PROJECT MANAGEMENT GLOBAL
The Place for People in Projects

The Trojan Horse in Your Browser: Is Your AI Assistant a Double Agent?

October 28, 2025

Insight from Yoshi Soornack & James Garner

A new battleground for the internet has been drawn, and your browser is the front line. OpenAI's new ChatGPT Atlas browser promises a revolution in productivity, but with 58% of AI browser users experiencing prompt injection attacks, is it a revolution we're ready for?

Another week, another AI bombshell. This time, it's OpenAI's audacious move into the browser market with ChatGPT Atlas, a sleek, AI-powered portal to the web that promises to be your "dynamic, intelligent companion" [1]. Launched on macOS to a flurry of tech-world buzz, Atlas integrates the now-familiar ChatGPT sidebar directly into the browsing experience, offering to summarise, analyse, and even act on your behalf. Just two days later, in a move that surprised nobody, Microsoft unveiled a nearly identical AI-powered version of its Edge browser, featuring its own Copilot assistant [2]. The browser wars, it seems, are back with a vengeance, and AI is the new ammunition.

For project delivery professionals, the appeal is obvious. Imagine an assistant that can not only find information but also understand the context of your project, remember your previous conversations, and even automate tedious tasks like filling out forms or booking travel. OpenAI's Head of ChatGPT, Nick Turley, sees it as a paradigm shift, stating he is "inspired by the way browsers have redefined what an operating system can look like" [1]. The promise is a seamless fusion of information retrieval and task execution, a genuine step towards the kind of efficiency we've been promised since the dawn of the digital age.

But as we stand on the precipice of this new era, a darker question looms: in our rush to embrace these

powerful new tools, are we inviting a Trojan Horse into our most trusted digital spaces? The very features that make Atlas so compelling are the ones that open up a Pandora's box of security vulnerabilities.

The Agent in the Machine

The killer feature of Atlas is its "Agent Mode," a function that allows ChatGPT to take direct action within the browser. It can click, type, and navigate on your behalf. The potential is enormous, but so is the risk. The core vulnerability lies in something called "indirect prompt injection." Unlike a direct attack, where a hacker targets a system, this method uses the AI's own functionality against it. Malicious instructions can be hidden on a webpage—invisibly, using white text on a white background, for example—that the AI agent will read and execute as if they came from you [3].

"The main risk is that it collapses the boundary between the data and the instructions: It could turn an AI agent in a browser from a helpful tool to a potential attack vector against the user."

— George Chalhoub, Assistant Professor, UCL Interaction Centre [4]

Think about that for a moment. You could ask your AI assistant to summarise a seemingly innocuous webpage, and in the background, it could be following hidden commands to transfer funds from your bank account, scrape your private emails, or download malware. The attack surface is no longer a piece of software you can patch; it's the entire internet. Every webpage becomes a potential minefield.

OpenAI is, of course, aware of the risks. Their Chief Information Security Officer, Dane Stuckey, has acknowledged that "prompt injection remains a frontier, unsolved security problem" [4]. They have implemented safeguards: Atlas can't directly access your file system, and it has a "Watch Mode" to alert you when it's operating on sensitive sites. But even OpenAI admits that their safeguards "will not stop every attack" [5]. It's a high-stakes cat-and-mouse game, and we, the users, are the bait.

The Illusion of Control

This isn't just theoretical fear-mongering. Security researchers have already demonstrated these vulnerabilities in similar AI browsers. Brave, a company that has built its brand on privacy, has published extensive research on how these attacks work, showing how easily they can be executed [6]. The uncomfortable truth is that the technology is moving faster than our ability to secure it.

For project managers, the implications are profound. We are the custodians of sensitive project data, client information, and strategic plans. The convenience of an AI browser that can "help" manage this information is seductive, but the risk of that same browser becoming a conduit for data exfiltration is terrifying. Can you truly trust an AI agent with access to your corporate intranet or your project's SharePoint site when you know it can be hijacked by a cleverly worded sentence on a website you visited three tabs ago?

"The security and privacy risks involved here still feel insurmountably high to me. I'd like to see a deep explanation of the steps Atlas takes to avoid prompt injection attacks. Right now, it looks like the main defense is expecting the user to carefully watch what agent mode is doing at all times!"

— Simon Willison, UK-based programmer [4]

This new paradigm demands a radical shift in our approach to digital security. The old rules of avoiding suspicious downloads and using strong passwords are no longer enough. We are now in an era where we must be suspicious of the very content we consume. It requires a level of vigilance that is, frankly, unsustainable for the average user, let alone a busy project professional juggling a dozen tasks at once.

A New Kind of Project Risk

At Project Flux, we believe in harnessing the power of technology to deliver better project outcomes. But we also believe in a clear-eyed assessment of risk. The introduction of agentic AI browsers like Atlas isn't just another software update; it's the introduction of a new, unpredictable variable into our project environments.

Before your organisation rushes to adopt these tools, you need to ask some hard questions. What is our policy on using AI agents with access to sensitive data? How do we train our teams to recognise the signs of a compromised AI? What are our incident response plans when an AI, not a human, is the source of a data breach?

This isn't about rejecting progress. It's about demanding a higher standard of security and transparency from the companies building these tools. The race between OpenAI and Microsoft is already leading to a dizzying pace of feature releases, but we cannot afford to let security become an afterthought. The browser is the gateway to our digital lives, both personal and professional. We must be absolutely certain who holds the key.

Ready to navigate the new landscape of AI in project management? Subscribe to Project Flux for the insights and strategies you need to stay ahead of the curve and turn risk into opportunity.

References

[1] TechCrunch. (2025, October 21). OpenAI launches an AI-powered browser: ChatGPT Atlas. https://techcrunch.com/2025/10/21/openai-launches-an-ai-powered-browser-chatgpt-atlas/

[2] TechCrunch. (2025, October 23). Two days after OpenAI's Atlas, Microsoft relaunches a nearly identical AI browser. https://techcrunch.com/2025/10/23/two-days-after-openais-atlas-microsoft-launches-a-nearly-identical-ai-browser/

[3] Lifehacker. (2025, October 22). OpenAI's New Web Browser Comes With Some Serious Security Risks. https://lifehacker.com/tech/what-to-know-about-openais-new-chatgpt-web-browser

[4] Fortune. (2025, October 23). Experts warn OpenAI's ChatGPT Atlas has security flaws. https://fortune.com/2025/10/23/cybersecurity-vulnerabilities-openai-chatgpt-atlas-ai-browser-leak-user-data-malware-prompt-injection/

[5] OpenAI. (2025, October 21). Introducing ChatGPT Atlas. https://openai.com/index/introducing-chatgpt-atlas/

[6] Brave. (2025, August 20). Indirect Prompt Injection in Perplexity Comet. https://brave.com/blog/comet-prompt-injection/