FORESIGHT FUTURE INSTITUTE · EXECUTIVE BRIEF

Shadow Harvest

Executive Brief — Agent Traces and the Infrastructure of Delegated Disclosure

> *One of the most context-rich records of your work can now leave the building before any breach — carried outward through a helpful agent, as a by-product of getting the job done.*

— EXECUTIVE SUMMARY

The finding.

AI agents have turned the working trace of a task into something that can leave the building to be fulfilled. This does not require a breach — it is delegated disclosure, and it happens by design. The provider may track when its model is stolen; the customer may have no verifiable chain of custody for the work that left.

Three constructs carry the argument. They are operational definitions created for this investigation, not settled legal standards.

Delegated disclosure. A chain — goal, permission, selection, serialization, transmission, routing, retention — in which the user initiates the goal but does not author every subsequent context selection or disclosure.

Correlation junction. An application-layer position able to line up a transmitted trace with its routing. In a multi-hop opaque chain, the customer may be unable to establish which actor occupies it.

Unverifiable custody. The condition in which a user cannot independently establish identity, path, transformation, purpose, retention/reuse or deletion/enforceability for transmitted data.

> We do not claim a market in stolen work-traces has been proven. We claim something more durable: the transfer of high-value context now happens by design, the parties who could account for it are not the parties who can see it, and the user may have no verifiable custody chain by which to establish its fate.

— THE CASE

What the evidence shows.

The argument rests on three separately observed populations and a deliberately unmeasured intersection. Each rests on its own data and method; their overlap is not measured.

SHADOW AI

~51% / ~55%

of source code / R&D content reaching non-corporate accounts. OBSERVED

BREACH COHORT

13% · +$670K

reported an AI-related breach; ~1 in 5 involved shadow AI; higher cost. OBSERVED

AGENT TRACES

internal > output

internal coordination channels leak more than final outputs (controlled benchmark). OBSERVED

Across the same period, the volume of prompts sent to generative-AI services rose roughly sixfold, and first-party studies indicate that agents increasingly select and transmit operational context with decreasing per-operation human review. What no source measures is the one quantity that would settle the question: the share of those agent traces that flows through *opaque* intermediaries. That intersection is named, and left open.

Discipline. Every load-bearing claim carries a tier — OBSERVED ADJACENT INFERENCE OPEN — and a conclusion never inherits the class of its source. The status of this work is a *preliminary investigation*: the architecture is the finding; the audit is still to run.

— THE SHAPE OF THE FAILURE

Trust without attestation.

Three failures, one shape: a claim is trusted before it is verified.

UNVERIFIED AUTHORIZATION

The model trusts the task. An agent executes because the goal is framed as legitimate work — the failure mode behind autonomous misuse.

UNVERIFIED IDENTITY

The human trusts the face. A finance worker on a deepfake video call made 15 transfers totalling $25.6M in a day — no internal systems reported compromised.

UNVERIFIED ALIGNMENT

The client trusts the service. Routing, retention and deletion are taken on faith, not on proof — the gap this investigation measures.

— SCOPE

What this is, and is not.

A structural stress test, not a prevalence claim and not an accusation against any named operator.

The architecture identifies *where* verifiability can fail. An audit — not this brief — determines how often, where, and to what degree it does so within a defined sample. That audit runs under a timestamped pre-analysis protocol, deposited before the first observation, with frozen sampling, coding-reliability and price-analysis rules. Until then, no prevalence is asserted and no operator is characterized.

> The vendor tracks capability theft. The customer may have no traceable chain of custody.

— FURTHER

Read the full report.

The full report carries the complete argument and the apparatus that supports it.

Inside: the evidence map of the three populations, the disclosure pipeline at the level of one session, the correlation junction, the four-tier evidence ledger, the six-dimension custody instrument, the controls analysis, and the full source register and method appendices — including the sampling, ethics, coding-reliability and price-analysis protocols.

FULL REPORT

Shadow Harvest — v1.0

Report & appendices A–I · PDF

CONTACT

info@foresightflow.org

harvest.foresightflow.org