ICS

T/GXDSL

团　　体　　标　　准

T/GXDSL —2026

人工智能大模型行业应用伦理治理指南

Ethical Governance Guidelines for Industrial Applications of Large Artificial Intelligence Models

（工作组讨论稿）

（本草案完成时间：2026-01-29）

2026 - - 发布　　2026 - - 实施

广西电子商务企业联合会　发布

目　　次

前　言

本文件依据GB/T 1.1-2020《标准化工作导则第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由广西产学研科学研究院提出。

本文件由广西电子商务企业联合会归口。

本文件起草单位：

本文件主要起草人：

本文件为首次发布。

人工智能大模型行业应用伦理治理指南

1 引 言

为深入贯彻落实国家人工智能发展战略及科技伦理建设要求，规范人工智能大模型（以下简称“大模型”）行业应用行为，防范化解技术应用引发的伦理风险、社会风险与安全风险，保障公民合法权益、社会公共利益与国家安全，依据《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》《生成式人工智能服务管理暂行办法》《科技伦理审查办法（试行）》等现行法律法规、部门规章及国家相关政策、标准，制定本指南。本指南作为国家层面人工智能伦理治理体系的重要组成部分，旨在为境内大模型研发、提供、应用及相关治理活动提供统一的伦理遵循、治理框架与实施规范，推动构建安全、可信、负责任的人工智能产业生态，支撑数字经济高质量发展。本指南兼顾行业自律与监管协同，相关组织应结合自身业务特点与应用场景实际，细化落实各项治理要求。

2 范围

本文件明确了大模型在各行业应用全生命周期（含数据获取与管理、模型研发与测试、部署与运营、安全监测与处置、评估与持续改进）应遵循的伦理原则、治理体系要求及具体实施规范，覆盖基础大模型、行业大模型及专用大模型的各类应用场景。适用于在中华人民共和国境内从事大模型研发、生产、提供、服务、应用、评估等相关活动的各类组织（以下统称“相关组织”），包括企业、高等院校、科研机构、第三方服务机构等。国家互联网信息办公室、工业和信息化部、科学技术部、教育部、卫生健康委员会、金融监督管理总局等行业监管部门，可将本文件作为监管执法、行业管理、合规检查的重要参照依据；行业用户在选用大模型产品与服务时，可参照本文件开展合规性与伦理性审查评估。

3 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 35273-2020 《信息安全技术个人信息安全规范》

GB/T 37988-2019 《信息安全技术数据安全能力成熟度模型》

GB/T 41867-2022 《信息技术人工智能术语》

GB/T 42571-2023 《信息技术人工智能机器学习模型与系统质量评估规范》

GB/T 44796-2023 《信息技术人工智能伦理风险评估规范》

《中华人民共和国网络安全法》（2017年6月1日起施行）

《中华人民共和国数据安全法》（2021年9月1日起施行）

《中华人民共和国个人信息保护法》（2021年11月1日起施行）

《中华人民共和国反垄断法》（2022年8月1日起施行，2024年修订）

《互联网信息服务算法推荐管理规定》（国家互联网信息办公室等四部门，2022年3月1日起施行）

《生成式人工智能服务管理暂行办法》（国家互联网信息办公室等七部门，2023年8月15日起施行）

《科技伦理审查办法（试行）》（科学技术部等十部门，2023年12月1日起施行）

《数字中国建设整体布局规划》（中共中央、国务院，2023年2月印发）

《新一代人工智能发展规划》（国务院，2017年7月印发）

《新一代人工智能伦理规范》（科学技术部等六部门，2024年修订）

4 术语和定义

GB/T 41867-2022、GB/T 44796-2023界定的以及下列术语和定义适用于本文件。

4.1 人工智能大模型

采用深度学习等核心技术构建，具备大规模参数（通常达到十亿级别及以上规模），可处理跨领域复杂任务，能够生成文本、图像、音频、视频等内容或提供决策辅助支持，涵盖基础大模型、行业大模型及专用大模型的人工智能模型统称。

4.2 行业应用

将大模型技术与特定行业领域（含金融、医疗、教育、制造、政务、传媒、交通、能源、安防、文旅等）的业务场景深度融合，提供产品、服务、解决方案或支撑业务决策的各类应用活动，包括线上线下各类交互场景。

4.3 伦理治理

为落实国家科技伦理要求，确保大模型技术开发与应用符合法律法规、社会公序良俗、道德伦理规范及人类共同价值观，由政府监管部门、相关组织、社会公众、行业协会等多方协同参与，建立并实施的组织架构、政策制度、流程机制、监督考核、责任追究等管理活动的统称。

4.4 可解释性

能够以人类可理解的方式，清晰描述大模型的决策逻辑、推理过程、输出结果依据及能力边界的程度，满足监管合规、风险管控、用户知情权保障及责任追溯等核心需求。

4.5 算法偏见

因训练数据缺陷（含样本不均衡、代表性不足）、算法设计偏差、场景适配不当、人为干预失当或环境因素影响等，导致大模型输出结果对特定民族、种族、性别、年龄、宗教信仰、健康状况、社会地位、地域等群体产生不公平、不公正对待的系统性偏差。

4.6 高风险应用场景

依据法律法规及国家政策规定，可能对国家安全、公共安全、社会稳定、公民生命财产安全或重大合法权益产生显著影响的大模型应用场景，主要包括但不限于金融信贷审批、医疗辅助诊断与治疗建议、司法辅助裁判、自动驾驶关键决策、公共安全防控、教育考试评价、重要政务服务等，具体场景由行业监管部门结合实际动态划定。

4.7 伦理风险评估

按照既定流程与标准，对大模型全生命周期各环节可能引发的伦理风险进行识别、分析、研判与分级，提出风险防控措施与应对方案的活动。

5 总体伦理原则

5.1 以人为本

坚持以人民为中心的发展思想，将保障人类福祉、尊重公民基本权利与自由置于核心地位。大模型技术研发与应用应服务于经济社会发展和民生改善，契合社会公序良俗，严禁利用技术损害公民人格尊

严、侵犯人身财产权利，坚决防范技术异化、滥用带来的社会危害，兼顾特殊群体（含老年人、残疾人、未成年人等）的使用需求与权益保障。

5.2 安全可控

将安全可控作为大模型发展的首要前提，全面落实总体国家安全观。强化技术安全、数据安全、内容安全与系统稳定性管控，建立全生命周期风险防控机制，具备风险识别、预警、处置、溯源与应急响应能力，确保大模型研发、部署、运营全过程始终处于有效管控范围内，坚决防范系统性、全局性安全风险。

5.3 公平公正

坚守公平正义的价值导向，全面排查、识别并主动消除算法偏见与歧视。保障不同群体平等享有技术发展成果，关注数字鸿沟问题，推动技术普惠应用，维护社会公平正义与公共利益，不得因任何法定禁止的歧视性因素对用户实施差别对待。

5.4 透明可释

提升大模型技术与应用的透明度，强化可解释性建设。在符合法律法规要求及商业秘密、知识产权保护原则的前提下，向监管部门、用户及相关利益方如实披露模型基本原理、能力边界、局限性、决策依据、数据处理规则及潜在风险，保障公众知情权与监督权，复杂场景应提供分层、易懂的解释内容。

5.5 隐私保护

严格遵守个人信息保护与数据安全法律法规，落实数据安全主体责任。在数据收集、存储、处理、传输、共享、销毁等全生命周期，采取必要的安全防护措施，遵循“最小必要”原则收集和使用个人信息，严禁非法收集、滥用、泄露、篡改用户隐私与敏感数据，保障数据安全与个人隐私不受侵害。

5.6 责任明确

构建“政府监管、企业主体、多方协同、权责统一”的责任体系。明确研发者、提供者、使用者、评估机构、服务机构等各方在大模型全生命周期中的伦理责任与法律责任，建立可追溯、可问责的责任追究机制，确保违法违规行为得到及时识别与惩处。

5.7 敏捷治理

适应大模型技术快速迭代、应用场景持续拓展的发展特点，建立动态、灵活、高效的伦理治理机制。结合技术创新进展、行业发展实践、监管政策更新及社会伦理共识，及时优化调整治理要求，精准识别和应对新型伦理风险，实现技术创新与风险防控的动态平衡。

6 治理体系要求

6.1 组织与领导

6.1.1 相关组织的最高管理者是本组织大模型伦理治理的第一责任人，对伦理治理工作的有效性、合规性承担最终责任，确保伦理治理所需资源（含人力、财力、技术）足额投入，推动治理制度落地执行与责任层层落实。

6.1.2 相关组织应设立专门的伦理治理机构（如伦理委员会、治理工作组等），成员应涵盖技术、伦理、法律、行业应用、公共事务等多领域专业人才，其中外部独立专家占比应不低于30%；高风险应用相关组织的伦理治理机构应接受行业监管部门的指导与监督，可根据业务规模与风险等级合理调整机构设置与人员配置。

6.1.3 伦理治理机构履行以下核心职责：制定并修订伦理治理策略与制度流程；组织开展全生命周期伦理风险评估与专项审查；监督伦理原则与治理要求落地执行；受理伦理投诉与争议处置；定期向最高管理者和行业监管部门报告治理工作情况；开展伦理治理能力建设相关工作。

6.2 制度与流程

6.2.1 相关组织应制定成文的《人工智能伦理治理章程》，明确伦理原则、组织职责、治理流程、问责机制与奖惩办法，按要求报行业监管部门备案；章程内容应符合本指南及国家相关法律法规、政策标准要求，并根据政策变化、技术发展及实践反馈及时修订完善。

6.2.2 建立覆盖项目立项、研发测试、部署运营、退役注销全生命周期的伦理风险评估与审查流程；对高风险应用场景，必须开展专项伦理审查，审查结果作为项目推进的前置条件，未经审查或审查不合格的，不得开展相关研发与应用活动；审查标准应结合行业特点细化制定。

6.2.3 建立健全内部举报与投诉渠道，明确举报方式、处置流程与反馈时限，对收到的伦理违规举报实行“首接负责制”，在规定时限内完成调查、响应与处置，并严格保护举报人的合法权益，严禁打击报复举报人。

6.3 能力建设与培训

6.3.1 相关组织应建立常态化伦理培训机制，定期对研发人员、产品经理、运营人员、审核人员、管理人员等开展人工智能伦理、法律法规、监管要求及本组织治理制度的培训，提升全员伦理意识、风险识别能力与合规操作水平。

6.3.2 培训覆盖率应达到100%，关键岗位人员（如伦理审查人员、高风险场景运营人员、数据安全管理人员等）每年接受专项培训学时不少于行业监管规定的最低标准，培训考核结果纳入个人绩效考核；相关组织应将培训计划、实施情况及考核结果定期报送行业监管部门，留存培训记录备查。

7 数据管理伦理要求

7.1 数据来源合规

7.1.1 训练数据、微调数据及推理数据的收集、使用必须具备合法依据，严格遵守《数据安全法》《个人信息保护法》等法律法规及相关标准要求，严禁使用来源非法、侵犯知识产权、包含违法不良信息或危害国家安全、公共利益的数据。

7.1.2 建立健全数据来源评估与审核机制，对数据的合法性、代表性、公平性、时效性、完整性进行全面审核，留存审核记录备查，留存期限不少于3年；涉及跨境数据收集、传输、使用的，应符合国家数据出境安全管理相关规定，完成必要的安全评估与备案程序。

7.1.3 收集个人信息的，应遵循“合法、正当、必要、诚信”原则，明确告知收集目的、方式、范围及使用规则，取得个人单独同意或符合法律规定的其他情形（如法定职责、公共利益等）；严禁以强制授权、捆绑授权、默认同意等方式规避用户同意义务，不得过度收集个人信息。

7.2 数据质量与偏见管控

7.2..1 建立完善的数据质量管控体系，对训练数据进行清洗、去重、校验、脱敏等处理，确保数据真实、准确、完整、有效，杜绝虚假数据、无效数据、冗余数据进入训练流程，保障数据质量满足模型研发与应用需求。

7.2.2 对训练数据开展全面的偏见检测与标注，覆盖民族、种族、性别、年龄、宗教信仰、健康状况、社会地位、地域等关键敏感属性，偏见检测覆盖率应达到100%，形成完整的偏见检测报告，明确偏见类型、来源及影响范围。

7.2.3 采取数据增强、样本平衡、算法修正、人工校验等技术与管理措施，主动降低数据集中的已知偏见；关键敏感属性的数据分布偏差度应合理控制在较低范围，高风险应用场景应进一步严格管控，确保偏见风险处于可接受水平。

7.3 数据安全与隐私保护

7.3.1 落实数据分级分类管理要求，根据数据重要程度、敏感级别划定保护等级，对核心训练数据、敏感个人信息、模型核心参数等实行最高级别的安全保护，采取加密、脱敏、访问控制、安全审计、行为监控等技术措施，符合GB/T 37988-2019、GB/T 35273-2020等相关标准要求。

7.3.2 在模型训练与推理过程中，优先采用联邦学习、差分隐私、隐私计算等隐私增强技术，最小化原始数据暴露风险；严禁未经授权向第三方共享训练数据、用户敏感信息或模型核心参数，确需共享的，应进行安全评估并采取必要的安全防护措施。

7.3.3 建立健全数据安全应急处置机制，针对数据泄露、篡改、丢失、滥用等安全事件制定专项应急预案，明确应急响应流程、责任分工与处置措施，定期开展应急演练，确保事件发生后能够快速响应、有效处置，并按规定及时向行业监管部门及相关利益方报告。

8 模型研发伦理要求

8.1 算法设计公正

8.1.1 在模型架构设计、算法选择与优化阶段，充分评估算法的伦理影响与社会后果，优先选择有助于促进公平、减少偏见、防范风险的技术方案，严禁设计具有歧视性、排他性、危害性的算法模块。

8.1.2 在关键算法模块中嵌入偏见监测与缓解机制，建立算法偏见动态评估模型，定期对算法公平性进行测试、验证与优化，确保高风险应用场景算法无系统性偏见，算法输出结果符合公平公正原则。

8.1.3 严禁利用算法实施垄断、不正当竞争等违法行为，不得通过算法操纵市场价格、侵犯消费者权益、损害同行合法利益或实施其他危害市场秩序与公共利益的行为。

8.2 可解释性增强

8.2.1 加大可解释性技术研发与应用投入，针对高风险应用场景，必须具备事前预警、事中干预、事后解释的全流程可解释能力；对金融信贷、医疗诊断、司法辅助等关键决策场景，应向用户、监管部门及相关方提供清晰、易懂、可验证的决策解释，保障决策过程可追溯。

8.2.2 建立模型研发全过程追溯机制，完整记录和保存模型各版本的训练数据来源、关键参数设置、架构设计方案、性能评估结果、伦理审查意见、优化修改记录等信息，留存期限不少于5年，确保研发过程可追溯、可审计、可复核。

8.3 安全与鲁棒性

8.3.1 研发阶段应开展全面的安全测试与验证，包括对抗性测试、漏洞扫描、压力测试、边界测试等，评估模型抵御恶意输入、诱导生成有害内容、规避安全管控、对抗性攻击等风险的能力，及时发现并修复安全漏洞与隐患。

8.3.2 内置多层级内容安全过滤机制，结合规则引擎、智能分类、人工审核等多种方式，对模型生成内容进行严格管控，违法不良信息过滤准确率应满足行业监管与安全运营要求，高风险应用场景应达到更高的过滤标准，确保输出内容合法合规、安全可控。

8.3.3 建立模型版本管理机制，对模型的开发、测试、升级、迭代、退役等全流程实行严格的安全评估与审批管理，严禁未经安全测试、伦理审查的模型版本投入使用，留存版本管理全流程记录备查。

9 部署应用伦理要求

9.1 应用准入与评估

9.1.1 大模型部署应用前，应开展全面的伦理影响评估与合规审查，评估内容包括但不限于：对国家安全、公共安全、社会稳定的影响；对用户权益、个人隐私的保护情况；对行业生态、市场秩序的影响；算法公平性与偏见风险；环境资源消耗；应急处置能力等；评估报告应按要求报行业监管部门备案，高风险应用场景评估报告需经第三方机构复核。

9.1.2 严格落实应用准入管控要求，严禁在法律法规明令禁止的领域（如制毒、赌博、恐怖主义宣传、危害国家安全等）及可能严重危害国家安全、公共安全、社会稳定、公民生命财产安全的场景中部署应用大模型；高风险应用场景应取得行业监管部门的准入许可或备案批复后，方可开展应用活动。

9.1.3 建立应用场景动态评估与风险复评机制，对已部署的应用定期开展伦理风险复评，频次根据应用风险等级合理确定，高风险应用场景复评频次不低于每半年一次；发现重大风险隐患的，应立即暂停应用并组织整改，整改合格并经复核后，方可恢复运行。

9.2 用户告知与知情同意

9.2.1 向终端用户清晰、显著告知其所使用的服务由人工智能大模型驱动，不得隐瞒人机交互身份，严禁以人工智能生成内容冒充人类创作内容误导用户，确保用户能够清晰区分人机交互场景。

9.2.2 以通俗易懂的语言和显著的方式，向用户披露服务的基本原理、主要功能、能力边界、局限性、潜在风险及用户数据的处理规则、使用范围、安全保护措施等信息，披露内容应置于服务界面显著位置，便于用户查阅与理解，不得使用模糊、晦涩的表述规避告知义务。

9.2.3 在涉及敏感个人信息处理、重大权益影响（如金融决策、医疗诊断建议、就业评估等）等场景，必须获取用户的明示同意，同意方式应符合法律法规要求，严禁以默认同意、捆绑授权、强制授权等方式规避用户同意义务，用户有权随时撤回同意并注销相关服务。

9.3 人机协同与责任保留

9.3.1 在司法辅助、重症诊疗、自动驾驶、公共安全防控、重大工程决策等关键决策领域，严格坚持“人类监督、最终决策”原则，大模型仅作为辅助工具，不得替代人类作出最终决策，明确人类决策者的核心责任。

9.3.2 建立明确的人机交互规程，结合应用场景特点，明确人类干预、接管的条件、方式与流程，配备具备相应专业能力的人员负责监控模型运行状态，确保在紧急情况下能够快速介入处置，防范因模型失误引发的风险。

9.3.3 对大模型辅助作出的决策，建立多层级复核机制与责任追溯机制，明确人类决策者、相关组织、技术提供者等各方的责任边界，留存决策全过程记录，确保决策失误或造成损害时可依法依规追溯问责。

10 安全监测与内容治理

10.1 运行监测

10.1.1 相关组织应建立7×24小时不间断的系统运行监测与内容安全监测机制，实现对模型输入、输出、运行状态、网络安全、数据安全等全流程实时监测，按要求接入国家网络安全监测预警体系及行业监管监测平台。

10.1.2 设置异常情况自动告警与熔断机制，对模型生成违法信息、深度伪造内容、大规模偏见输出、系统异常运行、数据泄露等异常情况，明确告警阈值与响应流程，告警响应时间应满足快速处置需求，重大异常情况应立即启动熔断措施并暂停相关服务，及时开展应急处置。

10.1.3 建立监测数据留存机制，监测记录保存期限不少于6个月，重大安全事件、伦理违规事件相关监测数据保存期限不少于3年，监测数据应真实、完整、可追溯，满足监管审计与责任追溯需求。

10.2 内容审核与处置

10.2.1 配备与业务规模、风险等级相适应的人工审核团队，审核人员应具备相应的专业能力、伦理素养与法律法规知识，经培训考核合格后方可上岗；高风险领域内容人工复审率必须达到100%，其他领域可结合风险等级合理确定复审比例，确保审核质量。

10.2.2 建立违法违规和不良信息快速处置流程，明确处置标准、责任分工与时限要求，具备对违规生成内容的屏蔽、过滤、溯源、下线、销毁等能力，重大违规内容处置应符合行业监管时限要求，确保违法违规信息及时清除，防范扩散传播风险。

10.2.3 建立便捷、高效的用户举报反馈机制，公布举报渠道（含线上、线下）、处置时限与反馈方式，在接到用户举报后48小时内作出受理响应，一般举报事项处置完毕后应及时向举报人反馈结果，重大举报事项按规定向行业监管部门报告。

11 评估、审计与持续改进

11.1 定期伦理审计

11.1.1 相关组织应每年至少开展一次全面的内部伦理审计，高风险应用相关组织应每半年开展一次；同时可委托具备国家认可资质的第三方机构开展独立伦理审计，审计范围应覆盖全生命周期伦理治理要求的落实情况，审计结果作为行业监管评级、准入许可的重要依据。

11.1.2 审计报告应全面涵盖本指南各项要求的落实情况、存在的问题、风险隐患、整改措施及完成时限，高风险应用相关组织的审计报告应按要求报送行业监管部门；相关组织应向社会或利益相关方摘要公开审计结果，接受公众监督，公开内容应符合商业秘密与隐私保护要求。

11.2 绩效评估指标

11.2.1 建立科学合理的伦理治理关键绩效指标（KPI）体系，至少包含：伦理审查通过率、偏见投诉率及整改率、用户知情同意获取合规率、安全事件发生率、举报处置及时率与办结率、审核准确率、培训覆盖率与考核合格率等核心指标，指标标准应符合行业监管要求并结合业务特点动态调整。

11.2.2 每季度对KPI完成情况进行评估，评估结果纳入相关部门和人员的绩效考核，与奖惩、晋升、岗位调整等直接挂钩；高风险应用相关组织应将KPI评估结果定期报送行业监管部门，接受监管考核与指导。

11.3 持续改进

11.3.1 相关组织应建立闭环改进机制，根据伦理审计结果、KPI评估情况、用户反馈、技术进步、法律法规更新及监管政策调整，及时修订完善伦理治理策略、制度流程、技术措施与操作规范，持续提升伦理治理水平，确保治理要求与技术应用、行业发展、监管需求保持同步。

11.3.2 鼓励相关组织积极参与国家及行业伦理标准制定、技术创新与最佳实践分享，加强国际交流合作，提升我国人工智能伦理治理的国际话语权与影响力；行业监管部门、行业协会应定期总结推广优秀治理案例，搭建交流平台，引导全行业规范、健康发展。

12 附则

本标准由广西电子商务企业联合会负责解释。本标准自发布之日起试行，试行期为一年。试行期满后，根据实施反馈情况进行修订和完善。各相关单位可依据本标准制定具体的实施细则。若本标准与国家新颁布的法律法规或强制性标准有不一致之处，应以国家法律法规和强制性标准为准。本标准所引用的规范性引用文件如有更新，其最新版本适用于本标准。广西电子商务企业联合会将根据技术发展和应用需求，适时组织对本标准的复审与修订工作，以保障其持续的先进性和适用性。本标准的有效实施，有赖于各级医疗机构、主管部门、技术服务商和各相关方的共同努力，通过规范智慧医院数据互联互通共享技术，推动医疗健康数据资源有效整合与安全共享，提升医疗服务质量和效率，促进智慧医院建设规范化发展，为推进健康中国建设提供技术支撑。