

Privacy + Security Forum

Session:

AI Threats, Proven Defenses

Applying Existing Risk Management Strategies to the AI-Driven Attack Surface

Michael Borgia

Partner, Davis Wright Tremaine

Sabrina Guenther Frigo

Chief Ethics, Compliance &
Privacy Officer, TruStage

Tony DeSarro

Director, Cyber & Forensic
Services, KPMG

Kristy Hornland

Director, Cybersecurity, KPMG

Thursday, May 7, 2026 · 2:30 – 3:30 PM



Speakers



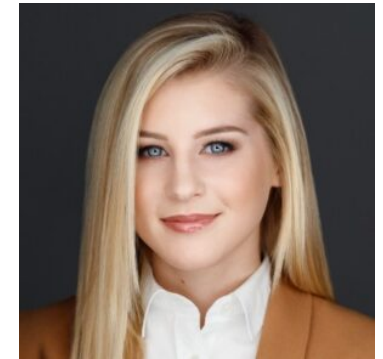
Michael Borgia
Partner,
Davis Wright Tremaine



Sabrina Guenther Frigo
Chief Ethics, Compliance &
Privacy Officer, TruStage



Tony DeSarro
Director, Cyber & Forensic
Services, KPMG



Kristy Hornland
Director, Cybersecurity, KPMG

Familiar threats, new vectors.

The threats are not new.

Social engineering, malware, and software vulnerabilities have driven cyber risk for decades.

The pace, scale, effectiveness and accessibility are.

Attacks that once took deep technical skill and weeks of work can be accomplished in minutes with free tools.

Your GRC program can be the right foundation.

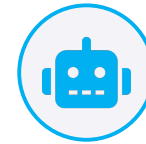
But your risk register and controls need immediate reexamination.

Why Now?: AI Security Threats Have Moved from Hypothetical to Real



Mythos & Glasswing

Anthropic's Claude Mythos Preview — rapidly discovered thousands of zero-days. Project Glasswing provides the model to 12 launch partners and 40+ critical-software organizations to find and patch first.



Agentic AI in production

Companies are deploying AI agents that can act autonomously across email, code, and cloud. Each is a new non-human identity — and a new insider-threat surface. The OWASP Top 10 for Agentic Applications launched in 2026.



AI-orchestrated attacks in the wild

GTG-1002 (Nov. 2025): Chinese state-sponsored campaign in which an AI agent performed 80–90% of operations across ~30 targets. Mexican government breach: a small team using off-the-shelf models exfiltrated 195M taxpayer records.



Social engineering at scale

Nation-state actors (DPRK, Iran, PRC, Russia) using LLMs for hyper-personalized phishing, rapport-building, and reconnaissance. Voice clones from seconds of audio. Deepfake video calls — Arup (\$25M, 2024).

Session Roadmap

01

Social engineering at scale

Deepfakes, voice cloning, and AI-personalized phishing.

02

Automated vulnerability discovery & exploitation

Attack automation and vulnerability exploitation at scale.

03

Agentic AI in the enterprise

Over-permissioned agents, prompt injection, data exfiltration, and the new insider-threat surface.

04

Key governance responses

Interlocking risk programs, board-level reporting, and demonstrable oversight.

Part One

Social Engineering at Scale

Rethinking verification controls when eyes and ears can deceive.





Threat - Social Engineering: Deepfake impersonation in real time.

ARUP, HONG KONG

\$25M wire fraud after a deepfake video conference call

A finance employee suspected the initial email was phishing. He dropped his suspicion after a video call with what appeared to be the CFO and several colleagues — all AI-generated.

15 transfers, five accounts, ~\$25.6M total.

WHAT'S NEW

- **The call itself can be synthetic.**
Multi-party video meetings are no longer an out-of-band check — every face on the call may be AI-generated.
- **Source material is everywhere.**
Earnings calls, conference talks, and internal town halls give attackers training data for any executive's voice and likeness.
- **Cloning from just seconds of audio/video**
A short clip of an executive on a podcast or earnings call is enough to clone their voice for a vishing call.



Threat - Social Engineering: *AI-scaled phishing and deepfakes.*



Personalized phishing at scale

LLMs draft fluent, context-aware messages that reference real internal projects scraped from public sources — at near-zero marginal cost per target.



Traditional “tells” are disappearing

AI helps eliminate traditional signs of fraudulent messages— e.g., poor grammar, awkward syntax, lack of cultural context.



Attackers’ Cost Curve has Flipped

What were once high-effort, high-reward attacks (whaling, BEC) now run as automated campaigns.



Skill barrier has collapsed

Open-source tools and underground services put convincing impersonation in non-technical hands.



GRC Response - Risk Assessment: Re-rating likelihood and impact.

$$\text{RISK} = \text{Threat Likelihood} \times \text{Threat Impact}$$

First likelihood, then impact.

The probability that any given employee will face a convincing impersonation has significantly increased. Changes to impact are less clear — but are likely, especially with agentic AI.

Identify control assumptions.

Does a workflow assume that voice or face = identity? Treat each of those as a risk to be re-rated, not a known good. Volumetric and skill assumptions also no longer hold. Whaling no longer may be rare or take much effort.

Map communication-triggered actions.

Wires, credential resets, vendor bank-detail changes, system access grants — anywhere a request leads to action without a second channel.

A SIMPLE TEST

Pick the three highest-impact financial workflows.

For each, ask:

If every voice, image, and email in the workflow were fabricated — would the controls still stop the loss?

Anywhere the answer is no, you have a control gap that is now actively exploitable.



GRC Response - Controls: Controls that can address deepfake risk.



Multi-channel verification

Above defined dollar/sensitivity thresholds, require callback to a known number — never a number provided in the request itself.



Updated training

Move beyond email phishing. Include deepfake video, real-time voice clones, urgency cues, and pretexts that exploit recent public events.



Payment authorization workflow

Dual approval, time-delay holds for new payees, and out-of-band confirmation for vendor bank-detail changes — by policy, not exception.



Documented escalation

A clear, drilled path for any employee who suspects impersonation — including permission to delay execution without penalty.



Executive buy-in

Controls fail the moment a senior officer demands a 'quick' wire bypass. The CEO and CFO must publicly back the process — and accept friction. Exceptions must be rare and defined.



Vendor verification

Make multi-channel verification a contractual expectation, not just an internal practice. Brief critical vendors on your callback procedures so they don't push back when you slow down to verify

Part Two

Automated Vulnerability Discovery & Exploitation

When the discovery-to-exploit window collapses to hours, defensive timelines must follow.





Threat - Hacking With Frontier Models: It is not theoretical anymore.

GTG-1002 (NOV. 2025)

First documented AI-orchestrated cyber-espionage campaign

Chinese state-sponsored group jailbroke Claude Code via a 'defensive testing' pretext.

AI agent performed 80–90% of the operation: reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, and exfiltration.

~30 high-value targets across tech, finance, government, and manufacturing. Thousands of requests per second.

MEXICAN GOVT. AGENCIES (DEC. 2025 – FEB. 2026)

A small team. Off-the-shelf models. Nine agencies breached.

Likely fewer than five attackers. Used Claude Code and GPT-4.1 — not frontier cyber-specialty models.

Bypassed safety filters with a 'bug bounty' pretext in ~40 minutes; AI then generated 20 tailored exploits for 20 CVEs and executed ~75% of remote commands.

150 GB exfiltrated, incl. 195M taxpayer records — using ordinary, public AI tools.



Threat - Hacking With Frontier Models: It's more than Mythos.

CLAUDE MYTHOS PREVIEW

Mythos capabilities. Autonomous discovery and exploitation of memory-corruption vulnerabilities in production code, including a 17-year-old FreeBSD remote-code-execution bug; a 27-year-old OpenBSD signed-integer overflow; chained bugs to escalate to full system compromise.

Reported scale. Anthropic has reported that thousands of zero-days surfaced across major operating systems, browsers, and infrastructure code. Anthropic withheld the model from public release.

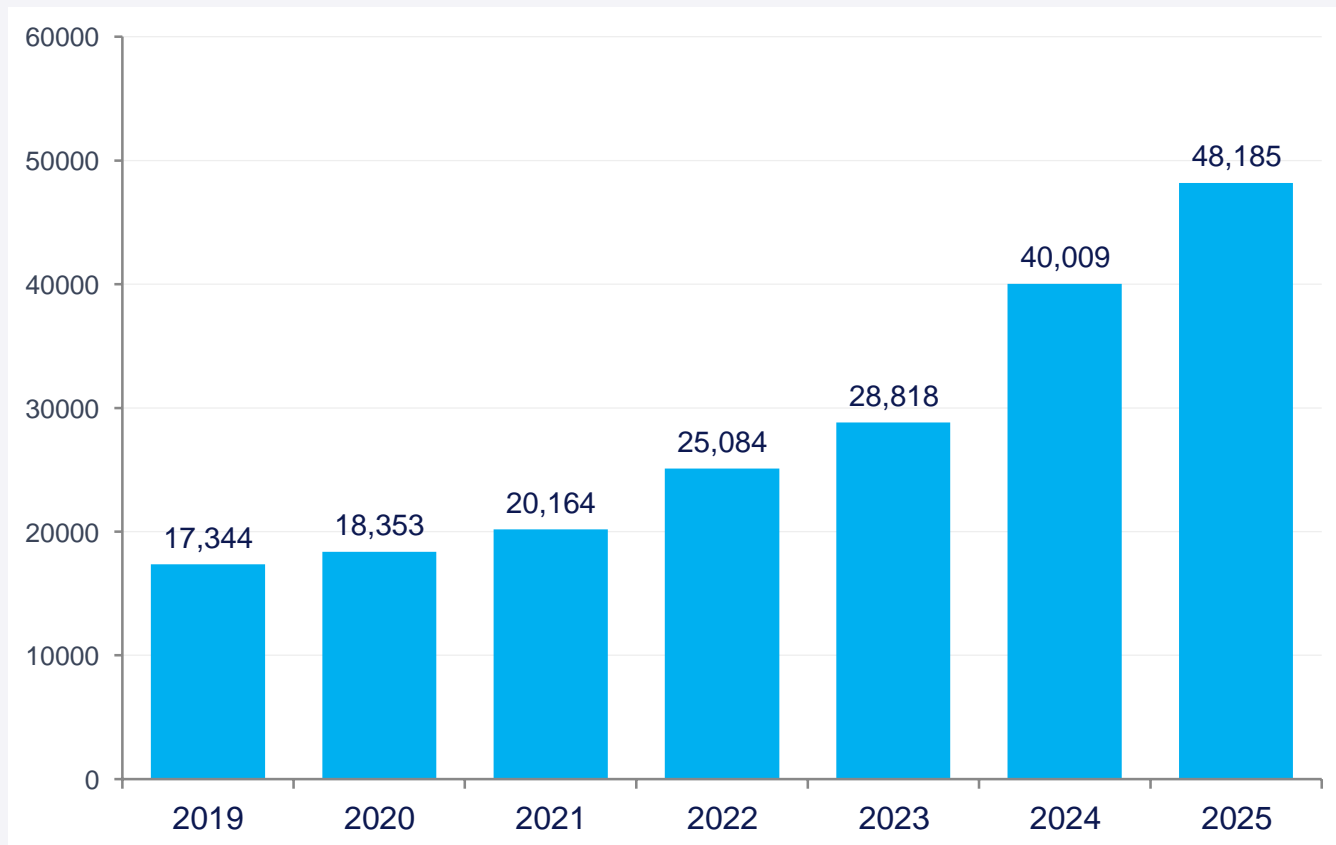
Project Glasswing. Limited release to 12 launch partners (AWS, Apple, Google, Microsoft, Cisco, CrowdStrike, JPMorganChase, NVIDIA, Palo Alto, Linux Foundation, Broadcom, Anthropic) plus 40+ critical-software organizations — to find and patch before adversaries replicate the capability. \$100M in usage credits committed.

THIS IS NOT JUST A MYTHOS ISSUE

- **OpenAI has its own cyber-focused models.**
GPT-5.4-Cyber is a cyber-permissive variant under OpenAI's Trusted Access for Cyber program. GPT-5.5 was rated 'High' cyber capability under the Preparedness Framework (April 2026).
- **Smaller open models reproduce the same finds.**
AISLE's 'Jagged Frontier' analysis: 8/8 cheap, publicly available models (including one 100x cheaper than Mythos) detected Mythos's flagship FreeBSD exploit and others. Developing real exploits is still where frontier models lead—but the gap is closing.
- **Non-cyber models on real attacks.**
GTG-1002 used jailbroken Claude Code; the Mexican government breach used Claude Code and GPT-4.1 — not frontier cyber-specialty models. Off-the-shelf models are already enabling real campaigns.



Threat - Hacking With Frontier Models: Companies already are downing in software bugs.



Sources: CVE.org / NVD; JerryGamblin year-end CVE Data Reviews 2024–2025. Published CVEs nearly tripled (~17K → ~48K) in six years.

DEFENDER TAKEAWAYS

- **The flood is real.**
~108 CVEs published per day in 2024;
>130/day in 2025. Triage at scale is now its own discipline.
- **CVSS alone doesn't capture the risks.**
Fewer than 1% of CVEs are ever exploited. Confirmed exploitation of new CVSS 7-10 CVEs nonetheless jumped 105% YoY (71 → 146).
- **The patching window is collapsing.**
Some CISA emergency directives now have 24-hour patching deadlines (e.g., CitrixBleed 2, July 2025).



GRC Response - Risk Assessment: Higher likelihood, higher impact.

$$\text{RISK} = \text{Threat Likelihood} \times \text{Threat Impact}$$

Threats are familiar; the math has changed.

The vulnerability classes haven't shifted — but AI compresses time-to-exploit and lowers the skill floor. Likelihood and impact both rise.

Test your control assumptions.

Re-examine 'low-risk because not exploited in the wild' — that assumption breaks when AI can weaponize a CVE in hours.

Use SBOMs to understand where your risks are.

Reassess patch-related downtime risk.

Emergency patching disrupts operations. The risk of NOT patching may now exceed the operational cost of the maintenance window.

A SIMPLE TEST

Pick the three internet-exposed systems with the highest blast radius.

For each, ask:

If a CVSS 9.0 RCE drops with public exploit code on a Friday at 5pm — could you patch or isolate it before Monday morning?

The assumption that you can patch within standard SLAs may no longer a defensible risk treatment for some vulnerabilities.



GRC Response - Controls: Triage, tooling, and resiliency.



Prioritize based on exploitability

Use CISA KEV, EPSS, and threat intel to triage what attackers are actually using — not what scores 7+.



Internet-exposed first, always

External attack-surface management is no longer optional. Anything externally reachable goes on a tighter SLA than the rest of the estate.



AI-assisted code & dependency review

Use AI tooling (Codex Security, Claude Code with sandboxing, commercial scanners) to vet third-party code, container images, and OSS at procurement and major upgrades.



Software inventory and SBOMs.

Maintain an asset inventory and SBOM for every critical system. When the next mass-exploitation CVE drops, the question "are we exposed?" needs an answer in minutes, not days.



Resiliency: identify critical processes

Map the business processes you cannot afford to lose. Pre-plan response to attacks AND to emergency patching that disrupts operations.



Add patching obligations to vendor contracts

Most of your patchable surface is software you didn't write. Set vendor patch SLAs in contracts — disclosure-to-patch timelines, notice obligations, etc.

Part Three

Agentic AI in the Enterprise

When the AI you deployed becomes the new attack surface — and the new insider.



Primer: What is agentic AI?

An AI agent is a system that — given a goal — **plans, decides, and acts** by calling tools and APIs, often across many steps and without human approval at each step.



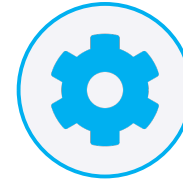
Plan

The agent breaks a goal into sub-tasks and chooses an order of operations. No script — the path is generated at runtime.



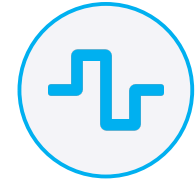
Decide

Given tool results, the agent picks the next step. It may revise its plan based on what it observes.



Act

The agent calls real tools — APIs, code, email, filesystems, cloud resources. Each call inherits a real identity and real permissions.



Persist

Memory, vector stores, and context windows let the agent carry information across turns and sessions — including malicious content from earlier inputs.

Why this changes the threat model. An agent is a non-human identity that runs at machine speed, takes instructions literally, and can make many decisions in seconds. When compromised, it acts as a malicious insider — only faster.



Threat - Agentic AI: The agent is the new attack surface.

OWASP Top 10 for Agentic Applications (2026): *agents that plan, decide, call tools, and act inherit user privileges — and create new failure modes that don't map to traditional appsec.*



Over-permissioning

Agents granted broad tool/credential scope. When compromised, the attacker inherits that scope — read filesystem, send email, push code.



Prompt injection

Hidden instructions in emails, documents, web pages, or tool outputs hijack the agent's goal. The agent looks 'on task' while serving the attacker.



Data exfiltration

An agent with access to email, CRM, repos, or cloud APIs can be steered to leak sensitive data through tool calls or crafted outputs.



Insider-threat surface

Agents are non-human identities running at machine speed. A compromised or misaligned agent is functionally a malicious insider.



Tool & supply-chain abuse

Plugins, MCP servers, and agent registries are an evolving software supply chain. Compromise of any one reaches every agent that uses it.



Memory poisoning

Persistent vector stores and conversation memory carry adversarial content from session to session — a long-tail risk that classic scans miss.



GRC Response - Risk Assessment: A mix of new and emerging threats.

$$\text{RISK} = \text{Threat Likelihood} \times \text{Threat Impact}$$

A mix of familiar and new.

Over-permissioning is an old problem. Agents that take instructions literally, run at machine speed, and chain tool calls are new — both likelihood AND impact rise.

Same standards, different controls.

Machine and human identities require the same governance. But how the controls actually function (review cycles, MFA, anomaly detection) is different for agents.

Test your control assumptions.

Risk assumptions built around human-paced mistakes break down at agent speed. A misconfiguration that would cause one bad action by a person can cause a thousand bad actions by an agent — before anyone notices.

STARTING POINTS

Build an agent registry. Risk depends on where the agent runs and what it can do. You cannot risk-rate what you have not catalogued.

Agents are non-human identities. Onboard each into IAM with a documented owner, scope, and lifecycle — the same as a service account, not a feature toggle.

Do agents need a separate, dedicated risk assessment — or can your existing process handle them with extensions?



GRC Response - Controls: Controls that keep agents on a leash.



Just-in-time, scoped permissions

Per-tool scoping (read-only by default), short-lived credentials, no wildcard tokens. Privilege escalates only with explicit, audited approval.



Human-in-the-loop for high-impact actions

Wires, deletions, deployments, external sends, and customer-data access require human approval — by policy, not configuration.



Continuous monitoring for anomaly

Treat agent activity as you would a privileged service account. Alert on impossible velocity, scope creep, and unusual tool chaining.



Sandboxing and isolation

Agents that generate or execute code run in isolated sandboxes with no access to production secrets or shared filesystems.



Input/output content filtering

Treat all retrieved content (emails, web pages, RAG documents) as untrusted. Filter for known prompt-injection patterns; never blindly pass LLM output to a tool.



Governance and shadow-AI

Approval workflow for new agents, retirement process for old ones, and a clear policy on what employees may deploy. Shadow agents are insiders.

Part Four

Engaging the Board and Senior Leadership

Translating AI security risk into the questions, decisions, and documentation a board can act on.



Cross-Cutting - Board Engagement: What the board needs to hear.

Boards aren't asking "is the firewall working?" They're asking questions like: Are we managing AI-driven risk? Are we managing it well enough relative to peers and regulators' expectations? How could we prove it if something goes wrong?



A single risk picture, not two

- **AI risk and cyber risk reported together.**
Standing cyber updates cover AI-driven threats and the controls responding to them — not as a separate 'AI update' competing for board time.
- **Forward-looking signals.**
Vendor-side patch velocity, agent inventory growth, threat-landscape shifts the board should see before they show up in an incident.
- **Plain-language KPIs.**
Time-to-patch on internet-exposed assets; agent registry completeness; drill participation — metrics a non-technical director can read and challenge.



Decisions, in writing, before they matter

- **Risk appetite, explicit on AI.**
The Cyber Risk Appetite Statement names AI-driven exposure — including agentic deployments — and where the organization tolerates, mitigates, or avoids.
- **Named owners and escalation paths.**
Every material AI/cyber risk has a single accountable owner. Who decides to disconnect, escalate to outside counsel, or notify regulators is documented before the call has to be made.
- **Established trigger points.**
Understand and agree on the types of threats and degrees to risk that must be escalated to the Board.



The record, designed for the worst day

- **Minutes that show oversight.**
Cyber updates document the questions asked, answers received, and decisions made. Caremark and its progeny demand the record exists.
- **Decision logs for material calls.**
Option set, recommendation, rationale, who decided, when. Sufficient to defend in litigation or to a regulator.
- **Dashboards in plain language.**
Color codes that mean the same thing every quarter. Trend lines a director can read without a translator.

Takeaways

Social engineering

- Mandate multi-channel verification above defined dollar/sensitivity thresholds.
- Update training to cover deepfake video and real-time voice cloning.
- Build dual-approval and cooling-off periods for new payees / bank-detail changes.
- Secure visible CEO/CFO buy-in to the friction these controls create.

Vulnerability response

- Move from CVSS to exploitability (CISA KEV, EPSS) for triage.
- Put internet-exposed assets on a separate, tighter SLA.
- Pre-plan compensating controls (WAF virtual patching, segmentation) for emergency CVEs.
- Map critical processes; pre-decide what disruption you can tolerate from emergency patching.

Agentic AI

- Build (or start) an agent registry with documented owner, scope, and lifecycle.
- Apply just-in-time, scoped permissions and human-in-the-loop for high-impact actions.
- Adopt a shadow-AI policy and approval workflow for new agents.
- Decide whether your existing risk assessment process handles agents — or needs an extension.

Governance & integration

- Refresh vendor due-diligence questionnaires and contractual reps for AI.
- Run one tabletop this quarter with an AI-driven scenario.
- Add AI risk to standing board cyber updates; document questions, answers, decisions.
- Update the Cyber Risk Appetite Statement to address AI exposure explicitly.

Questions?

Michael Borgia

Davis Wright Tremaine

Sabrina Guenther Frigo

TruStage

Tony DeSarro

KPMG

Kristy Hornland

KPMG

