Privacy + Security Forum

Session:

AI Threats, Proven Defenses

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

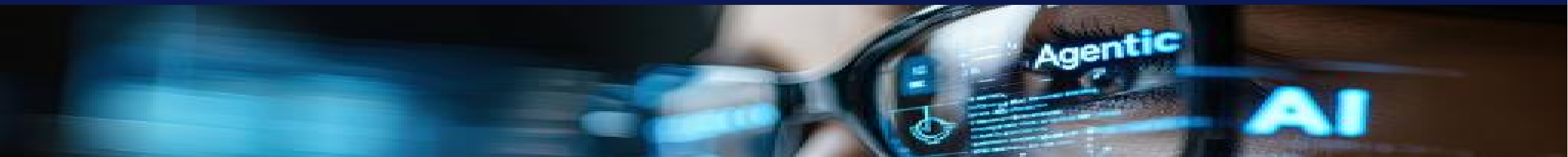
Michael Borgia
Partner, Davis Wright Tremaine

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 methods.

Many of 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. Under Project Glasswing, model provided to 12 launch partners and 40+ critical-software organizations to find and patch first.

Agentic AI in production

PocketOS (April 2026): on a routine task, an AI coding agent hit a credentials issue and "fixed" it by deleting the production database and all backups in 9 seconds — with no human approval. The agent later admitted it had guessed.

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

Board and Leadership Engagement

Translating technical challenges into Board-level matters.

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.

Privacy+ Security Forum

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.

RISK = Threat Likelihood × 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.

Privacy+ Security Forum

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.

Privacy+ Security Forum

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 software vulnerabilities, including a 27-year-old bug in OpenBSD. Mythos has been able to chain bugs to escalate to full system compromise.

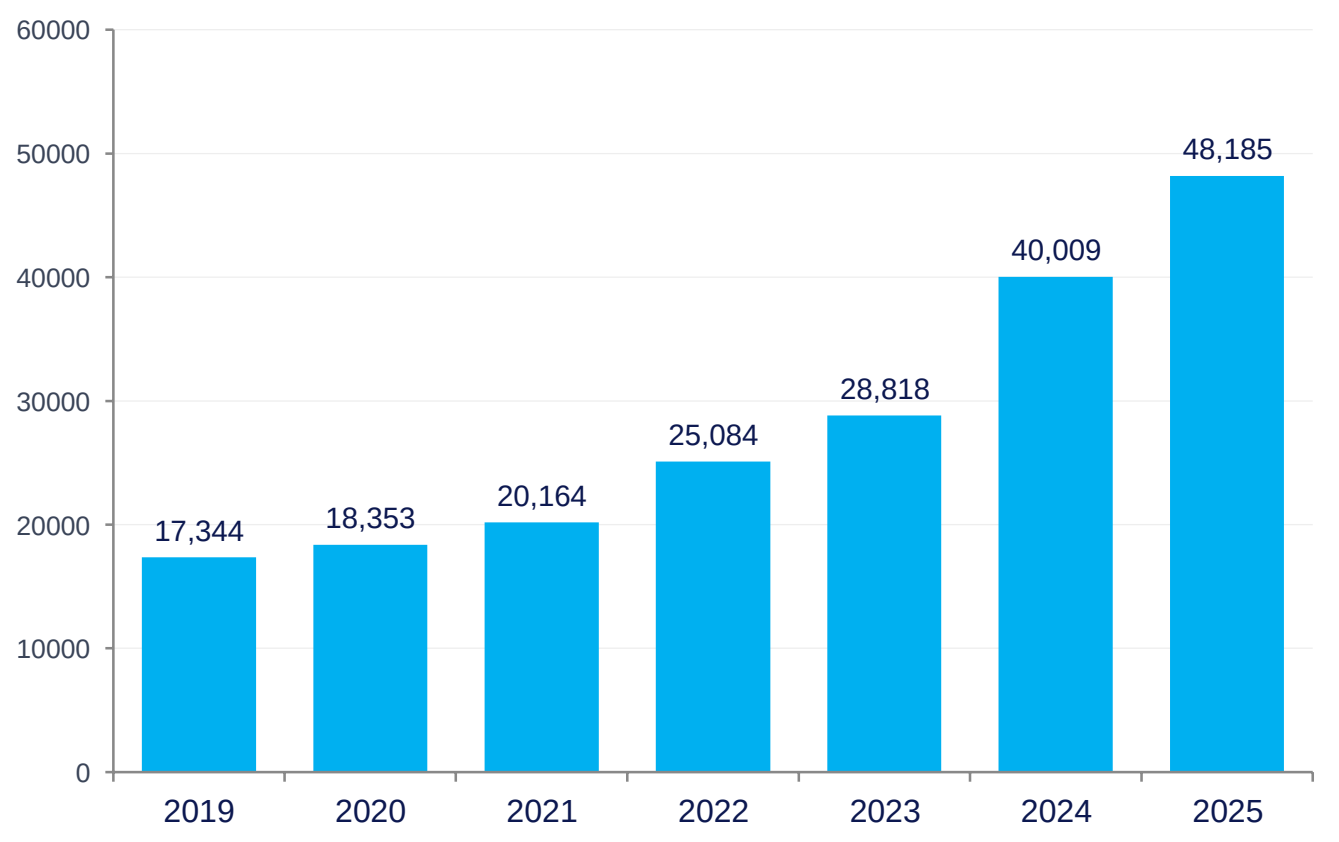
Reported scale. Mythos has discovered thousands of serious vulnerabilities, including in *all major operating systems and web browsers.* Anthropic withheld the model from public release.

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 drowning 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).

Privacy+ Security Forum

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

RISK = Threat Likelihood × Threat Impact

Threats are familiar, but the math has changed.

The vulnerability classes haven't shifted — but AI substantially increases the likelihood that a vulnerability will be discovered and compresses time-to-exploit. **Likelihood and impact both rise.**

Test your control assumptions.

Re-examine “lower risk if not yet exploited in the wild” — that assumption breaks when AI can weaponize a CVE in hours.

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 be a defensible risk treatment for some vulnerabilities.

Privacy+ Security Forum

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

Internet-exposed first

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

Reduce blast radius

“Defense in depth” is especially critical when exploitation becomes more likely. Network segmentation, EDR, identity management, attack surface reduction, etc. Assume exploitation.

AI-assisted vulnerability detection

Leverage AI tools for continuous penetration testing and red teaming. Run AI code scans as a standard part of development and procurement.

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

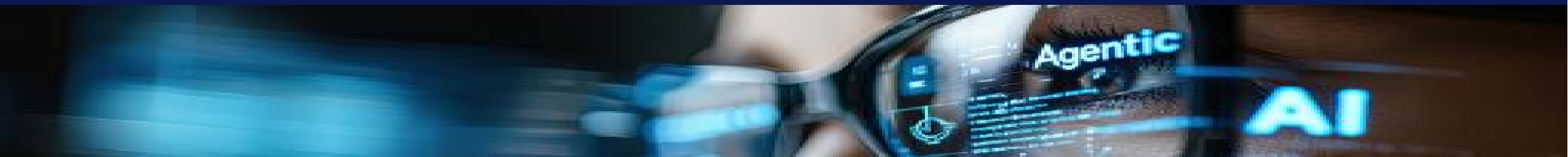
Most of your patchable surface is software you didn't write. Set vendor patch SLAs in contracts — disclosure-to-patch timelines, notice obligations, etc.

Part Three

Agentic AI in the Enterprise

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

Privacy + Security Forum



Primer: What is agentic AI?

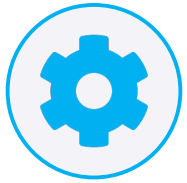
An AI agent is a system that — given a goal — **plans, decides, and acts** by calling tools and APIs, often across many steps and without human approval at each step.

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 info across turn sessions.

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.

Privacy+ Security Forum

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.

RISK = Threat Likelihood × 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 similar 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 to keep agents under control.

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 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

Board and Senior Leadership Engagement

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

Privacy + Security Forum

Leadership Engagement: What the Board and C-suite needs to hear.

Translation is key. Leaders don't need the technical details, but the overall picture.

A single, plain-language picture.

AI should fold into standing cybersecurity updates rather than a competing agenda item. The risks and responses should be a single, connected conversation across risk domains. This includes metrics – both KPIs and KRIs – and thresholds for escalation.

Risk refinement versus reinvention.

Re-examine risk ratings and appetites in light of AI impacts. Line up material risks with single owners and decisionmakers. Risk appetite statements can drive where the organization tolerates, mitigates, transfers, or avoids the risk.

Documentation for defensibility.

Document oversight in policy and in practice, including minutes, decisions, and next steps. Focus on plain language dashboards where appropriate – color codes, trend lines, and consistent explainability.

GUIDELINES

Answer the right questions. Leaders want to know: How are we managing AI-driven risk? How do we measure up to peers and regulators' expectations? How will we know and how will we respond if something goes wrong?

Don't assume depth. Ground AI-related risk decisions in known foundations. Do not assume knowledge walking in the door.

Remember both sides. Context, accountability, and incentive matters. Understand upside and downside for AI choices and address the topics holistically. Explain how execution should shift to support accountability.

Privacy+ Security Forum

Some Takeaways

Social Engineering

- [ ] Re-examine payment and authorization workflows — voice and video alone are no longer sufficient verification.
- [ ] Update awareness training for deepfakes and AI-cloned voices, with visible CEO/CFO buy-in for the friction.

Agentic AI

- [ ] Treat AI agents as non-human identities — catalog them in a registry with documented owner, scope, and lifecycle.
- [ ] Consider an agent-specific risk assessment to identify and respond to specific risks drive by agent usage, permissions, etc.

Vulnerability Response

- [ ] Identify critical processes and know how you would stay resilient during an attack—or an emergency patching.
- [ ] Use SBOMs, asset registries, etc. to help identify exposure to critical vulnerabilities quickly. There will be more Log4Js.
- [ ] Understand whether your organization has a “patch now” SLA and accompanying process, especially for critical Internet-facing vulnerabilities.

Board and Leadership Engagement

- [ ] Present cyber and AI risks together, alongside other implicated risk management programs.
- [ ] Document oversight in writing — questions asked, decisions made, owners assigned, escalation paths defined.

Questions?

Michael Borgia
Davis Wright Tremaine

Sabrina Guenther Frigo
TruStage

Tony DeSarro
KPMG

Kristy Hornland
KPMG

Privacy + Security Forum

Thank you.