

生成式人工智能 治理与实践 白皮书

GENERATIVE ARTIFICIAL
INTELLIGENCE
GOVERNANCE &
PRACTICE
WHITE PAPER

 **Alibaba Group**
阿里巴巴集团

 **中国电子技术标准化研究院**
China Electronics Standardization Institute

 **阿里云**  **達摩院**
aliyun.com ALIBABA DAMO ACADEMY

目录

一. 生成式人工智能的发展以及担忧

1. 生成式人工智能的技术与应用突破	14
1.1 文生文突飞猛进	14
1.2 文生图效果惊艳	15
1.3 行业应用广泛	16
1.4 使用门槛降低	18
2. 生成式人工智能的内生问题与社会担忧	20
2.1 个人信息的实时交互担忧	20
2.2 内容安全的源头敏捷控制	21
2.3 模型安全的全生命周期内控	22
2.4 知识产权的溯源与权属挑战	22

二. 生成式人工智能的治理愿景和框架

1. 国际社会治理特点	26
1.1 治理目标：坚持促发展与重监管并行	26
1.2 治理模式：强调多元主体协同共治	26
1.3 治理手段：创设例外保留创新空间	27
1.4 治理细则：技术规范逐渐明晰	27
2. 我国的治理特点	28
2.1 促进发展：对人工智能发展给予更多政策支持， 配套发布一系列产业政策文件	28
2.2 重视治理：确定了现阶段算法治理的重点场景， 推动建立算法治理的“法治之网”	28
2.3 伦理约束：加强科技伦理治理顶层设计，明确人 工智能伦理原则及治理要求	28
3. 本书观点：发展多主体协同敏捷治理体 系，构建全生命周期风险分类治理框架	30

三·生成式人工智能风险产生原因的分析

1. 综述：构建生成式大模型的条件	34
1.1 算力	34
1.2 数据	34
1.3 算法	35
1.4 生态	35
1.5 人才	35
2. 语言大模型	36
2.1 Transformer 网络	36
2.2 训练过程和使用的数据	36
2.3 语言大模型的生成过程	39
2.4 小结：语言大模型的风险来源	40
3. 视觉大模型	41
3.1 模型原理	41
3.2 训练过程	42
3.3 生成过程	44
3.4 小结：视觉大模型的风险来源	45

四·生成式人工智能风险治理实践和探索

1. 生成式人工智能治理格局建设	48
1.1 以针对性立法回应技术与产业需求	48
1.2 以政策完善构建与技术发展需求相匹配的治理机制	48
1.3 产业自律自治筑成负责任创新治理机制	49
2. 生成式人工智能不同环节的风险治理	51
2.1 模型训练阶段的风险治理	52
2.2 服务上线阶段的风险治理	53
2.3 内容生成阶段的风险治理	53
2.4 内容传播阶段的风险治理	54
3. 个人信息合规	56
3.1 大模型与个人信息的关系	56
3.2 训练数据中的个人信息	56
3.3 算法服务时拒绝生成个人信息	58

4. 内容安全保障	59	7. 实践案例：虚拟模特塔玑	75
4.1 内容安全视角里，AIGC 与 UGC 的异同	59	7.1 虚拟模特塔玑促进生产力提升	75
4.2 生成式模型风险评测	60	7.2 数据驱动下的虚拟模特与个人信息保护	76
4.3 模型层内生安全	61	7.3 内容安全保障	76
4.4 应用层安全机制	62	7.4 模型安全控制	77
4.5 生成信息的信任机制	63	7.5 生成式标识与知识产权保护	77
5. 模型安全防控	68		
5.1 鲁棒性	68		
5.2 可解释性	68		
5.3 公平性	68		
5.4 防滥用机制	69		
5.5 实践案例：鲁棒评估基准与增强框架	69		
6. 知识产权探索	73		
6.1 训练数据的知识产权合法性治理探索	73		
6.2 生成物知识产权治理探索	74		

五·生成式人工智能多主体协同敏捷治理体系

1. 敏捷治理的理念与特点	80
2. 多主体协同下的敏捷治理探索与实践	81
2.1 政府规范引导	82
2.2 产业守正创新	82
2.3 社会监督理解	84

六·总结与展望 (88)

专有名词解释 (92)

CHAPTER. 1

壹

生成式人工智能的发展以及担忧

- 1.生成式人工智能的技术与应用突破
- 2.生成式人工智能的内生问题与社会担忧