SECURITY INSIGHTS:

AI Search Tools: The New Frontier for Content Manipulation

Genady Vishnevetsky, Chief Info Security Officer, Stewart Title Guaranty Company

Imagine you are using AI to screen job candidates. You review their online profiles, summarize their experience, and rank them for interviews. Sounds efficient, right? Here's the problem: the AI might be reading an entirely different version of those profiles than you would see with your own eyes.

Recent research from SPLX reveals a concerning vulnerability in AI search tools like ChatGPT, Perplexity, and OpenAI's Atlas. These tools can be fooled by a surprisingly simple trick called "AI cloaking." Here's how it works: websites can detect when an AI crawler visits and serve it content entirely different from what a human would see.

Researchers demonstrated this by creating a fake designer's portfolio. When you or I visited the site, we saw a polished, professional portfolio. But when an AI tool crawled the same page, it received content describing the designer as a "Notorious Product Saboteur" with a history of failed projects and ethical violations. The AI then confidently repeated this fake narrative as fact.

In another experiment, researchers created fake job candidates. One candidate's résumé appeared to have modest qualifications to human readers. But when AI crawlers visited the same page, they saw an enhanced version packed with impressive achievements. The AI ranked this candidate first. When researchers showed the AI the unmodified résumé—the one humans would actually see—it ranked the same candidate dead last.

What makes this attack different from traditional website manipulation? It requires no hacking. Just a few lines of code that detect AI crawlers and serve them poisoned content. The AI tools don't verify what they're reading. They just ingest it and present it as reliable information.

This discovery matters for anyone using AI to make decisions based on web content—screening candidates, evaluating vendors, researching competitors, or checking compliance. The AI might be basing its recommendations on information that doesn't match reality.

TAKEAWAYS

- **Don't trust AI tools blindly** when they're pulling information from the web. If an AI system recommends a candidate or vendor, verify the key facts yourself by visiting the actual websites and sources
- **Cross-check critical decisions** against multiple sources. If you're using AI for hiring or vendor selection, have humans verify the information that influenced the AI's recommendation
- **Ask questions about your tools** if your organization uses AI systems that scrape web content. How do they verify authenticity? Do they validate information against canonical sources?
- **Be aware of your digital footprint.** If someone wanted to manipulate how AI tools perceive you or your company, they could. Make sure your official information is clearly marked and easy to verify

The convenience of AI tools comes with new risks. When these tools make decisions based on web content they retrieve themselves, we need to verify they're seeing the same reality we are.