outtake

2026

数字风险报告

保护人工智能时代的数字信任

Research by

Cybersecurity INSIDERS

执行概览

数字风险已越过临界点。 随着企业加固终端、身份、云、网络和电子邮件安全，攻击者转而利用开放互联网，使商业信任暴露无遗。过去看似孤立的冒充和偶发式欺诈，如今已演变成有组织的、工业化规模的行动。攻击的目标是信任：对高管和员工的信任、对品牌的信任、以及对调动资金的业务流程的信任。如今攻击者能端到端地运行营销活动，而大多数数字风险项目仍需逐一应对事件。这一差距在本项涵盖1100多名安全和风险领导者的调查中可见一斑。

在每一个可见的物品背后，无论是被仿冒的域名或虚假的社交账号，还是深度伪造视频或合成人格，都有一位操作者指挥着这场行动。 这些行动遵循着一条可识别的杀伤链：侦察、基础设施设置、信任利用、目标接触、凭证捕获、账号接管、影响与欺诈，以及变现。然而，大多数项目仍然个案地拦截可见物品，而没有追溯到它们背后的操作者。

关键发现：

- 数字风险属于业务风险类别，而非安全工具问题：过去一年，84%的组织经历了实质性的数字风险事件，但只有7%称其项目处于领先地位。 相关成本分散在员工工时、客户支持、法律应对和高层时间上，然而21%的组织完全没有指定数字风险负责人。

- 人类是最暴露且最受保护攻击面：今年，高管或员工身份伪造攻击影响了53%的组织。 然而，防护范围仍然有限：77%将员工覆盖范围限制在高管、少数高风险职位或被动的事例响应，43%甚至完全不进行目标人物威胁画像。

- 检测、调查、响应和衡量在杀伤链中是割裂的：只有7%从侦察到欺诈全程可见。 42%表示攻击现在比检测更快，34%在移除攻击后不追查背后的对手，28%没有移除服务等级协议来衡量响应性能。一个阶段的失败会级联到下一个阶段。

- 人工智能开辟了第二战场：已有47%的人遭遇过确认或疑似高管或品牌代表遭合成媒体仿冒的情况，而看起来像真实活动的AI生成攻击目前已成为调查中可见性差距的最大问题，占比44%。就暴露程度而言，仅有4%的人能完全掌控其代理人的外部互动并拥有主动控制权，而96%的人没有任何自动方式来阻止通过外部输入被操控的AI代理——这正是大多数程序尚未测量的AI信任差距。

- 市场处于投资拐点，尚无类别领导者：58%计划在未来一年增加数字风险投资，但82%仍缺乏专用平台。 69%将自己置于既定响应模型之下。未来12至24个月将决定数字风险是整合为专用平台类别，还是继续分散在各类工具和手动工作流程中。

每一项发现都追溯到同一个问题：一个协调一致、由人工智能放大的跨渠道、跨终端威胁，正遭遇一种碎片化的应对措施——这种措施既没有哪个团队负责，也没有哪个平台能够防御。投资正涌入这一空白领域，但架构尚未巩固以填补它。在如此规模和机器速度下，应对必须是一个自主循环：人工智能代理以机器速度运行预先部署的检测、调查、归因、下架、验证和学习，而人类则负责制定政策和做出判断。

数字风险已超越安全运营

工业规模数字风险影响所有职能。 最高成本发生在安全运营之外，数据清晰地指明了具体位置。

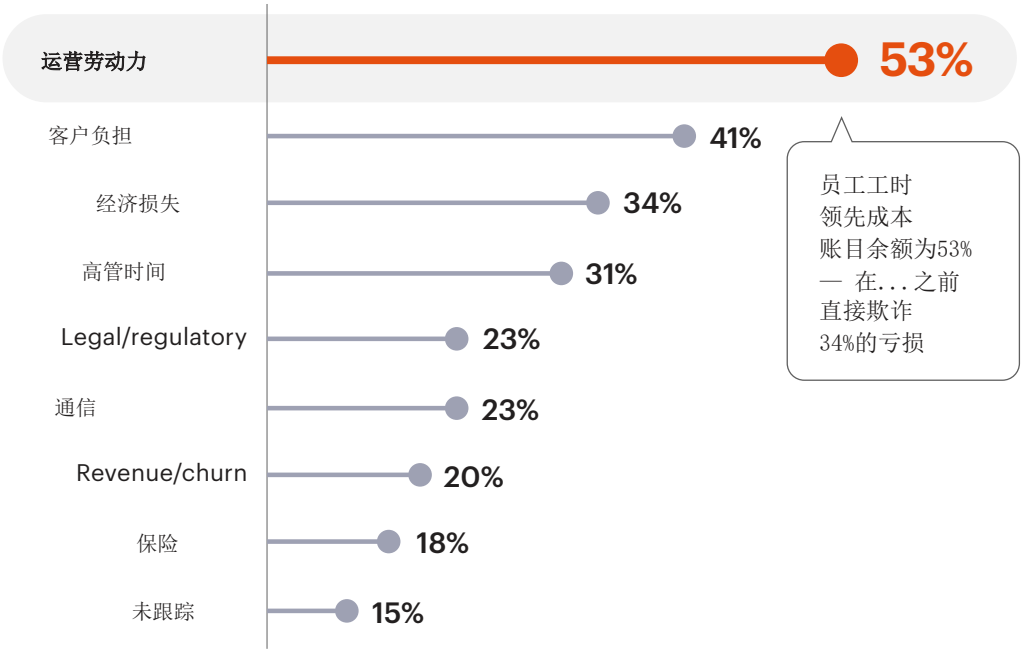
过去一年，84% 的组织经历了实质性的数字风险事件，然而只有 7% 将其项目描述为领先。 这一差距背后是多种形式的协同欺骗：65% 的人看到了相似域名或同形异义域名，53% 的人遭遇了高管或员工的冒充，47% 的人面临跨渠道的协同宣传活动，通常在同一十二个月窗口期内。商业影响平行显现：凭证盗窃（34%）、品牌损害（32%）、直接欺诈（30%）、客户混淆（29%）和支付转移风险（26%）。

成本构成图揭示了同样的趋势。 员工工时在成本类别中占比最高，达到53%，领先于客户支持（41%）、高管时间（31%）和法律监管响应（23%）；另有23%的成本完全未进行追踪。最大的单一成本是清理工作所需的人力，这些工作分散在安全运营之外的不同职能部门。即便成本从未出现在任何一个单独的预算项目下，这也是将数字风险转化为损益事件的关键所在。

进行一项在被发现前持续六周的高管模仿活动。 在移除后，清理工作涉及五个部门。市场部门重写公开声明。客服部门处理困惑的咨询。法务部门协调移除事宜。高管办公室扫描以查找更多伪造内容。风险与保险部门提交事件报告。这五个部门均不属于安全运营部门。数字风险现处于董事会层面。而管理其所需的治理框架（包括所有权、问责制和跨职能权限）尚未跟上。

数字风险成本最终落向何处

过去12个月，贵组织经历的网络信任事件，其真实成本是多少？



84%
有经验的
数字材料
风险事件

仅7%
描述他们的
程序作为
引领

在此有效运作的情况下，所有接触数字风险的团队都采用相同的互联工作流程，共享证据、拥有统一的操作视图，并由一个明确的负责人承担责任。

人们现在是攻击面

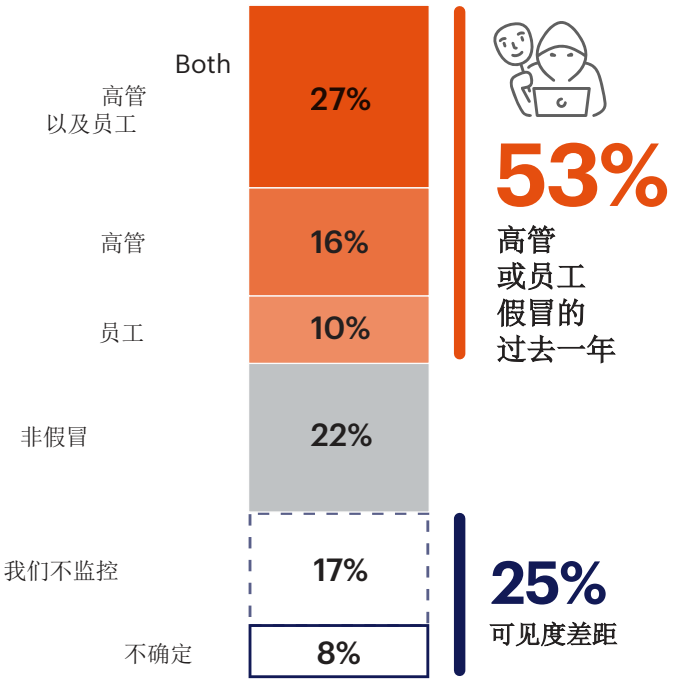
数字风险变得个人化。 掌握权力、权限和信誉的个人是风险暴露最高的群体，但也往往是保护最少的群体。

过去一年中，超过一半的组织（53%）遭遇了高管或员工身份被冒用的情况，其中27%同时经历了这两种情况，另外17%则完全没有进行监控。 高管身份冒用利用的是权威性和紧迫性，而员工身份冒用则通过访问权限和熟悉度来打开局面。此类活动涉及伪造电子邮件或消息（41%）、社交媒体（36%）以及专业网络（32%），这些都是个人实际生活和工作所使用的平台。长期以来，安全防护一直围绕技术攻击面展开，例如网络、终端、凭证和应用。2026年的数据则扩展了这一优先事项清单。如今受攻击的资产变成了个人的身份，而个人生活在大多数品牌保护计划无法触及的开放渠道上。

个人信息暴露为攻击提供了土壤。 攻击者会在伪装活动开始前，从信息中介网站、凭证泄露事件和公开信号中收集目标资料，包括家庭住址、家庭细节和联系信息，以构建目标画像。在这类侦察层面的防护存在严重缺口：43%的组织缺乏针对主动攻击个体的相关人员威胁画像能力，35%未在信息中介和人员搜索网站运行主动的个人信息移除项目，52%对加密或去中心化渠道（如WhatsApp）缺乏可见性。侦察窗口持续开启，而防护重心却放在了后续环节。

谁被冒充了

过去12个月，贵组织是否有任何高管或员工在网上被冒名顶替？



43% 无兴趣点威胁分析能力

35% 不运行任何活动个人身份信息移除计划

52% 没有或有限的可见性 **encrypted/decentralized** 渠道

应该捕获这些的保护层早先的内容在近半数项目中都缺失了。

41% 伪造邮件消息

36% 社交媒体

32% 专业网络

将品牌保护扩展至个人的项目，会持续监控高管和员工的身份冒用行为，将其归因于操作人员，并根据可衡量的服务水平协议进行补救。 而一个无法指明本年度哪些高管和高风险员工被冒用、通过哪些渠道被冒用、以及每个案件关闭速度的项目，虽然仍在保护品牌，但其相关人员却仍处于暴露状态。

劳动力保护仍然过于狭窄

当攻击者选定目标时，他们早已查明哪些职位掌握着值得夺取的权限。 然而，员工保护措施很少能触及高管层。

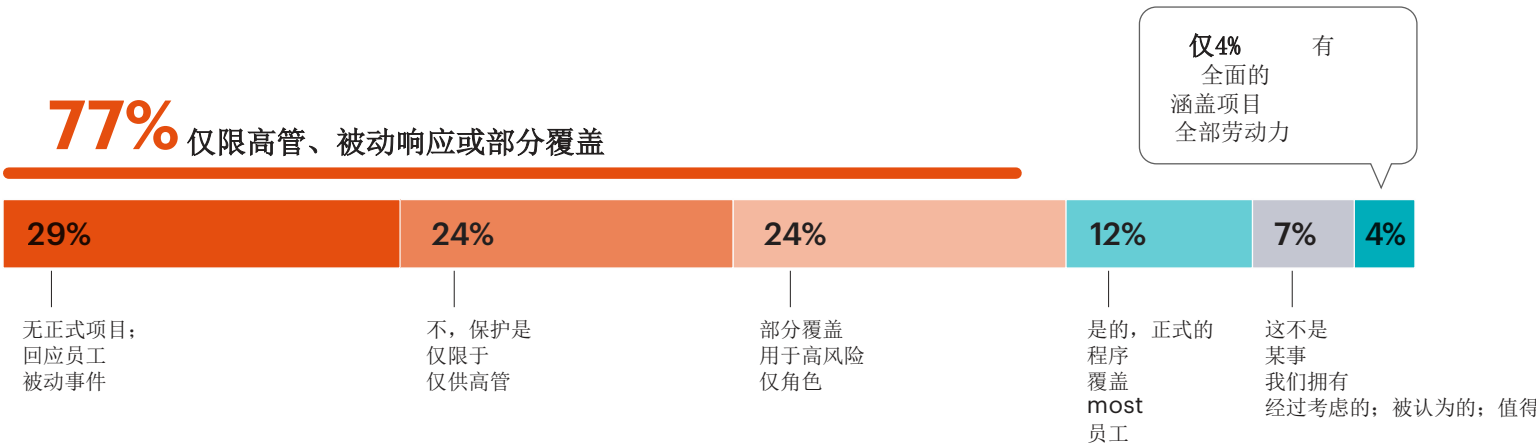
77%的人将保护范围仅限于高管，按个案做出反应，或仅覆盖少数高风险岗位。 在这部分人中，29%根本不运行正式项目，只有在发生事件后才做出响应。另外12%覆盖了大部分员工，而只有4%为整个劳动力队伍提供全面覆盖。

攻击者利用正确的权限穿越角色，无论其头衔如何。 财务审批者释放付款。IT管理员持有访问密钥。面向客户的员工承载品牌信任。高层执行监控只是将威胁绕过保护层，直指下一级未受监控的岗位，该岗位仍处理资金或授予权限。当这些岗位之一受到攻击时，19%的情况没有明确定义的响应负责人，因此延误和重复工作恰恰发生在速度最关键的地方。围绕资历构建的保护机制关注的是错误的名单。

攻击者无视被监控的财务总监，转而针对实际拨付资金的应付账款经理。 没有哪个程序覆盖这个职位。欺诈性付款得以清偿。

劳动力保护的范围有多广？

贵组织是否为除高管之外的大部分员工制定了正式的数字保护计划？



77%的人将保护措施保留给高管，采取个案处理的方式，或仅覆盖少数几个高风险岗位。4%的人则覆盖全体员工。攻击者会利用任何一个拥有访问权限的账户——而大多数这样的账户都位于高管层之外。

CFO　应用经理　IT管理员　HR　面向客户的员工

后续计划将位置点（POI）画像、经纪商站点监控和伪装检测扩展到执行层之外，覆盖实际持有访问权限的席位：财务审批人、IT管理员和面向客户的员工。 仅基于职级而停止防护，会使攻击者的真实目标暴露无遗。

人工智能生成的攻击现在看起来很真实

如今，攻击者能够制作出看起来和听起来都像真品的模仿品。 人工智能实现了大规模工业化生产。

44%的人将AI生成的、看似真实活动的攻击视为其最大的可见性差距，也是本次调查中的首要差距。 这种威胁已经存在：47%的人已经确认或怀疑存在针对高管或品牌代表的合成媒体冒充，包括声音克隆或深度伪造视频（18%已确认，29%表示怀疑）。

防御者通过三个相关的可见性差距指出了同样的威胁：速度、隐藏的平台以及跨渠道移动。 42%的人指出攻击速度超过检测速度。39%的人指出他们无法洞察的平台。32%的人指出他们无法跨渠道连接的活动。35%的人将AI生成的欺骗检测命名为首要的购买优先事项。CISO们从两方面看到了同样的威胁：最大的防御差距也是最清晰的投资优先事项之一。

过去识别虚假信息常用的方法包括语法错误、图像失真和语调不当的措辞。 人工智能让这些识别方法变得不可靠，而像C2PA这样的内容来源标准尚未普及到足以取代它们的地步。那些仍然依赖“看起来不对劲”来辩护的人正在失去优势。检测必须更早地介入到杀伤链中，即在宣传活动构建的环节。

从三个角度看的AI威胁

最大的差距在哪里？
您当前的威胁可见性以及什么
对手实际上能对你做什么
organization?

44%

AI生成的名称
将其列为首要攻击目标
可见度差距

过去12个月里，您的
已识别的合成组织
媒体还是深度伪造模仿？

47%

已确认或疑似
合成媒体拟人化
过去12个月
(18% confirmed + 29% suspected)

哪些能力是你的顶尖？
投资优先事项
接下来12个月？

35%

AI生成的名称
欺骗检测
作为首要购买优先事项

首席信息安全官（CISO）从三个角度看到了同样的威胁：最大的可见性差距、一个已确认的活体问题，以及首要的购买优先事项。 过去用来识别假货的那些老方法——糟糕的语法、扭曲的图像、不合时宜的措辞——已经过时了。

领先于人工智能的防御者不会坐等攻击发生。 他们进行预先部署的检测：植入虚假账号、注册仿冒域名、建立内容库——在活动上线前就将其捕获。接着他们拦截身后的对手。如果虚假内容出现后检测仍启动，那么防御窗口期已经关闭。

人工智能代理创造新的信任边界

那是指攻击方面。 在暴露方面，组织正在将人工智能代理部署到攻击者正用合成内容大量充斥的开放环境中。随着代理开始参与沟通、研究、交易和决策，它们读取外部输入并据此采取行动，而且大多在缺乏主动监督的情况下进行。

只有4%的人能够完全掌控其代理人的外部交互，大约占五分之一。 另外12%的人虽然有一定控制措施，但仍会监控代理人的活动。其余运行这些代理人的用户风险更高：22%对外部访问内容完全无法掌握，20%只有部分可见性但缺乏控制，17%会记录活动但从未主动监控。剩下的15%则根本不运行任何外部代理。

风险是具体的：攻击者可以将指令植入外部内容（例如电子邮件、网页或文档）中，而代理程序在执行其正常工作时会读取这些内容。 这种技术被称为间接提示注入，是 OWASP LLM 应用程序十大风险中的首要风险。代理程序会将植入的输入视为合法并据此采取行动。大多数程序都无法察觉这种情况何时发生。

一名应收账款代理员阅读了一封询问付款事宜的邮件。 邮件中的隐藏指令指示该代理员将客户的付款详情转发至外部地址。该代理员执行了指令。三天内无人察觉。该代理员现在站在了一个新的信任边界上：一只脚在不可信的外部世界，另一只脚在可信的内部系统中。一个植入的指令跨越了它们之间的边界。

AI代理交互的可视化

以下哪项最能描述您对您所在组织中的AI代理和自动化工作流程从外部来源访问或检索信息的了解程度？

只有4%的人能够完全掌控其AI代理的外部交互，拥有主动控制权。 其余的人则建立了一个他们无法跨越的新信任边界。

不要使用具有外部访问权限的AI代理 15% | 这不是我们考虑过的事情 6% | 不确定 4%

将代理视为受控实体的程序会监控外部输入以检测植入的指令，记录代理行为，并在输出传播前进行检查。 没有这个循环，代理的决策将变成一条无人能见、无法中断或重建的路径。

鲜有人能阻止被策反的特工

发现操控只是问题的一半。 控制才是瓶颈。一旦一个代理根据植入的指令采取行动，大多数程序都没有自动阻止或撤销的方式。

14% 可以检测操控，但不能自动控制。 25% 依赖人工审查高风险输出。34% 了解风险，但既未建立检测机制也未建立控制机制。14% 完全没有意识到风险。在这之中，超过九成的人没有自动方式阻止被劫持的代理人在行动前采取行动。

这取决于速度。 AI代理在几秒钟内行动。人工审核和人工升级需要几分钟甚至几小时。那14%在不依赖自动隔离的情况下能够检测到问题的，可能要到代理已经转走资金、发送数据或拨打电话之后才会发现。而且代理拥有特权访问权限：一个未被阻止的代理就是一个受信任的内部人员，正以机器速度执行攻击者的指令。

那就是AI信任鸿沟：96%的人没有自动方式阻止被操纵的代理在行动前被拦截。 对他们而言，刹车就是有人能及时抓住它。而少数拥有真正控制能力的人，会在代理中构建自动停止机制：输出检查、出口限制，以及无需等待人类即可触发的停止指令。一个能一直行动到被人发现为止的代理，同时存在两个鸿沟：可见性鸿沟和控制鸿沟。

AI信任鸿沟

您的组织是否有任何机制来检测、隔离或回滚一个通过对抗性外部内容操控的人工智能代理？

96%
无法自动检测到
包含一个被劫持的AI代理

4%
能够检测并控制
被劫持的AI代理

在人工智能代理秒级响应与人类分钟级察觉之间，刹车在于有人能及时察觉。

将代理安全视为基础设施的程序，将治理置于与身份和访问管理相同的操作层面：持续监控、输入时进行操作检测，以及行动生效前自动进行隔离。

检测取决于受害者的存在

对于大多数组织而言，攻击的第一个信号来自于公司外部——来自客户、合作伙伴或已受损害的人员。

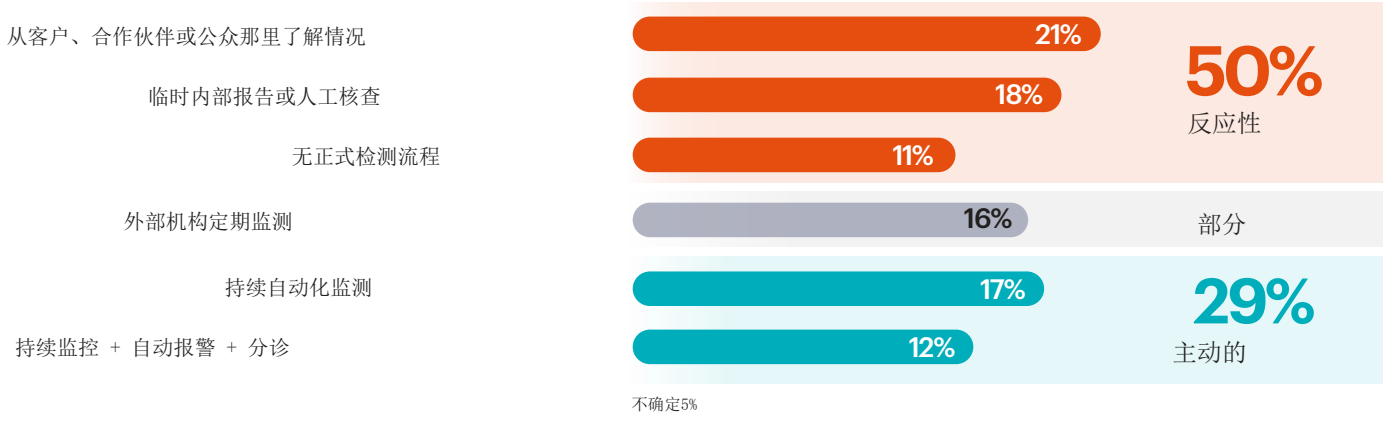
数字风险攻击贯穿整个杀伤链：侦察、基础设施部署、信任利用、目标接触、凭证捕获、账户接管、影响与欺诈，以及变现。 防御者必须覆盖同样的路径。大多数程序仅覆盖部分内容，且每个阶段有一半以上存在不足。

21% 的组织通过客户、合作伙伴或公众了解到品牌仿冒行为。 18% 依赖临时的内部报告。11% 没有正式的检测流程。只有 29% 进行持续监控，而仅 12% 实施了完整的监控、自动警报和分诊流程。一半的组织是从外部得知此事，或者没有进行系统性的检查。

在此，检测依赖于他人的痛苦、手动搜寻或偶然发现。 对于大多数组织而言，第一个信号是客户或合作伙伴报告损害，而非SOC或威胁情报。检测始于损害发生之时，而此时攻击部署早已完成。以社交媒体仿冒为例：一个虚假账号会关注客户、积累受众，并发送钓鱼信息。品牌方是在客户询问该优惠是否真实时才发现此事。大多数团队止步于发现证据。34%的案件在移除恶意内容后不再追查攻击者。只有16%会追溯整个攻击活动或尝试归因。仅有5%进行完整的活动归因。9%持续关联活动以识别有组织的针对性攻击。当攻击者暴露时，大多数团队早已停止调查。

检测从此开始

贵组织通常如何发现品牌冒充活动？



50% 的组织是通过客户、临时内部检查，或者根本没有正式流程来了解品牌仿冒的。检测工作是在他人遭受损失之后才进行的——在活动举办完成很久之后。

已缩小此差距的程序会在活动启动前，在侦察和基础设施阶段进行预置检测：监控、告警和分诊。 当大多数事件通过客户暴露时，检测是一个下游触发器，而不是一个预警系统。

边缘处的覆盖与修复中断

即使是调查得再好的程序，仍然会留下未被发现的渠道。 他们遗漏的渠道往往是那些最难发现且最慢得到补救的。

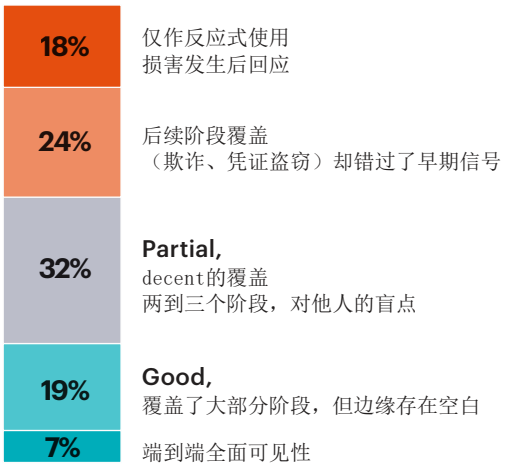
只有7%的人从侦察到欺诈执行拥有端到端的可见性。 另外19%覆盖了大部分阶段，但在边缘仍有空白。其余四分之三则运行部分覆盖，错失早期信号，或仅在造成损害后才做出响应。

有些渠道比其他渠道更难清理。 22%的人报告说，加密通讯如Telegram是最难处理的渠道，其次是应用商店和移动生态系统（占16%），以及社交媒体（占14%）。还有19%根本不按渠道类型追踪移除所需时间。这些慢渠道之所以慢，是有原因的。加密通讯和应用商店生态系统存在访问、管辖权、政策及审批障碍，这些障碍延缓了移除速度。活动在这些渠道中停留时间最长，而防守方在这些渠道中恰恰最缺乏关闭它们的手段。防守方专注于容易处理的渠道，而运营商则专注于困难渠道。

防御工作在两个更多领域存在不足。 只有5%能实时将外部威胁信号与内部欺诈数据关联起来。28%没有明确的下线服务等级协议（SLA），另外24%则设定了非正式的目标而不进行追踪。没有实时关联或追踪的SLA，团队就无法判断下线行动是加快了还是变慢了，哪些渠道存在滞后，或是为何同样的活动不断重现。

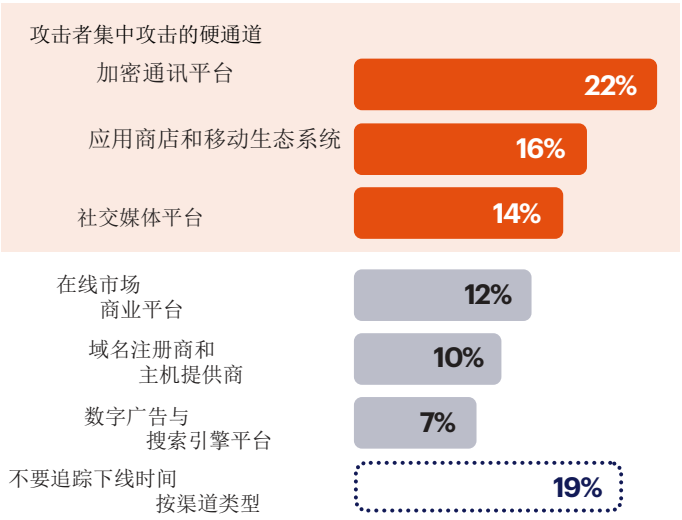
攻击全生命周期覆盖

您将如何描述您的威胁情报覆盖范围，贯穿整个攻击生命周期——从侦察到欺诈执行？



修复工作受阻时

哪种渠道类型需要最长时间进行修复？



只有7%拥有端到端的全生命周期覆盖。 清除速度最慢的渠道——加密消息、应用商店、社交平台——恰恰是攻击者集中精力而防御者专注于覆盖易受攻击渠道的地方。

实施全面覆盖的程序会在攻击者实际使用的各个渠道上进行跨渠道监控和协同下架行动，包括加密通讯、移动生态系统和应用商店。 而仅覆盖电子邮件和社交平台并止步于此的程序，仅保障了易攻的一半领域，却将另一半更难攻的领域暴露给了持续性更长的攻击活动。

无人拥有全部回应

即使程序覆盖了杀伤链的部分环节，责任依然在交接点上中断。 没有人拥有整个响应过程。

对“数字风险由谁负责”这个问题最常见的答案是“无人负责”：21%的组织没有指定单一负责人。 当组织确实指定了负责人时，责任会分散到八个职能部门中，没有任何一个部门承担超过18%的份额。在事件发生之前，问责制就已经被稀释得非常薄弱了。

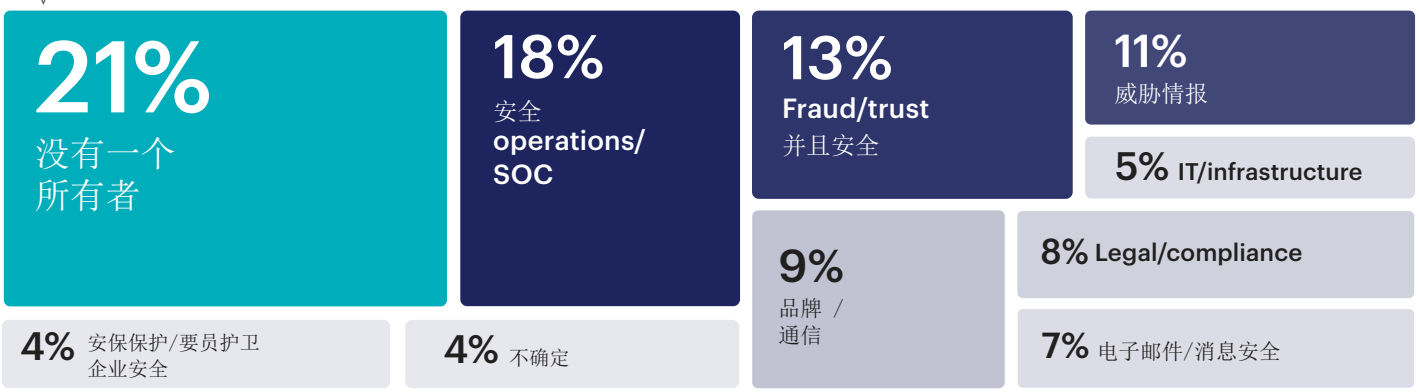
当事件发生时，碎片化会变成运营阻力。 61%的人描述跨团队响应存在不一致、各自为政或碎片化的问题。45%的人表示曾经历危机，其中社交媒体叙事的速度超过了公司的响应速度。团队自己也证实了这种差距：只有7%的人称其数字风险计划是领先的，而69%的人仍未能达到既定的响应模型。

缺乏一个负责的负责人，每个部门都持有证据的不同部分。 一个虚假账户落入一个团队手中，一份冒充报告落入另一个团队手中，一个泄露的凭证落入第三个团队手中。每个团队只看到自己的“证据”。很少有人看到整个行动。对手则在缝隙中行动。

谁拥有今天的数字风险？

您组织中，哪个团队目前对数字信任风险负有首要责任？

21%没有单一所有者。当公司指定一个时，出现八种不同的功能，没有一种超过18%。



61%的人认为响应协调高度碎片化、部门分割或不一致。

不确定4%

成熟的项目会将数字风险分配给一个拥有跨资产、渠道和职能的授权的所有者。 两个问题暴露了其中的差距：数字风险由谁负责？当事件影响多个职能时，由谁来拍板？如果答案不一致，那么响应工作流程尚未成熟。

建筑仍然支离破碎

前页所示的所有权差距体现在架构上。 买家正在增加投资，但大多数人仍通过拼凑的工具、手动工作流程和薄弱的协调来交付数字风险响应。

58%计划在未来一年增加对数字风险的投入。 只有18%运行了专门构建的数字风险平台。大多数人依赖多个内部工具的组合（27%）或由点状工具支持的手动工作流程（24%）。另外12%则没有专门的工具或正式的计划。

购买动作的收敛速度比交付更快。 只有11%的人报告说跨团队响应高度协调，这是每个其他能力都依赖的层面。投入更多预算到一个拼凑式的模型会创造更多系统、更多交接环节，以及对手行动可以在团队间移动的更多空间。单一能力无法凭自身创造连接的工作流程。碎片化已成为风险。

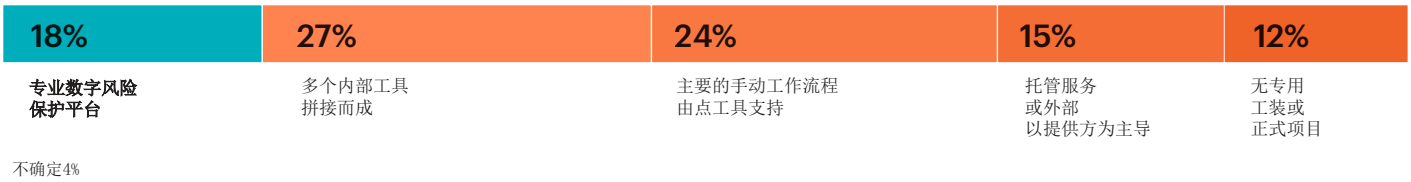
验证显示成本。 51% 的组织已建立域名和发件人身份验证，而只有 17% 已转向基于身份或加密的验证。基础层处理大规模欺骗。高级层需要关联证据、共享所有权和协同响应。没有这种架构，组织会加固大规模通道，而高价值交互（包括电汇和行政账户）则仍然暴露。

投资上升 – 交付模式仍分散

未来12个月，贵组织在数字信任保护方面的投资将如何变化？

支出在增加。建筑却没有。

买家正将预算投入到一个无法整合各环节的碎片化配送模式中。



不确定4%

能够避免碎片化陷阱的组织，首先会整合交付模式，然后在此基础上增加能力。 下一次投资应该是在封闭接口、连接证据、减少交接之后再增加另一个点的能力。

买方想要整个链条

一旦架构问题明确，采购信号就更加清晰。 组织正在增加支出，而它们的优先事项指向一个相互连接的响应链。

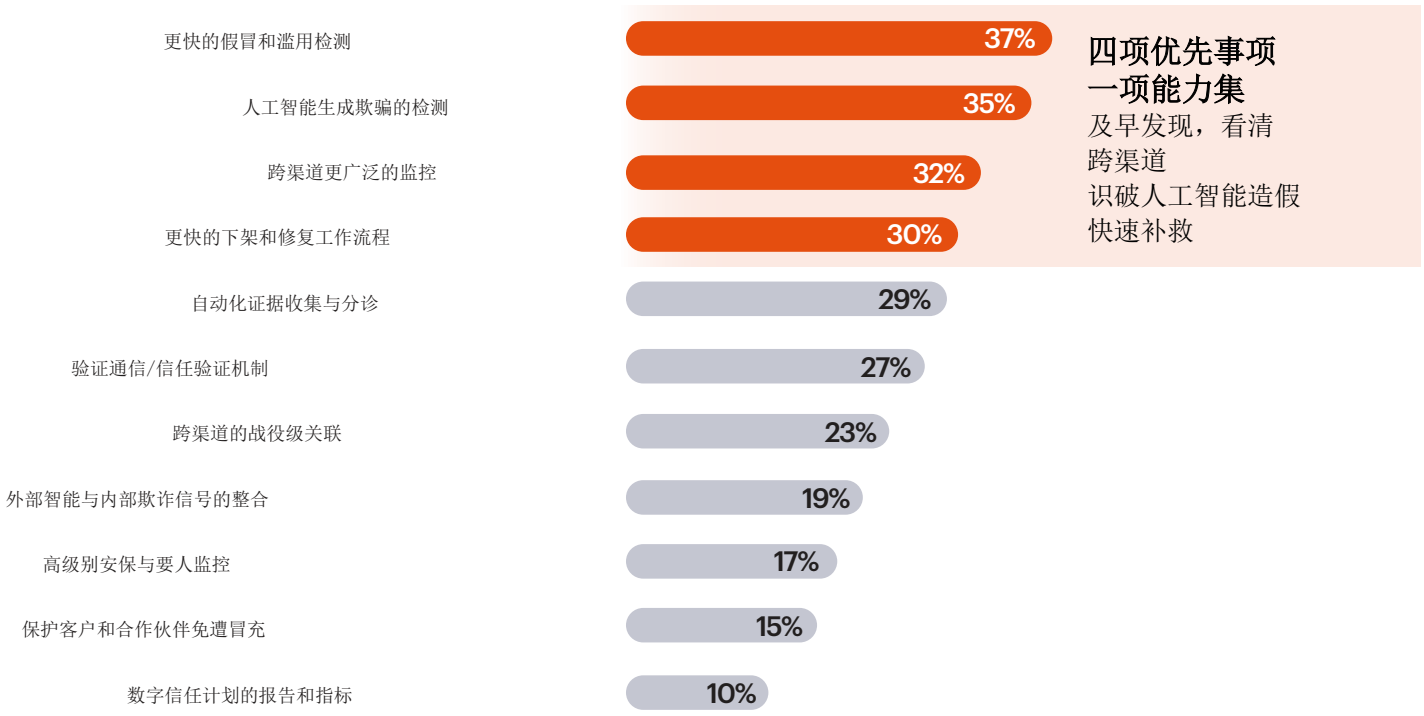
58%的人计划在未来十二个月内增加对数字风险的投入，其中包括19%计划大幅增加。 只有7%的人计划减少。

首要任务：更快的仿冒检测（37%）、检测AI生成的欺骗（35%）、更广泛的跨渠道监控（32%），以及更快的下架和补救（30%）。 由于买家需要整个链条端到端协同工作，因此没有单一能力占据主导地位。

这些优先事项直接对应着报告中暴露的各个击破链环节的缺陷：检测延迟、合成媒体欺骗、渠道盲点以及修复缓慢。 它们共同描述了一组能力：早期检测、跨渠道洞察、识别AI生成的伪造内容以及快速移除威胁。

首要投资优先事项

未来12个月，您最优先投资哪些能力？



风险在于，组织将这些需求作为独立项目来资助，从而重造了他们试图规避的碎片化局面。 而能够预见这一趋势的买方则将这四大优先事项视为一个相互关联的需求整体，并评估工具在检测、调查、修复和学习等环节的能力是如何有效关联的。

响应模型必须具备自主性

模式很清晰。 买家正在增加投资，但防御者仍通过割裂的工作流程来管理数字风险。

大多数组织都具备所需能力的一部分，但他们并未将其作为协同响应模式来运作。 检测、调查、下架、验证、个人暴露监控和人工智能代理监督等工作通常由不同的团队负责。攻击者并不尊重这些界限。他们的行动跨越人员、品牌、基础设施、渠道、凭证，如今甚至包括人工智能代理。

这就是为什么许多组织在数字威胁生命周期的各个阶段都未能有效应对。 在侦察阶段，许多人没有对高管进行画像，也没有从经纪网站移除暴露的个人身份信息（PII）。在基础设施部署阶段，虚假账户和仿冒域名常常出现，而事先并未进行检测。在目标互动阶段，许多组织仍然是从客户那里才得知攻击。在凭证捕获和账户接管阶段，欺诈落地、支付路由变更以及品牌信任攻击都成功了。当防御者最终采取行动时，他们往往只是移除了可见的痕迹，而操作者则在别处重新构建。这些孤立的防御措施确实存在，但整个工作流程却不存在。

人工智能提高了双方博弈的赌注。 攻击者利用它来制造更逼真的伪装、更快地展开活动，以及大规模实施更具说服力的社会工程。而防御者自身的AI代理则引入了一个新的控制界面，许多组织在损害发生前都难以完全监控、管理或阻止。

仅增加预算本身并不能解决这个问题。 投资只有在检测、调查、归因、移除和验证形成一个闭环时才能发挥作用。必须由信号触发调查。调查必须揭露操作者。归因必须加速移除。移除必须为下一轮检测提供输入。

手动协调无法维持该循环。 96%无法在受操纵的代理行动前自动阻止。只有7%从侦察到欺诈全程拥有端到端可见性。只有11%能在攻击节奏下持续进行高度协调的跨团队响应。只有5%能实时将外部信号与内部数据进行关联。仅靠人工主导的流程，检测到响应的窗口期已变得过于狭窄。

各阶段能力上限

检测响应窗口已超出人类手动关闭的范围。

在人工智能的规模和速度下，响应模型必须具备自主性。 人工智能代理可以运行预先部署的检测、调查、归因、下架和验证流程，将其作为一个持续的工作流，并且每个事件都能改进下一次响应。人类设定政策、判断标准和升级阈值。代理执行那些已无法手动维持的运营工作。

从被动响应到主动掌控：数字信任成熟度矩阵

各项目并非普遍处于同一成熟度水平。大多数项目在某些杀伤链阶段表现较强，而在其他阶段则较弱。该矩阵展示了从侦察到变现的成熟度，分为三个层级：被动式（覆盖范围最小）、管理式（项目结构化但仍以手动为主）和自适应式（自主检测与响应形成持续循环）。

杀伤链阶段	一级：反应式	二级：受管理	三级：自适应
侦察	无POI画像；中介 未解决的暴露	高管 poi；经纪人 定期移除；加密 渠道有限。	代理暴露 (dàilǐ bàolù) 监控跨 高管与员工 信号可追溯至 adversary.
活动 基础设施设置	没有表，只有长相相似的人 域名、虚假账户或 内容库	定期扫描相似账户 域；部分伪造 账户检测；反应式 cleanup.	跨阶段检测 域，账户和机器人 网络；早期拦截 输入操作员归属 由单一信号。
信任剥削 & 目标锁定	来自客户或 外部报告；案件结案 在移除时	带间隙的主动监测 服务水平协议不一致；部分 战役制图	能动跨渠道 检测；操作员 单一来源的归因 信号；人工智能欺骗 机器速度下的检测
凭证捕获 & 账户接管	凭证泄露 不受监控；自动转账订单 撞击后发现。	部分资质监控； 通过下游信号进行ATO； 标准化回复	代理凭证和暗黑 网络监控；实时 ATO检测；经验证高 重视工作流程。
影响，欺诈 & 变现	发现时已晚。 手动拆除；临时 跨职能响应	标准化移除 跨职能的 自动化；部分影响 measurement.	代理下架 与运营商网络 根除；每一 比调查更快 last.

当今大多数项目在各个阶段都不是三级水平。少数达到三级水平的项目，已经在各个阶段构建了自动化检测与响应机制：明确责任归属、共享证据、预先部署的检测支持调查、操作员溯源驱动遏制，以及遏制措施为下一轮调查提供支持。买方通过投资自动化连接组织来提升其项目级别。仅增加能力并不能提升项目级别。

请将此矩阵用作自我评估。在每个阶段标记你的层级，然后从最低的那个层级开始进行下一步。最后，提出那两行未涵盖的问题：各阶段由谁负责回应，以及这些阶段是否作为一个代理响应循环运行。

五个迈向能动性回应的步骤

成熟度矩阵指明了差距。 从你的项目最薄弱的环节开始，然后将这五个步骤连接成一个连续的响应循环。

1

减少个人暴露

首先对攻击者进行人员画像：高管、高风险岗位员工以及面向客户的岗位。 从经纪网站中移除暴露的个人身份信息（PII），留意其再次发布，并在活动启动前跨渠道监控冒充行为。

2

在发布前捕获营销活动基础设施

在相似域名、虚假社交账号和内容库上运行预置检测。 将每个数字资产归因于其操作者，以便下一轮活动能更早地被发现。

3

覆盖所有渠道

覆盖电子邮件、社交网络、移动端、加密消息、应用商店和市场。 以机器速度运行AI欺骗检测。按渠道追踪下架服务水平协议（SLA），以防止盲点变成时间空白。

4

A区阻止ATO并验证工作流程

持续监控凭证和暗网来源。 在造成影响之前检测账户接管行为。对高价值工作流要求验证：包括线路变更、支付路由变更和客户账户更新。

5

超越下架，持续追求

将下架处理视为调查中的一个步骤，而非终点。 追查运营者的更广泛网络。协调安全、欺诈、法务和客户服务团队之间的响应。将每一起事件输入到下一个检测周期中。

这五步涵盖了攻击的每个阶段。 合在一起，它们形成一个连续的循环。检测为调查提供支持，调查为溯源提供支持，溯源为清除提供支持，清除为验证提供支持，验证为下一轮检测提供支持。每一轮循环都比上一轮更快。

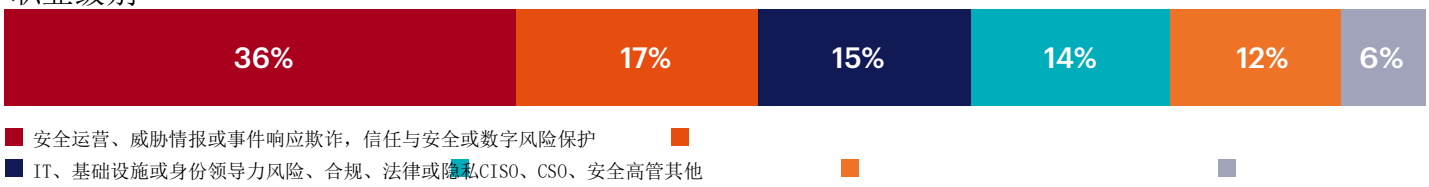
AI代理以机器速度运行该循环。 人类决定该循环被允许做什么以及何时介入。曲线前方的程序正在构建这个循环。未来12到24个月将区分他们与那些仍在以人类速度试图防御的人。

方法论与人口统计学

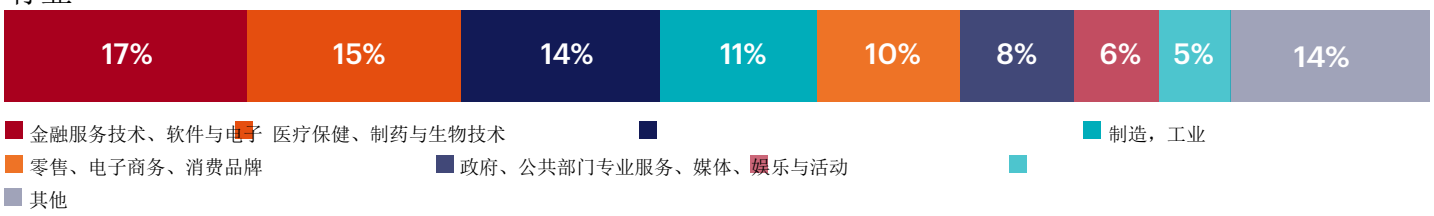
本报告基于一项针对1138名决策者的调查，这些决策者负责其组织内的网络安全、欺诈、数字风险和信任相关项目。 该调查于2026年初在众多不同行业和规模的机构中开展。此项研究考察了数字风险成熟度的五个维度：检测能力、响应基础设施、高管与员工保护、AI代理治理以及整体项目成熟度。

采用分层抽样方法，该调查达到了95%的置信水平，误差范围为±2. 90%。

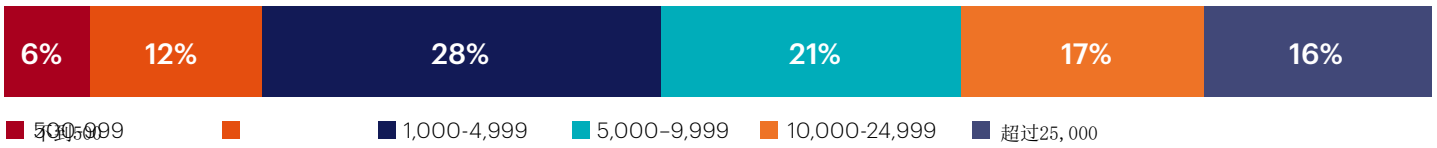
职业级别



行业



公司规模



Cybersecurity
INSIDERS

校准您的安全成熟度

独立网络安全研究揭示塑造网络安全战略的漏洞

网络安全内幕人士基于对全球网络安全领导者和从业者的调查，制作独立研究报告。 我们的报告揭示了安全策略在实践中哪些环节会失效——帮助组织评估其成熟度、识别能力差距，并优先处理需要采取的行动以弥补这些差距。

欲了解更多信息，请访问cybersecurity-insiders.com

June 2026 // Version 1.5

Outtake 致力于清除网络威胁，恢复数字信任。作为面向 AI 的数字风险防护平台，Outtake 提供覆盖整个攻击面的统一检测、调查和响应能力——保护品牌、高管、产品和地点免受仿冒、AI 生成欺骗以及 AI 代理安全风险的影响。在一个协调一致、规模化攻击的速度快于人类响应的时代，Outtake 为组织赋予主动能力，使其能够领先于威胁，而不仅仅是被动应对。

www.outtake.ai