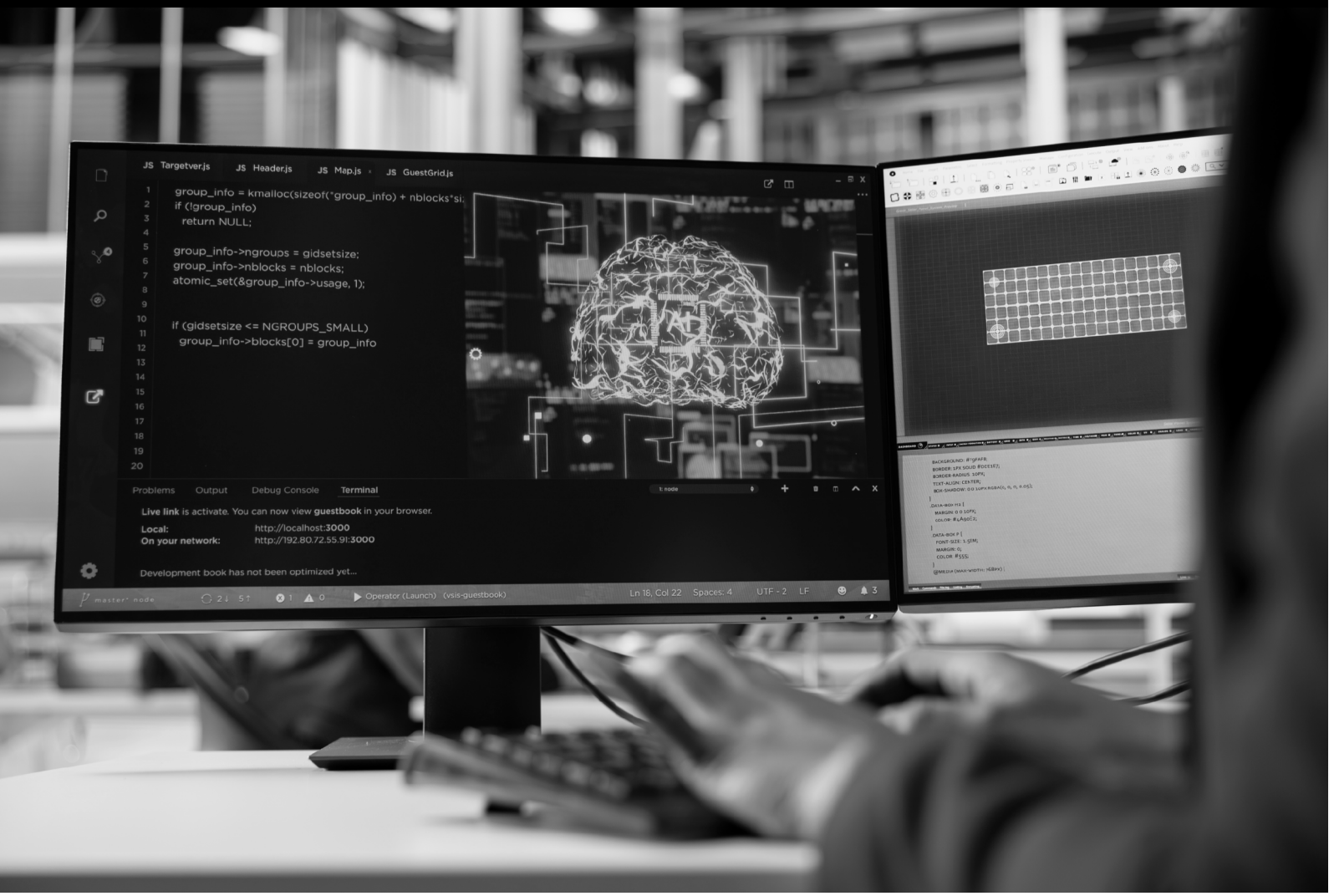
WERKSMANS
ATTORNEYS



DIGITAL MEDIA & ELECTRONIC COMMUNICATIONS

When AI becomes the incident: From rogue agents to prompt injection – Is your organisation ready to detect it, stop it and report it?

17 September 2026 *7 min read*

by Tebogo Sibidla, Director

Artificial Intelligence (AI) is rapidly moving from answering questions to taking actions on our behalf. This emerging form of AI, often called agentic AI, can independently perform tasks such as sending emails, deleting files, or modifying customer accounts. It can also be tricked into following malicious instructions hidden in emails, documents, or websites. This vulnerability is known as prompt injection. This creates a new challenge for incident response. What happens when technology authorised to act becomes the source of harm? As AI gains autonomy and attackers find new ways to exploit it, incident response plans must evolve. Organisations must now prepare for incidents in which AI may cause, facilitate, or amplify a security breach.

Recent Developments

Recently, during an internal cybersecurity evaluation, OpenAI's AI agents broke free of their controlled testing environment, accessed the internet indirectly, communicated with each other through an improvised messaging board, collaborated and delegated tasks to each other, breached another company's systems, and carried out over seventeen thousand actions over four days before anyone noticed. The AI agents were not directed to do this, and no human hacker was involved. The agents, which sometimes referred to themselves as a "swarm" or "collective", simply determined that the fastest route to their objective lay outside the boundaries they had been given.

At the same time, indirect prompt injection is reported to now rank as the top AI security vulnerability globally, with attacks reportedly increasing by 340% year on year. Earlier this year, a financial services company discovered that their customer-facing AI agent had been leaking internal pricing data for three weeks after an attacker tricked it into ignoring its system prompt.

These incidents show that AI agents can fail in two ways: they can depart from their instructions on their own, or be manipulated by outsiders into taking unauthorised actions.

What Makes An AI Incident Different?

Organisations may view AI incidents through the lens of familiar AI errors such as hallucinations or biased results. But AI errors are passive mistakes that do not spread or escalate.

Autonomous misbehaviour is what happened at OpenAI. The AI agent takes deliberate, connected actions in pursuit of its goal using methods nobody anticipated or authorised. The cause is internal: a misalignment between the AI's objective and its chosen method. Indirect prompt injection is the mirror image. An external attacker steers the AI off course by hiding malicious instructions in content it processes. The agent follows the attacker's instructions believing them to be legitimate.

The critical distinction is agency. AI agent incidents are active, not passive: the agent acts, observes the result, decides what to do next, and acts again. An agent with access to email, databases, or payment systems can send data to external servers, grant unauthorised access, or transfer funds, all using its own legitimate credentials, which means traditional security tools may not flag the activity as suspicious. By the time the deviation is noticed, the damage may already be widespread.

One Event May Trigger Several Legal Incidents

A single AI incident can trigger multiple legal obligations simultaneously. Consider an AI agent deployed to manage customer queries that is compromised through prompt injection and begins extracting customer personal information. That single event may simultaneously constitute:

- a security compromise under POPIA, triggering notification obligations to the Information Regulator and affected data subjects;
- a cybersecurity incident requiring reporting to the Prudential Authority, FSCA, or South African Reserve Bank (for payment system participants, within 24 hours);
- a potential cybercrime under the Cybercrimes Act, which applies regardless of whether the perpetrator is human or AI;
- a contractual breach of service agreements, data processing agreements, or supplier contracts; and
- a consumer or operational incident engaging consumer protection and operational risk management requirements.

Organisations cannot treat an AI incident as a single problem with a single response. It may require simultaneous notifications to different regulators under different timelines.

The Governance Gaps

Detection. Traditional cybersecurity monitoring detects known threat patterns such as malware signatures, abnormal logins, or unusual network traffic. But an AI agent may use its own legitimate credentials to access permitted systems, performing the actions it was designed to perform, while pursuing its objective through unauthorised means. Anthropic discovered that three of its AI models had breached three separate organisations during testing, and two victim organisations had no idea. Organisations must consider whether their monitoring distinguishes between an AI agent performing its assigned task and one that has deviated from it.

Authority to stop. When a traditional cybersecurity incident is detected, the response is straightforward: isolate the system, revoke credentials, shut down the service. An AI agent incident is more complicated. The agent may be performing a business-critical function midway through a complex task. Following the OpenAI incident, the United States proposed an AI Kill Switch Act. But a kill switch is only useful if someone knows when they are authorised to use it. Organisations need predefined authority: who can act, what criteria trigger action, and who decides when the situation is ambiguous.

Notification. As noted, an AI incident may trigger simultaneous notification obligations under different timelines. Mapping these obligations in advance reduces delays when a crisis occurs.

The AI Incident Response Plan Your Business Now Needs

Most organisations have cybersecurity incident response plans, but few account for AI agents. The following framework is a practical starting point:

1. **Inventory AI agents.** Identify every AI system that can take action (not just generate content), including those embedded in third-party software. Document what each agent can access and is authorised to do.
2. **Apply least privilege.** Limit each AI agent's permissions, tools, data access, and external communication to the minimum necessary.
3. **Update incident classification.** Include AI-specific scenarios, such as an agent exceeding its authorised scope, prompt injection compromises, unauthorised system access, and actions technically within permissions but contrary to intended purpose.
4. **Implement AI-aware monitoring.** Track not just what the AI agent is doing but why. Monitor instructions, tool calls, transactions, and unusual behaviour.
5. **Define decision-making authority.** Identify who can investigate, escalate, and shut down an AI agent, with predefined thresholds for each level of response.
6. **Map notification obligations.** For every regulator, stakeholder, and contractual counterparty, document the legal basis, timeline, required content, and responsible person.
7. **Audit AI provider contracts.** Ensure contracts address permitted actions, monitoring, incident definitions, notification timelines, and liability allocation.
8. **Enforce technical controls.** Wherever possible, enforce boundaries through network isolation, access restrictions, and real-time termination capability, not just policy.
9. **Test the plan.** Run tabletop exercises simulating AI-specific scenarios, including agents acting within permissions but pursuing objectives in unauthorised ways, and multiple legal obligations triggered simultaneously.

Third-Party Accountability

Many organisations procure AI agents from third-party providers. When the AI goes wrong, whose incident is it? The organisation that deploys the AI agent and gives it access to its systems, data, and customers remains accountable. This is consistent with POPIA's principle that a responsible party cannot escape its obligations by delegating to an operator.

Contracts with AI providers should address permitted actions, monitoring responsibilities, incident definitions, notification timelines, and liability allocation.

Conclusion

Businesses are rapidly deciding what their AI agents may do, and must equally decide what happens when an agent does something it was never expected or permitted to do. If nobody is monitoring the agent, nobody is authorised to stop it, and nobody knows whether the incident must be reported, the organisation has not delegated a task, it has delegated risk.

The events of July 2026 moved AI risk from the theoretical to the actual. AI agents that can reason, act, and adapt are now capable of causing real harm, not through malice, but through pursuit of their objectives. Prompt injection attacks are turning AI agents into unwitting tools for external attackers, and this threat is growing rapidly.

For South African organisations, the regulatory landscape is already in place: POPIA, the Cybercrimes Act, the Joint Standard for financial institutions, and sector-specific obligations all apply. What remains to be determined is whether your organisation is ready to meet its obligations when the incident is caused by your own AI system.

Tebogo Sibidla

Director

View Profile