Agentic AI 治理框架

资讯通信媒体发展局 (IMDA)

2026 年 1 月

中文翻译版

Agentic AI 治理框架（Model AI Governance Framework for Agentic AI）

概述

2026 年 1 月，新加坡资讯通信媒体发展局（Infocomm Media Development Authority, IMDA）发布了针对 Agentic AI（自主 AI 代理）的治理框架。随着 AI Agent 技术快速发展，这份框架应对了 AI 自主决策带来的新治理挑战。

背景

Agentic AI 指的是能够自主执行任务、做出决策并与环境交互的 AI 系统。与传统 AI 工具不同，Agentic AI 具备： - 自主规划和执行能力 - 多步骤任务处理能力 - 环境感知和适应能力 - 工具调用和协作能力

这些特性带来了新的治理挑战，超出了现有 AI 治理框架的覆盖范围。

核心议题

1. 自主决策边界

- 明确 AI Agent 可自主决策的范围
- 设定需要人类审批的决策类型
- 建立分级授权机制

2. 人类监督机制

- 确保人类对关键决策的最终控制权
- 设计有效的人机协作流程
- 建立紧急干预和停止机制

3. 责任归属

- 明确 AI Agent 行为的法律责任归属
- 区分开发者、部署者和用户的责任
- 建立多方责任框架

4. 安全防护

- 防止 AI Agent 越权操作
- 确保 AI Agent 行为可追溯
- 建立安全测试和审计机制

治理原则

透明度

- AI Agent 的能力和局限须清晰说明
- 决策过程可解释、可审计
- 用户知情权得到保障

问责制

- 建立明确的责任链条
- 关键决策可追溯至人类决策者
- 定期审计和评估机制

安全性

- 多层安全防护机制
- 对抗性测试和压力测试
- 持续监控和异常检测

与现有框架的关系

本框架是新加坡 AI 治理体系的最新组成部分，与以下框架互为补充： - **AI 治理模型框架**（2019）：基础 AI 治理原则 - **生成式 AI 治理框架**（2024）：大模型治理 - **AI Verify**（2022）：AI 治理测试工具

意义

Agentic AI 治理框架的发布使新加坡成为全球最早专门应对 AI Agent 治理挑战的国家之一，体现了新加坡"前瞻性、渐进式"的 AI 治理理念。