

人工智能价值对齐的向善困境与纾解路径

■广西师范大学 蒋裕澄、刘彦萍、冯丽茹

人工智能价值对齐是确保技术安全可控的核心议题,“向善”是价值对齐的伦理指向。通过剖析价值对齐在价值基准模糊、技术逻辑的结构性错位、治理结构滞后与场景适配缺失四个层面的向善困境,揭示向善困境的生成根源在于价值基准的社会性分化、技术权力的资本偏向、工具理性的过度凸显与系统演化的涌现性约束,并提出确立分层价值共识框架、推进可解释技术与多方评测机制、健全法治化治理体系与嵌入人类在场的决策机制的四维纾解路径。

一、人工智能价值对齐向善的多重困境

人工智能价值对齐的“向善”伦理指向在实践过程中遭遇多重困境,这些困境体现于价值基准、技术逻辑、治理结构与场景适配四个层面。

(一)价值基准模糊

向善的实践前提在于明辨“善”的具体所指,而人们至今未能达成普遍共识。人类价值极为复杂且庞大,由于个体所处环境各异,每个人形成的价值观也千差万别,因此并不存在统一的标准。功利主义以最大幸福为尺度,义务论以普遍法则为准绳,德性伦理学以完善人格为目标,诸范式间不存在通约的元标准。价值判断随历史条件变迁而不断调整,价值对齐的规范性基准也始终处于流变之中。

(二)技术逻辑的结构性错位

在技术实现层面,价值对齐主要沿两条路径展开。自上而下的规则驱动路径将伦理原则编码为算法符号,造成事实与规范之间的逻辑断裂。从“如何行为”的描述性命题,无法推导出“应当如何行为”的规范性要求,编码过程必然损耗伦理判断固有的智慧。自下而上的数据驱动路径中,训练数据固有的社会偏见被模型不加批判地内化为“人类偏好”,价值对齐因而退化为既有偏见的算法性强化。而当混合进路试图引入规则加以矫正时,规则与数据之间的冲突裁决机制却始终缺位。

(三)治理结构滞后

人工智能系统产生价值偏离时,责任归属难以确定。技术开发者将责任归因于算法的自主学习特性,系统部署者将责任归咎于自动化决策的超预期表现,算法本身不具备承担法律责任的资格。“技术先行、伦理滞后”的结构性失衡中,技术演进速度持续领先制度更新节奏,现有治理框架侧重于数据安全等技术议题,缺乏系统性的伦理评估指标。

(四)场景适配缺失

不同应用场景所要求的价值序列存在差

异。医疗场景中生命健康权往往居于优先序列,而内容推荐场景中言论自由与信息真实性之间的张力则成为主要矛盾。一种统一的价值对齐方案无法通约于所有场域,而针对每个场景单独设定伦理约束又将导致标准碎片化与监管套利。此外,开放世界中的伦理情境具有无穷尽性,训练数据的静态属性与真实世界价值的动态演化之间形成了一种“规范开放性”的张力,使得任何预先编码的伦理规则都面临情境适用性的局限。

二、人工智能价值对齐向善困境的生成根源

人工智能价值对齐的向善困境植根于社会生产方式、资本权力、技术权力与系统演化动力学之中。

(一)价值基准的社会性分化

价值观念根植于特定的社会经济基础,是对特定社会生产方式与利益关系的反映。由于GAI的研发、设计与运行高度依赖海量算力、庞大资本与顶尖人才,资源的集中垄断便将对齐何种价值以及如何定义对齐的决定权赋予了少数掌握强大算法与算力的科技巨头与数字资本寡头。不同社会群体不仅在价值偏好上存在差异,在根本利益上更存在不可调和的对立。回避对价值对齐的社会根源追问,其成效便只能停留在表象性的局部调整。

(二)技术权力的资本偏向

算法从特征提取到目标函数设定,每一环节涉及价值判断的技术表达。在资本逻辑下,用户活跃度、商业转化率等效率目标直接关联资本增值而获得优先地位,公平性、可解释性等公共价值因无法直接转化为利润而被系统性边缘化。并且资本运作往往以“技术中立”加以遮蔽,使得公共价值的系统性退让被包装为纯粹的技术效率追求。

(三)工具理性的过度凸显

“当工具理性被推崇到极致,价值对齐实践便‘异化’为一场围绕利益指标的核心技术竞赛。”以效率与可计算性为核心的工具理性,因其量化特征便于与资本逻辑及技术指标相耦合,在资源配置中获得优先地位;与之相对的价值理性,所强调的伦理与人文关怀等准则,因难以度量而始终被边缘化。

(四)系统演化的涌现性约束

当模型的参数规模达到数千亿乃至万亿级别时,其内部决策过程的复杂程度已远超人类理解力的上限。即便我们能够在训练阶段通过人类反馈强化学习等方式注入价值偏好,但模型在对开放世界中无限多样的输入组合时,其行为空间仍然存在着大量未经测试且不可预测的“暗域”。同时,价值对齐所依赖的“训练、反馈、修正”循环,在模型的涌现行为面前始终处于滞后状态。模型规模与能力持续增长的同时,约束其行为的各种技术规约机制的相对效力却在衰减,价值对齐有效性反而越难保障。

三、人工智能价值对齐向善的纾解路径

针对上述困境及其根源,纾解路径需从价值、技术、治理、场景四个维度协同推进,形成系统化的应对方案。

(一)确立分层价值共识框架

针对价值基准模糊及其社会根源,可确立分层价值框架。以马克思主义关于人的自由全面发展为规范性标准,以阿玛蒂亚·森的可行能力框架为操作尺度,将价值目标转化为可评价的指标体系,具体表现为考察系统运行是否有利于拓展人的自主性、丰富人的能力结构,增进人的实质性福祉。规范性标准与操作尺度构成价值目标与评价指标的互补关系。在操作层面,价值对齐需与现行算法分级分类监管相衔接。具体而言,以维护人类基本尊严作为不可逾越的伦理底线,对应《生成式人工智能服务管理暂行办法》中的强制性审查规定。在底线之上,则建立包容性价值差异空间,允许多元价值偏好通过开放参与机制展开对话与协商。同时,共识体系保持动态性,持续吸纳新诉求。当价值分歧无法协调时,以人的自由全面发展为最终仲裁判断。

(二)推进可解释技术与多方评测机制

针对技术逻辑错位及其工具理性过度凸显的资本偏向根源,以提升价值决策过程的透明性作为技术研发的核心方向。算法作出具有伦理后果的决策时,应能追溯其价值偏好的来源、还原其权衡推理的过程,说明其判断的理据。可解释性打破资本以技术复杂性回避问责的策略。建立多方参与的价值评测机制,将独立第三方伦理审计嵌入算法备案年审制

度,评测数据广泛吸纳不同群体价值表达,防止商业巨头垄断解释权。

(三)健全法治化治理体系

针对治理迟滞及其制度化根源,出路在于从主体自觉转向刚性制度约束。事前确立价值影响评估法定程序,评估报告须经独立伦理委员会审查;事中建立常态化伦理监测机制;事后明确算法损害法律责任归属,由获益主体承担举证责任倒置。将中华人民共和国科学技术部发布《新一代人工智能伦理规范》中的原则转化为可度量、可审计、可追责的技术标准,完成从软性倡导到刚性约束的升级。同时,必须正视资本增值逻辑与人类道德秩序的关联。将算法公平性评分作为企业享受税收优惠与政府数据公开合作的前置准入条件,平衡工具理性的过度凸显。

(四)嵌入人类在场的决策机制

针对场景适配缺失及其涌现性约束根源,在关键伦理决策环节保留人类的主体地位。在自动驾驶、自动诊疗等高风险场景中,当AI系统对当前情境的概率判断置信度低于预设阈值时,系统必须强制将最终决策权交还人类。这是对“在价值判断领域最终责任主体只能是人类”原则的制度确认。同时,为确保人类决策者不至于因信息过载或时间压力而做出次优判断,AI系统提供多方案评估报告,人类决策者在充分知情的基础上做出最终选择,并对该选择承担说明义务与事后可追责性。决策记录应纳入审计范围,以形成可回溯的问责链条。

四、结语

人工智能价值对齐的“向善”之路,既是技术命题,更是关乎人类文明走向的价值命题。本文在剖析四重困境及其根源的基础上,提出四维协同的纾解路径。当前,全球人工智能治理格局正加速重塑,我们应立足中国实际,推动构建兼具中国特色与包容性的人工智能伦理体系,并在持续迭代优化中使智能时代真正实现科技向善、造福于人。

本文系2026年广西师范大学马克思主义学院研究生教育创新计划项目“人工智能价值对齐的向善困境与纾解路径”(项目编号:XYCS2026151)的研究成果。

新质生产力赋能宁国市经济高质量发展

■中共宁国市委党校 缪思曼

党的二十届四中全会对发展新质生产力作出了重要部署,强调要加快高水平科技自立自强,引领发展新质生产力,建设现代化产业体系。近年来,宁国市深入实施创新驱动发展战略,加快培育新质生产力,并与县域传统生产力的新质化转型协调并进,赋能宁国市经济高质量发展。

一、新质生产力赋能宁国市经济高质量发展的现实困境

(一)企业创新能力提升的瓶颈制约

一是中小企业创新意愿和能力不足。宁国市虽拥有中鼎、保隆等领军企业,但多数中小企业依然存在研发投入不足的问题。2025年规模以上工业企业平均研发强度为2.36%,作为长三角新兴产业聚集地重要支点城市,与周边同级别县市仍存在差距,如江苏宜兴近两年这一数据都超过3%;且技术人才匮乏,本科以上学历技术人员占比较少,对自建创新平台、关键核心技术攻关的参与度较低。部分企业仍依赖传统生产模式,对数字化改造样板、绿色技术应用等投入意愿不强,存在“等靠要”心态。

二是科技成果转化“最后一公里”梗阻。尽管合肥工业大学宁国市汽车产业研究院等校地平台已落地,但高校科研成果与企业实际需求匹配度不足,实验室技术向产业化产品转化周期长(平均2年至3年)。市科技创新综合服务中心的检验检测、中试熟化功能尚未完全激活,缺乏专业技术经纪人队伍,2025年全市技术合同成交额仅占GDP的1.1%,低于全省县域平均水平。

(二)新兴产业与未来产业培育的现实挑战

一是产业链协同配套能力有待加强。汽车零部件产业虽已形成集群,在密封件、减震

件等领域具备竞争力,在“三电”核心领域已实现部分关键零部件的本地化生产,形成了较为完整的配套产业链,但在高端核心部件(如动力电池电芯、电机总成、电控系统等)方面仍存在较高的对外依赖度。中德智造产业园与整车企业特别是新能源车企的战略合作尚处于初期,供需对接活动的实际订单转化率不高,未能形成“龙头企业+配套企业”的共生生态。

二是未来产业技术与市场协同发展不足。低空经济、“人工智能+”等领域面临技术成熟度不足与市场需求培育滞后的双重挑战。青龙湾通用机场的飞行培训业态年培训量不足300人次,无人机零部件招引尚未落地实质性项目;“人工智能+”场景仅在少数领军企业试点,AI技术对生产效率的提升幅度不足。

(三)区域协同与要素保障的深层矛盾

一是长三角一体化中的“虹吸效应”与配套困境。在与合肥、南京和杭州都市圈协同发展,宁国面临高端人才、优质项目被核心城市虹吸的压力,近五年引进的硕士以上人才中,30%流向省会或苏浙地区;同时,跨区域要素流动的硬件支撑存在协调障碍,如“宁国—临安”城际铁路这类关键交通设施的规划需省级层面协调,一定程度上制约了人才等要素的跨区域流动,进一步加剧配套困境;跨区域生态补偿机制尚未明确资金分摊比例,制约区域合作纵深推进。

二是土地与资金要素约束趋紧。虽然“亩均论英雄”改革释放了土地潜能,但全市工业用地亩均税收仍低于长三角先进县(市)平均水平,存在土地供而未供的情况,难以满足新兴产业项目落地需求。政府投资基金虽设立5亿元规模,但首投、领投作用尚未充分发挥,社会资本参与度低,耐心资本培育缓慢。

二、新质生产力赋能宁国市经济高质量发展的路径措施

(一)以科技创新为驱动,夯实新质生产力发展根基

一是精准化建设科技创新平台。一方面,差异化布局园区,聚焦汽车零部件、耐磨铸件等主导产业,在经开区建设专精特新产业园,设立“初创企业孵化区”和“成长型企业加速区”。另一方面,完善场景化服务配套,建议搭建共享实验室,整合领军企业和高校的检测设备,以远低于市场价格向中小企业开放,并进一步激活宁国特色产业研究院的公共服务属性;定期举办技术需求对接会,设立成果转化服务补贴制度,覆盖技术评估、专利代理等环节。

二是鼓励企业加大研发投入。通过政策激励,对新兴技术领域研发活动给予税收减免、财政补贴,降低企业研发成本,借助市场竞争机制倒逼企业提升研发投入,增强技术竞争力;通过管理优化,引导企业建立研发预算管理体系,加强研发过程监督与评估,提高资金使用效率。

三是推动科技成果转化与应用。建立企业创新能力评价体系,通过要素保障倾斜,对高创新能力企业优先保障用地、用能、金融等资源;基金与项目联动,以重大科技攻关项目突破关键技术,通过产业投资基金加速成果产业化;通过产学研嫁接,筛选高校院所科研成果与企业需求相匹配,组织靶向洽谈,推动科技成果转化落地。

(二)以产业升级为抓手,构建现代化县域产业体系

一是聚焦新兴产业培育。精准选型,依托汽车零部件产业基础,重点培育新能源汽车“三电系统”及智能底盘系统,如推动中鼎电池

冷却系统、天成电气新能源电容等核心产品扩产,打造全国重要的新能源汽车零部件产业基地。

二是实施传统产业智能化改造。针对耐磨铸件产业开展“环保+智能”双提升工程,对铸造企业实施烟尘收集系统改造和数字孪生产线建设,实现熔炼参数实时优化;完善人才与管理配套,每年开展数控操作工培训,联合安徽材料工程学院开设工业机器人操作专班;建立耐磨铸件产业云数据库,整合规上企业生产数据,实现订单、仓储、物流信息实时共享。

(三)优化要素保障,破解要素瓶颈

一是构建“三位一体”科技金融体系。财政资金引导,设立科技创新专项资金,对企业自建省级以上创新平台给予一次性奖励;金融产品创新,推广“知识产权质押贷”“研发投入贷”,对科技型企业贷款实行财政贴息;资本市场赋能,完善上市企业后备资源库,实施动态化和精细化管理,对科创板上市企业给予相应奖励,推动保隆、亚新科等企业登陆北交所。

二是实施高层次人才引育与产教融合计划。靶向引才,对“科技副总”等高层次人才给予年薪补贴,进一步完善人才居住、医疗、子女入学等方面的权益;本土育才,扩大“企业冠名班”规模,推行工学交替的培养模式,企业为学生提供实训津贴。

(四)深化区域协同,优化营商环境

一是聚焦新能源新材料打造“长三角创新飞地”与“产业协作联盟”。建议牵头组建“皖东南汽车零部件产业联盟”,联合宣城、芜湖等地企业共享研发设备,让高校在宁国研究院的检测平台向联盟企业开放,收费相应减免。

二是推进“高效办成一件事”改革。对工业项目推行“标准地+承诺制”,实现拿地即开工,最大程度压缩审批时限;建立“企业诉求智能响应平台”,整合12345热线、政务服务网等渠道,实现“诉求提交、分办、督办、反馈”全流程闭环,缩短办理时限。