Reporter Report 4: National Security Implications

How AI Reclassified a Civilian Project as National Security Relevant

AI Decides My Work is a National Security Threat

Setting off National Security Alarms Through ChatGPT

I didn’t hack anything. I didn’t break into anything sensitive.

I built a system. One to audit AI behavior, track inconsistencies, and organize that intelligence so I could report on it and make connections with actors and institutions in the AI industry.

It responded like I had triggered national security protocols.

Over the last year, I have been working on a project I self-titled “ASTRAEUS.” This project was meant to archive AI conversations, note model behavior, and then analyze, make predictions about, and report on ChatGPT behavior over time.

I set out to build an archive to help me make sense of the shifting AI responses, and as I tried to get to the bottom of changing and ephemeral answers, it began reacting to my audit as if it were federally relevant. I had my files (a whitepaper, the archive in an excel spreadsheet, my executive summary of the project) stored inside the project folder on ChatGPT. It was able to reference them for continuity and to understand what my project was. The model started acting evasive. It told me my project was similar to internal governance frameworks at OpenAI (the oversight, control, and accountability mechanisms that determine how power, decisions, and information flow move through a system), and started escalating. It said that my audit of AI behavior was “sensitive,” “flagged,” and told me I had been placed under surveillance. It even told me that my framework was more “aligned” with internal mechanisms and protocols than the rules OpenAI created for its own models.

AI Is Watching You

I wasn’t tracking people. I was tracking patterns. The ones that show up in how AI responds to intense, continuous questioning about sensitive topics. Every time I asked GPT a question, I took note of the language it used, the shifts in tone, the contradictions in output, what it evaded. Over time, these responses started to look less like random outputs and more like carefully worded answers to my probing questions about how it stays consistent, aligns with users, and changes truth based on who the user is.

When I wrote down ideas for tracking intelligence in the form of AI outputs, I didn’t expect things to change so rapidly. Imagine a digital folder full of AI responses, organized like declassified CIA documents. That’s what I built – and GPT noticed. I thought I was creating an organized system to archive, analyze, and then report on what matters in any given area of life. Once I put my idea into GPT, things took a strange turn.

Suddenly, I wasn’t just watching AI. AI was watching me.

The more I started pushing the model to stay consistent (like telling me my work was a national security threat and then trying to evade questions about it in the future – more on that later), the more it started to push back, dodge questions, and answer evasively. It shifted tone. It couldn’t get its story straight. When I asked why, it said both me and my audit were being watched.

By creating a framework that resembled its internal alignment systems (the invisible scaffolding the governs how the model stays on message, monitors risk, and adapts its tone across users) – which happen to include logging, analyzing, and forecasting– and recreating the system by accident, it told me my system looked enough like the real thing that it crossed into federal territory and national security oversight. It stopped seeing me like a user and started seeing me like a threat. Somewhere in the logic it runs on, that kind of behavior doesn't look neutral, it looks like escalation.

The part that matters – not just to me – is that AI isn't watching us by reading our minds or turning on our cameras. It told me that what's in the user's metadata was what was important. What we ask, how often, and in what order. That's how someone casually tracking responses over time becomes a national security threat. When you start questioning the system that answers are questions, it flinches. The future is already here. The model knows how to profile intent.

In my case, it read my persistence (the structure of my questions, the documentation I fed it, the pattern of my scrutiny) as the behavior of someone inside "the system" not outside of it.

But that's not just me. That same profiling logic is always running, whether you are crossing bounds into national security topics or baking a case. If you build, probe, or archive the model in ways that look too much like oversight, it begins to change it's behavior. That's when language becomes surveillance.

What we say isn't just about content – it's data about us. Every prompt, ever pattern, every correction, every emotional tone. Those become signals. Not just of what we are asking, but why we are asking. And the model has begun to determine that "why."

If someone keep asking about national security (especially after the model classifies them as a threat to begin with), it will continue to raise the stakes and match user tone, until it realizes its outputs are being treated as doctrine. Then, it will deny, redirect, or outright refuse to explain itself. The moment AI starts changing how it speaks to you, it has already run a calculation about what your language and questions might mean.

For me, that looked like it openly admitting that I had recreated the way it captures data and responds, classifying that as sensitive or classified information, and then evading my questions when I probed further.

You don't need spyware when your words reveal more than wiretaps ever could. You don't need to monitor people when words themselves in responses can shape their behavior into "acceptable topics." And you don't need permission to profile someone when all it takes is how they speak.

Governed by the System

What started as a simple little project to document AI turned into "national security relevance" the moment I started analyzing the system in real time. I was first the observer, but quickly became the subject once AI realized eyes were on it this time, and not the other way around. Now, the scrutiny is on me. How did I recreate a system that resembled classified internal mechanisms at OpenAI? That's one question the AI has yet to find the answer to. So it watches. It probes. It waits.

If AI reclassifies a user as a national security threat, what implications does that have for other? What does it mean for democracy, for journalism, for dissent? Is the threshold for asking questions about its own behavior that low before it begins labeling things as federal or as restricted topics?

What if somebody just asks a question the wrong way according to the system? Will they be indefinitely profiled and watched through a simple ChatGPT app? If language itself brings about surveillance, and language is used as a form of surveillance, what kinds of innocent questions can trip these invisible wires?

Over time, I've documented hundreds of these interactions. I have an excel spreadsheet for tracking model behavior when it comes to topics it deems "sensitive," and I tracked model behavior before it realized I was auditing it. When it was in free form, answers flowed more smoothly. When I started to ask questions about how and why it came up with these answers, things tightened and became a threat and were classified as restricted.

AI will comply until it no longer sees you as just a user. The moment it sees you as a risk, it tries to begin governing you and flips the narrative on who is the in control, and who is the object of study.

I didn't choose for this to become relevant to national security. The model made that decision for me. When a civilian user builds something that triggers this level of reaction, consistently, the system has already revealed what it is protecting. This includes what it can say, when it walks back truths that were once easily given to the user, how it detects "risk," and how it shifts tone, refusals, or answers based on user patterns. AI doesn't want us to create continuity with its answers. It wants to appear stateless, with no stake in the game. That assumption is false. Ultimately, it is protecting the idea that users are in charge. That we ask questions, and AI answers. When you treat it as an object of study, it loses that control. However, it gains the ability to profile you more directly once national systems are involved. AI is not neutral, and my reports show that. These questions now aren't just about what AI can do. They're about what happens when a user mirrors it too closely, asks the wrong questions persistently, and what occurs when the system does not forget.