数据安全面临旧风险的重塑和新风险的实现

尽管对当前方法能否有效保护 GenAI 存在相当大的怀疑，但企业在投资 DLP（数据丢失预防）、数据发现和分类以及 GenAI 的安全方面并未受阻。尽管 GenAI 应用的非确定性本质让安全和合规利益相关者感到不安，但已有的数据处理原则和实践有能力应对这些放大的攻击，并保护数据不被盗窃或丢失机密性。然而，GenAI 正在使更新的威胁和伤害成为可能并被放大。模型偏见、深度伪造和自动生成的叙述攻击可能会威胁到企业的声誉；数据完整性的缺失可能导致可信度的丧失和品牌的损害。幸运的是，数据安全控制不仅是 GenAI 采用的抑制剂，也可以是加速器，因为企业对数据的理解和控制随着成熟而增强。

为了更好地理解 GenAI 的数据安全问题，有必要基于信息安全风险的普遍理解原则和框架。数据安全建立在三个基本原则之上：机密性、可用性和完整性。从这个三元组中，信任、安全和隐私的其他原则随之而来。

建立在诸如 NIST 800-53 等标准上的风险管理框架描述了操作化控制的顺序，以最小化风险。企业的风险是漏洞、威胁、资产和影响的组合，与可能性相乘。通过泄露、拒绝服务或不可信的体验造成的伤害风险可以减轻，但永远无法完全消除。数据安全面临旧风险的重塑，和 GenAI 实现的新风险。理解和减轻威胁为消费者和企业提供了一条前进的道路，使他们更具韧性。

机密性风险被放大

GenAI 应用可能会显著放大现有数据机密性和完整性的风险。这包括个人助理、代理、私有语言模型和连接到公共“前沿”模型。在使用 GenAI 或机器学习之前，大多数信息检索或派生是通过确定性方式执行的。在关系型、半结构化和完全非结构化数据中，索引被查询以检索和根据查询对数据进行排名。

对这种确定性行为的评估和预防是许多数据分类或发现工具的基础。由于底层数据的变化导致索引发生变化，每个单独的查询将返回相同的结果，而不考虑先前提交的查询。理论上，如果能够在企业的全部资产中进行完全准确的查询，可以确定企业的敏感数据。

GenAI 在支持信息检索时不提供这种确定性。无论是仅使用公共大型语言模型，还是用内部检索增强生成（RAG）或私有内部 LLM 使用来补充，结果都不是确定的。安全团队想要了解用户是否特定访问了任何数据，将面临艰巨的任务。

GenAI 在尊重或区分代理方面面临挑战。在 GenAI 架构中，代理或助手必须清楚地理解用户以及特定访问信息的角色或上下文。虽然像 GPT-4o 或 Llama 3 这样的大型前沿模型的数据被认为是公共的，但企业信息不是，也不应该被自由共享或访问。在最低权限模型中管理这些动态权限将很困难。

新的机密性风险

鉴于在检索或生成信息时缺乏确定性，区分用户或代理意图变得更加困难。在所有 GenAI 使用中，尤其是在聊天或会话功能中，都将有一定程度的直接或间接提示注入。LLM 根据收到的用户交互来确定或预测要返回的信息；在同一会话中以不同顺序询问的多个交互返回不同的信息。

要求无害会议摘要的用户可能会从高度机密的演示文稿中提取材料。尽管 DLP 和第一代 GenAI 安全工具可能会监管无意的"提示注入"，但对抗性提示从 GenAI 应用中提取或篡改需要不断变化的检测规则、工作流程和补救措施。防范提示注入是一场猫捉老鼠的游戏，涉及内部和外部威胁行为者。

在这些非确定性情况下实现信息治理是困难的。对于要求删除特定个人身份信息的数据主体访问请求（DSAR），在理论上对于企业自有资产中的确定性系统可能更直接。与公共 LLM 结合使用，可靠地遵守任何给定的 DSAR 是一个重大挑战。公共和私有数据源的结合使得确定 PII 的血统成为一个挑战，如果该 PII 是从企业控制范围之外的系统检索的，但仍然出现在其系统中。

对于开始使用内部 LLM 的企业，存在模型数据盗窃和模型权重盗窃的问题。一般来说，LLM 是其模型和权重的总和。对于某些企业，模型的权重和偏见将是专有的；对于攻击私有 LLM 的对手来说，衍生或窃取模型的权重实际上窃取了企业商业模式的一部分。随着企业以不同的方式与政府或其他企业合作，这些组合系统使得确定性和正当程序难以证明。

难怪，根据 2023 年生成性 AI 商业趋势研究，67%的受访者认为当前的安全技术不足以应对 GenAI 带来的风险。这种情绪在 61%的大型企业受访者和 70%的 SMB 中共享。尽管处于数字化转型阶段的受访者对 GenAI 的风险保持警惕，但那些没有数字化转型策略的组织最担心（77%）。

新的完整性伤害

GenAI 独特地为数据完整性和可信度创造了新的损害机会。就像从普通和无辜到恶意尝试提取或泄露机密信息的提示连续体一样，从每次 GenAI 互动中也有准确性和可信度的连续体。因为 GenAI 不是确定性的，其绝对准确性和可信度无法证明或验证。

就像输入被监管以用于用户意图一样，输出必须由用户解释监管。与用户现实不符的 GenAI 结果被视为幻觉——然而，如果它们具有足够的可信度，幻觉可能被接受为真相。对于企业来说，可信度的轻微降低是微妙的，但随着时间的推移会累积。非确定性的本质还揭示了其他对完整性和可信度的伤害。常见的前沿 LLM，如 GPT-4o，不透露其权重，难以确定的偏见或其他滥用行为。

有了 GenAI，对手可以迅速生成更多的恶意幻觉。合成的、更改的或深度伪造的数据对维持企业可信度尤其有问题。图像、文本、内容或企业的其他虚假表现可能非常有害。更具挑战性的是，GenAI 系统正在摄入的许多内容已经被合成更改。像 iPhone 15 这样的当前手机使用一系列 AI 更改每个拍摄的照片。生成的文本或内容，包括今天的幻觉，可能训练明天的前沿模型。Reality Defender，2024 年 RSAC 创新沙箱竞赛的获胜者，利用 GenAI 检测生成或深度伪造的内容。

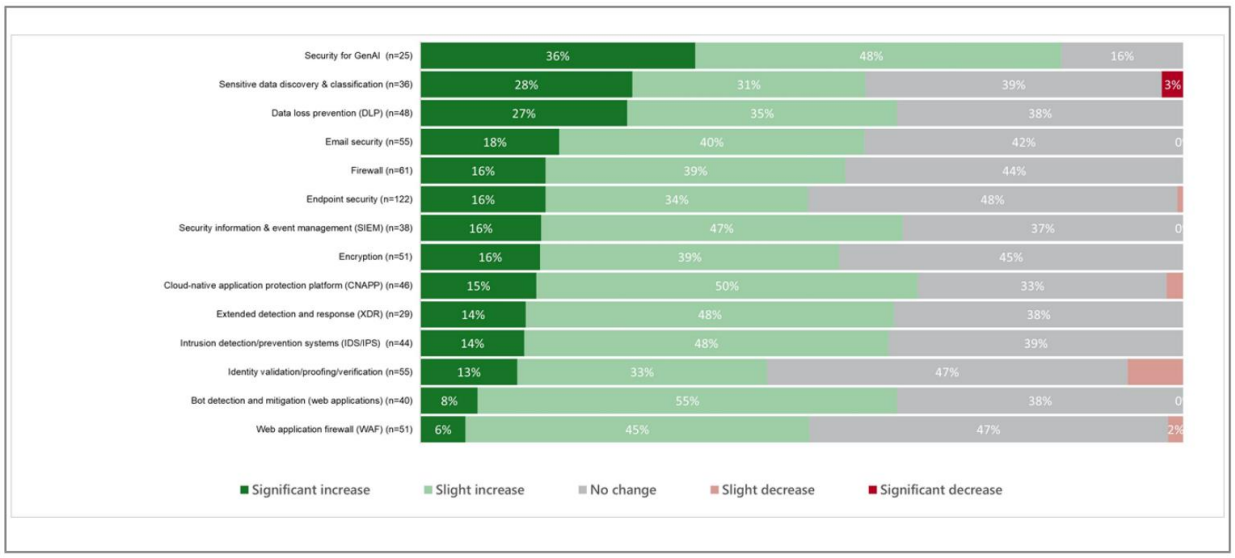
对于企业和政府来说，打击可能由 GenAI 创建或促成的虚假信息或错误信息是一个巨大的挑战。针对组织叙述的攻击——内部和外部通信和内容的语料库——对品牌感知、危机响应和国家选举产生了实质性影响。Blackbird.AI 是一家寻求防范错误信息或虚假信息下游伤害的初创公司。

新的控制，旧的控制

尽管 GenAI 架构呈现或放大了风险，但现有的数据处理原则和实践可以被应用来减少数据盗窃、篡改或其他妥协造成的伤害。根据技术路线图，支出增长最大的三种技术是 GenAI 的安全、敏感数据发现和分类以及 DLP——72%的受访者预计启动或增加 GenAI 安全支出，59%将增加敏感数据发现和分类支出，52%将增加 DLP 支出。只有 3%的受访者表示他们将减少敏感数据发现的支出。

新型隐私增强技术正在获得新的势头，以防止模型盗窃和中毒，以及未经授权或机密数据进入 GenAI 工作流程。AWS 在数据清洁室和保密计算等领域宣布的改进应该有助于其 SaaS 提供商客户代表自己的客户安全地处理敏感数据。初创公司和更成熟的供应商将寻求利用匿名化、保密计算、令牌化、合成数据、加密使用（如同态加密）和多方计算的各个方面来保护模型和数据。希望在与其他 AI 代理、API 或公共 LLM 接口之前锁定其敏感数据的企业，将需要在数据效用和隐私之间取得平衡。

How will your spending on the following technologies change in the next 12 months?



数据来源：451 research

涉及 RAG、私有 LLM 和其他内部敏感数据存储的使用案例，DLP 和数据分类已变得更高优先级。以非确定性或不可预测的方式检索的机密数据类型对安全和合规团队来说是令人震惊的。曾经存在于默默无闻中的现有数据存储或类型，可以被代理和 API 检索，这些代理和 API 可以无限期地使用。立即数据分类的原则，以及管理归属和血统的上下文感知 DLP，将帮助安全团队最好地理解非确定性情况。

虽然风险和伤害类别存在挑战，但 GenAI 为进一步数字化转型和帮助企业实现其目标所提供的机会不容忽视。根据客户体验报告，43%拥有正式数据转型策略的企业正在寻求 GenAI 以增强客户服务或支持建议，而没有正式数据转型策略的企业中这一比例仅为 30%。寻求进一步现代化其数据姿态的组织可能对整体数据安全采用最为重要。数据安全和治理可能不是 GenAI 采用的约束，而是推动者，因为利益相关者共同实现他们的目标。