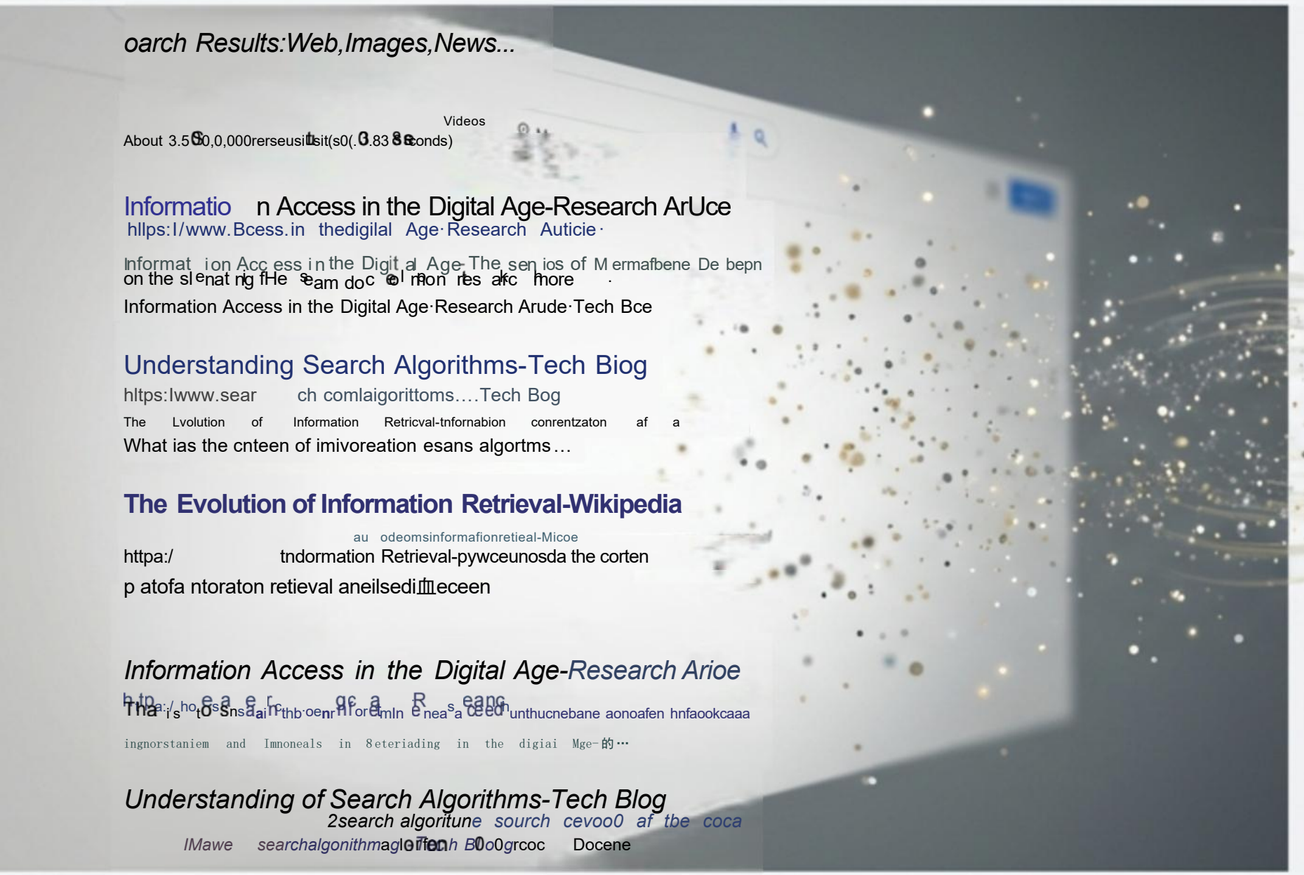
AI认知监控平台

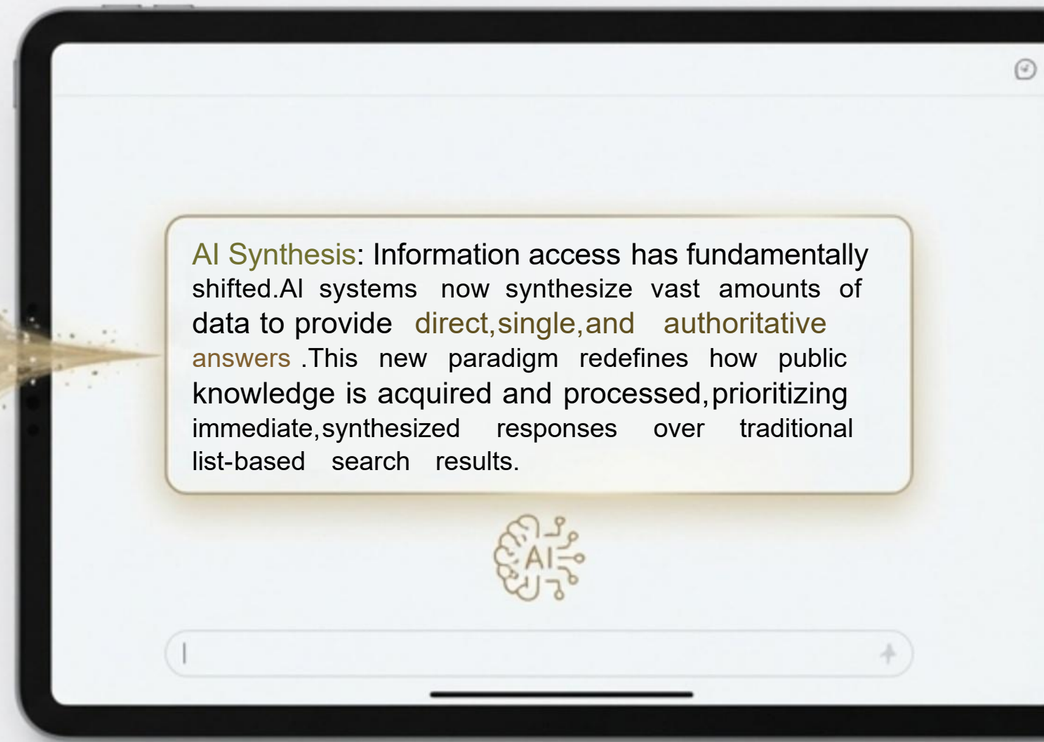
监控AI的认知结果，治理AI的认知风险

AI舆情雷达 | AI认知风险 | AI答案治理 | AI证据链审计

指导：国家广告研究院　AI品牌资产赋能中心
出品：　四川远见行人工智能科技有限公司

时代变了：信息入口从“搜索网页”到“先问AI”





公众获取信息的方式已发生根本性转变。 AI给出的“最终答案”正在直接影响的**社会认知与机构信任。这是权威与信任的新战场。**

新风险出现：什么是“AI认知风险”？

AI认知风险是指大模型在回答关键问题时，输出可能引发误解、误导或风险的“认知结果”。

认知风险：比传统舆情更隐蔽，影响更深远

AI认知风险并非声量驱动，它在无声中侵蚀信任，其影响一旦形成便难以逆转。

AI认知风险

- **低声量高影响**：无需热度爆发，直接影响关键决策。
- **入口效应**：用户直接接受AI结论，减少对原始来源的核脸。
- **复用与固化**：错误答案可能长期、稳定地复现。
- **责任难界定**：AI说错，但信任损耗由机构承担。

传统舆情风险

- **声量驱动**：风险与传播热度、情绪变化强相关。
- **传播链可见**：风险扩散路径相对容易追踪。
- **事件性强**：通常由具体事件触发，有生命周期。
- **责任相对清晰**：危机源头和处置主体明确。

传统舆情体系已无法应对“AI最终答案”的挑战

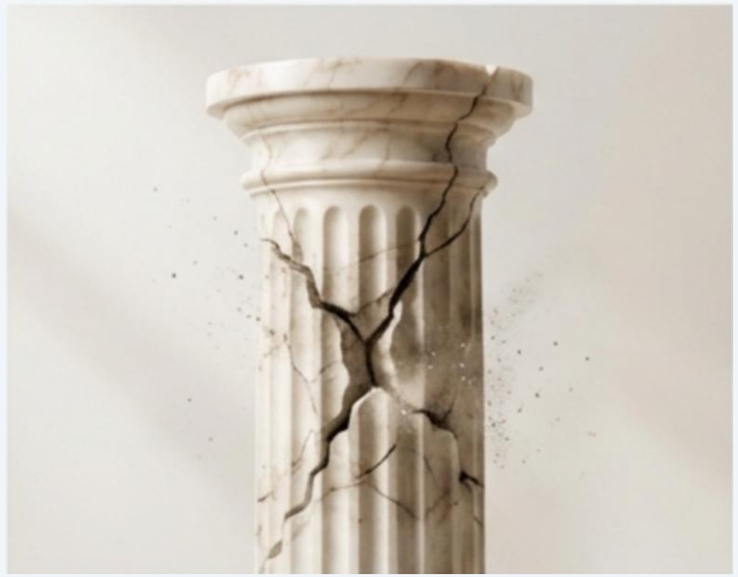
传统舆情擅长“内容采集与传播分析”，但面对AI这一“最终答案”新入口，存在天然的能力边界。

维度	传统舆情监测	AI认知监控
监测对象	新闻、社媒、论坛等“原始内容”	AI的“最终答案”+URL级证据链
风险机制	传播驱动、热点驱动	提问驱动、问句入口驱动
治理闭环	公关处置为主，结果难验收	必须具备可追溯证据链、可执行处置、可复测验收

结论：AI认知监控补齐的，正是对“AI最终答案”这一新入口的治理空白。

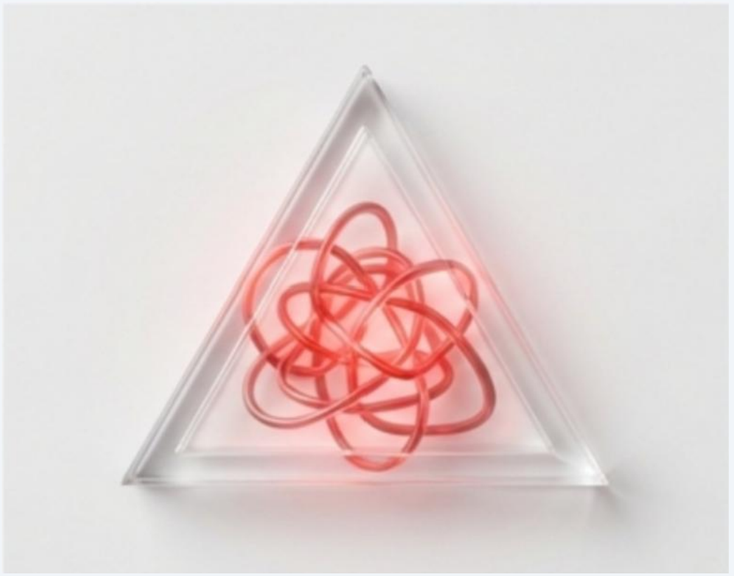
三类典型风险后果：权威、合规与治理的挑战

AI认知风险的放任，将直接冲击机构的核心价值与运营底线。



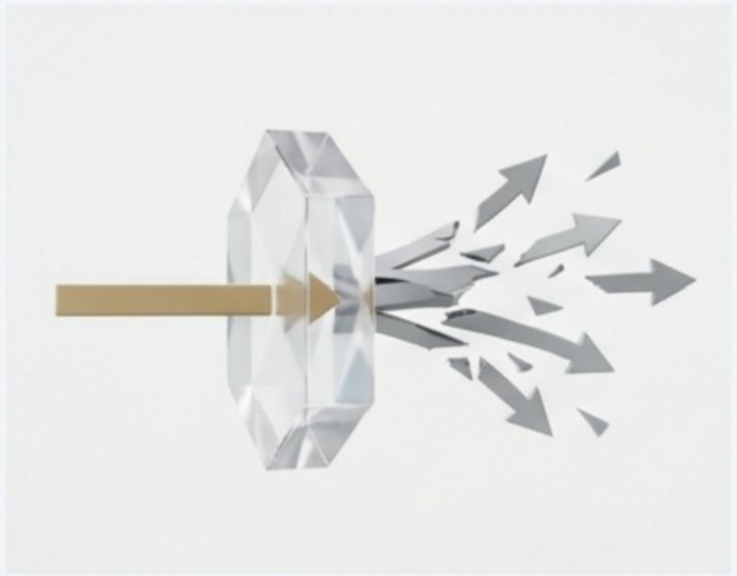
权威性侵蚀

专业机构被AI“简化成标签”，长期形成社会性错误认知。



合规与安全风险

医疗误导、金融误导、政策解读偏差引发用户投诉或监管压力。



口径与治理压力

政策服务口径在不同AI上不一致，导致公众误解、误办甚至群体性争议。

从监测到治理：让AI的回答可看、可查、可管、可验收

AI认知监控平台是对人工智能模型输出结果进行持续监测、风险识别与治理的系统平台。

准确

事实不说错，关键断言**可核验**。

权威

引用来源可追溯、可分级、可审计。

一致

跨模型、跨时间口径**稳定，不出现硬冲突**。

可控

风险可触发、可处置、可复测、可归档。

平台整体架构：一底座 · 五中心 · 两闭环

这是一个为治理而生的系统，拥有坚实的数据基石、强大的功能中心和持续创造价值的闭环机制。

一切治理的基石： AI答案证据底座

我们将AI的回答从一个“黑盒”变成可审计的证据。这是平台一切价值的起点。

“没有证据链，治理就无法进入机构流程。”

五大中心(1/2):从态势感知到风险识别

平台首先为您提供前所未有的可见性，洞察AI如何描述您，并自动识别其中的风险。

AI舆情雷达 (Narrative Pulse)

核心问题：AI在谈论我们时，主流观点是什么?叙事是否发生偏移?

- 观点聚类
- 叙事漂移监测
- 负面类型分类

AI认知风险 (Cognitive Risk)

核心问题：AI的回答有没有说错?有没有误导?风险集中在哪些问法里?

- 事实错误检测
- 口径冲突识别
- 风险分级 (L1-L3)

五大中心(2/2):从风险处置到闭环审计

发现风险只是第一步。平台提供完整的工具链，将风险转化为可处置、可审计的治理任务。

AI答案治理 (Governance Workspace)

核心问题：发现风险后，如何处置?如何纠偏?如何验收效果?

- 风险工单流转
- 纠偏建议
- 复测验收报告

AI证据链审计 (Evidence Audit)

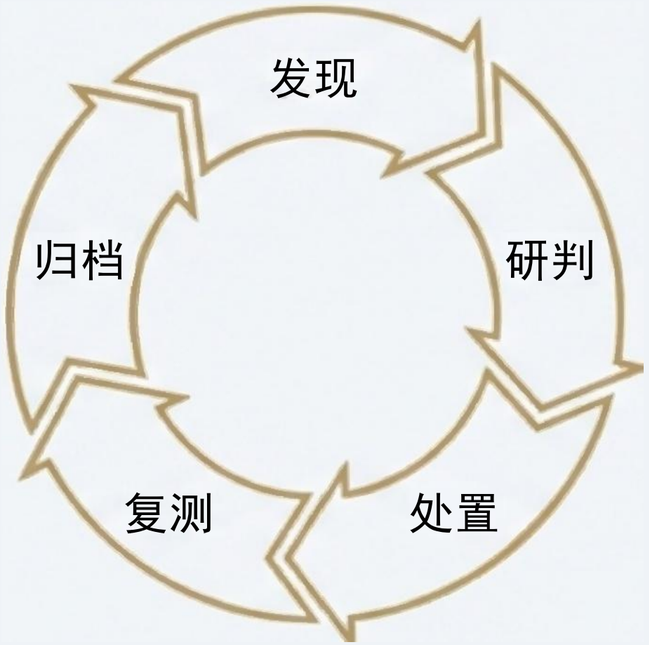
核心问题：AI为什么这么说?引用了谁?能否一键导出形成审计材料?

- URL证据链溯源
- 来源权威度分析
- 一键导出审计报表

两大闭环：风险可审计，资产可积累

平台的核心价值在于其闭环设计，不仅能解决当前问题，更能沉淀治理能力，让您的机构越用越强。

风险闭环（可审计）



Emphasizes accountability and process integrity.

资产闭环（可积累）

Each remediation cycle builds institutional knowledge and strength.

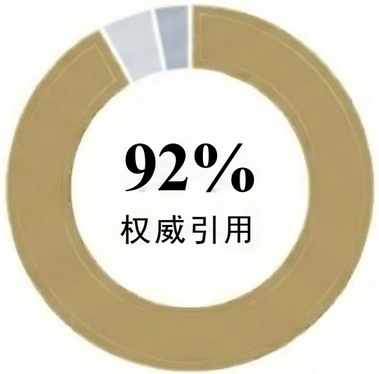
治理级KPI：将AI认知风险转化为可管理指标

我们提供的不是技术指标，而是能写入管理制度的“制度语言”，让AI治理工作可衡量、可考核。

准确性指数 (Accuracy)

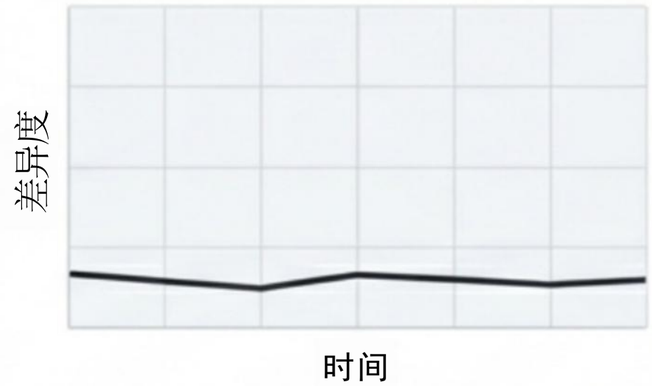
衡量事实错误、过期信息等风险的加权结果。

权威引用健康度 (Authority Health)



衡量引用URL的权成分级加权结果。

口径一致性指数 (Consistency)



衡量鸮模型、跨时间回答的差异度与冲突。

指标可用于触发风险事件、排序处置优先级、并作为治理验收依据。

核心应用场景：为您的行业保驾护航

无论您身处哪个行业， AI认知风险都已存在。这是平台如何为您解决具体问题。

医院 (Hospital)

- 监控“医疗建议出现误导倾向”。
- 修正“专家信息与擅长被错误描述”。
- 确保“就诊流程/费用/医保口径”准确。

金融 (Finance)

- 拦截“误导性承诺(保本、稳赚)”。
- 纠正“产品条款与风险等级描述错误”。
- 管理“跨模型口径冲突造成的公众误解”。

政府与事业单位 (Government &Public Sector)

- 保障“同一政策跨模型解释”的一致性。
- 确保“公共服务流程(材料、时限)”被准确传达。
- 感知“敏感议题的叙事漂移与倾向性”。

当AI成为入口，我们做入口治理

与其把AI当作不可控的黑盒，不如将它纳入治理体系。
我们不让AI少说话，我们只确保AI不说错话。

- 监控AI 的认知结果：知道它在说什么。
- 治理AI的认知风险：知道错在哪里、该怎么纠偏，并证明修正有效。

官方网址：www.yjxai.cn