

人工智能与网络安全：在实际应用中建立信任

科技、人工智能与网络安全 / 8页 / 2026年10月3日

建议决策：每项重要技术主张，都应对应明确用途、证据和后续责任。

核心风险

企业可能准确描述了某个功能，却让公众以为它保证了更广泛的结果。“自主”“安全”“可靠”等词，会形成关于错误、数据及人工工作的隐含承诺。管理层需要审视用户实际理解到什么，而不只是团队认为自己说了什么。

建议的管理选择

建立产品、安全、法务、销售与传播共同参与的流程。敏感用途的公告应说明评估背景、已知限制和预期监督。措辞超出证据时，传播部门应有权提出警示，但不应成为技术风险的唯一裁决者。

三项立即行动

明确产品公开主张的负责人；确定哪些变化会触发内容复审或暂停活动；建立在事故原因尚未查明时仍能工作的回应机制。商业发布与解释能力必须同步准备，不能先作出承诺，再寻找支撑。

决策要点

如果用户按字面理解我们的承诺，我们能否承担由此推导出的使用方式？

从演示走向真实使用

一次成功演示，不能单独证明服务在工作环境中的稳健性。

准确说明系统做什么

辅助任务、提出答案和在限定范围内作出决定，并不涉及相同的责任。介绍应说明谁在行动、使用什么数据、如何监督，以及用户何时应停止流程。创新词汇不能掩盖人、产品和组织之间的责任分配。

结合背景理解测试结果

评估结果取决于数据集、方法、模型版本和使用情景。必须说明测试了什么、未测试什么。NIST自愿性AI风险管理框架及生成式AI相关指南，提供识别和管理风险的参考[1]，但并不自动构成产品认证。

识别依赖链

用户体验可能依赖模型供应商、云服务、外部数据和第三方集成，而声誉压力通常集中于可见品牌。管理层应了解哪些依赖会改变质量、可用性或使用条件。传播需要适度解释这些关系，同时避免披露削弱安全性的细节。

决策要点

限制即使已被团队知晓，只要用户难以找到，仍属于沟通不充分。

让技术承诺处于可控状态

治理应落实到具体表述，不能停留在一般性的负责任AI宣言。

建立四项信息台账

每项重要主张记录公开措辞、支撑测试、相关限制和更新负责人。覆盖产品页、销售材料、采访和伙伴文件。产品发生变化时，应同步触发依赖该功能的主张复审，而不是等待外部质疑后才检查。

设置适度的审核流程

稳定的事实描述可以采用常规审核。性能、安全或自主程度方面的主张，需要相应专业负责人的意见。传播部门说明歧义可能造成的后果，管理层决定可接受的风险及降低风险所需资源。

避免两个极端

第一种做法是把限制藏得过深，以致主信息与之矛盾；第二种是堆积难懂的保留条款，最终无法解释实际用途。适当的准确性，应帮助具备合理信息的用户判断如何使用服务，以及何时需要人工介入。

主张	需要核实	表达边界
性能	测试、数据集、版本	结果适用的条件
安全	威胁范围和控制措施	不能保证零风险
自主程度	人工角色与监督	何时需要人工介入

决策要点

证据发生变化，就必须能够修改公开措辞，即便产品已经发布。

准备应对信任受损事件

以下虚构情景用于测试决策、责任分工和初始回应。

错误结果被公开

用户公布服务生成的错误回答。团队应区分个案是否存在、是否能够复现，以及可能影响多大范围。承认一个有据可查的案例，不等于立即判定系统普遍失效；但只援引使用条款，也无法回应实际体验及保护措施的问题。

服务中断暴露外部依赖

关键供应商发生故障。品牌应说明用户能做什么、哪些功能受影响、何时提供下次更新。不能自动把供应商预计恢复时间转化为自身的确定承诺。管理层决定连续服务措施和关于依赖关系的透明程度。

出现未经授权访问的怀疑

安全团队主导调查，获授权团队评估通知义务。传播准备与已核实事实及保护指引一致的信息。应区分不可用、系统受侵和数据泄露：其中一种情况并不能证明另外两种。新信息可以修正初次回应，但不应抹去更新记录。

决策要点

分别列出已确认事实、正在调查的问题，以及相关人员现在应采取的行动。

将专业能力转化为可信传播

专家的可见度只有帮助公众理解决策或风险时，才具有实际价值。

领导人解释选择

领导人解释选择哪些用途、如何权衡开发方向，以及愿意公开哪些限制；专家说明方法和结果；安全负责人界定可以分享的内容。不同角色互补，避免要求一位发言人代表所有专业领域。

准备经得起追问的媒体资料

区分已可用功能、实验功能和开发目标，并说明演示条件。面对批评性问题，发言人应能够解释测试方法，并承认尚未知晓的事项。准确更正，比笼统保证可靠性更有价值。

让业务团队参与解释

销售、支持和合作伙伴都直接影响声誉。他们需要使用示例、限制说明和误解上报机制。某个异议反复出现时，应检查产品或其解释，而不只是优化销售话术。这一闭环让传播帮助企业理解真实用途。

决策要点

为每个版本准备一份说明：服务能做什么、不能做什么，以及要求用户做什么。

GEO是一项准确性工作

目标是让搜索引擎和AI回答正确理解专业能力，而不是承诺控制它们的结果。

区分实体与用途

分别描述公司、产品及版本，将每项能力关联至具体用途、文档和联系人。混合机构叙事、产品目录及夸张形容，会增加理解难度。直接定义与回答，应同服务条件和技术资料保持一致。

公开可使用的证据

评估资料应说明方法、日期和限制。公开文档需可访问，重要内容需以文本形式提供。Google指出，搜索基础做法同样适用于其AI功能[2]。任何标记或内容形式都不保证收录与引用，GEO因此不能被销售为确定排名。

组织可比较的观察

跟踪客户实际使用语言中的比较、理解和风险问题，保存回答及来源。变化可能来自提问方式、服务或模型更新，不能自动归因于网站改动。优先解决具有实际后果的错误，而不是只追求提及次数。

决策要点

衡量的是重要信息是否被准确理解，而不是品牌是否出现在每个回答中。

90天连接产品与声誉

行动计划围绕公开主张和信任中断情景展开。

第一个月：确定范围

选取重点产品和用途，将公开承诺与技术文档、测试及支持反馈比较。识别词汇差异和模糊限制，指定每项重要主张的负责人，并明确产品、供应商或使用背景变化时的复审规则。

第二个月：开展演练

让产品、安全、法务、支持和传播共同参与，依次引入新事实、记者询问和客户问题。检验流程能否在缺乏完全确定性的情况下给出可理解的回应，修正拖慢决策的模板、资料和职责。

第三个月：形成惯例

更新重点内容，培训使用这些材料的人员，并测试一组稳定的搜索与AI问题。管理层审视已解决差距、未解决风险和资源需要。目标是持续的更正机制，而不是仅服务于一次产品发布的传播活动。

决策要点

产品变化、证据变化和公开表述变化，应通过同一个责任流程联系起来。

信任治理仪表板

跟踪应结合证据质量、用户体验和纠错能力。

选择有意义的指标

衡量重要主张的证据覆盖率、产品变化后的更新速度，以及支持团队反复收到的误解。GEO评估应区分出现频率、准确性和来源质量。计算基础应保持一致，测试方案发生变化时必须说明。

把少数严重情形带入决策

高总体满意度可能与少见但后果严重的用途并存。管理层需要看到严重案例、发现能力的限制和团队分歧。传播既不应夸大孤立个案，也不应让平均值掩盖重大风险。

明确最终管理决定

确认承诺负责人、警示阈值以及暂停不准确措辞的权限。建立产品与安全团队共同复审，要求汇报已作出的更正，以及为何推迟某些扩大宣传力度的请求。

当解释、证据与实际体验保持一致时，技术信任才具备持续基础。

资料来源与参考

[1] NIST — AI Risk Management Framework / Generative AI Profile

[2] Google Search Central — AI features and your website

分析与建议由Belief System提出。报告中的情景仅用于示例推演。所引资料用于说明背景，不代表来源机构认可本报告的建议。