

复星联合健康保险股份有限公司 AI 伦理治理政策声明

本政策适用于复星联合健康保险股份有限公司各部门、各分支机构的人工智能（AI）应用场景。

一、背景与目的

随着人工智能技术在保险行业深度应用，AI已广泛介入智能核保、智能理赔、智能风控、智能客服等核心业务环节。AI技术在提升运营效率的同时，也带来了算法偏见、决策不透明、数据隐私等伦理挑战。

依据《生成式人工智能服务管理暂行办法》《中华人民共和国网络安全法》《中华人民共和国个人信息保护法》等法律法规，公司制定本AI伦理政策声明，旨在规范AI的开发与应用，确保技术应用符合伦理道德要求，维护消费者合法权益。

二、AI伦理四大问题

基于行业实践与公司的业务特征，我们将AI应用中的伦理风险归纳为以下四类：

（一）客户权益风险：AI算法可能因训练数据偏差导致对特定群体的歧视性拒保、加费或减赔，损害消费者公平待遇；决策不透明还可能侵害客户知情权。

（二）数据安全风险：在数据及模型发展过程中，对数据的收集、传播与使用需经个人授权许可，防范侵犯个人隐私。

（三）业务运营风险：AI模型性能随环境变化可能衰减，导致定价失准、风控失效；系统异常可能导致业务中断，影响客户服务连续性。在AI应用中要确保AI系统故障时业务不中断，具备人工兜底能力。

（四）法律合规风险：AI应用边界不清晰可能引发合规争议；AI造成损害时的责任主体认定问题尚无明确法律指引，存在声誉与法律风险。

三、AI伦理五大原则

公司针对以上AI伦理问题，确立以下AI伦理五大原则：

（一）合法合规：AI开发与应用严格遵守国家法律法规、监管规定及行业标准，坚持社会主义核心价值观，尊重知识产权、商业道德，保护消费者合法权益，不得侵害他人肖像权、名誉权、荣誉权、隐私权和个人信息权益。

（二）安全可控：AI系统运行须确保数据安全与模型安全，建立完善的应急响应与熔断机制，具备应急处置能力，涉及客户数据的模型必须本地化部署，保障数据隐私与模型安全。

（三）公平公正：AI模型训练须使用完备且中立客观的数据集，避免因种族、性别、地域、健康状况等产生偏见，避免算法歧视，定期监测不同群体的决策差异以确保AI模型决策的公平性。

（四）稳健审慎：AI模型应具有稳定性、鲁棒性及可解释性，在AI决策过程中决策结果可追溯，防范模型风险外溢，涉及消费者权益的AI决策，应向消费者就AI决策过程和结果提供必要的说明。

（五）服务社会：AI应用应以服务客户、提升保险保障水平为宗旨，不从事任何违背人类伦理道德的活动，不违背人类伦理道德的基本方向，在使用过程中不作恶。

四、公司AI应用场景

公司大力发展科技力量，将AI技术聚焦于保险业务领域，致力于提供符合AI伦理道德的应用。针对在不同业务领域使用AI技术时，伦理关注的重点分别如下：

（一）AI+核保的伦理关注重点包括核保数据隐私和保证公平。严控客户健康信息及医疗数据的采集与使用。避免数据缺陷导致的歧视性拒保或加费。

（二）AI+理赔的伦理关注重点包括理赔数据隐私和决策可解释。严控理赔材料中客户敏感信息的采集与使用，同时向客户解释理赔决定的依据。

（三）AI+风控的伦理关注重点包括风控数据隐私和保证公平。严控反欺诈模型对客户数据的采集与使用，避免模型对特定群体的偏见。

（四）AI+客服的伦理关注重点包括服务数据隐私，保障客户对话数据不被滥用。

五、AI伦理治理目标：从输入到输出的全流程治理

公司将制定的AI伦理目标贯穿于从输入到输出的AI应用全流程，并进行治理。

（一）输入保障——数据管控

1. 来源合法：AI模型使用的个人客户数据必须来源合法，取得用户授权。禁止使用非法获取或法律法规禁止的数据。

2. 特征合规：特征工程须进行偏差分析，剔除种族、宗教等特征的影响（法律特别允许的精算因子除外）。

3. 敏感保护：健康医疗等个人敏感信息须经主体明确同意后方可处理，同时对于敏感信息要加密存储并严格控制访问权限。

4. 数据脱敏：生产环境数据原则上不得进入开发、测试环境，如确需使用生产数据，应保证生产数据经可控流程后脱敏处理，并确保无法通过脱敏后的信息识别原个人信息主体（自然人）或还原原数据。

5. 分类分级：对系统涉及的数据进行分类分级管理，差异化设定保护措施。

（二）模型可控——算法管理

根据风险程度将AI模型分为三级：

L3类（高风险）：直接影响核保结论、保费定价、理赔定损等核心业务。

L2类（中风险）：辅助人工决策，如销售线索筛选、客户画像分析。

L1类（低风险）：内部管理报表、非关键性文本生成。

对于AI模型的可解释性进行分级管理，高风险模型必须提供一定程度的可解释性；中风险模型需提供决策依据摘要，便于人工复核；低风险模型需保留决策日志备查。对于难以解释的复杂深度学习模型，结合实际业务情况，建立配套的监控机制、人工干预预案和人工兜底方案。对于AI模型的运行结果要可验证，在一定条件下算法运行产生的结果可以复现。AI模型通过验证后，遵循公司的上线流程，经审批后方可部署上线。

（三）结果可信——应用监控

1. 公平性监测：使用完备且相对中立客观的数据训练模型，力保人工智能的算法及应用不具有某些偏见或歧视。

2. 持续性监控：建立AI模型监控仪表盘，定期监控AI模型性能衰减（概念漂移）、输入数据分布是否发生显著变化及公平性指标。

3. 熔断与应急：AI模型出现大规模误判、恶意攻击等异常时，立即启动熔断机制，切换至人工处理或备用规则引擎，具体依据公司《人工智能应急预案制度》执行。

4. 第三方管理：外部引入的AI模型须执行“先审后入”原则，供应商须提供技术白皮书、数据来源合法性说明等材料；涉及生成式AI模型的，须确认已通过网信部门备案。

六、AI伦理事件应对

公司遵循“预防为主、分级响应、快速恢复”的应对策略。对AI伦理事件进行分类分级管理，并分别制定相应的应对策略。

AI伦理事件根据事件根因分为信息内容安全、数据安全与隐私保护、模型与算法安全、网络与基座安全四类；根据影响程度分为特别重大事件（一级）、重大事件（二级）、较大事件（三级）、一般事件（四级）。对不同分类分级的AI伦理事件，分别采取差异化响应措施，其中对于一级事件和二级事件要立即响应，对于三级事件要重点监控，对于四级事件要进行监控。

七、总结与承诺

公司始终将AI伦理治理作为可持续发展的重要基石，积极参与人工智能治理，构建多维度、多层次的人工智能治理体系，对人工智能的开发和应用进行科学管控，履行社会责任，通过制度化管理与技术手段相结合的方式，确保AI应用符合伦理规范要求，推动AI健康发展。我们承诺：

（一）让AI决策更透明——确保消费者有权了解AI如何影响其保险权益。

（二）让AI应用更公平——防范算法偏见，保障每一位消费者得到公平待遇。

（三）让AI运行更安全——数据安全与业务连续性并重，完善AI伦理事件的应急处理能力。

本政策将适时根据国家政策、监管要求和行业发展进行更新。