企业网 D1Net
企 业 I T 第 1 门 户

2025年第六届全国医药大健康CIO大会
AI+数据赋能业务场景

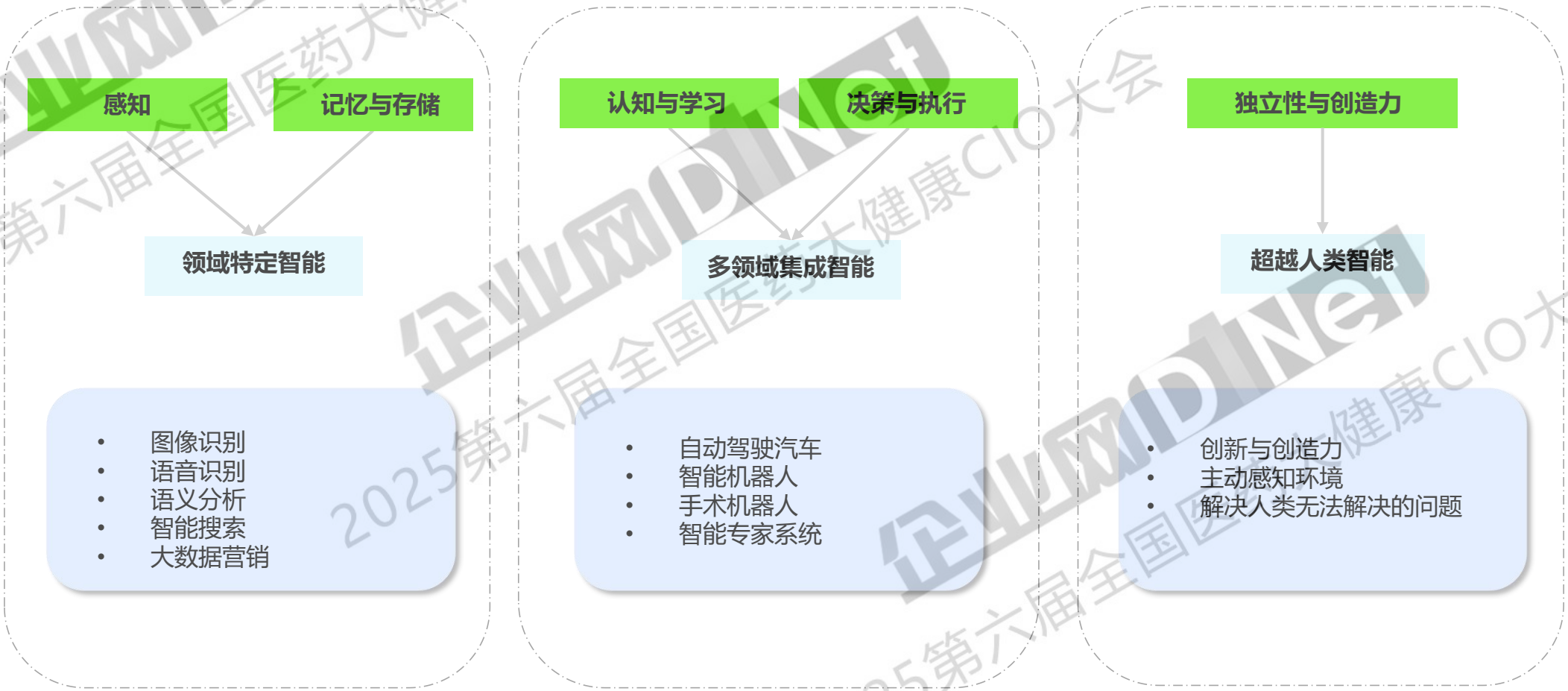
# 负责任的人工智能治理模型建设

宣讲人：Great Gu　公司：Haleon

# 1. AI介绍

# 人工智能的类型





智能的飞跃：AI技术正以惊人的速度发展。

游戏领域的突破：AI不仅学会了游戏规则，更在策略和反应速度上超越了人类玩家。

人类智慧的挑战：在棋类、电子竞技等游戏中，AI已经成为不可小觑的对手。

1997年5月11日

2011年2月16日

2016年3月15日



# 通用人工智能的里程碑



**2018年 6月**
***GPT-1***

**2019年 2月**
***GPT-2***

发布时限制访问

**2020年 6月**
***GPT-3***

在规模、性能和功能方面进行了重大改进

**2020年 7月**
***ChatGPT-1***

优化了对话交互

**2021年 9月**
***ChatGPT-2***

**2022 年 7月**
***ChatGPT-3***

开发了推理机制和模型，以更好地理解对话的前后文

**2022年 11月**
***ChatGPT-3.5***

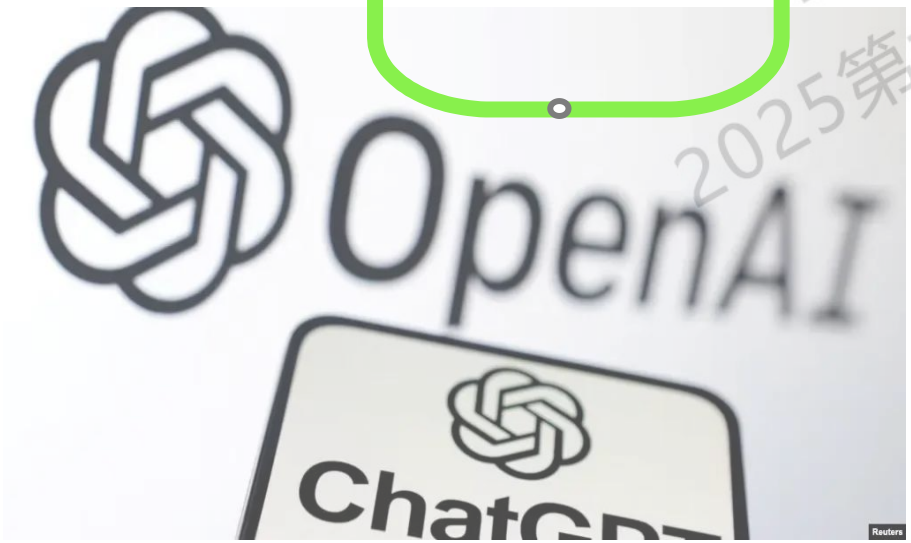
**2023年 3月**
***ChatGPT-4***

**2024年 5月**
***ChatGPT-4o***

实现了音频、视觉和文本可以在实时中进行推理

**2024年 9月**
***OpenAI-o1***

在处理复杂推理任务方面取得了更显著的进展



# 人工智能的应用领域



## AI+医疗健康

- WOW健康+
- 39AI全科医生
- 佳医康健康咨询小程序
- 开心健康-健康咨询
- 左手医生
- 诺华E时代

## AI+其他

- 有道翻译
- B612咔叽
- 文心一言
- 通义千问
- 讯飞星火

○ ○ ○

**新型商业模式**
加速拓展新型创收机会

**运营效率**
提高效率和标准化流程

**资源加速**
在其他领域产品和服务增长上提高效率

**增强体验**
提高客户服务和定制化的能力

# AIGC并不完美；我们需要在风险和机遇之间找到平衡点。

**技术前提**
需访问内部的结构化和非结构化数据

**新技能发展**
使用批判性思维有助于减少歧义

**知识产权和法律**
不断变化的监管环境和有关内容所有权的问题

**潜在的固有偏见**
现有的信息源存在差异

**准确度的限制**
准确的回答需要基于大量的统计文本分析

**隐私问题**
不同级别的受保护信息可能会对人工智能模型产生限制

**误传和误用**
以非预期的方式使用生成式人工智能工具

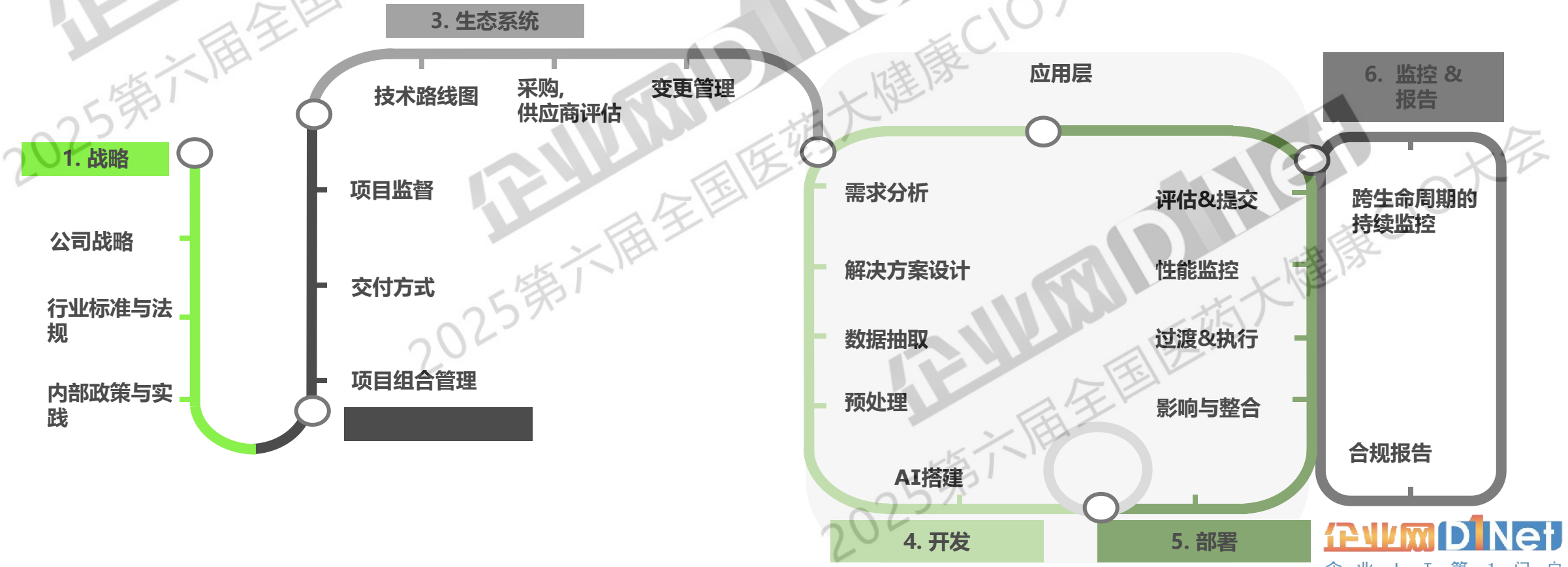
# 什么是负责任的人工智能(RAI)?



随着对人工智能相关风险的认识和审查增加，构建负责任的技术已成为各行业组织中的首要关注点。

在全球范围内，负责任的人工智能正从“最佳实践”成熟为推动系统级变革和培养信任所必需的高级原则和指导。监管机构也在关注这一点，倡导围绕人工智能的监管框架，包括增加数据保护、治理和问责措施。

**RAI端到端企业治理框架：**



# 为什么选择负责任的人工智能?



在这个人工智能越来越多、对这些人工智能系统的可信度要求越来越高的世界里，企业需要 RAI 来帮助降低风险。

## 日益增长的人工智能使用

各行各业越来越多的组织和行业采用人工智能来提供产品和服务。

**现状:**

- 越来越多的数据
- 越来越多的模型
- 越来越复杂

## 日益严格的审查要求

**对负责任和透明的人工智能的呼声更高:**

- 新兴的监管要求
- 消费者对于知情权的需求
- 不断变化的员工的期望

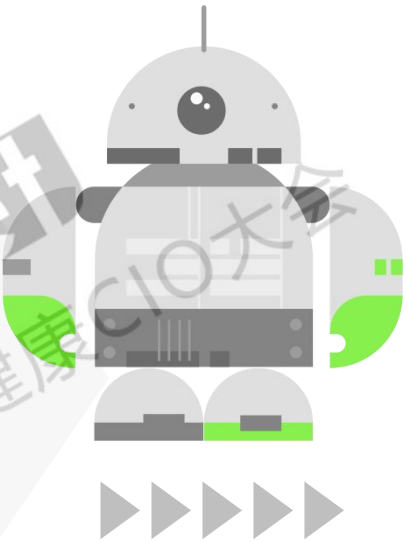
## 日益增加的风险

**组织和公共层面的风险**

- 经济风险
- 社会与环境风险
- 企业声誉和监管风险

**系统级风险**

- 模型和系统风险
- 数据风险
- 用户风险
- 法律、合规和流程风险



# 负责任人工智能的四个维度



负责任的人工智能是一种让人工智能系统获取信任的方法。其核心为数据科学，从战略到执行都受到关键指导原则的约束。

## 战略

**数据与AI的道德伦理**
考虑使用数据和人工智能的道德含义，并将它们编入组织的价值观

**政策与法规**
预测并了解关键的公共政策和监管趋势，以调整合规流程。

## 控制

**管理**
实现对三道防线系统的监督。

**合规**
遵守法规、组织政策和行业标准。

**风险管理**
扩大过渡性风险检测和缓解做法，以应对人工智能特有的风险和危害。

## 负责任的实践

**可解释性与可说明性**
使模型决策透明化。

**可持续性**
最大限度地减少对环境地负面影响

**鲁棒性**
实现高性能、高可靠性的系统。

**偏见与公平**
定义和衡量公平性，并更具标准测试系统。

**网络安全**
增强系统的网络安全。

**隐私**
开发可保护数据隐私的系统。

**物理安全**
设计并测试系统，以确保物理安全。

## 核心实践

**问题公式化**
确定你正在解决的具体问题，以及它是否需要AI / ML解决方案。

**验证**
评估模型性能并继续迭代设计和开发以改进指标。

**标准**
遵循行业标准和最佳实践。

**监控**
实施持续监控以识别偏差和风险。

# 构建负责任AI：共同使命



组织内各个部门的不同利益相关者必须相互协作，以一致、有效地应用负责任的人工智能实践。

**三道防线**

*企业层*

**外部**

监管机构
外部审计师

外部审计
监管要求和法律行动
外部标准制定

**第三道防线**

内部审计
伦理学家

内部审计审查
道德审查

**第二道防线**

领导层
合规和法律
安全和隐私
独立审查小组
SME模型审批者

标准和规范
独立审查和质疑
最终批准
法律、数据和隐私管理要求

*应用层*

**第一道防线**

模型开发人员
模型所有者
运营支持

风险评估
模型采购/开发
数据收集和使用
流程文档

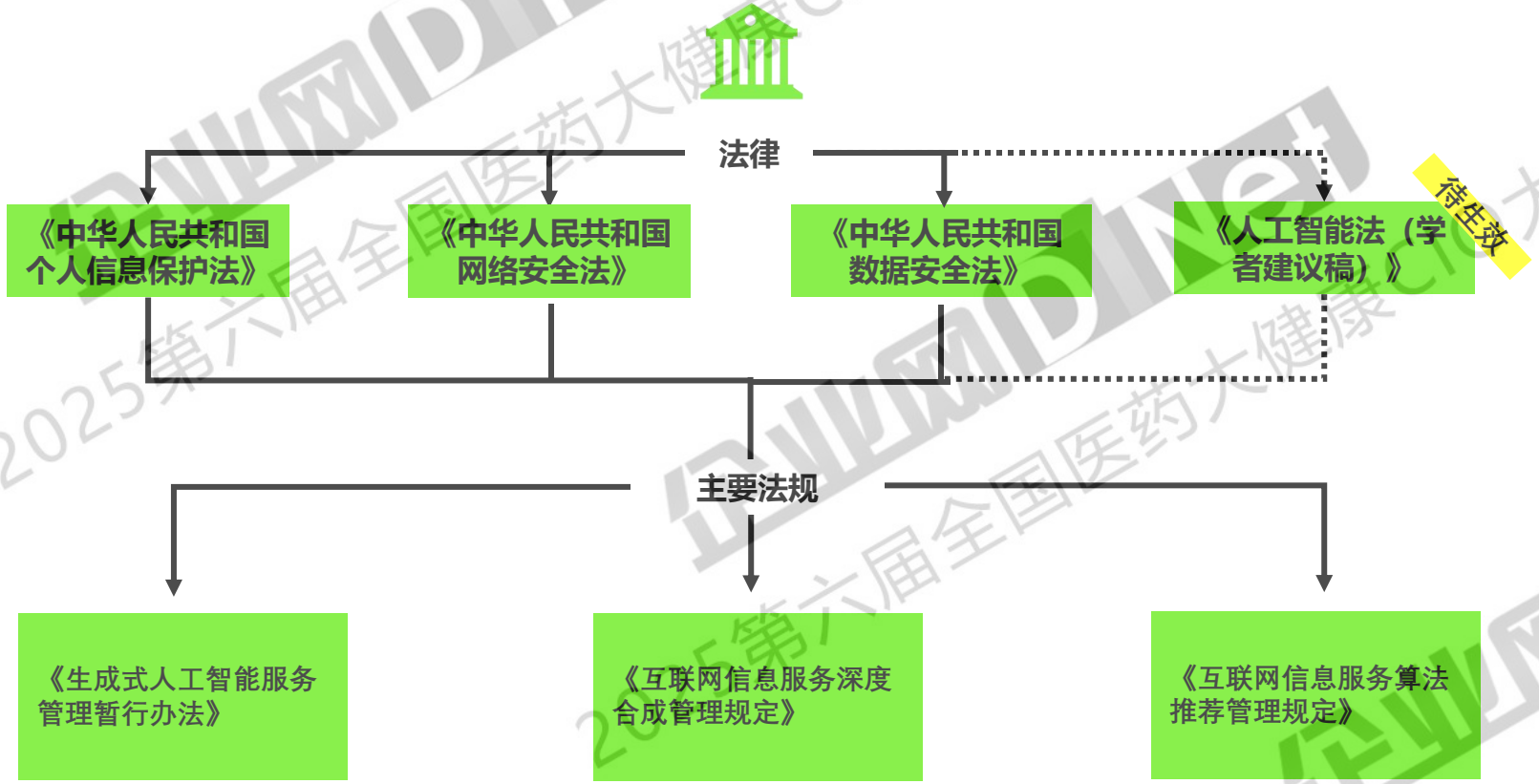
基础设施
部署
持续监控

# 2. 中国AI监管形势

# 中国AI法律体系框架



从 2021 年到 2022 年，中国成为第一个对人工智能的一些最常见应用实施详细的、具有约束力的规定的国家。这些规定构成了中国新兴人工智能治理制度的基础。2023 年和 2024 年初，更多法规和标准的出台，进一步塑造了中国的人工智能监管格局。




- 贯穿整个生成式人工智能产品生命周期的**合规要求**
- **加强安全防护措施**，防范潜在的攻击和滥用行为

- 建立**实名认证制度**和**内容标识机制**
- 实施**健全的内容审查机制**

- **贯穿算法推荐服务**整个生命周期的**合规要求**
- 建立**算法备案**监管体系

**其他支持性法规/要求**

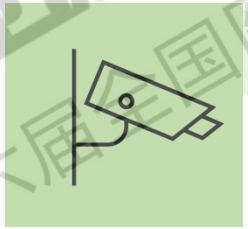
- 《生成式人工智能服务安全基本要求》
- 《互联网信息服务管理办法》
- 《网络安全标准实践指南——生成式人工智能服务内容标识方法》
- 《关于规范和加强人工智能司法应用的意见》
- 《网络信息内容生态治理规定》
- 《具有舆论属性或社会动员能力的互联网信息服务安全评估规定》
- 《信息安全技术 个人信息安全规范》
- 《人工智能安全标准化白皮书 （2023版）》
- 《人工智能安全治理框架》
- 《网络安全技术 人工智能生成合成内容标识方法（征求意见稿）》
- 《人工智能生成合成内容标识办法（征求意见稿）》

# 部署AI系统时应考虑的因素



人工智能技术有可能彻底改变我们的生活和工作方式，其潜力几乎是无限的——但要充分利用并保持所创造的价值，就需要关注并管理伴随而来的风险。

## 模型风险



与人工智能系统本身的训练、开发和性能有关的风险，包括概念的合理性、输出的可靠性和监管的有效性。

## 训练数据风险

算法设计：遵守基于种族、信仰和国家以及伦理和道德原则

训练数据：注意来源、质量、安全性和匹配度。

## 算法风险

合规性：确保算法备案符合当地法律法规的要求。

透明度：备案过程中需要提供算法的工作原理、使用的数据和预期用途等信息。

## 科技伦理评估风险

进行伦理审查，确保AI系统的使用不会侵犯人权或造成不道德的结果。

## 生成内容风险

知识产权：生成的内容可能存在侵犯了他人的版权。

偏见和歧视：生成的内容可能反映了训练数据中的偏见，导致歧视某些群体。

## 安全风险

数据泄露：敏感数据可能被未经授权的第三方访问。

数据篡改：数据在传输或存储过程中可能被恶意篡改。

# 针对生成式人工智能服务的“双备案制”监管体系

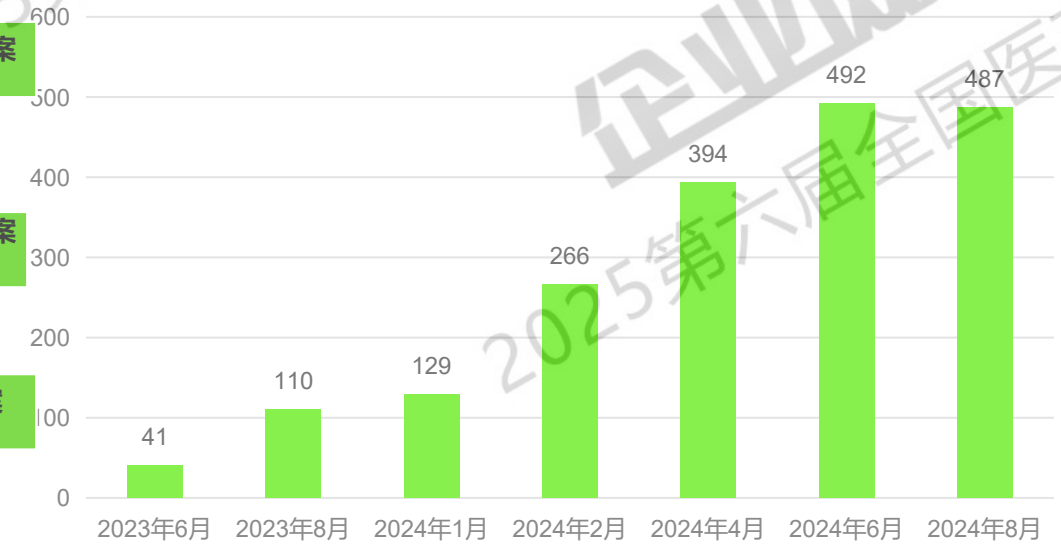


**“算法备案”**

具有舆论属性或者社会动员能力的生成式AI服务提供者、算法推荐服务提供者、深度合成服务提供者（当涉及深度合成技术时，还包括深度合成服务技术支持者）应当就AI所使用的算法在提供服务之日起十（10）个工作日内通过“互联网信息服务算法备案系统”进行备案。

**通过备案应用举例：**

- 个性化推送备案
- 调度决策类备案
- 图片生成备案




截至2024年8月，已通过“算法备案”的AI应用数量为 ***1,919***

**“大模型备案”**

生成式AI服务备案是网信办针对具有舆论属性或社会动员能力的生成式AI服务在算法备案外设定的另一备案要求。目前生成式AI服务备案的主要对象为提供大语言模型技术的生成式AI服务提供者。

**截至2024年8月：**

| | |
|---|---|
| 已备案模型数量 | *188* |
| 已登记应用数量 | *26* |

**通过备案模型举例：**

- 大模型备案

# 执法案例说明 - AIGC服务违法违规处罚

*来源：https://mp.weixin.qq.com/s/ayHC7MpG6Jpiogzp-opQFw
**来源:北京互联网法院（2023）京0491民初11279号民事判决书
***来源：广州互联网法院（2024）粤0192民初113号



2024年上半年，重庆市网信部门涉及依法查处违规从事生成式人工智能服务的网站：

## 违规提供AIGC服务 – 未备案、未经安全评估等

针对“灵象智问AI”“重庆哨兵拓展迷”等网站**未经安全测评备案、违规提供生成式人工智能服务**问题，依法对运营主体开展执法约谈，责令立即停止相关服务；对**未经安全评估上线**提供Chat-GPT生成式人工智能信息服务的“南川区蓉城网络科技工作室”依法开展执法约谈，**责令立即关停相关服务。**

## AIGC内容安全 – 传播法律法规禁止发布信息

针对“重庆酷酷跑网”“律师服务网”“开山猴AI写作网”运营主体**未尽到内容审核义务**，致使网站出现法律、行政法规禁止发布、或传输的信息内容，依法对3个网站运营主体分别作出**行政处罚**；对属地某App**履行主体责任不到位**，致使平台账号发布法律、行政法规禁止发布的信息内容，按照《中华人民共和国网络安全法》对其运营主体作出**罚款10万元的行政处罚。**

## 数字版权侵权与被侵权

### ” AI文生图 “著作权被侵权纠纷*

2023年2月24日，原告李某使用AI图片生成软件“Stable Diffusion”通过输入提示词的方式生成古装少女图片，后将该图片以“春风送来了温柔”为名发布在小红书平台，并标注为“AI插画”。但在2023年3月2日，原告发现被告刘某通过百家号账号发布了名为《三月的爱情，在桃花里》的文章，文章里使用了该图片，并且去除了该图片原有的水印。随后，李某以侵害作品著作权和信息网络传播权为由将刘某起诉。2023年11月27日，北京互联网法院对上述案件做出了一审判决，主要内容为被告人刘某赔礼道歉，并赔偿经济损失500元。


### AIGC平台侵权案**

2024年2月26日，广州互联网法院近日生效了一起生成式AI服务侵犯他人著作权判决。Tab（化名）网站生成的奥特曼形象与原告奥特曼形象构成实质性相似。该案认为，被告（某人工智能公司）在提供生成式人工智能服务过程中侵犯了原告对案涉奥特曼作品所享有的复制权和改编权，并应承担相关民事责任，要求被告平台立即采取相应技术措施防止生成侵权图片，并赔偿原告经济损失及合理开支10,000元。

2023年8月，多位知名画家发现在小红书上上传的原创作品，未经授权被小红书非法投喂给自家AI模型软件Trik AI，并且出现大量侵权图例证据。2024年6月20日北京互联网法院在线开庭审理了该案件，目前该案件正在进一步审理中。

# 3. 负责任的人工智能治理模型建设

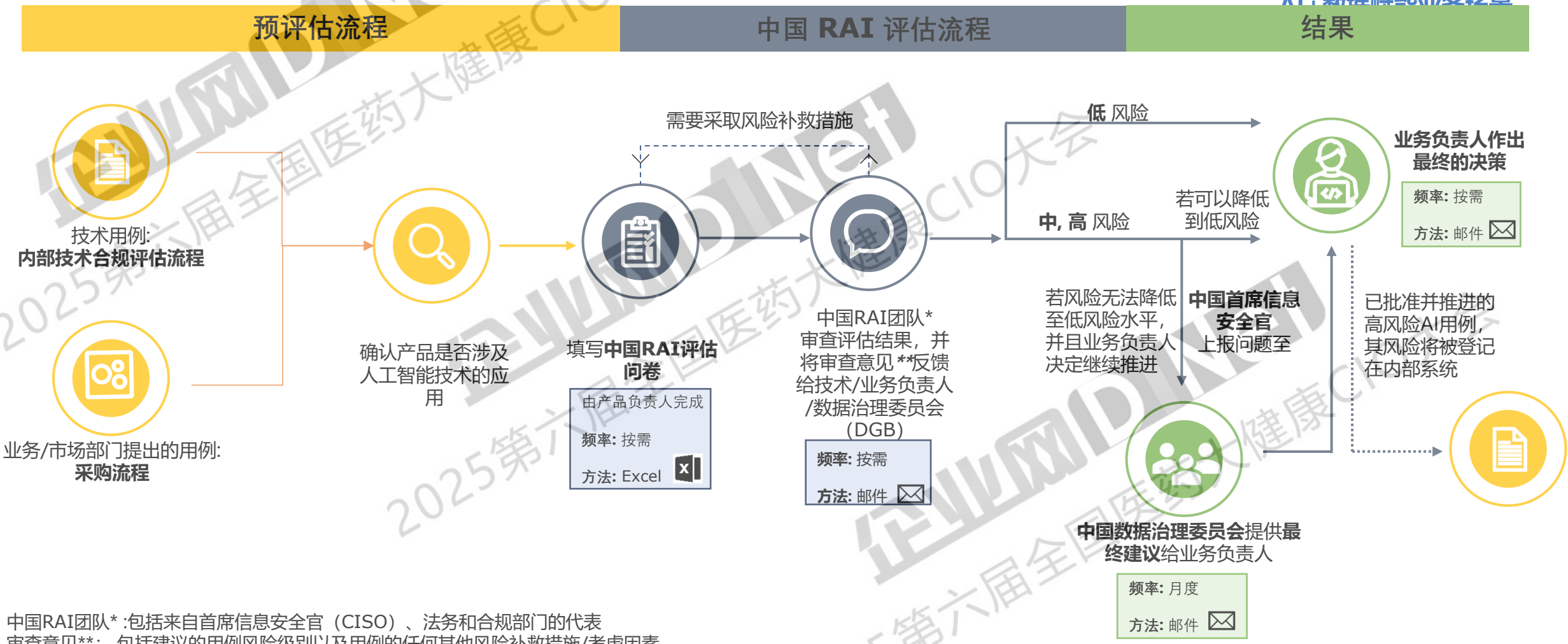
# 建立负责任的人工智能评估



| 用例使用所在地： | 适用评估流程及其评估问卷类型： |
| --- | --- |
| 中国大陆 | **中国RAI评估流程*+中国RAI评估问卷** |
| 中国大陆 & 全球总部 | 中国 & 全球总部的RAI评估流程<br>+<br>中国 & 全球总部的RAI评估问卷 |
| 全球总部（除中国大陆外） | 全球总部的RAI评估流程 + 全球总部的RAI评估问卷 |

# 负责任的人工智能评估流程






中国RAI团队* :包括来自首席信息安全官（CISO）、法务和合规部门的代表
审查意见**： 包括建议的用例风险级别以及用例的任何其他风险补救措施/考虑因素

# 中国RAI评估问卷方法总结



**中国RAI评估问卷的主要原则**

- 满足中国**监管**要求；
- 与全球评估问卷结构**保持一致**；
- 通过范围问题，针对实际用例**定制**评估问题；
- 具有**可操作性**的评估结果和**自动化**风险评级；
- **内容丰富**的帮助文本，提供更清晰的指导。
- 针对复杂控制点/任务的额外**指导/准则**，例如算法备案、数据标注、生成内容测试、安全评估。

此人工智能应用是否被xx使用?

此人工智能应用是否涉及第三方（若涉及，请明确与第三方的合作形式）？

此人工智能应用是否涉及机器学习训练?

此人工智能应用的目标使用场景?

此人工智能应用的模型类型?

此人工智能应用是否具有舆论属性?

此人工智能应用是否涉及以人为研究对象的生命科学和医学研究?

1.1 针对人工智能应用可能涉及的训练数据及/或生成内容的**知识产权**问题，是否已在本项目组中指派了专人管理?
1.2 请确认第三方供应商已签订**服务协议**（含标准知识产权条款）且其已经书面承诺（已签署《**生成式人工智能承诺书模板**（赫力昂非生成式人工智能服务使用者）》?
2 您是否**与第三方供应商签署了合同**，并在合同中对于**人工智能相关服务的安全合规责任**进行了约定?
3 **第三方供应商**是否已开展相关人工智能服务**安全评估**并具备在中国市场运营的**相应资质**?
4 你是否已经根据相关要求对人工智能系统的**生成内容进行测试**，并理解该测试是如何确保该生成内容的高质量、安全性及符合国家标准?
5 您是否对训练数据进行**规范标注**并开展安全评估，以确保训练数据的合法性、安全性、多样性、可追溯性符合要求?
6 对于提供具有舆论属性或社会动员能力的人工智能应用，您是否在应用系统上线前或发生主要功能变更时，开展了**全面的安全评估**?
7 针对人工智能应用，您是否建立了基于用户实名管理的**服务监测机制**（包括对输入内容质量的管理）及**事件响应机制**，并同时建立接受公众或者使用者**投诉举报**的**渠道和处理规则**以及时限?
8 对于提供具有舆论属性或社会动员能力的人工智能应用（如生成式人工智能、算法推荐、或深度合成服务），是否已按相关监管要求完成**备案手续**?
9 对于人工智能应用生成内容，是否进行了**显著的标识**，以防引发公众混淆或误导?
10 您是否建立了相应的**信息公开机制**（包括在用户交互界面或服务协议中向用户告知相关人工智能服务的详情、风险、局限性、备案编号等信息）以确保人工智能服务的透明度?
11 人工智能应用是否已按相关监管要求完成**科技伦理审查**?

- ❑ 针对每个评估问题的自动化风险评级
- ❑ 针对整体AI用例的自动化风险评级
- ❑ 针对相应的控制点需收集相关证据
- ❑ 关于改进措施的详细指导和说明

## 风险评级机制

对于现有的控制检查类问题，我们采用了最终统一风险判定方式。
如所有的问题回答均为低风险，则总体风险判定为**低风险**。
如有一个问题出现了高风险，则总体风险判定为**高风险**。
其他类的回答组合均为**中风险**。

# 中国RAI补充文档



中国 RAI 补充文档是根据全球 RAI 政策/标准与中国本地法规要求之间的差距分析制定的，作为仅适用于中国市场的补充内容。

## 全球RAI文档

**全球RAI政策**

**全球RAI标准**

***原则:***

- 问责制和人工干预
- 透明和可解释性
- 公平和减少歧视
- 隐私和尊重
- 安全和可靠
- 可信的科学

## 中国RAI补充文档

**针对RAI政策/标准的中国补充要求**

***原则:***

- 问责制和人工干预
- 透明和可解释性
- 公平和减少歧视
- 隐私和尊重
- 安全和可靠
- 可信的科学
- 合法性和合规性

**中国RAI工作指导书**

- 算法备案
- 生成内容安全评估
- 数据标注
- 安全评估
- 科技伦理风险评估