CYBER STASH

AI and Cybersecurity

Separating Fact from Fiction in the Age of Autonomous Risk

Controlled Enablement. | Defensible Innovation. | Resilient Security.

May 2026

cyberstash.com

SECURITY ADVISORY

CYBER STASH

The Noise, the Reality and the Strategic Question

Artificial Intelligence has become one of the most overused phrases in cybersecurity. It is used to sell products, justify budgets, attract investors, alarm boards, impress customers and explain almost every new development in the threat landscape. The result is predictable: the discussion has become noisy.

At one end of the debate, AI is described as a revolutionary force that will soon make traditional cybersecurity obsolete. At the other end, it is dismissed as another cycle of vendor hype, no different from the inflated promises previously attached to machine learning, automation and "next-generation" security platforms.

Both views are incomplete.

The more important truth sits between them. AI has not made cybersecurity impossible. It has not removed the need for patching, identity control, logging, segmentation, secure development, incident response or human judgement. But it has changed the economics of cyber operations. It is reducing the time, effort and specialist knowledge required to perform tasks that previously demanded scarce expertise. It is accelerating vulnerability discovery, exploit development, reconnaissance, phishing, code analysis, malware troubleshooting, fraud preparation and defensive triage. It is also introducing a new class of enterprise risk: AI systems that do not merely generate content, but observe, reason, decide and act.

The more profound shift is that AI is moving from the edge of the organisation into its operating core. What began as content generation is rapidly becoming enterprise integration. AI systems are being embedded into business processes, software platforms, finance workflows, development pipelines, operational technology environments and third-party supply chains. They are being connected to internal data, granted access to enterprise applications, given privileges to query or modify systems, and authorised to assist with decisions that affect customers, employees, finances, operations and security.

This transforms AI from a tool into an actor. An AI assistant may help an employee write an email. An AI agent may read internal data, interact with systems, modify configurations, raise tickets, make recommendations, trigger workflows, submit changes, approve actions or communicate externally. The second scenario is not simply a productivity tool. It is a digital actor with access, context, permission and operational consequence.

That distinction is where many organisations are currently underprepared. Once AI has access, context and permission to act, it must be governed as part of the enterprise control environment. It requires identity, least privilege, monitoring, auditability, human oversight and clear accountability. Without those controls, organisations may unknowingly create a new class of privileged digital worker — one that is fast, useful and scalable, but also vulnerable to manipulation, excessive trust and unintended consequences.

The Australian Signals Directorate's Australian Cyber Security Centre has warned that agentic AI systems are increasingly operating across critical infrastructure and defence sectors, and that they introduce risks through their use of tools, memory, external data, planning workflows and execution privileges. The same guidance warns against granting agentic AI broad or unrestricted access, particularly to sensitive data or critical systems.

This advisory is written for organisations that need to move beyond simplistic AI narratives. It is not a call to panic, and it is not an argument for delaying adoption. AI will become part of how organisations operate, compete, defend themselves and make decisions. The strategic question is whether organisations can adopt AI with enough visibility, governance and resilience to benefit from it without quietly expanding their attack surface.

The Cybersecurity Debate Has Shifted

For many years, the cybersecurity industry has assumed that the most advanced offensive capabilities were limited by human scarcity. Elite vulnerability researchers, exploit developers, malware authors and intrusion operators were difficult to find, expensive to retain and slow to scale. That scarcity was itself a form of defence. Even when software contained serious defects, finding and weaponising them required time, judgement and specialised skill.

AI is eroding that assumption.

The most important recent signal is not that one AI model performed well in a benchmark. It is that multiple frontier models are now demonstrating material cyber capability uplift, including vulnerability research, exploitation and multi-step attack simulation. Anthropic reported that Claude Mythos Preview could identify and exploit zero-day vulnerabilities across major operating systems and web browsers in controlled testing, including a now-patched 27-year-old OpenBSD vulnerability and complex exploit chains involving modern browser defences.

Mozilla's experience gives the claim more weight. The Firefox team reported that Claude Mythos Preview helped identify 271 vulnerabilities fixed in Firefox 150 during an initial evaluation. Mozilla's conclusion was nuanced: the model did not appear to discover a new category of vulnerability beyond human comprehension, but it dramatically accelerated the discovery of defects that would previously have required scarce human expertise.

This nuance is critical. The story is not that AI has become magic. The story is that AI is turning difficult but human-comprehensible work into work that can be repeated, scaled and accelerated.

The UK AI Security Institute reached a similar conclusion. Its evaluation of Claude Mythos Preview found that the model represented a step change in cyber performance and could complete parts of multi-step attack simulations under controlled conditions. However, it also noted meaningful limitations, including failure in an operational technology-focused range and the fact that simulated environments lacked many of the active defenders, defensive tools and alerting conditions present in real organisations.

This is where the facts and fads begin to separate.

The factual development is that AI can now perform increasingly sophisticated cyber tasks. The exaggerated claim is that this automatically translates into reliable, autonomous compromise of hardened enterprise environments. It does not. Real-world attacks still require context, persistence, infrastructure, stealth, timing, access and operational judgement. But the barrier to entry is falling, and the cost of iteration is collapsing.

That is enough to change the threat model.

The Real Impact of AI Is Economic Before It Is Technical

Cybersecurity leaders should resist the temptation to view AI only as a technical capability. Its deeper impact is economic.

AI reduces the marginal cost of analysis. It reduces the time required to test hypotheses. It allows attackers and defenders to run more experiments, review more code, generate more lures, analyse more logs, simulate more paths and operationalise more knowledge. In cybersecurity, where advantage often belongs to the side that can learn and adapt faster, this matters enormously.

Google's Mandiant reported in M-Trends 2026 that the mean time to exploit dropped to an estimated negative seven days. In practical terms, this means exploitation was being observed, on average, before a vendor patch was available — collapsing the window defenders traditionally relied on to assess, test and remediate vulnerabilities. Mandiant also reported that attackers are already abusing large language models inside compromised environments, while cautioning that most breaches still stem from fundamental human and systemic failures rather than AI itself.

This is the right strategic lens. AI is not replacing the old threat landscape. It is accelerating it.

The organisations most exposed are not necessarily those facing the most futuristic attacks. They are often those still operating vulnerability management, identity governance, third-party assurance, patching, security monitoring and incident response at a pace designed for a slower era. If exploitation is moving faster than remediation, and AI further accelerates reconnaissance and exploit development, then the traditional "find, prioritise, approve, patch later" model becomes increasingly fragile.

In an AI-accelerated threat environment, delay itself becomes part of the attack surface.

This does not mean every vulnerability will be exploited immediately. It means the likelihood of timely exploitation is increasing as AI reduces the technical complexity, manual effort and specialist expertise that previously limited how quickly attackers could analyse, prioritise and weaponise exposed weaknesses.

Defenders may still gain some time from the fact that attackers also face operational limits. If AI dramatically increases the number of systems that can be discovered and compromised, adversaries must still decide which victims are worth progressing through the kill chain towards persistence, lateral movement, data theft, extortion or disruption. That backlog may occasionally work in the defender's favour. But it is not something to rely on. No organisation knows where it sits on an attacker's list, or whether today's ignored foothold becomes tomorrow's breach.

AI Has Not Replaced Cybersecurity Fundamentals — It Has Changed the Cost of Weakness

One of the most dangerous fads in the current AI discussion is the belief that AI creates an entirely new cybersecurity universe. It does not. Most successful intrusions still depend on familiar weaknesses: exposed systems, poor identity controls, weak monitoring, excessive privileges, unpatched software, insecure third-party access, inadequate segmentation and delayed response.

AI changes the pressure applied to those weaknesses.

- A phishing campaign becomes easier to personalise.
- A vulnerable service becomes easier to analyse.
- A misconfigured cloud environment becomes easier to understand.
- A leaked codebase becomes easier to review.
- A help desk workflow becomes easier to manipulate.
- An executive's writing style becomes easier to imitate.
- A defensive alert backlog becomes easier for both sides to mine for signal.

The Australian Cyber Security Centre's April 2026 update on frontier AI models reached a similar conclusion: strong cybersecurity fundamentals remain an important foundation, but the new age of AI requires more than simply "getting the basics right". Organisations must also consider how to use AI defensively, minimise attack surfaces, patch promptly, implement layered architectures and prepare incident response for faster-moving threats.

This is the practical balance boards need to hear. AI does not make established security controls irrelevant.

AI turns yesterday's tolerated delays, exceptions, underinvestment and accepted risks into tomorrow's exploitable weaknesses. In an AI-accelerated threat landscape, old assumptions about risk tolerance may no longer hold.

Chatbots to Agents: The More Important Enterprise Risk

Most organisations began their AI journey with content generation: drafting emails, summarising documents, producing marketing copy, analysing policies or assisting developers. These use cases carry risk, especially around confidentiality, intellectual property, privacy and accuracy. But they are not the end state.

The direction of travel is agentic AI.

Agentic AI systems combine language models with tools, data sources, memory, planning and permissions. They do not merely answer questions. They can act toward a goal. That may include querying databases, reading email, generating code, interacting with SaaS platforms, raising tickets, updating records, conducting analysis or triggering business workflows.

This changes the control model. A chatbot can be wrong. An agent can be wrong and then do something about it.

The OpenAI and PwC collaboration around AI agents for the office of the CFO is a useful market signal. The collaboration is focused on finance workflows including planning, forecasting, reporting, procurement, payments, treasury, tax and the accounting close. These are not low-value use cases. They sit close to capital allocation, financial control, decision-making and enterprise risk.

That is why agentic AI must be treated differently. An AI agent with access to financial systems, procurement workflows, customer records, cloud infrastructure or code repositories is not just software. It is a semi-autonomous operational participant. It requires identity, ownership, least privilege, audit logging, control points, testing, rollback and incident response.

The ACSC's agentic AI guidance is explicit on this point. It states that agentic systems expand the attack surface through tools, external data sources and memory, and that malicious actors may exploit components to conduct attacks such as executing scripts or sending unauthorised emails. It also highlights privilege compromise, scope creep, identity spoofing and agent impersonation as key risks.

This is the moment where AI risk stops being a technology policy issue and becomes an operating model issue. The real question is not whether the organisation should use AI, but how safely it can give AI access, authority and autonomy.

That requires more than an AI policy. It requires clear ownership, identity controls, privilege boundaries, audit trails, human approval points, vendor assurance, incident response pathways and board-level visibility. Without that shift, organisations risk building a powerful new operating layer that is fast, useful and scalable — but also opaque, over-trusted and difficult to control.

The Personal Agent Layer: OpenClaw, Claude Desktop and the New Trust Boundary

OpenClaw and Claude Desktop are different products, but they point to the same strategic shift: AI is moving from a chat interface into a personal agent layer.

OpenClaw is a local AI agent environment designed to connect a language model to tools, services and host resources so it can perform tasks on behalf of the user. Claude Desktop is Anthropic's desktop application for Claude, and its significance extends beyond chat because it can connect to external tools and data sources through the Model Context Protocol, or MCP.

MCP is best understood as the connector layer for AI agents. It standardises how AI applications retrieve context, invoke tools and interact with external systems. That makes agents more useful, but it also creates new pathways through which sensitive data, permissions, actions and trust can be exposed.

This is where the issue becomes more important than the product names. As MCP servers, desktop extensions, browser agents and local agent runtimes become easier to install, the barrier to adoption falls. Employees no longer need to be highly technical to connect AI to files, browsers, code repositories, SaaS platforms, calendars, email or local workflows. What once required integration work can increasingly be enabled through plug-ins, extensions or local connectors.

OpenClaw provides a useful case study. The ClawJacked vulnerability showed how a malicious website could connect to a locally running OpenClaw gateway and take control of the agent, turning ordinary browser interaction into a pathway for agent compromise. Other concerns around exposed instances, malicious skills, token theft and local file access reinforce the same point: once an agent is connected to the local environment and user credentials, it may create a larger blast radius than an ordinary desktop application.

Claude Desktop raises a related but distinct concern. The issue is not that Claude Desktop is the same as OpenClaw. The issue is that desktop AI tools increasingly create pathways between browsers, local applications, files, extensions, MCP servers and external systems. Even where behaviour is not malicious, expanding browser-to-desktop integration can increase the attack surface if users, administrators and security teams do not understand what has changed.

The advisory lesson is practical. OpenClaw, Claude Desktop, browser agents, MCP servers, developer agents and similar tools should not be treated as ordinary desktop utilities. They should be treated as access brokers between human intent, enterprise data and operational systems.

For high-risk use cases, personal agents should not run on primary workstations with broad user credentials, unrestricted file access or sensitive SaaS sessions. Organisations should use approved deployment pathways, isolated environments, managed identities, least privilege, restricted tool access, controlled extensions or MCP servers, logging, endpoint monitoring and clear rules for what data and systems agents may access.

Personal agents are not another desktop utility. They are becoming a new operating layer between users, data, applications and business decisions.

They will enter the enterprise through approved platforms, SaaS updates, developer tools, browser extensions, third-party workflows and employee experimentation. The leadership question is no longer whether this can be stopped. It is whether the organisation can define the trust boundaries before agents are connected to sensitive data, privileged systems and critical workflows.

Organisations need to decide now where agents are allowed, where they are restricted, what they can access, what they can do, and which safeguards must exist before they touch Crown Jewels, financial approvals, customer data, source code, security tooling or administrative platforms. Low-risk use should be enabled. High-risk use must be controlled. Otherwise, personal agents will become unmanaged shadow automation — powerful, productive and embedded before the organisation is ready to govern them.

Shadow AI Is Becoming the New Operating Reality

For many organisations, the first instinct will be to manage AI adoption through restriction. That response is understandable, but increasingly unrealistic.

AI is not entering the enterprise only through formal transformation programmes, approved platforms or board-endorsed initiatives. It is already arriving through employees, SaaS tools, developer environments, productivity platforms, customer service workflows, finance systems, marketing teams, security operations and third-party suppliers. In many cases, AI capability is being introduced through ordinary software updates, embedded features or individual experimentation before risk, security and legal teams have had time to assess the implications.

This makes shadow AI more difficult to control than traditional shadow IT. Traditional shadow IT was often about convenience: an employee using an unsanctioned file-sharing tool, project management platform or messaging application to get work done faster. Shadow AI is different. It does not simply provide another place to store, share or process information. It multiplies output.

It helps employees write, analyse, code, summarise, research, translate, design, troubleshoot, automate and decide. It can compress days of work into hours, allow small teams to perform work that previously required specialist resources, and make complex tasks feel newly accessible. That productivity gain changes the psychology of risk acceptance.

When a tool saves time, employees may tolerate the friction of approval. When a tool materially increases their capacity, the incentive to use it becomes much stronger. AI does not merely make work more convenient. It can make work feel newly possible.

This creates a governance challenge that cannot be solved by policy alone. If the approved pathway is too slow, restrictive or disconnected from business needs, employees and business units will find alternatives. They may use personal accounts, unapproved tools, browser extensions, embedded SaaS AI features or third-party platforms that sit outside central visibility.

Over time, this can turn shadow AI from an exception into a preferred operating model, particularly where teams believe the business benefit outweighs the perceived risk. That does not mean organisations should accept uncontrolled AI use. It means they need to understand why it happens.

Shadow AI is often a signal that the business has found a capability gap faster than governance has found a safe path to enable it.

The organisations that respond well will not be those that simply say “no”. They will be those that provide safer, approved alternatives that are useful enough for the business to adopt. Otherwise, AI adoption will still happen — but without visibility, without consistent data protection, without audit trails, without vendor assurance and without clear accountability.

In this sense, the risk is not simply that employees will use AI. The greater risk is that AI becomes deeply embedded in how work gets done before the organisation understands where it is being used, what data it touches, what decisions it influences, and what dependencies it creates.

Crown Jewels Matter More in an AI-Accelerated Enterprise

If shadow AI is the operating reality, then the strategic response cannot be to treat every AI interaction as equal.

Organisations need a sharper sense of priority.

This is where defining the organisation's Crown Jewels becomes paramount. Not every system, dataset, workflow or AI use case carries the same level of risk. The priority should be to identify the assets, data and processes that would cause the greatest harm if exposed, manipulated, automated incorrectly or placed under the control of an over-permissioned AI system.

This includes sensitive customer data, intellectual property, financial systems, privileged administration platforms, security tooling, operational technology, critical business workflows and decision-making processes that affect customers, money, safety, trust or regulatory obligations.

In an AI-accelerated environment, organisations cannot defend everything with equal intensity. Nor can they afford to treat every AI use case as either harmless or unacceptable. The practical response is to concentrate governance and security effort where the consequences are highest.

That means asking sharper questions.

Who or what can access the Crown Jewels? Which AI systems can interact with them? What actions can those systems perform? Can they only observe and recommend, or can they modify, approve, trigger or communicate? What is the plausible worst-case impact if the AI system is misused, manipulated or wrong? Could it expose sensitive data, alter a critical configuration, approve an unauthorised transaction, disrupt operations or influence a material business decision? If that scenario occurs, what is the second line of defence — human approval, transaction limits, compensating controls, monitoring, rollback, segregation of duties or incident response — and, after those safeguards are considered, is the residual risk acceptable to the organisation?

This matters because AI is likely to increase the cost and complexity of managing cyber risk. More systems will become connected. More data will be exposed to automated workflows. More decisions will be influenced by machine-generated outputs. More third-party platforms will embed AI capabilities by default. The attack surface will not only expand through new tools, but through the deeper integration of those tools into ordinary business processes.

The answer is not to spread security effort thinly across every possible AI interaction. It is to prioritise the areas where compromise, manipulation or misuse would create material business impact.

The strategic response is controlled enablement. Organisations need to provide approved AI pathways that are useful enough for the business to adopt, secure enough for risk teams to trust, and flexible enough to evolve as the technology changes. This requires more than policy. It requires identity controls, data classification, access boundaries, monitoring, vendor assurance, human approval points and executive visibility over the highest-risk AI use cases.

This is where AI governance becomes practical. It is not about stopping innovation. It is about ensuring that the organisation knows which assets matter most, which AI systems can touch them, what those systems are allowed to do, and how the business remains in control when automation begins to move faster than traditional oversight.

The Governance Gap Is Becoming the Strategic Risk

The adoption curve is moving faster than the governance curve.

That gap is becoming one of the defining risks of enterprise AI. Not because organisations are wrong to adopt AI, but because the benefits are becoming too significant to ignore. AI can compress analysis, accelerate decision-making, reduce manual effort, improve customer engagement, support employees, enhance security operations, assist software development and allow small teams to produce outcomes that previously required far greater resources.

For many businesses, the upside is not theoretical. AI offers a way to increase productivity, unlock capacity, reduce operational bottlenecks, improve service delivery and compete with larger or better-resourced organisations. In cybersecurity, it can help triage alerts, summarise incidents, analyse logs, support threat hunting, accelerate vulnerability assessment and reduce the time between detection and response.

This is why AI risk cannot be managed simply by slowing adoption. The business case is too compelling. The more realistic challenge is to make AI adoption safe enough, visible enough and accountable enough that organisations can capture the benefit without absorbing unmanaged risk.

APRA has called for a step-change in AI-related risk management and governance across banks, insurers and superannuation trustees. Its observations warn that governance, risk management, assurance and operational practices are failing to keep pace with the scale, speed and complexity of AI adoption.

This warning should not be read as relevant only to financial services. Regulated industries often reveal the control problems that later become mainstream expectations. The same pattern will affect healthcare, education, government, critical infrastructure, legal services, managed service providers and any organisation embedding AI into high-value workflows.

The governance gap has several forms.

First, many organisations do not know where AI is already being used. Employees may use public tools. Business units may procure SaaS products with embedded AI. Vendors may introduce AI features through product updates. Developers may rely on coding assistants. Security teams may use AI-enabled triage tools. Finance, HR and customer service teams may trial agents with limited central oversight.

Second, many organisations govern AI as if it were only a data privacy problem. Privacy matters, but it is not enough. AI also raises questions of integrity, authorisation, explainability, operational resilience, vendor dependence, model drift, supply chain exposure, output trust and incident response.

Third, many organisations lack a risk-tiered AI adoption model. A low-risk writing assistant should not be governed the same way as an autonomous procurement agent, a security operations agent, a code deployment agent or an AI-enabled operational technology system.

The organisations that succeed will not be those that avoid AI risk altogether. They will be those that understand which risks are worth taking, which risks must be controlled, and which risks should not be accepted under any circumstances.

The goal is not to remove risk from AI adoption. That is neither realistic nor commercially sensible. The goal is to ensure that the risk being taken is visible, deliberate, proportionate to the business benefit and supported by controls that can withstand scrutiny.

The organisations that succeed will be those that turn AI governance from a document into a living control system — one that enables innovation, protects the Crown Jewels, supports accountable decision-making and evolves as quickly as the technology itself.

Mythos Moment: Why AI Vulnerability Discovery Changed the Conversation

For many years, the AI cybersecurity debate was mostly theoretical. Security leaders were told that AI would eventually help attackers discover vulnerabilities, generate exploits and automate parts of the kill chain. The concern was plausible, but often abstract. It lacked a defining moment that forced boards, regulators and defenders to reassess the speed at which offensive and defensive cyber capability could change.

The emergence of Claude Mythos Preview changed that conversation.

Anthropic positioned Claude Mythos Preview as part of Project Glasswing, a controlled initiative designed to give selected partners access to advanced AI capability for finding and fixing vulnerabilities in foundational software systems. Anthropic described the focus areas as local vulnerability detection, black-box binary testing, endpoint security and penetration testing of systems that represent a large portion of the world's shared attack surface.

The most important signal came from Mozilla. As part of its collaboration with Anthropic, Mozilla reported that Firefox 150 included fixes for 271 vulnerabilities identified during an initial evaluation using an early version of Claude Mythos Preview. Mozilla's own framing was significant: for a hardened target, a single serious bug would previously have been a major concern, and the discovery of so many at once forced the question of whether defenders can continue to keep up using traditional methods alone.

This does not mean AI has suddenly discovered a new category of vulnerability beyond human comprehension. That is not the lesson. The more important lesson is that AI appears increasingly capable of accelerating work that previously depended on scarce, highly specialised human expertise: reviewing code, identifying unsafe logic paths, testing assumptions, generating proof-of-concept behaviour, triaging findings and helping defenders understand where exposure exists.

The UK AI Security Institute's evaluation adds further weight to this conclusion. Its testing found that Claude Mythos Preview represented a step up over previous frontier models in a landscape where cyber performance was already improving rapidly. In controlled evaluations where the model was explicitly directed and given network access, AISI observed that it could execute multi-stage attacks on vulnerable networks and discover and exploit vulnerabilities autonomously — tasks it described as taking human professionals days of work.

The trend also appears broader than one model. AISI later reported that GPT-5.5 was one of the strongest models it had tested on cyber tasks and was the second model to solve one of its multi-step cyber-attack simulations end to end. Its evaluation stated that GPT-5.5 reached a similar level of cyber performance to Claude Mythos Preview, suggesting the shift is not isolated to a single vendor or model release.

This is why the Mythos moment matters. It did not prove that AI can reliably compromise any well-defended enterprise environment. It did not make human expertise irrelevant. It did not eliminate the need for validation, context, triage or judgement. But it did provide credible evidence that frontier AI can meaningfully compress the time and skill required to perform serious cybersecurity work.

That changes the strategic equation.

If defenders can use AI to find and remediate vulnerabilities faster, the defensive upside is enormous. But if similar capabilities become available, leaked, replicated or approximated by adversaries, then the same capability becomes an offensive accelerator. This is the dual-use reality at the centre of AI cybersecurity. The tool that can help secure critical software can also make it easier to discover where that software is weak.

The advisory conclusion is not that organisations should panic. It is that the assumptions underpinning vulnerability management, secure development, penetration testing, exposure management and incident response must be revisited. The world in which only a small number of highly skilled humans could perform advanced vulnerability discovery is beginning to change. That does not guarantee every weakness will be exploited, but it does increase the probability that weaknesses will be found, understood and prioritised faster than many organisations are prepared to respond.

Fact versus Fiction: What AI Can and Cannot Yet Do in Cybersecurity

The AI cybersecurity debate is suffering from two opposing distortions. One side treats AI as a near-magical offensive capability that will make defenders powerless. The other side dismisses it as another wave of inflated vendor claims. Both positions miss the point.

AI is neither magic nor marketing noise. It is a capability multiplier.

The first fiction is that AI has replaced the attacker. It has not. Real-world cyber operations still require access, infrastructure, timing, persistence, stealth, decision-making and operational security. A model may assist with reconnaissance, exploit development, phishing content, code review or malware troubleshooting, but successful compromise still depends on context and execution. In controlled environments, frontier models are showing meaningful capability, but those environments are not the same as live enterprise networks with defensive tooling, alerting, segmentation, identity controls, user behaviour, incident response and business complexity.

The second fiction is that AI makes every user an elite hacker. It does not. What it does is reduce the gap between intent and capability. A less experienced actor can use AI to understand a vulnerability more quickly, write better lures, troubleshoot code, interpret logs, generate scripts, research a target or adapt existing tooling. This does not instantly create elite tradecraft, but it does lift the baseline. Tasks that once required specialist skill become more accessible, and tasks that once took hours or days can be compressed into minutes.

The third fiction is that AI creates cyber risk only when attackers use it. That view is too narrow. Some of the most material risks will come from organisations embedding AI into their own environments without sufficient visibility, control or accountability. AI systems connected to email, documents, cloud platforms, finance systems, ticketing tools, development pipelines or security platforms can become part of the enterprise control fabric. Once they can access data, call tools, trigger workflows or modify configurations, they are no longer just producing advice. They are influencing operations.

The fourth fiction is that traditional security controls are obsolete. The opposite is closer to the truth. AI makes established controls more important, not less. Identity, least privilege, patching, segmentation, logging, secure configuration, third-party assurance, data protection and incident response remain foundational. The difference is that weak implementation becomes more costly when attackers and internal systems can move faster.

The fifth fiction is that AI adoption can be controlled through prohibition. That may work for narrow, high-risk cases, but it is not a sustainable enterprise strategy. AI is being embedded into ordinary software, SaaS platforms, developer tools, business workflows and third-party services. Employees and business units will use it because the productivity gains are real. The practical question is not whether AI can be kept out entirely, but whether the organisation can provide approved pathways that are useful enough to compete with uncontrolled alternatives.

The facts are more nuanced and more important.

AI can accelerate vulnerability research, but findings still require validation. AI can improve phishing and social engineering, but fraud still depends on trust relationships, process weaknesses and human decision points. AI can assist malware development, but deployment, persistence, evasion and monetisation still require operational maturity. AI can support security operations, but poor telemetry, weak playbooks and unclear ownership will still limit response. AI can analyse large volumes of data, but if the underlying data is incomplete, poisoned, unauthorised or poorly governed, the output can be misleading or unsafe.

This is why organisations need a balanced view. Overstating AI capability can lead to fatalism, poor investment decisions and unnecessary fear. Understating it can lead to complacency. The better position is to recognise AI for what it is becoming: a fast-improving capability layer that changes how work is performed, how expertise is accessed, how attacks are prepared and how defence can be scaled.

The Blurred Frontier: Why Today's Fiction Can Become Tomorrow's Capability

The most useful way to understand AI in cybersecurity is not as a replacement for human capability, but as a force that reduces friction. It reduces the friction of learning, testing, writing, analysing, translating, coding, summarising and iterating. For attackers, that means greater scale and faster experimentation. For defenders, it means the possibility of faster triage, better detection logic, more efficient vulnerability analysis and improved exposure reduction.

That said, AI should not be underestimated.

One of the lessons from recent frontier model progress is that capability does not always improve gradually; it can arrive in sudden, material jumps. In cybersecurity, that matters because today's unrealistic claim can become tomorrow's operational capability.

The boundary between fact and fiction is therefore becoming increasingly difficult to draw.

This is not because every claim about AI is credible. Many claims remain exaggerated. Some are vendor positioning. Some are based on laboratory conditions that do not translate neatly into real-world enterprise environments. Some assume autonomy where, in practice, models still require scaffolding, human oversight, tool access, validation and correction. But dismissing these developments because they are imperfect would be a strategic mistake.

Cybersecurity is especially sensitive to capability jumps because small improvements can have disproportionate impact. A model does not need to replace an expert to create risk. It only needs to help more people perform more technical work faster than before. It does not need to autonomously compromise a mature enterprise to matter. It only needs to improve reconnaissance, exploit understanding, phishing quality, malware adaptation, vulnerability triage or social engineering at scale.

This is where the line between fact and fiction becomes blurry. A claim may be overstated today but directionally correct tomorrow. A capability that requires human guidance today may become increasingly autonomous. A task that is unreliable in one model generation may become repeatable in the next. A defensive breakthrough in vulnerability discovery may also become an offensive accelerator if similar capabilities become broadly available, replicated or misused.

For leadership teams, the implication is clear: AI cyber risk should not be assessed only against what models can do today. It must also be assessed against the trajectory of capability improvement, the rate of adoption, the degree of enterprise integration and the consequences if assumptions are wrong.

That does not mean organisations should plan around science fiction. It means they should avoid building security strategies that depend on AI remaining limited, expensive, unreliable or accessible only to experts. Those assumptions are already weakening.

The more mature position is to treat AI as an accelerating variable in the threat landscape. It changes how quickly adversaries can learn. It changes how cheaply they can experiment. It changes how easily non-specialists can perform technical tasks. It changes how deeply software platforms can automate business workflows. And it changes what defenders must monitor, govern and secure.

In this environment, the question is not whether every AI claim is true. The better question is which claims are becoming true fast enough to affect today's risk decisions.

That is where organisations need to focus. Not on hype. Not on denial. But on the shrinking distance between what AI can do now, what it may soon be able to do, and what the business is already allowing it to touch.

The New Attack Surface: Data, Models, Agents, Tools and Supply Chains

AI risk is often discussed as if it begins and ends with the model. That is too narrow.

The model is only one part of the AI system. The real attack surface includes the data used to train, tune and operate the system; the prompts and context that shape its behaviour; the retrieval systems that feed it knowledge; the tools and APIs it can call; the identities and permissions it uses; the memory it retains; the vendors and cloud services that support it; and the workflows that turn its output into action.

This is why AI security must be treated as an extension of enterprise security architecture, not as a standalone technology control.

The Australian Cyber Security Centre's guidance on agentic AI makes this point clearly. Agentic AI systems rely on tools, external data sources and memory bases to interact with their environment and expand their capabilities. Each of those components can introduce vulnerabilities across an interconnected attack surface. ACSC also warns that external data sources can enable indirect prompt injection attacks, and that broader access to computing infrastructure can allow malicious actors to exploit system components to execute scripts or send unauthorised emails.

This matters because agentic AI blurs boundaries that security teams are used to managing separately. Information flows between AI and non-AI systems. AI systems may read from one environment, reason over another, and act through a third. They may interact with business systems, development tools, security platforms, cloud APIs and third-party services. The more connected the AI system becomes, the harder it is to isolate AI-specific risk from broader cyber risk.

Data becomes a first-order security concern. AI systems depend on data for learning, context and operation. ACSC's AI data security guidance states that data security is critical across the AI lifecycle and that machine learning models learn their decision logic from data, meaning an attacker who can manipulate the data can also manipulate the logic of the AI system.

That changes how organisations should think about data protection. It is not only about preventing disclosure. It is also about preserving integrity. A sensitive dataset exposed to an AI system can create confidentiality risk. A poisoned or manipulated dataset can create decision risk. In AI systems, data quality, data authority and data lineage become security concerns.

The supply chain also expands. ACSC's AI and machine learning supply-chain guidance notes that AI/ML systems can improve efficiency, inform decisions, streamline processes and improve customer experience, but they introduce unique supply-chain risks when pre-trained models, third-party datasets and external components are used. It identifies the AI supply chain as including training data, models, software, infrastructure, hardware and third-party services.

This is a major shift for vendor assurance. It is no longer enough to ask whether a supplier has ISO 27001, penetration testing or encryption. Organisations need to understand whether AI features are being used, what models are involved, what data is processed, whether customer data is used for training, where data is stored, what third parties are involved, how model changes are controlled, how incidents are reported, and whether the supplier can provide meaningful transparency over AI-related risks.

The attack surface also includes trust itself. AI-generated content, synthetic media and automated decision support can erode confidence in what people see, hear and read. This has direct cybersecurity implications for executive impersonation, payment fraud, social engineering, incident communications, public trust and brand manipulation.

The new attack surface is therefore not simply "AI". It is the combination of AI with data, identity, tools, autonomy and trust. That combination is powerful, but it is also where the risk concentrates.

The practical implication is that AI security must be lifecycle-based. It must begin before procurement or deployment, continue through integration and operation, and remain active as models, data sources, workflows and vendor capabilities change. AI systems are not static assets. They evolve through data, context, configuration, permissions, user behaviour and supplier updates.

Why Boards Should Care: AI Risk Is Now Business Risk

AI risk is no longer a technical issue that can be delegated entirely to IT, security or data teams. It is becoming a board-level business risk because it affects productivity, competitiveness, resilience, fraud exposure, regulatory expectations, third-party dependence, customer trust and the organisation's ability to make reliable decisions.

The upside is significant. AI can improve speed, reduce manual effort, accelerate software development, strengthen security operations and allow smaller teams to produce outcomes that previously required larger budgets or specialist resources. Organisations that fail to adopt AI intelligently may lose operational leverage, market relevance and defensive advantage.

But the risk is equally material. AI can expose sensitive data, influence decisions, automate flawed workflows, amplify fraud, weaken accountability and create new pathways for compromise. When AI is connected to systems that affect money, customers, safety, operations or trust, the consequences are no longer confined to technology.

This is why regulators are paying attention. APRA has called for a step-change in AI-related risk management across banks, insurers and superannuation trustees, including clearer risk appetite, stronger oversight, accountability, timely patching and cyber hygiene. Although APRA's focus is financial services, the lesson applies more broadly: regulated sectors often reveal the control expectations that later become mainstream.

Boards should care because AI changes both risk and opportunity. Treating AI only as a threat will slow the business. Treating it only as innovation will increase exposure. The mature position is to treat AI as both: a source of competitive advantage and a source of material operational risk.

The board-level questions are therefore simple but important. Where is AI already being used? What data does it touch? What decisions does it influence? Which systems can it access? Are high-risk use cases owned, logged and governed? Do AI agents have distinct identities and privilege boundaries? Are critical workflows protected by human approval points? Are vendors contractually accountable for AI-related risks? Has the organisation tested plausible failure scenarios?

Most importantly, AI must be governed in proportion to business impact. A writing assistant and an autonomous finance agent should not be treated the same way. A customer-service summarisation tool and an AI-enabled operational technology workflow do not carry the same risk. A cyber triage assistant and an agent that can modify firewall rules or cloud configurations require different levels of control.

The board-level issue is not whether AI is safe or unsafe in the abstract.

The issue is whether the organisation understands where AI creates value, where it creates exposure, and whether the residual risk is acceptable after safeguards are applied.

Boards do not need to become AI engineers, but they do need enough literacy to challenge assumptions, demand visibility and ensure that AI adoption is deliberate rather than accidental.

The organisations that get this right will be those that move with discipline — capturing the productivity and defensive upside of AI while maintaining control over the systems, data and decisions that matter most.

The CyberStash Advisory Position

The CyberStash position is that the sensible response to AI is neither fear nor blind enthusiasm. It is controlled enablement: adopting AI deliberately, governing it proportionately, and securing the systems, data and decisions that matter most.

AI is now part of the cyber threat landscape, but it is also part of the defensive answer. It can help organisations identify vulnerabilities earlier, analyse threats faster, improve detection logic, summarise incidents, accelerate secure development, support compliance, reduce analyst fatigue and make expert knowledge more accessible.

The organisations that benefit most will not simply be those that adopt AI the fastest. They will be those that adopt it deliberately, govern it proportionately, and understand where the risk is worth taking. That is the leadership challenge. AI introduces risk, but avoiding AI also carries risk. Organisations that move too slowly may lose productivity, competitiveness, innovation capacity and defensive advantage. Organisations that move too quickly, without visibility or control, may embed powerful systems into critical workflows before they understand what those systems can access, influence or change.

The next phase of cybersecurity will reward organisations that can do five things well.

1. They can see where AI is being used.
2. They can identify the Crown Jewels that AI must not expose, manipulate or control without safeguards.
3. They can govern AI according to risk, not hype.
4. They can secure the data, models, tools, agents and integrations that make AI useful.
5. And they can maintain human accountability over decisions that materially affect security, safety, money, customers, operations or trust.

The impact of AI has only begun to scratch the surface.

The technology will become more capable, cheaper, more embedded and harder to separate from ordinary software. Its adoption will not be limited to technical specialists. AI will allow non-technical users, business teams, suppliers and adversaries to perform increasingly technical tasks with confidence and speed.

The strategic question for leaders is no longer whether AI will change cybersecurity. It already has.

The question is whether organisations can capture the value of AI while maintaining control over the systems, data and decisions that matter most.

Those that succeed will not be the ones that treat AI as a side project, a policy issue or a technology experiment. They will be the organisations that make AI governance part of the way they operate: visible, risk-based, secure by design, accountable and continuously improving. In an AI-accelerated threat landscape, resilience will not come from avoiding change. It will come from adapting faster, governing smarter and ensuring that the business remains in control as machines become more capable of acting on its behalf.

If AI has not forced an organisation to reconsider its project priorities, investment roadmap and cyber defence strategy, then it is probably being treated too narrowly. AI is not another tool to be absorbed into business as usual. It should change what leaders accelerate, what they pause, what they redesign and what they defend first.

In cybersecurity, that distinction matters. Defence strategies built for a slower, more human-limited threat environment must now be tested against a world where employees, suppliers, software platforms and adversaries can all use AI to move faster, scale further and perform increasingly technical tasks. The organisations that adapt early will not simply use AI more effectively; they will be better prepared to defend against those who misuse it.

About the Author

Loris Minassian

Founder & CEO, CyberStash

Loris Minassian is the Founder and CEO of CyberStash, an Australian-owned cybersecurity company focused on managed detection and response, cyber resilience, threat intelligence, exposure reduction and practical security advisory services.

With more than 25 years of experience across cybersecurity, technology leadership and enterprise risk, Loris works with organisations to strengthen their security posture, improve visibility, reduce exposure and prepare for a rapidly changing threat landscape.

His work focuses on the intersection of cybersecurity, artificial intelligence, operational resilience and business risk — helping leaders separate hype from reality and make security decisions that are practical, defensible and aligned to business outcomes.

> *AI will not remove the need for cybersecurity fundamentals. It will make weak fundamentals more expensive.*

CyberStash Advisory Services

CyberStash helps organisations assess AI-related cyber risk, identify critical exposure points, strengthen governance, secure high-risk workflows and improve defensive readiness in an AI-accelerated threat landscape.

CyberStash can assist with:

- AI cyber risk assessments
- Crown Jewels and exposure mapping
- Agentic AI governance reviews
- AI-enabled threat modelling
- Security architecture and control validation
- Managed detection, response and exposure reduction

cyberstash.com | info@cyberstash.com