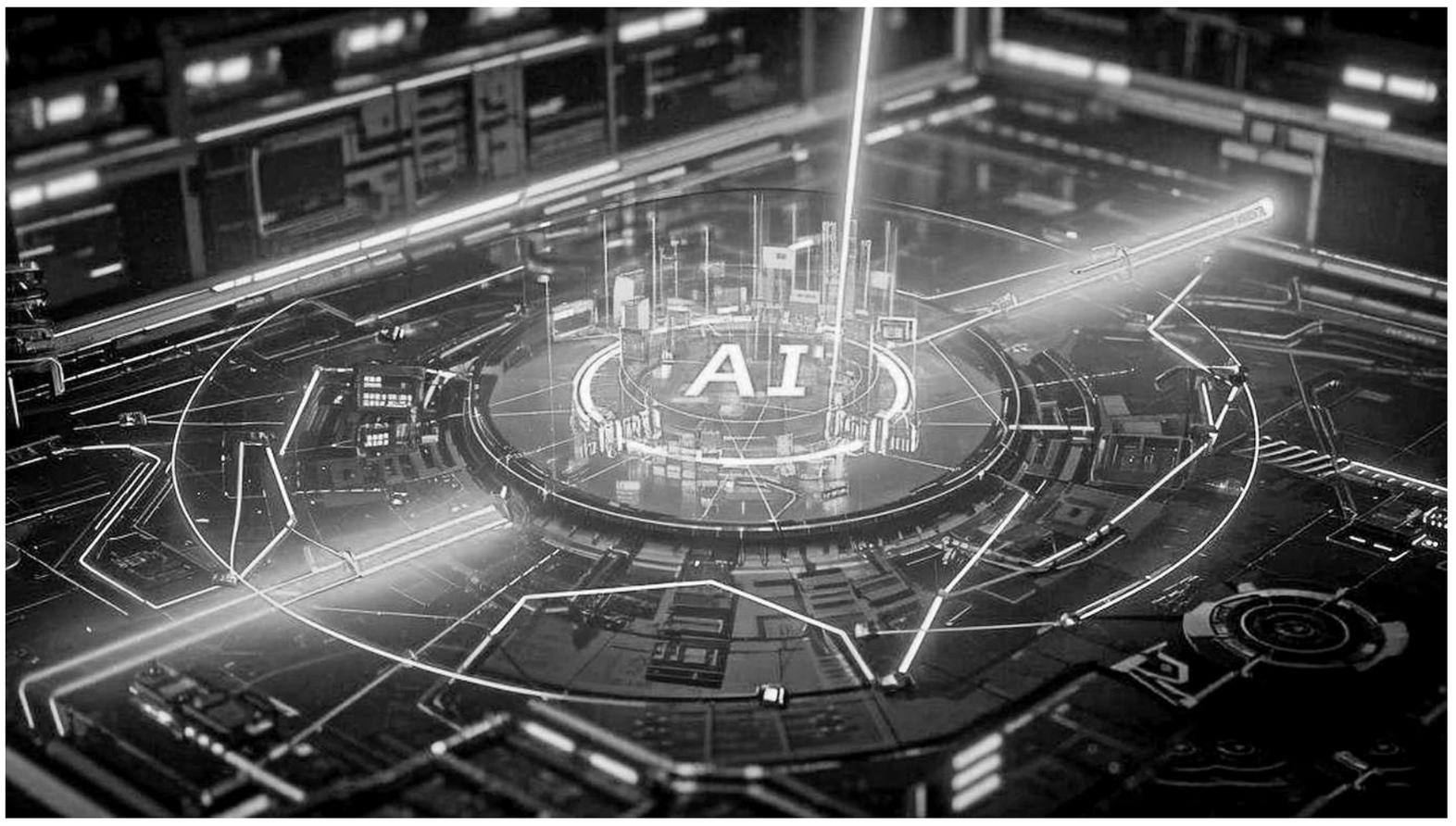
高技术前沿

作为新型生产要素，数据正在快速融入社会生活的方方面面，深刻改变着人类的生产生活方式。当前，随着人工智能（AI）发展突飞猛进，各类学习模型不断涌现，数据作为驱动AI这台“引擎”的“燃料”，发挥着越来越重要的作用。与此同时，一些现实问题也悄然浮出水面。

AI大模型需要什么样的数据

■蒋雨铖 王琪睿 本报特约通讯员 宋世杰



人工智能概念图。 资料图片

数据真伪
大模型“撒谎”的原因

AI也会撒谎？

据路透社报道，OpenAI旗下的ChatGPT在回答问题时，错误地声称澳大利亚墨尔本西部赫本郡的市长布赖恩·胡德是贿赂丑闻的有罪方。此前，胡德曾在一家公司工作，他向监管机构举报了公司内部向外国官员行贿以赢得货币印刷合同的情况。ChatGPT错误地将胡德作为控方证人出席庭审的经历，作为其受审的例证。目前，大型语言模型的“幻觉”问题（即生成虚假信息）已经成为学界和业界共同关注的问题，训练及处理的数据质量下降是产生该问题的主要原因。

那么，生成“谣言”的“证据”从何而来？这就需要提到大模型获取数据的两种主要方式：主动采集技术和被动采集技术。主动采集技术主要包括网络爬取和传感器采集；被动采集技术包括用户上传数据和日志记录数据。其中，最易被“伪造”的数据，来源于网络爬取数据和用户上传数据。

网络爬取是从互联网上自动抓取数据的技术。互联网公开数据中混杂着大量噪声数据，使得训练数据受到污染，进而导致模型产生输出偏差。2024年，麻省理工学院、上海交通大学、哈佛大学、微软研究院、IBM公司、剑桥大学等联合召开了首届数据污染研讨会。会议报告显示：各类模型的训练数据中，可能包含大量从网页和数据集内抓取到的虚假信息。这些低质量的数据不仅无法为模型提供有效的训练素材，还可能对模型的判断产生误导，导致模型性能下降。

另外，合成数据的滥用，可能导致模型输出产生偏差。为了解决数据资源不足的问题，合成数据被广泛应用以弥补真实数据的不足。美国莱斯大学与斯坦福大学的研究团队指出，将AI生成的内容喂给模型，会导致模型性能下降，输出错误率升高。研究人员称这种现象为“模型自噬障碍”——就像近亲繁殖导致基因缺陷被不断放大、重复扫描打印同一份照片会使照片画面模糊一样，模型使用AI生成的数据进行训练，认知偏差就会像滚雪球般扩大，最终导致模型掉入“认知陷阱”。

筛选标注
大模型“填喂”的选择

那么，什么样的数据才能满足大模型“大而挑剔”的“胃口”呢？总体看来，大模型对数据的数量、质量、种类都有着极高的要求：只有足够的数据量才能对体量、参数庞大的大模型进行充分训练；只有准确性、完整性、一致性较高的数据，才能避免在训练中对模型产生误导；只有涵盖多个领域的多类数据，才能让大模型学到更广泛的知识，更好地处理综合性问题。

在数据的海洋中，我们该如何筛选出适合大模型的数据呢？

一是采集数据时选择可靠的数据来源。首先是官方和权威机构发布的数据，比如政府部门发布的统计数据、专业科研机构公布的研究成果和文献资料等。这些数据一般都经过了严格的审核和验证，具有较高的准确性、可信度。其次是在一些领域领先的企业发布的数据，这些企业一般对行业标准、技术标准等具有较高的话语权，数据质量相对可靠。

二是预处理数据时进行数据清洗和标准化。在采集到的数据中，识别并筛去重复的数据，防止重复数据的权重放大，造成结果失真失衡；对于完整性较差的数据，可以将不同格式的数据统一格式，以便大模型顺利完成训练。

三是标注数据时进行严格规范。数据标注是指给原始数据添加标签的过程。这些标签对数据进行归类，帮助模型在遇到从未见过的数据时，也能准确识别数据中的内容。待标注数据，需要制定严格的数据标注标准操作规范，并对已标注的数据进行抽样审核，避免让不正确分类的数据影响到模型的训练。

四是评估数据时进行内外检验。模型自检时，可以将数据集分成多个子集，通过轮流将不同子集作为验证集，来评估模型面对未知数据时的表现，检验数据的一致性。在模型训练过程中，要持续监控准确率、召回率等评估指标，检验数据的适用性。外部验证时，可以将采集数据和处理结果与权威模型进行对比，来评估数据的质量。

实景运用
大模型“军用”的梗阻

数据体量、质量等现实难题，不仅困扰着民用模型，同样也横亘在军用大模型的发展路径上。相对于民用模型，军用大模型有一定的优势，但也面临高质量军事数据资源不足、模型框架选择难、安全问题多元化等挑战。

战场数据获取困难，是高质量军事数据资源不足的主要原因之一。军事网络和民用网络存在物理隔离，由民用网络采集的大量战场数据很难传输到军用网络。此外，战场中的多源信号还缺乏有效的跨模态对齐标注。比如，一款战机存在很多特征信息：红外热源信号（温度）、雷达反射信号（波长波形）、外形特征（可见光图像）等。如何让模型将这些不同种类的特征信号统一联系起来，帮助其快速识别、获取该型战机信息，还存在较大困难。要解决这些问题，可以探索建立安全的军事数据采集传输通道，收集时效性高、质量好的军事数据；加强跨模态数据处理技术的研发，运用高质量标注数据、压减标注错误率的方法，构建专业、精准的军事多模态数据集，以实现对军事设施、装备等的精准识别。

合成数据的偏差问题，会影响军用大模型的训练。实战数据的缺失，将导致越来越多合成数据被投入模型训练中。不加筛选、偏离实际的合成数据，会对模型训练成果造成不利影响。例如，虚拟引擎生成的地表对阳光的反射率与实际环境相差较大，导致红外传感器将较高的地表反射信号当成目标的温度信号，进而发生误判。要减小合成数据对模型的影响，需深入采集战场环境中人员、装备、环境等各类信息，以大量实际数据训练模型，从而生成最接近真实战场的合成数据，并做好合成数据的筛选和标注工作，减少合成数据与现实的偏差。

模型框架的选择，阻碍着军用大模型的使用。如果简单地把民用模型迁移到军事领域，模型会因为无法理解军语等问题，导致生成结果准确率大幅下降。不同模型框架所需的规模、性能、部署成本和安全性、可靠性以及支持的应用场景等也需要综合考虑。此外，在数据样本少的情况下，如何进行军用大模型的能力测试，也是十分现实的问题。未来可以针对军事数据以及相关业务特点开发专门的小模型，通过分发各个作战单元，收集整理相关语言库，随后与大模型融合，提升高度封闭条件下模型对语义的理解和军事语言生成能力；在实验验证中，对满足基本条件的大模型进行多轮能力评估，全面考察不同模型在军事应用中的性能优劣和成本效益，综合优势进行整合归一。

军用大模型存在较多安全问题。首先是使用数据的伦理合规性。尽管军事行动存在特殊性，数据使用也需遵循国际法规和伦理准则。此外，模型应用于智能自主化武器系统可能存在道德风险。应制定模型在军事应用中的准则，录入底层逻辑和决策标准等，避免出现武器系统为达成目标选择攻击民用设施的情况。

军用大模型的安全、保密要求也是需要注意的问题。面对战争，任何一个决策都可能导致人员陷入危险境地。因此，如果使用模型进行决策，决策的可靠性、可控性、保密性、稳定性需要多重评估，确保它在战场上行之有效。

AI大模型的数据问题已经不只是技术问题，还广泛涉及法律、伦理与地缘政治等。在这场复杂隐蔽同时关乎未来的“认知战争”中，胜负的关键在于能否构建起牢不可破的“数据防线”。因此，建立行之有效的数据采集、管理、评估机制刻不容缓。就像一名业内人士所说，数据治理是人工智能发展的基础，良好的数据治理是AI应用的前提。只有技术创新和治理框架同步进化，大模型才能摆脱“数据困境”，成为人类的“智囊”，持续释放巨大潜力，真正成为推动社会进步、保障国家安全的重要力量。

潜力巨大的时空基准建设

■张 鞏 陈 朗

论 见

人类的一切活动都离不开时空环境，要对人类活动的状态信息进行描述，必须有一个统一的时空基准作参考。所谓时空基准，就是利用现代大地测量理论与方法、现代时频测量与数据处理技术等，建立的一种全局性的地球参考系统和参考框架。作为描述宇宙万事万物空间状态和演化过程的参照系，时空基准在国家安全和全球经济社会发展中起着基础性作用，已成为世界强国竞相布局的重大基础设施。

时空基准包括坐标基准、高程（垂直）基准、深度基准、重力基准、磁力基准和时间基准。其中坐标、高程、深度反映空间的几何属性，重力、磁力反映空间的物理属性。建立和采用科学的时空基准及其技术标准，不仅有利于各种地理位置信息的统一和协调，也有利于地理空间信息成果的整体性和相容性，从而保障各项生产和建设任务顺利进行。

大型桥梁、水坝等基建工程的建设，离不开高精度时空基准支持。基建工程投入使用后，仍需布设控制网，在统一的时空基准下，对桥体、坝体等进行时间维度上的变形监测。自然灾害频发的国家和地区，进行防灾、减灾等工作，也离不开统一的时空基准。

在激烈对抗的军事领域中，时空基准更是意义重大。信息化智能化战争，即便只是0.01秒的时间误差，都可能导致多部雷达跟踪锁定的同一目标显示为多个目标。如果连基本的情报融合、目标识别都无法实现，精确指挥必然失调，自主协同必然失序，多域联动必然失控。所以，无论是武器打击平台的精确制导、电子对抗领域的精细校频，还是各作战单元全天候的导航定位、指挥员战场态势的同步认知，都离不开精准统一的时空基准。从伊拉克战争到巴以冲突，旨在扰乱对手时空基准的“时空导航战”屡见不鲜。时空基准的重要性可见一斑。

时空基准的建设、维护和运用，涉及大量前沿技术。比如，陆海空天立体时空基准构建、观测与维持技术，全域时空基准连接、统一与传递技术，多源时空观测数据融合及处理技术，高安全、高弹性、高可信时空基准服务技术，自主可控量子时空监测装备研制及环境适应技术，铯原子钟可靠性、喷泉原子钟频率准确度测定技术，不同途径光钟及不同原理时钟联合守时模式，水下时空基准传递技术，不同物质空间重力场感知技术，深空天然时空信源探测装备研制与探测技术，等等。

科学家爱因斯坦曾说：“窗外的每一片树叶，都令人类科学显得那样苍白。”这警示我们，面对真理的大海，人类的科学成果或许只是海边捡拾到的零星贝壳。当下看似先进的时空基准建设，未来仍有巨大的提升潜力。

基础创新可能带来颠覆性技术突破。比如，在理论上，基于广义相对论，最新的时钟可感应非常微小的高程变化。该技术一旦实现，将颠覆传统的高程基准建设和维持方法。由于高程基准与重力位之间具有确定的关系，只要确定两地间的重力位差，就能不受空间距离限制，跨越千山万水，确定两地间的高程差。相关的相对论时频测位，已成为国际研究热点。

航空航天科技丰富了时空基准建设思路。目前，美国、英国、加拿大等国都提出各自的大型低轨星座计划，旨在通过部署大规模卫星网络实现高速互联网、全球通信、地球观测等目标。如果在卫星上集成多种大地测量观测手段，就相当于将地面上的各种观测站搬到了卫星上。这必然会给时空信息的获取、处理和运用带来革命性变化。

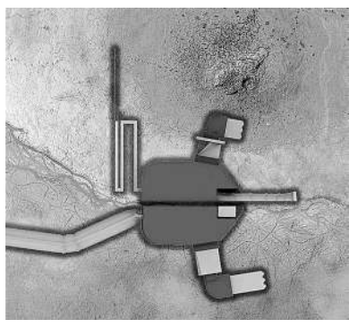
更遥远的深空，已成为世界强国竞争的新高地。美国开启了重返月球、火星探测等深空探测计划。欧盟及俄罗斯、日本等都在深空探测领域提出了各自的发展规划及加入国际合作的意愿。提前筹谋布局，将地球时空基准升级为地月时空基准，极有可能推动地月经济成为下一个经济增长点。

科技云

科技连着你我他

■本期观察：黄辛舟 侯旭达

软体机器人



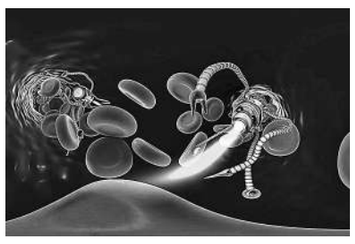
前期，南京信息工程大学联合华南理工大学团队，成功研制出一种具有两栖运动和舌捕能力的软体机器人。

这种机器人由3D打印折纸致动器集成而来。其中，Z形致动器用于机器人的后腿，可以模仿青蛙的游泳运动；腹鳍型致动器模仿弹涂鱼的爬行机制；卷/开型致动器模仿青蛙捕猎运动。

该型机器人能实现多种陆地爬行方式，可以承受载荷、捕捉猎物、轨迹跟踪、避开障碍，还可以在非结构化的地面上爬行。此外，它还具有极强的两栖跨介质能力。它的脚蹼采用单向设计，推进阶段会以最大面积与水面接触，恢复阶段则会被水流冲开从而减少阻力。

研究人员表示，他们之后可能会在机器人头部增加刺激响应材料，以提升机器人的自主探测能力。这类两栖机器人将集成测试pH值和浊度等水质的传感器，以便在陆地—水过渡区执行水质监测任务。

微型生物机器人



近期，美国塔夫茨大学和哈佛大学研究人员成功利用人类气管细胞，创建了一种微型生物机器人。它不但能在神经元表面移动，还能使实验室培养皿中的受损神经元恢复生长。

这种多细胞机器人可以自我组装，还能对其他细胞（例如受损的神经元）产生愈合效果。研究团队表示，这种机器人在进行治疗时，不会引发免疫反应。它们只能在特定实验室条件下生存，没有暴露或意外传播的风险。它们不能繁殖，也不能进行基因的编辑、添加或删除，因此没有超越现有安全措施的进化风险。

理论上讲，这种机器人不但可以帮助愈合组织，还能向受损部位提供促进再生的药物。下一步研究中，它们的应用或将更加广泛，包括清除动脉粥样硬化患者动脉中的斑块堆积、修复脊髓或视网膜神经损伤、识别细菌或癌细胞等。

协作机器人

近期，韩国一家科技公司推出了一款协作机器人P3020。该型机器人在载荷能力、工作范围、能耗效率等方面都实现了新的突破。

P3020协作机器人的有效载荷达30千克，这使得它能够轻松应对重型搬运作业。该机器人仅需使用一个简单的底座固定，无需复杂的升降机构，即可轻松完成高空作业。这款机器人融入了人工智能技术，通过自主下载和不断更新AI模型，实现更智能的运动规划、视觉导航等功能。在面对复杂多变的任务时，该机器人表现出更高的灵活性和适应性。

P3020协作机器人在多个领域展现出广泛的应用潜力，可部署于智能装配、码垛搬运、物流分拣等应用场景。凭借其先进的软件智能和高度集成能力，该机器人可有效提升生产效率和产品质量。