人工智能安全与可信技术白皮书

White Paper on AI Safety and Trustworthy Technologies

正式研究展示文件

本文件聚焦人工智能安全、可信 AI 和负责任技术应用，整理 DSRI 在成果评审和奖项评议中采用的关键观察维度。文件强调风险识别、评估记录、人工复核、模型边界和组织责任，不使用未经验证的案例包装成果。

发布日期：2025年11月26日
发布机构：数字科学促进与研究协会（ADSPR） / 数智研究院（DSRI）

一、问题背景

人工智能系统的能力快速提升，使研究评价从“是否有效”进一步转向“是否可控、可解释、可复核”。在学术成果、系统工具和产业化方案中，安全边界和责任机制逐渐成为基础要求。

二、可信技术的核心要素

安全边界

明确模型适用范围、禁止场景、输入输出限制和异常处理规则。

可解释与可追踪

保留数据处理、模型训练、参数调整和结果复核记录，支持评审和复盘。

人工监督

在高风险任务中设置人工复核节点，避免自动化系统直接替代专业判断。

方向	重点	建议材料
模型安全	鲁棒性、对抗风险、异常输出	测试记录、边界说明
数据治理	来源、授权、隐私保护	数据说明书、处理流程
责任机制	人工复核、版本管理、反馈渠道	复核记录、更新日志

三、评审关注点

DSRI 建议从技术可靠性、数据合规性、对抗风险、偏差控制和应用责任五个方向评价 AI 相关成果。材料应说明测试方法、失败情形、更新机制和风险缓释措施。

可信 AI 不是单一技术指标，而是技术、流程和治理共同构成的结果。

四、组织实施建议

研究团队和项目负责人可建立模型卡、数据说明书、评估记录表和版本归档制度。对于公开发布的工具，应提供使用限制、错误反馈渠道和更新说明。