生成式 AI 治理框架

资讯通信媒体发展局 (IMDA)

2024 年 1 月

中文翻译版

生成式 AI 治理框架提案（Proposed Model AI Governance Framework for Generative AI）

概述

2024 年 1 月，新加坡资讯通信媒体发展局（Infocomm Media Development Authority, IMDA）发布了《生成式 AI 治理框架提案》，这是全球较早的专门针对生成式 AI（Generative AI）的治理框架之一。

九大治理维度

1. 问责制（Accountability）

- 明确生成式 AI 系统全生命周期的责任归属
- 模型开发者、平台运营者和用户各负其责
- 建立有效的投诉和申诉机制

2. 数据治理（Data Governance）

- 训练数据的合法获取和使用
- 数据质量保证和偏见检测
- 个人数据保护和隐私合规

3. 可信开发与部署（Trusted Development and Deployment）

- 模型开发过程的透明度
- 安全测试和评估标准
- 部署前的风险评估

4. 事件报告（Incident Reporting）

- 建立 AI 事件报告机制

- 及时通报安全漏洞和滥用情况
- 跨机构信息共享

5. 测试与保证（Testing and Assurance）

- 模型性能和安全性的持续测试
- 第三方审计和评估
- 红队测试（Red Teaming）

6. 安全（Safety）

- 防止生成有害内容
- 建立内容安全过滤机制
- 应对对抗性攻击

7. 内容来源（Content Provenance）

- AI 生成内容的标识和溯源
- 数字水印和元数据标注
- 防范深度伪造（Deepfake）

8. 使用者素养（User Literacy）

- 提升公众对生成式 AI 的认知
- 推广负责任的 AI 使用实践
- 培养批判性思维能力

9. 辅助措施（Supporting Measures）

- 建立行业标准和最佳实践
- 推动国际合作和标准对接
- 支持创新与监管的平衡

治理方法

多利益相关方参与

- 政府、产业界、学术界和公民社会共同参与
- 广泛征求公众意见
- 国际经验借鉴

"沙盒式"治理

- 在受控环境中测试治理方案
- 渐进式推进，边试边调
- 在创新与安全之间寻求平衡

与全球框架的比较

新加坡的生成式 AI 治理框架在以下方面具有特色： - 更强调实用性和可操作性 - 采用多利益相关方的协商模式 - 平衡创新促进与风险管控 - 与 AI Verify 等测试工具紧密结合

意义

该框架体现了新加坡在 AI 治理领域的持续创新，从 2019 年的基础 AI 治理框架到 2024 年的生成式 AI 专项框架，反映了新加坡治理体系随技术发展持续演进的特点。