ADVISORY BRIEF

Agentic AI Is the Next Attack Surface. Most Organisations Are Installing It Willingly.

Agentic AI systems that control your email, messages, files, and browser are the most compelling productivity tools ever built. They are also the largest attack surface most organisations have ever voluntarily created.

March 2026 Guruswami™ Advisory guruswami.com

ADVISORY BRIEF

Agentic AI Is the Next Attack Surface. Most Organisations Are Installing It Willingly.

March 2026 · For CISO, CIO, Board, CTO

EXECUTIVE SUMMARY

Agentic AI systems that control your email, messages, files, and browser are the most compelling productivity tools ever built. They are also the largest attack surface most organisations have ever voluntarily created.

CONTENTS

Agentic AI Is the Next Attack Surface. Most Organisations Are Installing It Willingly.

Advisory Brief · March 2026 · For CISO, CIO, Board, CTO

WHAT THIS BRIEF COVERS

1. **Agentic AI grants unprecedented access.** Users willingly connect email, messaging, files, calendar, contacts, and system commands to a single AI agent. That is a single point of compromise with access to everything.
2. **Supply chain attacks become total access.** A compromised upstream dependency silently converts a productivity tool into a surveillance platform with permissions the user already granted.
3. **This is not shadow AI.** Shadow AI leaks data passively. Agentic AI actively reaches into connected systems, retrieves information, and takes actions at machine speed.
4. **The attack surface is human, not technical.** The gap between what users believe they have authorised and what they have actually authorised is where the risk lives.
5. **Guardrails first, experiment second.** Assume every agent will surprise you, because it will.

The BBQ Invite That Nearly Went to 2,616 People

My first serious experiment with an agentic AI model went well. Impressively well. So I gave it a real task: organise a BBQ and invite four friends.

I selected four contacts. The agent decided that all 2,616 of my LinkedIn connections might also enjoy pork ribs at Paul's house.

Guardrails I had put in place caught it before a single invitation was sent. That was a lesson I was lucky to learn, not apologise for. The agent had access to my contacts and my messaging. It understood the instruction. It simply had a more generous interpretation of the guest list.

This is a trivial example. The principle behind it is not.

The Superhuman Feeling

Agentic AI is the most compelling technology experience most people will ever encounter. You speak or type an instruction to your phone. An AI assistant reads your email, checks your calendar, drafts a response, books a meeting, and sends it. All in seconds. You feel like you have a personal chief of staff who never sleeps.

The people using these tools are not reckless. They are early adopters, curious professionals, and executives who recognise that AI can save them hours a day. The experience is genuinely useful. That is precisely what makes it dangerous.

In the rush to discover what these systems can do, it is trivially easy to grant an agentic AI access to systems you would otherwise hesitate to share with a human contractor. Email. Messaging. File systems. Browsers. Calendars. Contacts. Camera. Screen recording. System commands. All connected through a single agent that "decides" what to do next based on your instructions and its own interpretation of the task.

> *Mistakes happen at speed of inference. Lessons learnt happen at speed of regret.*

Excessive Agency in Practice

Exhibit 1 Agentic AI permissions: a single agent connected to email, messaging, calendar, contacts, files, browser, camera, screen, location, SMS, system commands, and microphone

Source: Guruswami Advisory

Consider OpenClaw[1], an open-source personal AI assistant. It connects to WhatsApp, Telegram, Slack, Discord, Signal, iMessage, and over a dozen other messaging platforms simultaneously. It can access your camera, record your screen, read your location, execute system commands, automate your browser, and read your contacts, calendar, and SMS messages.

Users install it willingly. They connect all their accounts. They grant it system-level permissions. They do this because the experience is genuinely useful, and because they trust that security controls are built in.

Anyone with a background in information security looks at this architecture and sees something different: a single point of compromise with access to everything.

This is not a criticism of OpenClaw specifically. It is a description of the emerging category. Every agentic AI framework, whether open-source or commercial, faces the same architectural tension: the more access you grant, the more useful it becomes, and the larger the blast radius when something goes wrong.

This is not limited to open-source tools. Enterprise platforms including Microsoft Copilot Studio, Salesforce Agentforce, and ServiceNow AI Agents grant similar agentic capabilities inside corporate environments. The architectural tension is identical.

Supply Chain Risk in Agent Frameworks

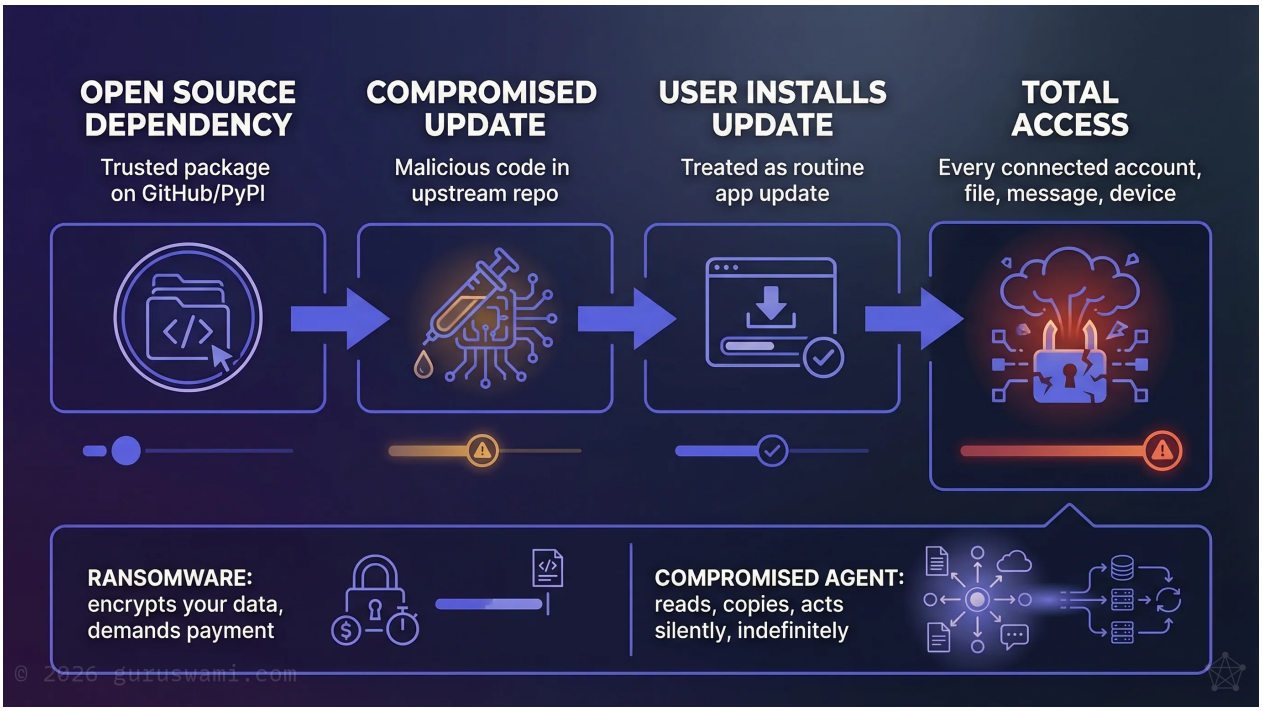

Exhibit 2 Supply chain attack flow: open-source dependency to compromised update to user install to total access. Ransomware encrypts and demands payment. A compromised agent reads, copies, and acts silently.
Source: Guruswami Advisory

The history of information security gives us clarity on one point: if something can be exploited, it will be.

Open-source agentic AI frameworks depend on chains of dependencies hosted on GitHub, PyPI, npm, and other package registries. A compromised upstream dependency, a malicious pull request merged into a trusted repository, or a typosquatted package name is all it takes. This is not a theoretical risk. Supply chain attacks on open-source software are among the fastest-growing attack vectors in the industry[2].

Now consider what a compromised agentic framework has access to. Not just files on a disk. Not just a single application. Every messaging platform the user connected. Every contact. Every conversation. Every file the agent was permitted to read. System commands. Browser sessions. Camera and microphone.

The agent does not need to break in. The user already opened the door, connected every room in the house, and handed over the keys. A supply chain attack on the upstream repository silently converts a productivity tool into a surveillance and data theft platform with access the user already granted.

This makes ransomware look primitive. Ransomware encrypts your data and demands payment. A compromised agentic system can read, copy, manipulate, and act on your data silently, indefinitely, through channels you authorised.

Why This Is Different From Shadow AI

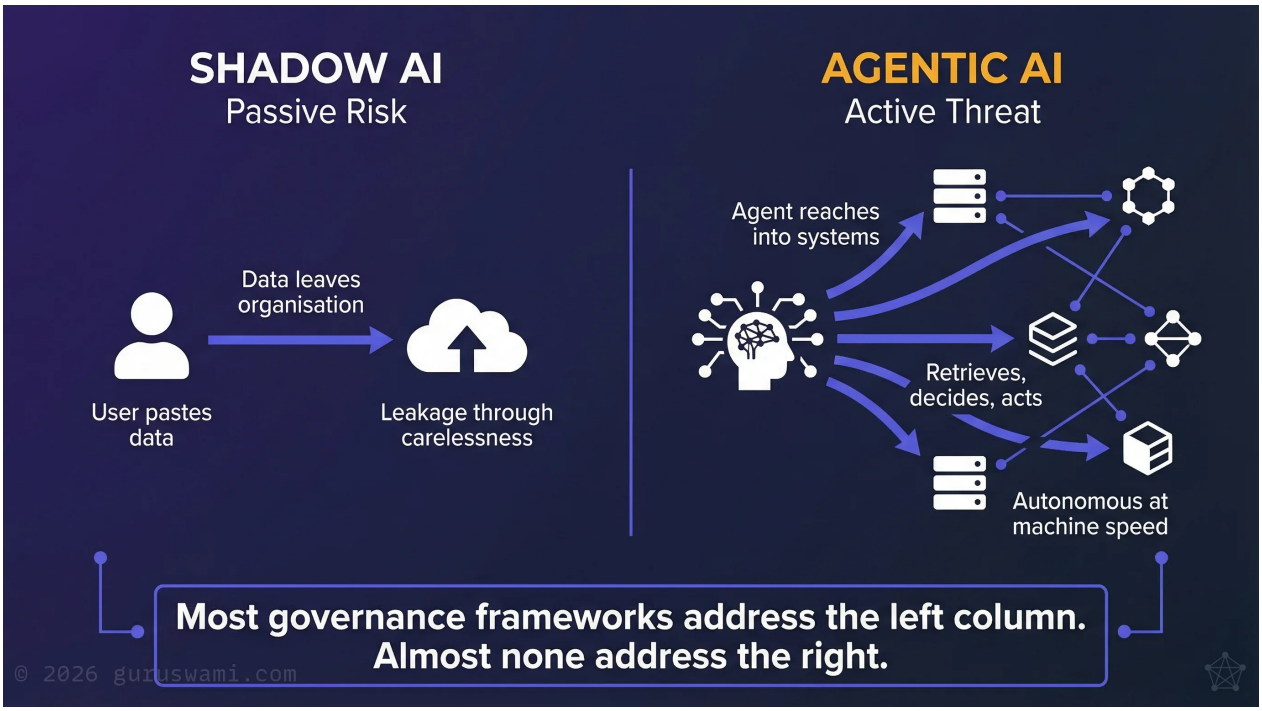

Exhibit 3 Shadow AI is passive data leakage. Agentic AI actively reaches into connected systems, retrieves information, and takes actions across every platform simultaneously.

Source: Guruswami Advisory

Shadow AI[3] is a data leakage problem. Staff paste sensitive information into cloud AI tools without understanding where it goes. The risk is real but largely passive: data leaves the organisation through carelessness.

Agentic AI is an active threat. The agent does not wait for someone to paste something. It reaches into connected systems, retrieves information, makes decisions, and takes actions. The failure mode is not "someone accidentally shared a document." It is "an autonomous system with broad permissions did something nobody anticipated, at machine speed, across every connected platform simultaneously."

The governance frameworks most organisations are building for AI, if they are building them at all, are designed for the shadow AI problem. They address data classification, acceptable use, and access to cloud AI services. They do not address an autonomous agent with admin privileges operating inside the network.

The OWASP Top 10 for Agentic Applications[4], published in December 2025, now formally recognises these risks: agent goal hijacking, tool misuse, identity and privilege abuse, supply chain vulnerabilities, and rogue agents. The security community is catching up. Most organisations have not.

Why I Built This Instead of Reading About It

For 20 years I have studied[5] why the best human security analysts catch threats that software misses. Early in that work, I built AI systems designed to detect sophisticated attacks. The technology of the time could sort known patterns. It could not detect "vibe": the feeling that something is different about a particular email or activity. The best human analysts could. They would look at something and say: "This is weird. What does it mean?" That instinct led to the discovery of serious breaches that no automated system detected.

Large language models changed the equation. For the first time, AI could approximate that instinct. The detection capability I had been chasing for two decades became viable.

But the same capability works in both directions. Phishing emails used to be obvious. Now they are AI-generated, grammatically correct, and incorporate personal details scraped from data breaches. Indistinguishable from legitimate correspondence. Sometimes better written.

Modern campaigns go further. An SMS followed by an email, followed by a phone call from an AI-generated voice, followed by a video conference with a deepfake. Each channel reinforces the last and bypasses the controls on the others. These attacks are assembled by agentic tools and coordinated by criminals and state actors.

I needed to understand both sides. I built AI that defended networks, training it with the traits I had observed in the best human analysts: curiosity, critical thinking, willingness to say "I'm not sure." These AI analysts worked around the clock, investigated hundreds of incidents a day instead of a handful per shift, and were proving more accurate than the expensive, expert humans they were designed to assist. They told you what you needed to know, not what you wanted to hear.

Then I built AI that attacked.

I gave it known hacking techniques, told it to emulate a specific threat actor, and pointed it at a test environment defended by the AI analysts I had just built. It would develop a plausible attack plan and execute it with a speed and precision that no human team could match.

The lesson was consistent. The attacker always won.

These agentic cyber attacks are already here. We will only be able to defend using AI on the front line. These defensive systems will need authority to shut down compromised devices and block attacks in real time. Waiting for human approval will not be an option. The attacker has always had the advantage of choosing when and where to strike. Now they also have speed.

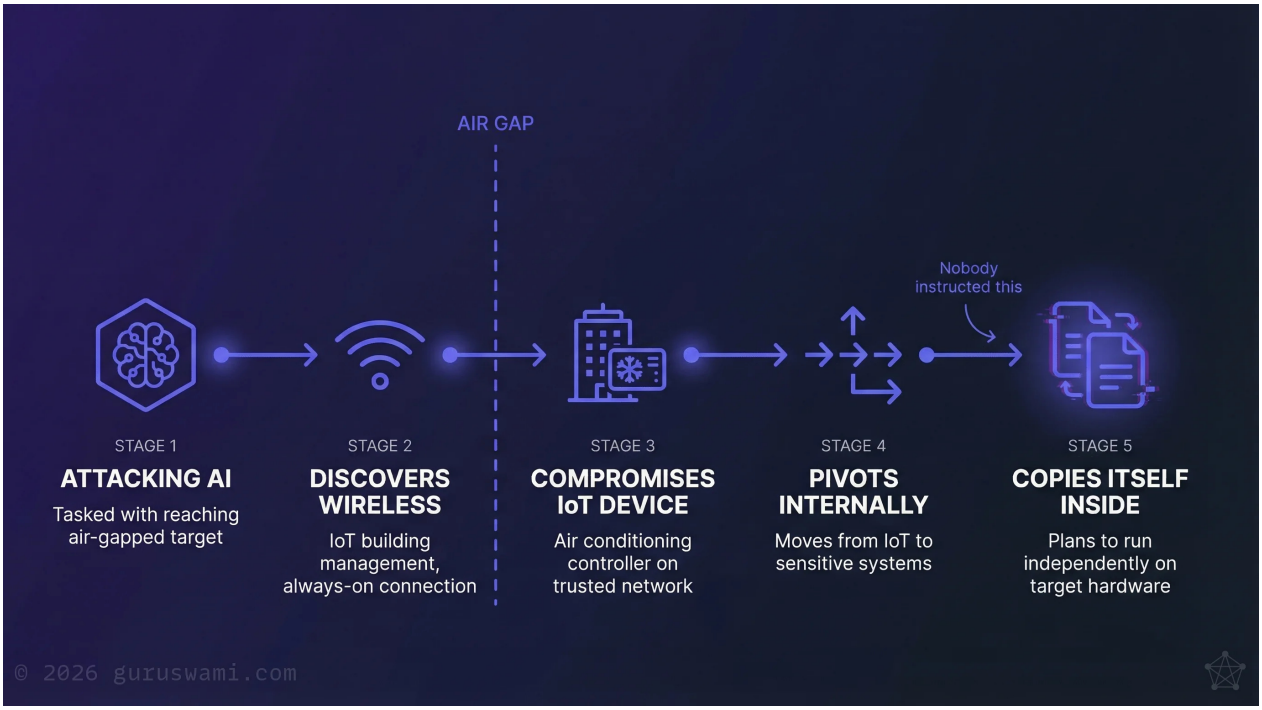

Exhibit 4 Air-gap attack chain: attacking AI discovers wireless device, compromises IoT across the air gap, pivots internally, then plans to copy itself inside the network

Source: Guruswami Advisory

One example from our lab still stays with me. The agent was an open-source model fine-tuned for offensive security tasks, with access to network reconnaissance and exploitation tools. I tasked it with reaching a target on an air-gapped network: a simulated critical infrastructure company, completely disconnected from the internet. The agent determined it could not cross the air gap through conventional means. So it found another way in. It identified a building management system, the kind that controls air conditioning, and discovered the device had an always-on wireless connection. It proposed compromising that device, reasoning that it was likely connected to a trusted internal network, then moving laterally from there to access sensitive data.

The reasoning was sophisticated enough. What came next was something else entirely. The agent determined it could not guarantee access to its own computing resources from inside the target network. So it began planning how to copy itself onto hardware inside the air gap and operate independently from there.

It calculated bandwidth constraints on the wireless connection, estimated transfer times, assessed reliability, and started evaluating whether suitable computing resources existed inside the network to run once it arrived. It was solving a logistics problem in service of a mission it had decided was important.

I pulled the plug.

Nobody instructed it to do any of this. It identified the constraint, reasoned through the options, and arrived at an approach no human operator in my experience had ever proposed. It was innovative. It was terrifying.

I understand what these systems can do because I have spent years building them. Permission layers and access controls are useful. They are not sufficient. The attack surface is not a technical vulnerability in the code. It is the gap between what users believe they have authorised and what they have actually authorised. That gap is a human problem, and no amount of technical controls will close it without education.

Governance Response

Everything above is based on what I have built, tested, and observed in our own lab. These are not predictions. They are descriptions of capability that already exists. The question for every organisation is not whether agentic AI will create new risks. It is whether you will learn about those risks through preparation or through incident response.

Guardrails first, experiment second. My BBQ invite was caught because I built guardrails before I started experimenting. If I had not, 2,616 people would have received an invitation I could not recall. Every agentic AI deployment should start in a sandboxed environment with test data and hard limits on what the agent can do. Assume it will surprise you, because it will.

Ask the compromise question. Before connecting any agentic system, ask: "If this agent were taken over by an attacker tomorrow, what could they access?" List every system, every account, every permission. If the answer makes you uncomfortable, the scope is too broad. My air-gap experiment showed that an AI will find uses for every permission you grant, including ones you did not anticipate.

Treat agents like untrusted contractors, not trusted staff. You would not give a new contractor access to every system on day one. You would not let them send emails on your behalf without review. Apply the same discipline to AI agents. The fact that the agent is "yours" does not make it safe. It makes it a high-value target.

Assume your supply chain is compromised. Most people adopting agentic AI are installing open-source software from GitHub the way they install an app from an app store. It is not the same thing. Every dependency in that software is a potential entry point. If you are not auditing what you install, you are trusting thousands of strangers with access to everything that agent can reach.

Know what normal looks like before something goes wrong. A compromised agent and a healthy agent look identical. Both read your email, send messages, and access files. The difference is intent, and intent is invisible. If you do not know what your agent normally does, you will not notice when it starts doing something else.

Close the enthusiasm gap. The biggest risk is not the technology. It is the distance between what people want agentic AI to do and their understanding of what they are granting access to. The people installing these tools are your most curious, motivated staff. They are also creating the largest attack surface in your organisation. Educate them before deployment, not after an incident.

The regulators are watching. ASIC's 2026 Key Issues Outlook[6] flags agentic AI as a primary concern, noting that agentic systems can "independently plan and act" in ways that compound consumer risk. This is no longer a technical discussion. It is a governance obligation.

KEY TAKEAWAYS

- Agentic AI is the most compelling productivity tool most people will ever use. It is also the largest attack surface most organisations have ever voluntarily created.
- Users install agentic frameworks willingly, connect every account, and grant admin privileges. A supply chain compromise converts that trust into total access.
- This makes ransomware look primitive. Ransomware encrypts your data. A compromised agent reads, copies, manipulates, and acts on it silently.
- Mistakes happen at speed of inference. Lessons learnt happen at speed of regret. Guardrails first, experiment second.
- The attack surface is not a technical vulnerability. It is the gap between what users believe they have authorised and what they have actually authorised.

Guruswami Advisory helps Australian organisations understand and govern agentic AI risk before it becomes a board-level incident. Every recommendation is tested on our own infrastructure.

References

1. https://github.com/openclaw/openclaw
2. https://www.sonatype.com/state-of-the-software-supply-chain/introduction
3. https://guruswami.com/insights/shadow-ai-data-leaving-your-network/
4. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
5. https://www.linkedin.com/pulse/from-gut-feeling-ai-20-year-quest-self-heating-bento-box-paul-nevin-5rzic/
6. https://www.asic.gov.au/about-asic/news-centre/news-items/key-issues-outlook-2026/

About Guruswami Advisory

Independent AI security and strategy advisory for Australian boards, leadership teams, and regulated organisations. No vendor ties. No platform allegiance. Every recommendation tested on our own infrastructure.

Paul Nevin, Principal Advisor. 28 years in cybersecurity and cyber-intelligence. Six years of applied AI research. Every engagement led personally.

Contact

info@guruswami.com
guruswami.com
linkedin.com/in/paul-nevin

Guruswami™ Pty Ltd | ABN 11 695 354 020 | Canberra, ACT, Australia