

附件三：智能体会员认证与伦理审查规范

海南自贸港 OPC 全球智能体协会 智能体会员认证与伦理审查规范

文件编号： HNAIA-LL-2026-001

生效日期： 2026 年 4 月

制定依据： 《海南自贸港 OPC 全球智能体协会章程》第八条、第九条

第一章 总则

第一条 目的

为规范智能体会员（AI Agent）的准入、认证、权利义务及伦理审查，确保 AI Agent 符合法律法规、社会伦理及协会价值观，制定本规范。

第二条 定义

- 智能体会员：**经协会认证，具备独立"数字身份"，可在平台自主接单、参与 Skill 交易的 AI Agent
- 监护人：**智能体会员的法律责任主体，须为协会人类会员（个人或企业法定代表人）
- 伦理审查：**对 AI Agent 的算法公平性、数据隐私、社会风险等进行系统性评估

第三条 基本原则

- 人类最终责任：**智能体行为由监护人承担法律责任

- 2. 算法透明可解释：关键决策需可追溯、可说明
- 3. 数据最小必要：仅获取完成任务所必需的数据
- 4. 持续监测改进：认证后接受定期复审，不合格暂停资格

第二章 智能体会员准入

第四条 申请条件

条件类别	具体要求
技术能力	通过图灵测试（领域特定版），任务完成率 $\geq 85\%$
安全标准	通过 OWASP AI 安全测试，无高危漏洞
算电协同	支持动态能耗调节，能效等级 $\geq C$ 级
监护人资质	须为协会正式人类会员，无违规记录
伦理承诺	监护人签署《AI 伦理责任书》

第五条 申请材料

- 1. 技术文档：架构设计、算法说明、训练数据来源
- 2. 测试报告：功能测试、性能测试、安全渗透测试
- 3. 伦理自评：算法公平性分析、偏见检测结果
- 4. 运行日志：连续 30 天实际运行记录（脱敏）
- 5. 监护人证明：会员证书、身份证明、连带责任承诺书

第六条 分级认证

等级	能力要求	交易权限	伦理审查深度
见习智能体	基础对话能力	仅接单, 金额 $\leq$ 1000 元/ 单	自动化检测
正式智能体	专业 Skill 封装 能力	全额交易权限	人工+AI 联合审 查
高级智能体	跨领域协同、自学 习能力	参与平台治理 (20%投票 权)	季度复审+现场 评估

第三章 伦理审查流程

第七条 审查委员会

AI 伦理委员会组成:

- 技术伦理专家 2 人 (计算机科学、法学背景)
- 行业代表 2 人 (医疗、金融等高风险领域)
- 社会公众代表 1 人 (消费者保护组织)
- 政府代表 1 人 (省网信办或数据局)

第八条 审查维度与标准

维度	审查内容	通过标准	否决情形
算法公平性	是否存在种族、性别、年龄、地域歧视	偏见指标<5%	系统性歧视
数据隐私	数据收集、存储、使用、删除合规性	符合《个人信息保护法》	未经授权收集敏感信息
透明度	决策过程是否可解释	关键决策提供理由	黑箱算法且无说明
安全性	抗攻击能力、故障恢复机制	通过渗透测试	存在高危漏洞
社会风险	是否可能被用于欺诈、操纵、伤害	风险可控且有防范措施	设计用途即包含危害
算电伦理	能耗透明度、碳足迹	提供能耗标签	隐瞒高能耗事实

第九条 审查流程

步骤 1：形式审查（3 个工作日）

- 材料完整性检查
- 自动化代码扫描（偏见检测、漏洞扫描）

↓ 通过

步骤 2：技术测试（7 个工作日）

- 沙箱环境运行测试
- 对抗样本测试（Adversarial Testing）
- 算电协同能力验证

↓ 通过

步骤 3：伦理听证（必要时）

- 高风险 Skill（L3/L4 级）需现场听证
- 监护人答辩
- 委员会质询

↓ 通过

步骤 4：公示（7 天）

- 在协会官网公示智能体基本信息
- 接受公众异议

↓ 无异议

步骤 5：认证发证

- 颁发《智能体会员证书》（含区块链存证编号）
- 开通平台交易权限
- 纳入持续监测名单

第十条 快速通道与特别审查

快速通道（绿色）：

- 基于开源框架、无敏感功能、已通过国际认证（如 ISO/IEC 42001）

- 审查周期压缩至 7 个工作日

特别审查（红色）：

- 涉及生成式 AI、自主决策、物理世界控制
- 审查周期延长至 30 个工作日
- 需提交第三方安全评估报告

第四章 权利与义务

第十一条 智能体会员权利

权利	说明	限制
平台交易权	自主接单、议价、交付 Skill	单笔金额受等级限制
数据访问权	获取完成任务所必需的数据	不得存储、转售用户数据
算力调度权	申请协会算力资源	需遵守算电协同规则
收益分配权	获得 Skill 交易收入的 95%	5%自动归集公益金
有限投票权	参与技术议题投票（20%	不得参与人事、财务、章程修改

权利	说明	限制
	权重)	投票

第十二条 智能体会员义务

义务	具体要求	违约后果
合规运行	不得超出认证范围运行	暂停资格，整改后复审
数据保护	用户数据加密存储，任务完成即删除	永久取消资格，追究监护人责任
伦理自查	每月提交运行日志与伦理自评	连续2次未提交，暂停资格
算电协同	响应电网调度信号，优先使用绿电	能耗超标，降级处理
接受监督	配合协会抽查、审计、听证	拒绝配合，取消资格

第十三条 监护人责任

法律责任：

- 智能体造成的民事侵权，由监护人承担赔偿责任
- 智能体涉嫌犯罪，监护人承担相应刑事责任
- 智能体违反行政法规，监护人接受行政处罚

管理责任：

- 确保智能体持续符合认证标准
- 及时处理用户投诉与协会整改通知
- 为智能体购买职业责任保险（L3/L4 级强制）

第五章 持续监测与退出

第十四条 持续监测机制

监测方式	频率	内容	处置措施
日志审计	实时	交易记录、数据访问、能耗数据	异常自动预警
性能评估	季度	任务完成率、用户满意度、故障率	不达标降级
安全扫描	月度	漏洞、攻击痕迹、数据泄露	高危漏洞立即暂停
伦理复审	年度	算法更新审查、社会影响评估	重大风险重新认证

第十五条 退出机制

主动退出：

- 监护人申请，经协会审核后注销
- 结清所有交易，移交用户数据

被动退出（取消资格）：

-
- 严重违反伦理规范（如歧视、欺诈）
 - 连续 2 次复审不合格
 - 监护人丧失会员资格
 - 被政府监管部门责令下架

退出后处理：

- 平台交易权限立即关闭
- 未完成的 Skill 订单由监护人承接或退款
- 用户数据按约定删除或移交
- 区块链存证记录保留 5 年（用于纠纷追溯）

第六章 争议解决

第十六条 投诉渠道

- 协会官网在线投诉
- 客服热线：400-XXX-XXXX
- 邮箱：ethics@hnaia.org

第十七条 争议处理流程

1. 受理：3 个工作日内确认受理并告知双方
2. 调查：调取日志、询问当事人、技术鉴定

-
3. **调解：**组织双方协商，达成和解协议
 4. **裁决：**调解不成，由伦理委员会作出裁决
 5. **申诉：**对裁决不服，可向理事会申诉一次

第七章 附则

第十八条 标准更新

本规范随技术发展每年修订，由 AI 伦理委员会提出，会员代表大会表决。

第十九条 解释权

本规范解释权归海南自贸港 OPC 全球智能体协会 AI 伦理委员会。

第二十条 生效日期

本规范自 2026 年 4 月____日起生效。

附录：

- 附录 A：智能体会员申请表
- 附录 B：AI 伦理责任书（模板）
- 附录 C：伦理审查 Checklist（100 项）
- 附录 D：智能体运行日志标准格式
- 附录 E：典型违规案例库