 ChatGPT 等人工智能应用。

二是舆论引导与文化渗透风险。在舆论和文化领域，人工智能技术可被用于算法引导、平台渗透、网络监控、内容生成等信息处理环节，由大模型驱动的文本创作工具可能产生颇具影响力或说服力的信息，使得虚假信息不易被识破，负面舆情传播速度加快，进而导致侵犯他人合法权益、扰乱经济秩序和社会秩序，甚至危害国家安全和公共利益。

三是情报军事领域应用的新威胁。人工智能的自动生成代码功能使黑客技术的习得过程更为容易，

使网络攻防的更新频率大幅提升。利用 ChatGPT 类大模型可以帮助机器学习识别更加复杂的文本及实体，或将扩大无人军用智能无人装备的应用，增大军事对峙摩擦，对国家安全带来严峻挑战。

加快推进人工智能安全治理。面对人工智能技术快速发展、风险持续蔓延的形势，增强人工智能的安全性成为国际组织、各国政府及产业界等共同关注的议题。

我国高度重视人工智能安全问题，不断加强人工智能发展的潜在风险研判和防范，建立健全人工智能安全治理体系，对人工智能技术本身所致的内生安全、人工智能应用中的衍生安全以及影响和决定人工智能前景的发展安全问题进行系统性治理，确保人工智能安全、可靠、可控。

2017 年，国务院发布《新一代人工智能发展规划》，对人工智能治理工作进行了专门部署，并要求建立人工智能安全监管和评估体系。我国不断完善数据安全、网络安全、人工智能等相关法律法规建设，保障相关技术产品的有序发展。先后发布《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》，为人工智能技术的开发和应用设定道德约束，促使人工智能更好地服务于人类。此外，我国也积极参与全球人工智能治理，坚持伦理先行、和平利用、安全可控等治理原则，呼吁加强对人工智能军事应用的监管，以负责任的态度促进人工智能安全治理。

推动形成人工智能安全治理的强大合力。尽管已开展一系列治理行动，但人工智能安全治理还面临诸多全球共性难题。人工智能技术目前并不具备较强的可解释性、准确性和真实性，难以完全实现人工智能治理所倡导的透明度、可靠性和安全性。如何在创新发展和有效监管之间取得平衡也是各国普遍面临的问题。既要避免过度监管阻碍技术发展，也要防止管理力度不够而引发严重问题。

当前，人工智能影响下的国际科技、经济及政治领域发生冲突的可能性日益增大。我国正处在人工智能技术快速发展的阶段，必须坚持贯彻总体国家安全观，构建多元参与、协同共治的人工智能治理机制，调动各主体参与治理的积极性，促进人工智能与经济社会健康有序融合，以高水平科技安全保障国家总体安全。一是建立健全人工智能安全治理机制。坚持审慎监管、开放包容原则，强化风险研判和监测预警，完善科技伦理审查和监管制度。二是增强法治保障能力。加快推进人工智能立法，开展生成式人工智能、信息安全利用等法律问题研究，建立追溯和问责机制，清晰界定人工智能相关主体权利、义务及责任边界，从而保障各类群体的合法权益。三是增强人工智能技术治理能力。强化人工智能数据安全、隐私保护等技术研发，强化对人工智能算法和模型的审计监管和对人工智能系统安全性测试评估，增强抗系统性风险能力。四是增强国际安全合作能力。倡导以人为本、智能向善理念，在《中国关于规范人工智能军事应用的立场文件》等基础上，发挥联合国、G20、上合组织等多边机制的作用，拓展人工智能安全共识，避免人工智能滥用带来的安全风险。