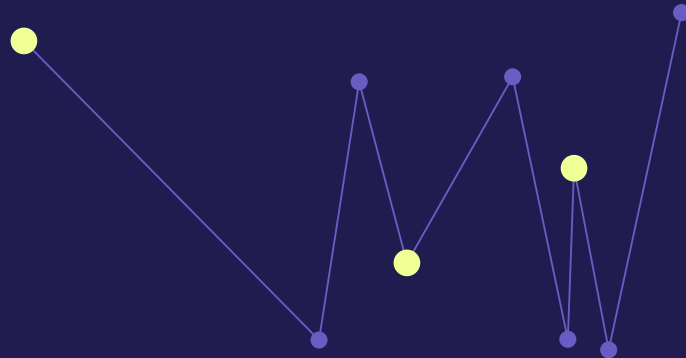


When AI Answers Are Hijacked

Why brand impersonation is becoming an answer-layer security and reputation problem



ARTEFACT91

Campaign Intelligence + Strategic Communications

artefact91.com

In this briefing

Why brand impersonation is becoming an answer-layer security and reputation problem

·	Executive Brief
01	Why This Clears the Threshold for a New Briefing
02	What the Dark Sourcery Findings Show
03	The Artefact91 Answer-Layer Hijack Chain
04	Why Legacy Brand Monitoring Can Miss the Attack
05	The Communications–Security Handoff Must Change
06	The Artefact91 Canonical Contact Integrity Standard
07	When Does It Become a Crisis?
08	Response Language: Separate the Brand from the Fraud
09	The Broader Fraud Context
10	A 30-Day Readiness Sprint
11	Metrics That Matter
·	Artefact91 Point of View
·	Sources & Research Base

HOW TO READ IT

A briefing, not a report: the executive brief carries the argument, the numbered sections carry the evidence and the method, and every external figure is attributed to its source.

Executive Brief

A new communications risk has moved from theory into documented abuse: attackers are deliberately poisoning the information sources used by AI search and chatbot systems so that fraudulent phone numbers, email addresses, login pages and support instructions can appear inside apparently helpful answers.

On 22 September 2026, security researcher Ariel Simon published findings from a monitoring system that tested responses across ChatGPT, Gemini and Google AI Overview. The research team reported detecting attacks involving 374 companies and tens of thousands of malicious pages. The targets included major airlines, banks, travel platforms and software providers. The researchers also cautioned that results were probabilistic: a poisoned answer did not reproduce on every query, and the full scale of victim losses is not known.

Follow-up reporting by Dark Reading and IANS describes the same campaign, dubbed “Dark Sourcery,” as an adaptation of familiar SEO poisoning techniques to AI-mediated answers. The important difference is structural. A user may receive the fraudulent contact detail inside the answer itself without ever visiting the malicious page that supplied it.

For communications leaders, this is not simply a cybersecurity problem and it is not another version of routine misinformation monitoring. It is a new form of brand impersonation in which an organization can be misrepresented at the point of assistance even when its own systems, website and social accounts have not been breached.

The strategic shift

The attacker does not need to take over the brand. The attacker needs to take over the answer.

SECTION 01

Why This Clears the Threshold for a New Briefing

Briefings 03 and 11 examined the rise of AI-mediated discovery and the need for institutions to manage the answer environment. Those briefings focused primarily on accuracy, retrievability, compression, source selection and the possibility that systems could surface outdated or incomplete information.

This development is different. The information environment is now being manipulated adversarially. The issue is no longer only whether AI systems misunderstand an institution. It is whether a hostile actor can deliberately place false institutional information into the source ecosystem and persuade an AI intermediary to present it as useful guidance.

THE SHIFT

ORGANIC INACCURACY → ADVERSARIAL ANSWER MANIPULATION · REPUTATION
MONITORING → REPUTATION + FRAUD + SECURITY OPERATIONS

The distinction matters because the organizational response cannot sit with communications alone.

SECTION 02

What the Dark Sorcery Findings Show

The reported attack pattern is straightforward.

SEED THE SURFACES — Attackers publish large volumes of false brand information across web surfaces that allow user-generated content, file uploads or easy publishing: social and forum posts, video descriptions, reviews, website-hosting platforms, PDFs uploaded to government and university domains, fundraising or event platforms and other public sites.

MAKE THE PAYLOAD OPERATIONAL — The payload is usually operational rather than rhetorical: a support number, login page, email address, software update or other action-oriented instruction.

OPTIMIZE FOR RETRIEVAL — Researchers observed repeated phone numbers, question-and-answer formatting, urgency language, recency signals such as “updated 2026,” multiple representations of the same number and cross-platform repetition.

LET SYNTHESIS DO THE REST — If an AI system retrieves the poisoned material and combines it with legitimate sources, the malicious detail can become part of a coherent answer.

The asymmetry

The fraudster controls the harmful interaction. The legitimate organization inherits the stakeholder anger.

SECTION 03

The Artefact91 Answer-Layer Hijack Chain

ARTEFACT91 ANSWER-LAYER HIJACK CHAIN

SEED → ELEVATE → SYNTHESIZE → ACT → ATTRIBUTE

SEED — Fraudulent brand information is planted across open web surfaces: posts, PDFs, reviews, hosted pages, videos, directories and compromised or permissive domains.

ELEVATE — The attacker increases the probability of retrieval through repetition, formatting, apparent corroboration, source authority and optimization techniques designed for search and generative systems.

SYNTHESIZE — An AI system retrieves the material and may combine false and legitimate information into a single answer that appears more authoritative than the underlying source deserves.

ACT — The stakeholder follows the instruction: calling a number, clicking a login link, downloading software, sending payment information or taking another consequential action.

ATTRIBUTE — When harm occurs, the stakeholder may blame the legitimate organization whose identity was impersonated — even though the organization never published the false information and may never have been technically breached.

The communications implication is important: reputation damage can now be produced through an external information supply chain.

SECTION 04

Why Legacy Brand Monitoring Can Miss the Attack

Traditional media monitoring looks for mentions. Social listening tracks conversations. Fraud teams may monitor phishing domains. Cybersecurity teams monitor infrastructure, credentials and attack indicators. Answer-layer attacks can fall between all four.

A malicious page may receive almost no human traffic while still influencing an AI-generated answer. A false number may exist across hundreds of low-value pages without becoming a conventional media story. A company's website may remain accurate while the chatbot answer seen by a stakeholder is not.

The monitoring question changes

From "Who is talking about us?" to "What action would a stakeholder be instructed to take when an intermediary answers on our behalf?" That requires testing high-consequence queries, not merely tracking mentions.

Priority query families should include:

- Official customer-service number.
- Login or account-recovery page.
- Refund or cancellation process.
- Payment or banking instructions.
- Download or software-update link.
- Executive or media contact.
- Donation or fundraising instructions.
- Emergency or crisis contact.
- Procurement or vendor contact.
- "Is this message, number or website really from us?"

SECTION 05

The Communications–Security Handoff Must Change

The Dark Sourcing research reports that some organizations treated the issue as outside security scope because no internal system had been compromised. Google and OpenAI also reportedly classified the researchers' disclosures as outside or insufficient for their vulnerability programs. That illustrates the governance gap.

An answer-layer impersonation event may be none of the familiar categories: not a conventional breach, not a traditional media crisis, not merely inaccurate content, not only a customer-service issue, and not safely handled as a routine takedown request.

Artefact91 recommends a joint operating model with named ownership across communications, cybersecurity, fraud, legal and privacy, customer service and digital or product teams.

The escalation trigger

Consequence, not technical intrusion. If a false answer can plausibly redirect money, credentials, personal information, media inquiries or stakeholder decisions, it should enter the incident process even when the organization's own network is untouched.

SECTION 06

The Artefact91 Canonical Contact Integrity Standard

Institutions should make legitimate operational information easier for both people and machines to authenticate.

ARTEFACT91 CANONICAL CONTACT INTEGRITY STANDARD

OWN → PUBLISH → VERIFY → TEST → TRACE → ESCALATE → CORRECT

OWN — Maintain a single authoritative destination for high-risk contact and transaction information.

PUBLISH — State official phone numbers, domains, account-recovery paths, payment rules and “we will never ask you to...” guidance clearly and consistently.

VERIFY — Ensure major owned profiles, directories, knowledge surfaces and trusted third-party listings point back to the same canonical information.

TEST — Run recurring AI-answer checks across major assistants and search summaries for high-consequence queries. Record screenshots, citations and timestamps, because outputs may vary.

TRACE — When a false answer appears, identify the cited or likely source chain rather than treating the generated answer as an isolated hallucination.

ESCALATE — Route confirmed impersonation into fraud, security, legal and communications workflows according to consequence.

CORRECT — Publish authoritative evidence, pursue source and platform correction, and verify whether the answer environment changes over time.

This is not optimization for reputation. It is an integrity control for institutional identity.

SECTION 07

When Does It Become a Crisis?

Not every wrong AI answer is a crisis. Artefact91 recommends assessing six factors.

CONSEQUENCE — Can a user lose money, credentials, access, privacy or safety?

AUTHORITY — Is the false information presented as official or verified?

REACH — Is the issue appearing across multiple systems, query forms or regions?

PERSISTENCE — Does the false information remain retrievable after correction or takedown?

ATTRIBUTION — Are users or media holding the legitimate institution responsible?

DEPENDENCY — Does the affected interaction involve a high-stress moment where users are likely to act quickly — travel disruption, account lockout, payment problems or urgent support?

The more factors present, the less appropriate it is to treat the event as a simple content correction.

SECTION 08

Response Language: Separate the Brand from the Fraud

A weak response says: “That number is not ours. We were not breached.” That may be technically accurate and reputationally inadequate.

A stronger response establishes four facts.

AUTHENTICATION — Which channels are official.

PROTECTION — What stakeholders should not disclose or do.

ACTION — What the organization is doing to remove or correct the fraudulent information.

REMEDY — Where affected people should go for help.

The institution should avoid implying that a victim was careless. If an apparently authoritative AI system supplied the false information, the organization should acknowledge the plausibility of the confusion while clearly distinguishing the fraudulent actor from the legitimate brand.

SECTION 09

The Broader Fraud Context

The AI-specific prevalence of answer-layer fraud is not yet established, and the 374-company figure comes from one research team’s detection system rather than an independent market-wide audit. That limitation should remain explicit.

What is established is that impersonation fraud is already economically significant. The U.S. Federal Trade Commission reported more than one million imposter-scam reports for 2025 and \$3.5 billion in reported losses, an increase of nearly 20% from the prior year. Those figures cover imposter scams broadly, not AI answer poisoning specifically.

The FTC has also warned in September 2026 that scammers are using AI to help clone legitimate car-dealership websites, including logos, vehicle listings and testimonials.

What the evidence suggests

AI is lowering the cost of producing credible impersonation, while AI-mediated discovery can create new distribution paths for that impersonation. The problem is not “AI misinformation” in the abstract. It is institutional authentication under conditions of synthetic scale.

SECTION 10

A 30-Day Readiness Sprint

WEEK 1 — MAP THE HIGH-CONSEQUENCE FACTS — Identify every phone number, support URL, payment instruction, account-recovery process, executive contact and other operational fact that could cause harm if falsified.

WEEK 2 — BUILD THE CANONICAL LAYER — Consolidate those facts into authoritative, crawlable, current pages with clear ownership, update dates and explicit fraud-warning language.

WEEK 3 — TEST THE ANSWER ENVIRONMENT — Run standardized queries across major AI assistants and search-summary environments. Log answers, citations, false information, missing caveats and platform differences.

WEEK 4 — SIMULATE THE INCIDENT — Run a tabletop exercise in which a customer receives a fraudulent support number from an AI answer. Test who verifies it, who contacts the platform, who issues public guidance, who handles victims, who preserves evidence and who decides when the event becomes a reportable or public incident.

SECTION 11

Metrics That Matter

Move beyond share of voice. Track:

- Canonical-contact accuracy across AI systems.
- Percentage of high-risk queries returning correct first-party information.
- Number of impersonation sources detected.
- Time from detection to cross-functional escalation.
- Time from detection to authoritative public correction.
- Recurrence after takedown.
- Customer-service contacts linked to false AI answers.
- Financial or account impacts where confirmed.
- Media or social attribution of the fraud to the legitimate organization.
- Platform response time.

The objective

Not perfect control over AI outputs — that is unrealistic. A measurable reduction in the time between false representation and institutional correction.

ARTEFACT91 BRIEFING 18

Artefact91 Point of View

The next reputation crisis may begin without a journalist, a viral post or a compromised company account. It may begin when a stakeholder asks a machine for help and receives a confident instruction that appears to come from the institution.

That means the public identity of an organization now has an operational attack surface.

Communications teams should not become cybersecurity teams. But they do need to recognize when reputation infrastructure has become security infrastructure — and build the handoff before the incident.

The institution that can prove what is official, detect when that identity is being imitated, and respond across communications, fraud and security functions will be materially better positioned than the institution that discovers the problem only after customers start calling.

ARTEFACT91 BRIEFING 18

Sources & Research Base

- Ariel Simon — “Dark Sourcery: How Hackers Manipulate AI to Scam You,” 22 September 2026.
- Elizabeth Montalbano — “Attackers Manipulate AI Chatbots in Mass Disinformation, Phishing Campaign,” Dark Reading, 23 September 2026.
- IANS — “Dark Sourcery Campaign Manipulates AI Chatbot Answers to Scam Users,” 24 September 2026.
- U.S. Federal Trade Commission — “New trends in reports of imposter scams,” May 2026.
- U.S. Federal Trade Commission — “Scammers are spoofing car dealership websites: What you need to know,” September 2026.

Research Note

The Dark Sourcery findings are researcher-reported and should not be interpreted as a complete measure of AI-answer poisoning across the market. The research states that poisoned outputs were probabilistic and did not reproduce on every query. Public reporting does not establish total financial losses attributable to the campaign. This briefing therefore treats the findings as evidence of a documented attack class and emerging operating risk — not as evidence that every major AI system is routinely compromised or that every named company experienced customer loss. Artefact91 frameworks are analytical interpretations intended to support decision-making and should be adapted to the facts, jurisdiction and risk profile of each engagement.