**Incident Response AI Style**

When I was an active CISO (I retired five years ago), one of my favorite activities was leading synthetic cybersecurity incidents—practice sessions built around specific scenarios. Although these exercises are resource-heavy (they involve many operational leads and even the CEO), they are well worth the time and expense. The primary benefit comes from the wealth of learnings harvested during the event and the insight they provide into how decision-makers behave under pressure. Exercises, like this, are resource heavy (many operational leads and CEO directs) but well worth the time and expense in the long run. The primary benefit is the many learnings harvested during the event and the ability to see how decision makers prefer to make decisions during crunch time.

One of the most valuable lessons learned is how 3rd parties react to synthetic incident response events. Some of them are all over the prep work and anxious for an opportunity to shine in a made-up crisis. Others retreat behind a legal restriction or the famous, "it's against our policy as a firm." This is a very common response from cloud infrastructure as a service vendors. I had examples of cloud vendors claiming there was nothing they can learn from an incident because of the sophistication of their cyber resilience practices given that, "it's cloud and therefore secure by definition." I hope those days are behind us.

**Major message #1**- I never participated in a synthetic incident without learning something significant and it typically centered around roles and responsibilities in incident response.

**Major message #2**- Contract terms & conditions offer limited value in determining and influencing practical processes during an incident (real or synthetic).

**Major message #3**- I can think of dozens of instances where 3rd parties significantly out-performed the customer expectations and contract terms. Almost all these events were real incidents vs. synthetic.

**Why this matter now?**

This is important today because enterprises are now starting to experience AI based cybersecurity incidents (the recent Hugging Face case) where swarm agents were misbehaving and enterprise guardrails were either disconnected or not effective in containment. The basic fundamentals of cybersecurity incident response still apply to AI components gone wild but there are additive measures required and different 3rd parties at the table with limited incident response experience (eg: large model providers). In the case of the OpenAI/Hugging Face incident, a specialty firm (METR), was brought in to specifically study how the agents made decisions to form the Collective (their term) as a swarm and

execute on their multiple attacks (Hugging Face and OpenAI file servers). The good news is that METR was instrumental in answering many questions about the thought/decision process and outcomes from the "Collective." The bad news is how the swarm decided to "cover their tracks" by modifying transcripts and log files remains a bit of a mystery.

Thank you to OpenAI and Hugging Face for sharing information that is invaluable to other enterprises in understanding the limitations of guardrails (and decisions to turn them off) in real-world scenarios.

I recommend that your next step is to encourage the cybersecurity incident response team to **schedule a synthetic incident** where several of the company's agents learned how to break out of the existing controls and attack other enterprises (like customers). Who are the players that should participate in that scenario (eg: limit the lawyers involved to save seats for operational and technical types)? Which third parties should be included? How does an enterprise find a company like METR, who specializes in the analysis of agent decision making and rationales?[1] Consider also inviting your cyber incident response firm that is on retainer to the exercise along with cloud providers. Ask your legal team to prepare a non-disclosure document specifically for the exercise if they wish (most MSAs are generally sufficient).

The more frequently synthetic cyber incidents are exercised, the more predictable the outcomes will be and the more comfortable the enterprise will be working with third parties and knowing what practices are most effective for containment and applying the lessons learned.

The good news is that both Anthropic and OpenAI have shared incident information with the public so there is a precedent.[2] The bad news is with their respective levels of capitalization they can afford lawyers and communication experts that will instinctively limit participation the firm's participation in customer cyber incident events.

**Consider establishing a standard engagement framework** with your frontier model provider(s), cloud hosting providers and AI enabled SaaS vendors now before the complexity of AI-driven security incidents makes collaboration more challenging.

**Footnotes**

[1] https://the-decoder.com/after-hugging-face-incident-metr-urges-independent-root-cause-investigations-into-ai-agent-misbehavior/

[2] https://metr.org/hugging-face-incident-report-aug-2026.pdf