着力构建可解释性模型

——“解读人工智能前沿技术趋势”

热点透视 rediantoushi

当人工智能(AI)机器人为你诊断疾病,你是否信任它的判断?当用AI模型来预测蛋白质的复杂结构,你能否理解其预测逻辑?如今,AI已经融入生活的方方面面,一个问题显得愈发关键:如何让人们理解并信任AI输出的结果。

构建可解释性模型是解决这一难题的重要途径。正如中国科学院院士张钹所言,如果不能建立可解释的AI理论,就会无法解决人们对AI技术的困惑和误解,这也将限制AI技术的发展。

什么是可解释性模型?它如何帮助人们理解AI?构建可解释性模型有哪些技术路径?笔者日前就上述问题采访了相关专家。

“看清”决策过程

“可解释性模型是指能够帮助人们理解其预测或决策过程的模型。可解释性意味着模型的运行逻辑和预测结果对人类是透明且易于理解的。”中国科学技术大学人工智能与数据科学学院教授王翔解释道。

王翔认为,可解释性模型应当具有四个重要特性。一是透明性,可解释性模型需要提供清晰的决策依据,让人们能“看清”模型根据输入数据做决定的全过程;二是一致性,可解释性模型的解释需要与人类已有知识一致,不与已知规律矛盾,例如医疗AI的诊断解释应与医学标准一致;三是语义性,可解释性模型的解释方式应易于理解,能通过自然语言或图示化形式呈现;四是因果性,可解释性模型能够明确说明驱动其预测的关键输入特征,用户能够通过调整输入数据,观察模型输出的变化,从而验证预测结果的可靠性和模型的局限性。

在AI影响力持续提升的今天,构建可解释性模型成了一项极为重要的任务。王翔认为,可解释性模型不仅能提升用户对AI系统的信任度,还有助于提升全社会对AI的接受程度,推动其在各领域广泛应用。

“可解释性模型还可以促进AI的公平性和伦理性,避免出现偏见或歧视。对于开发者来说,可解释性模型能提高AI的调试和优化能力,帮助开发者更好地理解和改进模型。可解释性模型也有助于提升AI安全性,能帮助监管机构和政策制定者更清楚了解AI技术,确保技术应用符合法律和伦理框架。”王翔补充说。

探索两种路径

事实上,并不是所有的模型都难以理解。王翔介绍,对于结构较为简单、透明的模型,例如线性回归或决策树等,人们往往能够直接理解输入与输出之间的关系。而对于复杂、高性能的神经网络模型来说,则需要借助相应的方法或工具来提高可解释性。

目前,提升模型可解释性的技术路径主要分为内在可解释性方法和事后可解释性方法。

内在可解释性方法是在模型设计阶段就融入可解释性需求,通过构建天然具备解释能力的模型,使其决策过程透明、直观。例如,深度学习模型中的注意力机制就是一种常见的内在可解释性技术,它通过显示模型关注的区域,帮助用户理解模型的行为。“透明性和实时性是内在可解释性方法的优势。这种方法特别适合对解释性要求高且需要快速响应的场景,如医疗诊断或金融决策。”王翔说。

事后可解释性方法则是在模型训练完成后,通过外部工具或算法分析模型的决策过程,无需修改模型本身。王翔介绍,事后可解释性方法的最大优势在于灵活性,它几乎适用于任何复杂的“黑箱”模型,但计算成本通常较高,尤其是单样本解释可能需要多次模型评估,不适合实时性要求高的场景。此外,事后可解释性方法仅能辅助分析模型行为,无法深入影响或改变模型本身的结构。

提升可解释性

在提升模型可解释性方面,学界正在积极进行探索。例如,王翔团队致力于建立可信赖的图基础模型。图基础模型是能够处理和分析各种复杂图数据的数学模型,它要处理的图数据可以是社交网络中的朋友关系、生物中蛋白质之间的相互作用、通信网络中的设备连接,甚至是人类大脑中的神经元连接等。传统的图神经网络通常通过聚合节点特征和拓扑结构信息进行学习,但实际图数据中往往包含冗余或噪声信息,这可能导致模型捕获与任务无关的特征。王翔团队提出了一种基于因果的内在可解释架构(DIR),有效剔除了干扰因素,保留了因果特征,大幅提升了模型的透明度和鲁棒性。

外滩大会上,观众在观看“蚁鉴2.0”的介绍 ■ 资料图

产业界的探索也取得了进展。例如,蚂蚁集团联合清华大学发布的AI安全检测平台“蚁鉴2.0”融入了可解释性检测工具。综合AI技术和专家知识,通过可视化、逻辑推理、因果推断等技术,“蚁鉴2.0”从完整性、准确性、稳定性等7个维度及20余项评估指标,对AI系统的解释质量进行量化分析,帮助用户更清晰验证与优化可解释方案。

商汤科技推出的自动驾驶大模型DriveMLM,则可以根据输入的信息给出自动驾驶每一步决策和操作时背后的逻辑和推理的原因,能够以自然语言解释自动驾驶的行为。也就是说,人们只需将图像、激光雷达信息、交通规则甚至是乘客需求“一股脑”丢给大模型,它就能给出驾驶方案,并解释为什么要这么开。商汤科技联合创始人王晓刚说,自动驾驶的挑战之一就是AI模型的可解释性不高,决策过程难以被人们理解,增强AI模型的可解释性能推动自动驾驶技术进一步发展和普及。

仍存技术难题

行业在构建可解释性模型上取得一定进展,但仍存在不少技术难点,需要研究人员进一步攻克。

首先,随着大模型性能、参数不断提升,其内部结构的复杂程度也在增加,这使得大模型内部决策机理变得越发难以理解,如何实现高度复杂模型的可解释性是重要问题。

其次,通常情况下,模型性能越强,可解释性就会越差,如何在模型的性能和可解释性之间找到合适平衡点,也是亟待解决的问题。王翔认为,未来,开发新的可解释性算法或将成为重要发展方向之一,新算法可以结合深度学习和逻辑推理等多种路径,实现模型高性能与解释性的更好平衡。

最后,目前没有统一标准来衡量模型的可解释性,如何准确评估并提升模型的解释性仍然是开放问题。“可以通过跨学科合作,结合认知科学、心理学等领域知识,共同定义和量化解释的标准,提高模型的可解释性。”王翔建议。 吴叶凡

新设备让粮食烘干更高效零破损

近日,中国农业科学院农产品加工研究所牵头研制的智能化零破损气吸循环式粮食烘干机在河北邢台、宁夏石嘴山、山西太原等地投入使用。

“该技术的核心是在粮食烘干过程中,让粮食随热风旋转上升,避免了机械撞击,将粮食烘干破损率降低到0.1%以下,实现了粮食烘干过程中的零破损。”1月2日,该技术研发带头人、中国农业科学院农产品加工研究所研究员邢福国在接受笔者采访时说。

能耗降低40%以上

“目前,我国粮食烘干技术虽然取得很大进步,但具有自主知识产权的核心技术仍十分匮乏。因此,我们下定决心攻克粮食干燥的核心技术。”邢福国说,传统烘干技术及设备存在三大瓶颈,即粮食破损率高、干燥速率低、烘干后粮食品质低。

邢福国面临的最大挑战,是如何跳出传统烘干机技术和机械结构,创造出全新的机械原理和机械结构。“除电机外,烘干机其余的设备、系统、核心部件、零部件都需要进行自主研发、设计、制模、试制、实验、制造。”邢福国说。

在无数技术路线、设计原理被否定后,“龙卷风”让邢福国和河北皓凯农业机械有限公司总经理梁凯产生了灵感。“被龙卷风卷起的杂物不就相当于粮食吗?”两人就这样开始了模拟实验,验证了气吸悬浮提升粮食的可行性。

提升烘干效率是团队要攻克的另一个难题。“烘干效率提升的关键是烘干全程使用热风。我们研发的烘干机使用了气吸技术,这种技术与传统烘干机使用的技术相比,热风与粮食接触范围提高了1倍,从而提高了热风的使用效率。”邢福国说,传统技术采用的是一次干燥,气吸技术则是二次干燥,粮食提升过程分为提升过程初级干燥和进仓二次干燥两个过程。

为此,团队创造了内外筒结构的全覆盖技术。“我们设置了内筒和外筒。内筒作为热风室,可让提升后的热风进入;内外筒之间是干燥室,可让热风自内向外穿过粮层热交换后排出机外,实现热风梯级高效利用。”邢福国说。

通过创新设计,团队提高了热风的使用效率和烘干效率,降低了能耗。“气吸循环式烘干机具备自然风风干功能,当自然环境温度高于30摄氏度,湿度低于40%时,所有粮食和油料作物各种种子均可自然风干,与传统烘干机相比,能耗降低了40%以上。”邢福国说。

该系列烘干机还采用模块化设计,配备智能控制和监测系统,实现了烘干操作的智能控制,温度、水分的实时监测,提高烘干精度和效率。

“我们团队还与农业农村部南京农业机械化研究所合作,揭示了大宗粮食的干燥特性,绘制了干燥动力学曲线,构建了基于该烘干机的热湿传递模型。”邢福国说。

适用于多种经济作物

随着我国城市化进程的加快、土地流转政策的推进、土地规模化经营的逐渐形成,粮食烘干机的市场需求量不断提升。种植500亩以上粮食作物的新型经营主体必须配备粮食烘干机才能保证粮食生产的安全和颗粒归仓。

如今,智能化零破损气吸循环式粮食烘干机不仅能处理玉米、小麦、水稻等三大主粮,还能烘干多种经济作物,如花生、花生果、油茶果、菜籽、大豆、油莎豆等。

“为解决对多种作物种子烘干适用性这一关键问题,我们创新研发了自动拨板下料控制系统。该系统设立了物料运转平台,并在平台内安装了拨板,该拨板可以将各种物料输送到扩大的下料口,实现一机多用。具体来看,我们通过采用智能控制调整风量和风速来实现各种物料烘干,同时增设电子调频。这样,不管颗粒大小、重量差异都可以实现提升、循环、干燥。”邢福国说。

基于该技术,由河北皓凯农业机械有限公司生产的系列粮食烘干机已经在多地使用。

数据显示,采用智能化零破损气吸循环式粮食烘干机烘干的小麦,品质优于自然晒干;烘干的稻谷,精米率比自然烘干提升2~3个百分点。“如玉米烘干后无破坏,将比使用烘干塔减损增收3%,每吨增值72元;玉米烘干后容重升一个等级,每吨可增收40元;烘干后玉米干净、光亮、无毛无小杂,每公斤可增收4分钱,每吨可增收40元。”邢福国算了一笔账,采用该技术,还可显著降低粮食真菌毒素含量,每吨综合增收约150元。

“应用该技术可降低粮食破损率约2%,相当于增收2%,这对于保障国家粮食安全具有重要意义。此外,该技术可降低粮食带菌量,降低真菌毒素污染水平,提升粮食品质,这对于保障国家食品安全和人民生命健康同样具有重要意义。”邢福国表示,他们将继续深入研究,创制出更多系列机型,满足不同产区用户需求。 马爱平

(上接A1版)5T设备都安装在铁道线路上,受天气、环境、温度等诸多因素影响,设备状态较为复杂。有时数据反馈不对了,翻山越岭到达设备安装点,却发现只是探测站机房空调故障,设备凉下来数据就正常了,虚惊一场;有时带着沉重的配件好不容易到了探测站点,却因为雨天雪天的原因,天窗点取消,无功而返;有时天窗临近结束,但还没有找到故障源头,急得一头汗……在跟点作业的过程中,李秀军深切感受到了和比武练习完全不同的作业状态。

特别是集中修的时候,作为施工负责人的李秀军,不仅要盯控施工安全,负责所有作业人员的住宿、吃饭、派车,还得注意作业人员的身体状态、心理状态。那段时间,年纪不大的他成了组里的主心骨,“有困难找秀军”是大家的口头禅。

2021年底,凭着扎实的基本功和丰富的实践经验,李秀军被聘任为技术科业务主管,主要负责全段范围内的5T及AEI设备的技术管理。

管理岗位与车间专职的工作性质大不相同,需要统筹协调、思路清晰。刚刚接手的李秀军有点摸不着头脑,面对每天的几十个甚至上百个电话,询问设备故障处理,询问厂家维修,常常是一团乱麻。一向“急性子”的他,突然“慢了下来”。

李秀军重新踏上“升级”之路,不断翻阅理论书籍,研究设备原理,探索故障处置方法,选取作业难度大、设备故障多的站点,逐个跟班作业,把大家的难题作为他的课题,记在随身携带的小本本上,时不时就要翻看翻看。随着“数字化”车辆段建设工作持续推进,他把全部精力投入到5T设备优化和促进重载技术创新发展工作上。

由李秀军制定的《铁路营业线施工安全管理简介》《铁路营业线施工标准化流程管理体系》《铁路营业线施工标准化指导手册》,实现了施工作业的全过程标准化管控;提出的大秦线THDS货车疑似抱闸三级预警标准,将一级预警数量压缩71%;提出THDS“抱死闸”预警标准,实现车轮抱死车辆的全路首例预报。同时,他组织先后完成TFDS沉箱盖板加热装置、TFDS相机盒雨刷装置、TFDS设备磁钢冗余、THDS盖板加热装置、THDS设备尖峰波形自动识别与提示等多项技术攻关成果,并全部投入运用;通过优化设备布局、完善设备功能及结构、深度挖掘设备探测数据潜在价值,实现了利用全路2%的5T设备,提供了全路近16%货车探测数据的能力,为大秦重载列车提供最优质的安全监控保障。

星光见证努力,时光见证不凡。担任技术科业务主管的4余年间,李秀军完成61套5T设备更新改造、152套设备大修、11个工程项目施工任务。李秀军凭借着“拼”的气势、“闯”的劲头和“钻”的精神,在实践中成长、在淬炼中前行,奋力投身大秦铁路重载事业发展,深刻诠释着新时代铁路人应有的本色,书写了不负韶华勇担当的青春之歌。

创新杂谈 chuangxinzatan

让科技成果不再『沉睡』该如何发力

■产健 张丽

2024年是上海国际科创中心建设十周年。2023年上海基础研究投入增长至2013年的3倍以上;科创板首发募资额和总市值位居全国首位;科技创新助力上海经济体量突破4万亿元大关……闯在基础研究“无人区”,创在体制机制改革生态田,育在未来产业新质生产力,上海加快从“建框架”向“强功能”推进,迈向科技强国建设的下一个十年。

科技创新是引领经济社会高质量发展的第一动力。唯有高效、顺畅地推动科技成果转化,才能让知识资本真正成为发展的不竭动力。

从全国来看,各地高校、科研院所和企业在成果转化方面取得了积极进展,但仍有许多优质成果因“机制不畅、政策不细、资金短缺”停留在实验室层面,成为“沉睡”的科技成果。如何在新时代拓宽科技成果转化渠道,让这些宝贵成果产生更大价值,已成为助力高水平科技自立自强、推动建设科技强国的关键命题。

提升供给质量,让原创成果更易“破茧成蝶”。成果转化能否落到实处,取决于“源头”能否持续输出高质量成果。结合产业痛点与科技前沿,探索“需求导向+潜力前瞻”双驱动模式。在科研立项时,产业与科研机构应建立定期对接机制,共享产业需求与技术发展的相关信息,对未来新兴领域预留一定资源。建立“迭代式”成果评价与孵化机制,引入动态评估,重视成果的早期市场验证与商用潜力,通过分层孵化为早期技术提供支持。此外,还应推动“企业前移+科研后移”耦合机制,让高校与企业在早期研发中同频互动,为技术方案设置“导向牌”和“红灯区”,加速科研成果的产业化进程。

完善政策体系,搭建从“纸面”到“落地”的衔接桥梁。科技成果转化涉及科研立项、经费管理、设备采购、成果评估、确权、收益分配等多个环节,只有环环相扣、扎实推进方能将纸面成果“落地生金”。当前,我国虽已出台多项促进成果转化的法律法规及配套文件,并在部分地区探索专门条例,但实践中,一些科研人员仍因缺乏明确操作指引,对政策执行的效果与可预见性心存顾虑。

为此,应从两方面发力:一方面,进一步简化、细化转化流程,充分运用现有改革举措,切实下放审批自主权,让高校、科研院所乃至科研人员“一看即懂、一操即成”。另一方面,加强监督问责和尽职免责相结合,给敢闯敢试者以明确制度支持,让他们敢于试错、勇于担当。只有打通“纸面”与“落地”之间的“肠梗阻”,才能让科研机构放心使用公共资源,把成果高效转化为现实生产力。

发挥专业力量,推动科技成果转化从“单打独斗”转向“生态共建”。成果转化常需技术与市场、科研与法律、实验室与投融资等多方对接。除政策助力外,更需打造协同联动的专业化服务生态。平台体系要系统布局。支持高校、科研院所建立“分级孵化”与“概念验证”中心,为早期技术提供小规模试验与市场检验;对成熟度较高的项目,支持与社会资本共建转化平台或加速器,提供从原型打磨到市场推广的全流程支持。人才与资本要深度协同。在技术经理人培养中,应兼顾知识产权布局、商业模式设计与投融资策略,打造“跨界型”技术经理人队伍;政府与社会组织也可通过“母基金+子基金”等模式,引入天使投资与“耐心资本”,配合专业评估机构严控风险,让潜力成果平稳跨越“死亡之谷”。

优化激励机制,让“能者多得”成为鲜明导向。科技成果转化的“主角”是科研人员。只有让他们敢想、敢拼、敢干,才会激发持续的创新创业热情。要在激励与约束上双管齐下,为科研团队、管理者等主体制定清晰的“利益航标”和“责任红线”。推广“技术股权+分期付费”模式,在成果确权、收益分配上充分赋权,鼓励拥有突破性成果的研发团队享有更大自主决定权,让科研人员真正享受创新的红利;并可试点“转化跟投制”,允许科研团队以部分科研经费或股权收益进行再投入,持续为后续研发与孵化赋能。在职称评审、项目审批、人才选拔中增设“转化业绩”独立指标,将转化成效与科研人员职业发展紧密挂钩,倒逼团队聚焦“真难题”攻关。同时,通过“尽职免责+责任约束”为基础研究、中试等阶段的失败或试错留足空间,赋予人员更多创新探索的勇气与魄力。

强化协同互动,让“沉睡”科技资产找准用武之地。科技创新从来都是学界、业界、政府多方深度融合的系统工程。唯有各方协力,潜能巨大的科研成果才能真正“物尽其用”。应倡导“跨主体共同体”思维。支持组建产业联盟和跨学科专家团队,完善“揭榜挂帅”机制,鼓励企业高管、技术骨干与高校科研团队组建柔性项目组,开展联合攻关和示范推广。深化“场景牵引+资本撬动”双轮驱动。在重点区域或行业扩大公共应用场景,允许技术团队快速“试水”并迭代升级;鼓励金融机构以“产业链金融+成果担保”模式介入,为潜在商业价值高的技术提供专项债或风险分担基金,挖掘更多耐心资本“投早、投小、投硬”,让更多高水平科研成果“破茧成蝶”、为社会生产生活带来福祉与利好。

“黑科技”为公园添“智慧”


当前,深圳正在全力推进人工智能先锋城市建设。智慧跑道、室外智能健身房、智能淋浴间、无人驾驶观光游览车、无人机送外卖……在深圳的各大公园里,新鲜有趣的人工智能应用随处可见,“黑科技”为公园增添“智慧”。图为在深圳莲花山公园拍摄的环卫机器人。 ■ 新华社记者梁旭摄