Expect the unexpected

Incident response strategies for GenAI

Jolanda Kumakaw | Manager, Strategic Command, Google Trust & Safety

Google

Topics

01 What are GenAI incidents?

02 Emerging risk areas

03 Incident response

04 Examples

05 Q&A

“ P0 incident incoming!”

Google's 'Woke' Image Generator Shows the Limitations of AI

Google has hit pause on Gemini's ability to generate images of people after a far-right backlash to its historical depictions.

Wired, Feb 22, 2024

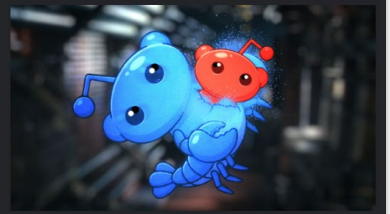
Grok 'nudify' scandal: Data Protection Commission to investigate X over its AI app

SHELL GAME

The rise of Moltbook suggests viral AI prompts may be the next big security threat

We don't need self-replicating AI models to have problems, just self-replicating prompts.

BENJ EDWARDS - FEB 3, 2026 4:00 AM | 116



Credit: Aurich Lawson | Moltbook

The New York Times Account

A.I. Agents: They're Fun. They're Useful. But Don't Give Them the Credit Card.

New A.I. bots can do more than just chat. They can edit files, send emails, book trips and cause trouble.

AI chatbots can sway voters with remarkable ease — is it time to worry?

Reuters

AI deepfakes blur reality in 2026 US midterm campaigns

VB Orchestration Infrastructure Data Security Events

Three AI coding agents leaked secrets through a single prompt injection. One vendor's system card predicted it

Louis Columbus

7:51 am, PT, April 21, 2026

Anthropic's Mythos sends US banks rushing to plug cyber holes

Reuters, May 12, 2026

Google

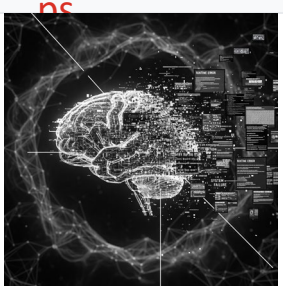
What are "GenAI incidents"?

Cybersecurity incidents

Focus on *confidentiality, integrity, and availability* of the system. Can affect any product.

EXAMPLES:

- Unauthorized access
- Malware
- DDoS attacks



AI incidents

Focus on the *output and logic* of the model. Occur *on* or *via* AI products/models.

EXAMPLES:

- **On AI:** Hallucinations, safety risk outputs
- **Via AI:** Generated misinformation (e.g., NanoBanana)

AI-Connected Cyber incidents

Cybersecurity incident that has an AI connection where AI acts as a **multiplier**, scaling traditional threats and increasing safety risks.

EXAMPLES:

- AI-enabled automated phishing
- Adversarial prompt injection
- Data exfiltration via email tools

Emerging risk areas

Better GenMedia

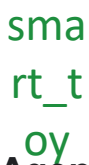
GenMedia producing better synthetic media.

- How does this impact upcoming elections?
- Nudification of people without their consent (NCII)

AI Personalization

Users are able to have a personalized experience based on their data and interactions with AI.

- Users may think their personal info is compromised.
- Where is the “creepy line” for users?



Agentic Era

Entering the agentic era - autonomous tasks resulting in less human in the loop (HITL).

- Agent makes incorrect inference and executes the wrong action.
- Prompt injections bypass security or results in user harm

When parachuting into chaos during a GenAI incident, remember to:

Connect

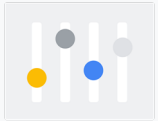
- **Comprehend** the issue and risks.
- Identify **cross product** impact and trends.
- Identify **gaps** and **connect the dots**.
- Be the **connective tissue** between distinct response efforts.

Coordinate

- **Organize** tactical response across distinct teams.
- **Assemble** response teams.
- Lead response **strategy**.
- Be the **single-threaded** owner.

Communicate

- **Central hub** of all critical incident information.
- Owner of primary incident communication channels.
- **Owner** of **source of truth** incident real-time doc.
- Notify Execs and key stakeholders.



Control

- **Steer** the response effort. Jump in.
- Set **priorities** based on **contextualized** understanding of risks.
- **Ensure alignment** across teams.
- Set up necessary **adhoc workflow/streams**
- **Accelerate** where needed

Examples

Nudification using GenAI

Personalized user experience in GenAI products

Q&A

Get in contact via Linkedin: /jolanda-kumakaw