

# Securing Enterprise Cognition

## A CISO's White Paper for the Generative-AI Era

Matthew Schwartz

### Contents

Abstract . . . . .	1
1 GenAI as a Force Multiplier . . . . .	1
2 The Contemporary Threat Landscape, 2023 – 2025 . . . . .	2
3 Regulation and Governance Pressures . . . . .	2
4 Technical Controls for Cognitive Security . . . . .	3
4.1 Cognitive-Security Maturity Model . . . . .	4
5 Operational Road-Map & Metrics . . . . .	4
6 Organizational and Ethical Considerations . . . . .	4
7 Forward Look 2026–2028 . . . . .	4
8 Conclusion . . . . .	5
Appendix A Reference Architecture (text description) . . . . .	5
Appendix B Supply-Chain Governance Toolkit . . . . .	5
References . . . . .	5

### Abstract

Generative artificial intelligence (GenAI) has eliminated the last practical barriers to industrial scale deception. A lone actor can clone an executive's voice, craft photoreal evidence, or mass-produce micro-targeted narratives in minutes-eroding judgement faster than any firewall can respond.

This white paper reframes *cognitive security* for the GenAI era, maps the 2023–2025 threat surface, and delivers an actionable playbook: a four-level maturity model, 36-month operational roadmap, and technical controls that outpace regulatory baselines. It is written for Chief Information Security Officers who must preserve authenticity, integrity, and brand trust in organisations adopting GenAI for competitive advantage rather than expanded risk appetite.

---

### 1 GenAI as a Force Multiplier

Before large language and diffusion models were publicly accessible, disinformation faced three friction points: expert labour, production latency, and crude audience targeting. GenAI eliminates each. A lone actor can prompt a foundation model to draft articles in multiple languages, generate supporting imagery, and clone an executive's voice in minutes. Open-source tool-chains such as Stable Diffusion and Eleven-Labs compress production time from hours to seconds. Fine-tuning on demographic micro-segments lets campaigns A/B-test narratives in real time, iterating messages until engagement peaks. The result is exponential, not incremental, growth in adversarial capacity. (newsguardtech.com)

---

## 2 The Contemporary Threat Landscape, 2023 – 2025

Table 1 Representative Incidents & Indicative Losses

Vector	Incident	Impact	Source
Synthetic Market Shock	AI-generated photo of a Pentagon explosion triggered algorithmic sell-off.	≈ \$500 million erased in minutes	reuters.com
Deep-Fake Election Interference	Voice-cloned tape alleged ballot rigging two days before Slovakia’s 2023 vote.	Poll swing ≥ 5 wired.com; misinfo-view.hks.harvard.edu	
Voice-Clone Wire Fraud	Deep-fake Microsoft Teams call convinced staff at UK engineering firm Arup to wire funds.	\$25 million	weforum.org
News-Farm Flooding	1 271 AI-generated “news” sites blend genuine wire copy with fabricated stories.	Brand-trust erosion; OSINT contamination	newsguardtech.com
Shadow-AI Data Leakage	Samsung engineers pasted confidential code into ChatGPT.	Undisclosed IP exposure; LLM ban	techcrunch.com

Each incident compresses the defender’s Observe–Orient–Decide–Act loop, demonstrating that perception — not perimeter — has become the primary attack surface.

## 3 Regulation and Governance Pressures

Regulators set the **deadlines**; security teams must still set the **controls**.  
The obligations below form the hard edges of today’s cognitive-security program.

Instrument	Core obligation	Clock-speed
<b>SEC Cyber-Incident Rule</b> (Item 1.05, Form 8-K)	Report any material incident—including deep-fake-enabled fraud—to investors within <b>4 business days</b> .	In effect since 26 Jul 2023
<b>EU AI Act</b>	Map systems to risk tiers, log GenAI outputs, run post-market monitoring for “high-risk” services.	Full compliance by 2 Aug 2026 (phased adoption began Aug 2024)
<b>ISO/IEC 42001</b>	Operate an <b>auditable AI-Management System</b> : policy, roles, risk register, incident playbooks.	Standard published Dec 2023; certification cycle ≈ 3 years
<b>C2PA Content Credentials</b>	Attach a <b>cryptographically signed provenance manifest</b> (“Content Credential”) to every outbound image, video, audio or PDF; quarantine unsigned media at ingress.	v1.4 spec; ISO ballot underway, broad adoption expected <b>2025</b>
<b>MITRE ATLAS AI-Incident Sharing</b>	Exchange TTPs and red-team telemetry across peer firms under NDA to accelerate counter-measures.	Live since Oct 2024; monthly intel drops

### What is C2PA?

The **Coalition for Content Provenance and Authenticity** defines an open, royalty-free

standard that bundles a JSON-LD manifest (capture device, edit history, signer ID) with an X.509/COSE signature. Any downstream tool—from e-mail gateways to SIEM rules—can verify that hash and reject tampered or unsigned content, giving enterprises a “TLS padlock” for images, video and audio.

Regulations establish the **minimum viable authenticity**. The next section turns these static mandates into dynamic technical controls.

## 4 Technical Controls for Cognitive Security

1. **Authenticity by Default.** Require cryptographically signed provenance (C2PA or equivalent) at every ingress point — e-mail, social feeds, threat-intel APIs. Unsigned media is quarantined, just as browsers distrust invalid TLS certificates.
2. **Multimodal Forensics in the SOC.** Deploy GPU-backed detectors that score temporal artifacts in video, spectral anomalies in audio, and compression fingerprints in images; pipe results to your SIEM so deep-fake alerts sit beside malware alerts.
3. **Secure Plug-in Supply Chains.** Maintain a signed registry for LLM tools. Gate admission on static analysis and pentest results; enforce runtime policy that allows tool calls only to whitelisted domains; log every invocation with arguments and return payloads.
4. **Context-Window Guardrails.** Insert an input-sanitation layer that strips homographs, zero-width characters, and known jailbreak tokens before user content reaches the model; add an output filter that diffs responses against protected concepts (e.g., unreleased earnings).
5. **Cross-Channel Verification Routines.** Automate out-of-band checks — text or secure-chat confirmations — for high-value approvals initiated by voice or video to defeat real-time deep-fake latency advantages.
6. **GenAI-Aware Red-Teaming.** Quarterly exercises should include adversarial prompt injection, data poisoning, and synthetic-persona campaigns against internal collaboration tools, with telemetry feeding rule-tuning and executive tabletop drills.
7. **Full-Stack Model Observability.** Log token-level anomalies — sudden sentiment spikes, profanity bursts — and correlate them with user IDs and plug-in calls for early-warning detection of covert influence.
8. **Zero-Trust for Information.** Extend zero-trust principles to content: no data object gains employee attention without proving identity, integrity, and least-privilege necessity.

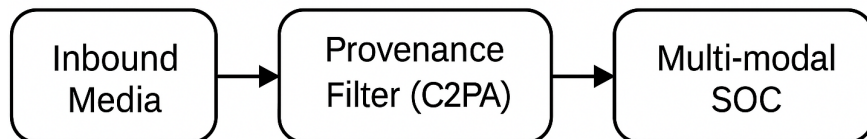


Figure 1: Content-provenance pipeline feeding multi-modal-SOC workflows

## 4.1 Cognitive-Security Maturity Model

Level	Characteristics	Typical Evidence
<b>Ad-Hoc</b>	Isolated GenAI pilots, no provenance checks.	Spreadsheet inventories; informal Slack guidance.
<b>Managed</b>	Provenance filter & basic deep-fake detection in SOC; supply-chain vetting policy drafted.	SIEM rules firing; signed plug-in repository.
<b>Measured</b>	KPIs reported quarterly; red-team drills include cognitive scenarios.	Dashboard of lead/lag metrics; tabletop-drill after-action reports.
<b>Optimised</b>	Cognitive-SOC fusion of brand sentiment, ATLAS feeds, and behavioural analytics; continuous control-tuning via ML.	MTTR < 30 min for synthetic-persona incidents; ISO/IEC 42001 certified.

## 5 Operational Road-Map & Metrics

Horizon	Key Actions	Lead Indicators	Lag Indicators
<b>0–6 Months</b>	Inventory all GenAI deployments; embed provenance checks in press releases and investor decks; add “synthetic-media fraud” injects to IR run-books.	% outbound media with C2PA; number of synthetic-media drills executed.	—
<b>6–18 Months</b>	Publish ISO 42001 or NIST AI-RMF profile; integrate credential validation at e-mail gateways and social-listening APIs; stand up signed plug-in repository.	Mean time to flag deep-fake asset; % vetted plug-ins.	Fraud-loss ratio YoY.
<b>18–36 Months</b>	Operate cognitive-SOC fusing behaviour analytics, brand sentiment, ATLAS feeds; negotiate provenance clauses in supplier contracts; report AI-risk metrics alongside cyber metrics.	MTTR for narrative incidents; % supplier contracts with provenance SLA.	Brand-sentiment delta post-incident.

## 6 Organizational and Ethical Considerations

Cognitive security dissolves silos: security operations, fraud prevention, corporate communications, and investor relations must share dashboards and escalation paths. Defensive use of synthetic media for red-teaming must respect privacy statutes and emerging “synthetic-persona” disclosure laws. Establish an internal review board to apply the same rigor to narrative tools that privacy offices apply to customer data.

## 7 Forward Look 2026–2028

- **Standards on the Horizon.** ISO provenance ballot (C2PA → ISO/IEC 62366), IETF Message Layer Security for real-time voice authentication.

- **Adversary Adaptations.** Watermark stripping, multimodal jailbreak chains, adversarial audio perturbations that evade liveness tests.
  - **Architectural Shifts.** Embedded provenance chips in mobile devices; SIEM vendors bundling cognitive-threat intelligence feeds; insurers requiring GenAI-risk attestations akin to SOC-2.
- 

## 8 Conclusion

Enterprises have hardened networks, devices, and clouds; now they must harden reality. GenAI industrialises deception, but it also equips defenders with provenance chains, multimodal forensics, and language-aware analytics unavailable five years ago. Compliance frameworks set the floor; genuine security arises when authenticity becomes a default property and narrative risk is scored in real time. CISOs who adopt these controls will keep revenue, reputation, and regulatory standing intact — even when the first breach targets belief itself.

---

## Appendix A Reference Architecture (text description)

**Figure 1.** Inbound Media → **Provenance Filter** (C2PA validation) → **Multimodal Forensics Stack** (image/audio/video detectors) → **SIEM + SOAR** enrichment → **Cognitive-SOC Dashboard** → Executive Communications & IR Teams. Integration points: GPU inference servers, policy engine, data-loss-prevention bus, brand-sentiment API.

---

## Appendix B Supply-Chain Governance Toolkit

### Vendor Questionnaire (excerpt)

1. Provide model card with training-data sources and alignment methods.
2. State ISO/IEC 42001 certification status and renewal date.
3. Describe provenance-tag adoption (C2PA or equivalent) for all generated content.
4. List incident-response SLAs for AI-related security events.

### Contract Clause Template

*Supplier shall notify Customer of any AI-security incident—including prompt-injection, data-poisoning, or model-integrity compromise—within 24 hours of detection and shall provide a post-mortem report within seven business days.*

---

## References

1. SEC, “Disclosure of Cybersecurity Incidents Determined To Be Material,” Statement of Erik Gerding, 21 May 2024. (sec.gov)
2. European Commission, “AI Act — Regulatory Framework,” accessed 24 Jun 2025. (digital-strategy.ec.europa.eu)
3. ISO, “ISO/IEC 42001:2023 — AI Management Systems,” summary page. (iso.org)
4. C2PA response to NIST, noting expected ISO adoption by 2025. (downloads.regulations.gov)

5. MITRE, "AI Incident Sharing Initiative," News Release, 2 Oct 2024. (mitre.org)
6. NewsGuard, "Tracking AI-Enabled Misinformation: 1 271 Unreliable AI-Generated News Websites," updated 5 May 2025. (newsguardtech.com)
7. Reuters, "Tech experts see rising threat of GenAI deepfakes," 10 Jul 2023. (reuters.com)
8. HKS Misinformation Review, "Beyond the Deepfake Hype: The Slovak Case," 22 Aug 2024. (misinfo-review.hks.harvard.edu)
9. Wired, "Slovakia's Election Deepfakes Show AI Is a Danger to Democracy," 28 Sep 2023. (wired.com)
10. World Economic Forum, "Lessons Learned from a \$25 Million Deepfake Attack," 4 Feb 2025. (weforum.org)
11. TechCrunch, "Samsung bans use of generative AI tools after internal data leak," 2 May 2023. (techcrunch.com)