编者按 当前，人工智能发展方兴未艾，大幅提升了人类认识世界和改造世界的能力，同时也带来一系列难以预知的风险挑战。为帮助读者更好了解人工智能，本版今起推出“解读人工智能前沿技术趋势”系列报道，分析技术变革深层逻辑，凝聚智能向善共识。

如何构建可信赖的AI系统

——“解读人工智能前沿技术趋势”系列报道之一

◎本报记者 **吴叶凡**

近期，国内外一些人工智能（AI）产品问答内容价值导向错误的新闻频上热搜。随着AI技术的发展，AI的价值导向问题逐渐引发广泛关注，“构建可信赖的AI系统”的呼声越来越高。日前在2024年世界科技与发展论坛期间发布的《2024年人工智能十大前沿技术趋势展望》，就列入了“人机对齐：构建可信赖的AI系统”。2024年世界互联网大会乌镇峰会也聚焦AI，释放清晰信号——拥抱以人为本、智能向善的数字未来。

什么是可信赖的AI系统？构建可信赖的AI系统有哪些路径？科技日报记者就以上问题采访了相关专家。

可靠稳定是关键

随着AI在社会生活和各个行业中渗透程度的加深，其决策和行为的影响范围也日益扩大。例如，在医疗、交通、金融等高风险领域，AI系统的决策影响着人们的生命、财产与福祉，一些错误决策可能直接威胁到人类生命或财产安全。康奈尔大学约翰逊商学院讲席教授丛林介绍，AI在金融领域的应用主要包括资产管理、资产回报预测、资产定价等。“我们希望金融领域的AI要准确。我们并不需要它有发散思维或特别有创造力，而是希望它能给我们准确的答案，或是具有一定的稳健性。”他说。

“确保AI系统可信赖，已经成为AI发展不可忽视的要求。这不仅是技术层面的改进，更是社会伦理与责任的体现。”中国科学技术大学人工智能与数据科学学院教授王翔认为，可信赖AI系统不仅能让技术更好地满足人类需求，还能有效防范AI误判和偏见可能引发的负面效应。可信赖的AI系统不但要有优秀的预测、生成、决策等业务能力，而且在透明度、公平性、可解释性、安全性等方面也要符合用户预期。

其中，可解释性是指用户应能够理解AI的行为和决策流程，以便增强用户对AI的信任，并更好地加以利用。公平性要求AI的决策不应受到偏见影响，避免形成对不同群体的歧视。安全性则是指AI系统在运行过程中不会带来安全隐患，并能在一定范围内控制自身行为，特别是在极端或意外情况下要能保护人类安全。“AI系统还需要具备可靠性和稳定性，这要求它的表现在复杂和变化的开发环境中也要始终如一，不轻易受到外部因素干扰。”王翔说。

人机对齐是前提

那么，如何确保AI系统可信赖？王翔认为，人机对齐与构建可信赖的AI系统之间关系密切。“只有具备人机对齐特质的AI系统，才能进一步实现可信赖的特性。”他说。

从概念上看，人机对齐是指确保AI系统在执行任务、进行决策时，其行为、目标和价值观能够与人类保持一致。“这就是说，AI系统在自我优化和执行任务过程中，不仅要高效完成任务，还要符合人类的伦理和价值体系，不能偏离人类设定的目标或带来不良的社会影响。”王翔进一步解释，“尤其是在涉及社会伦理和安全的场景中，确保AI输出内容与人类的价值观和道德准则相符，是人机对齐的核心意义。”

如果AI系统没有经过人机对齐的过程，即使具备强大的功能和智能，也可能因不符合人类的期望和价值观而导致信任危机或负面影响。“因此，确保AI系统在目标和行为上与人类保持一致是构建可信赖AI系统的重要前提。两者的结合不仅能提升AI的表现，还可为未来AI在各领域的广泛应用奠定基础。”王翔说。

确保AI以人为本、智能向善，完善伦理和法律框架是重要发力方向。王翔认为，技术的进步往往伴随着新问题的发生，因此需要设立法律边界和伦理准则，为AI的发展提供指导与约束。这不仅可以减少AI应用中潜在的伦理风险，还能使AI应用更加规范和安全。此外，建设可信赖的AI系统需要跨学科合作，哲学、伦理学、社会学等学科的参与能为AI的设计与发展提供更全面的视角。

2024年世界互联网大会“互联网之光”博览会上，观众在参观AI核心算力单元的设备。 新华社记者 **黄宗治**摄

技术优化是手段

构建可信赖的AI系统，还需要在技术层面和应用实践中不断探索和完善。王翔介绍了三种主要的技术路径。

一是数据驱动路径。王翔认为，数据质量和多样性是实现可信赖AI的基础。训练数据的多样性可以有效减少模型中的偏见问题，确保系统决策更加公平、全面。“只有在庞大的优质数据基础上构建的AI模型才能适应广泛的应用场景，降低在特殊或极端条件下出现偏见的可能性。”王翔说，数据的安全性也至关重要，尤其是在涉及个人隐私的领域，保障数据安全可以提高用户信任度。

二是算法驱动路径。王翔说，算法的优化与控制是实现可信赖AI的关键手段。在模型的设计阶段，开发者可以通过设置伦理规则、嵌入人类价值观等约束条件，确保系统在实际运行中符合社会准则。同时，设计透明的算法结构有助于提升模型的可解释性，便于用户理解其运行机制，并为未来的模型更新和优化打下基础。

三是奖惩引导路径。王翔说，通过合理设计奖惩机制，可以让AI在不断试错和学习过程中，逐渐形成符合人类价值观的行为方式。例如，可以在奖惩系统中设置反馈机制，当AI的行为偏离预期时施加相应惩罚，引导其在自我训练过程中符合人类期望。同时，奖惩机制需具备时代适应性，确保AI系统能在运行中持续更新并优化自身。

这三种技术路径的侧重点各有不同。王翔解释，数据驱动路径主要聚焦于通过高质量、多样化的数据源减少AI系统的偏见，提升系统的适用性；算法驱动路径更注重模型的设计和透明性，使系统在行为逻辑上更符合人类预期；奖惩引导路径则侧重于在AI自我学习和优化过程中提供有效指引和反馈，让系统逐渐趋向人类认可的方向。“不同路径相结合，可以为实现可信赖的AI提供更加丰富的技术支持。”王翔说。

要构建可信赖的AI系统，还需在实际应用中不断进行迭代和优化。“通过多次评估和测试，可以在不同环境和条件下验证AI系统的性能，确保其在现实应用中的表现符合人类预期。”王翔说。

鸢目智能边缘计算平台亮相

科技日报讯（记者**俞慧友** 通讯员**乔欢**）11月21日，记者从北京大学长沙计算与数字经济研究院（以下简称“北大长沙院”）获悉，该院联合北京中科航星科技有限公司自主研发的鸢目智能边缘计算平台不久前正式亮相。

鸢目智能边缘计算平台具有低延迟、高性能、强算力等技术优势，可支持目标识别、目标追踪、智能预警、智能蜂群等特种场景需求，具备满足装备环境要求的软硬件自主可控能力。

随着生成式人工智能（AI）的发展，端侧AI逐渐走入人们生活。如何在较低传输量的基础上让模型更智能、怎样在边缘侧高速运行高质量模型等问题，是AI模型进入终端时遇到的难题。

“如何在小终端上高速应用高精度视觉模型，是我们着力攻克的难题之一。”北大长沙院相关负责人介绍，团队通过AI算法仓库、跨平台高性能计算系统等三大核心技术，一栈式解决了复杂场景下一些国产硬件“不好用、算得慢”等问题。

鸢目智能边缘计算平台依托这三大核心技术，在目标识别、目标追踪、智能预警等场景中，充分展示了软硬件自主可控、轻小化高可靠模组、多场景高精度适配、高时敏低功耗支持、低成本可定制服务五大性能优势。技术加持下，平台的模型识别精度提升50%以上，处理时间较国外同类型平台提速20%左右。

目前，以鸢目智能边缘计算平台为核心，涵盖板卡、智算盒子、车载部件等多样化产品的体系已经形成，可提供包括软件、硬件、综合解决方案等在内的全栈服务。

浙江首个新型数字化智能变电站投运

科技日报讯（**夏洁 范金凯**记者**杨雪**）11月19日，浙江省首座新型数字化智能变电站——110千伏雷东变电站正式送电投运。该变电站位于浙江省杭州市萧山区靖江街道，主要服务于周边居民以及临空经济示范区。

作为国家电网新型数字化智能变电站试点之一，雷东变电站创新采用二次在线监测技术和二次回路智能标签系统。通过构建变电站二次系统数字孪生模型，雷东变电站实现了全站信息的三维可视化查询，不仅全面提高了智能变电站调试、运维和改扩建工程实施的效率和质量，也便于回路故障定位，可进一步保障变电站安全可靠运行。

“我们积极响应变电站智能运维和数字化转型的要求，结合浙江省数字化建设优势，创新应用了混合气体、智能二次标签、一体化事故油池等多项试点推广技术，打造具备推广应用意义的数字化智能变电站。”国网杭州市萧山区供电公司工作人员刘佐然说。

据介绍，国网杭州市萧山区供电公司为雷东变电站加装了二级能效节能型电力变压器、智能通风系统、精准送风系统等设备，既精准满足了设备运行环境要求，也降低了全站运行能耗。在精准送风系统的精细控制下，气体绝缘全封闭组合电器设备使用寿命可延长20%—30%，每年可减少二氧化碳排放约15吨。

国网杭州市萧山区供电公司员工正检查110千伏雷东变电站内设备。 **沈书廷**摄

法律行业有了基座大模型

◎本报记者 **代小佩**

11月15日，最高人民法院正式发布“法信法律基座大模型”（以下简称“基座大模型”）。这是国内首个法律行业AI基座大模型，也是国家级法律人工智能基础设施。

该模型有哪些技术创新点和应用前景，有什么研发难点？包括研发团队成员在内的多位专家学者对相关问题进行了解答。

破解模型训练难题

“当前，大语言模型已成为推动人工智能技术进步的关键动力之一，并广泛影响着社会各个方面，包括法律行业。”人民法院出版社总编辑余茂玉说，未来法治建设将更加紧密地与人工智能等先进技术结合，法律行业将迎来智能化变革，同时也存在一定的安全和风险挑战。

为更好迎接挑战，人民法院出版社按照最高人民法院部署，启动研发建设自主可控的基座大模型。在清华大学千亿参数通用大模型基座上，研发团队投入3.2亿篇共计3.67万亿字的法律文献、裁判、案例、观点等数据语料，经过数月的预训练、优化训练、监督微调和多轮测评，最终建成基座大模型。

在研发过程中，团队曾遇到诸多技术难题。人民法院电子音像出版社副社长石鹏是研发团队的核心成员。他介绍，最大的难题是如何使大模型的通用训练技术方法与法律行业相匹配。比如，要充分考虑如何体现法律数据的专业性、结构性、时效性等特点，如何利用已有的法律知识体系和图谱来增强训练，如何进行法律专业性和内容安全性评测等。

为解决这些问题，研发团队采取了多种方法。石鹏介绍，团队组织法律专业人员构建高质量的训练数据和指令集，并进行体系化模型迭代。团队还引入了最高人民法院“法信”平台全流程法律资源标注机制，以及历时10余年搭建的、包含18万法律知识体系编码的“法信大纲”，借此增强模型的知识理解和应用能力，提高模型的专业性和准确性。

“基座大模型的技术创新点，主要体现在探索如何将我国司法审判业务需求与大模型技术相结合，让大模型‘先通后专，通专结合’，最终实现对法律业务的可靠支撑。”清华大学计算机科学与技术系副教授刘知远说。

重塑审判业务流程

谈及研发建设基座大模型的意义，余茂玉说，这是落实总体国家安全观的具体措施，是推动法律行业新质生产力发展的创新引擎，也是助推审判工作现代化的有效路径。具体来说，基座大模型一方面推动现代科技与司法审判工作深度融合，积极探索人工智能技术赋能法律行业的路径，推进科技赋能公正司法，提升应用实效；另一方面促进规范人工智能技术应用，保障技术和数据内容安全可控，守住安全底线。

清华大学科研院院长、互联网司法研究院院长刘奕群认为，基座大模型是数字法院建设的重要基础设施，具有赋能并重塑审判业务流程的巨大潜力。

“未来，基座大模型有望在极大提升法律工作者工作效率的同时，更好地实现工作的规范化以及统一法律适用。”刘知远说，“随着大模型智能体技术的发展，每位法律工作者都可以拥有专属的智能体助理，用于承担机械的重复性劳动以及简单的创造性工作。”

刘知远认为，在不久的将来，法律智能体还可能会呈现多角色分工与多智能体协同的趋势。“当事人、律师、法官、法律学者都能通过属于自身角色的智能体助手，构建出一个更高质量、更高效的群体智能工作协同网络。”他说。

屏幕显示发光材料体系上新

科技日报讯（记者**洪敬谱**）11月21日，记者从维信诺科技股份有限公司（以下简称“维信诺”）获悉，由该公司研发的维信诺全新发光材料体系Foremost（以下简称“F1”）日前发布。

发光材料是直接影响屏幕显示效果的关键因素之一。F1在发光效率、器件寿命、视角色偏和低蓝光上，均实现了突破。相比原有的发光材料体系，F1发光效率提升10%，器件寿命提升22%，视角色偏改善50%，低蓝光改善10%。使用F1的屏幕品质更佳，低亮更细腻、高亮更均一。目前，F1已应用于一些手机旗舰产品等。

据悉，维信诺“有机发光显示材料、器件与工艺集成技术和应用”项目获2011年度国家技术发明奖一等奖。该公司也是首个主导制定柔性显示国际标准的中国企业。