AI SECURITY COUNCIL

Guardrails for Autonomous AI in Cyber Defense

Practitioner Insights from the AI Security Council

The Importance of Keeping Autonomous AI in Their Lane

Autonomous AI is reshaping the daily rhythm of cyber defense. Just as teams were growing accustomed to rigid automation and scripted actions, they're now adapting to systems that can analyze context, weigh options, and take limited action on their own. This shift brings clear advantages, including faster detection, reduced manual work, and operations that can keep pace with modern threats.

These benefits come with new responsibilities. Security leaders should be determining how far these systems should be allowed to act, who remains accountable when they make the wrong call, and what guardrails ensure safe operation without slowing progress. Autonomy can't be granted by default. It has to be introduced gradually as systems demonstrate reliability, safety, and alignment with the organization's risk tolerance.

AI should enhance the team's speed and precision while preserving stability and trust. This paper outlines the operational boundaries, accountability models, and oversight structures needed to adopt autonomous AI responsibly, enabling defenders to work faster and more confidently with clear controls in place.

This paper reflects the collective perspectives of 19 CISOs and security leaders who participated in the AI Security Council's second workshop on autonomous AI security.

Participating AI Security Council Members

Assi Ungar
Group CISO,
bolttech

Ben Ramduny
CISO
Eni Energy Netherlands
(Formally Neptune Energy)

Greg McCord
CISO
Lightcast

John Sapp
CISO
Texas Mutual

Matthew Meersman
CIO
ZTA Baltimore

Fred Descloux
Director of Data Privacy, Data Security and GRC
Nu Skin

Madhul Sachdeva
Founder, AI Security Specialist
HonestLabs

Matt Cerny
Director Cyber Security
Integra LifeSciences

Sebastian Kinnaird
Head of Information Security
Centrica Energy

Yousef Syed
Director of Information Security
Bayvision Limited

Andy Lenzen
Senior Manager
Upwork Inc.

Bradley Heath
Director of Cybersecurity Operations
SysArc

David Bosomworth
IT Director
Oakley Capital

Emilio Mazzon
Director CSA, Global Information Security and IT
SNC-Lavalin / Atkins Global

Jaya Agnihotri
Head of Cyber Assurance
ICBC Standard

Matthew King
Head of Security and Compliance
dais

Richard Marsden
Cybersecurity Architect
TDSynnex

Rob Emmerson
Global Head of Security Architecture
Delinian

Siva Inguva
Head of SaaS Security
PTC

From Automation to Autonomy

AI is moving beyond scripted workflows and simple rule execution, and this shift is most clearly felt in how teams are using it. As **John Sapp** noted, ***"Autonomy should start small, containment, or enrichment. Configuration changes still need people."*** That framing reflects where most organizations are today. They are comfortable letting AI handle tightly scoped, low-impact tasks, but they are not ready to hand over the keys to the environment. Nor should they.

> "Push AI to the edge of safety, but do not remove the guardrails."
>
> — **Assi Ungar**

Autonomy introduces a new separation between actions that AI can perform and decisions that still require human approval. **Matthew Meersman** described this as a change in how we think about control itself. ***"AI will change how we think about control, not as restriction but as resilience."*** In this model, autonomy grows only when the supporting processes, validation steps, and rollback plans reach the maturity needed to absorb mistakes.

Assi Ungar reinforced the speed pressure that makes autonomy attractive. Teams want AI to operate close to the edge of what is safe. ***"Push AI to the edge of safety, but do not remove the guardrails."*** The aim is not full independence. It is safe acceleration. AI is allowed to help where the blast radius is small and where its actions can be undone without business disruption.

Fred Descloux captured the mindset required to scale responsibly. ***"Treat autonomy like privilege escalation. It should be requested, not assumed."*** Autonomy has to be justified, tested, and monitored. It earns its place by proving reliability, not by being available.

Guidance: Begin with small, low-impact forms of autonomy. Automate what can be reversed, monitor every action, and expand autonomy only after consistent evidence of safety and reliability.

Defining the Blast Radius

The clearest boundary for autonomous AI is not the task it performs but the impact it can have if it chooses incorrectly. **Ben Ramduny** summarized this principle directly. ***"If the system can't back out of a bad decision, it's not ready to run alone."*** Teams across the workshops echoed this view. Reversible actions such as isolating a host, blocking a domain, or disabling an account are strong early candidates for autonomy because mistakes can be corrected quickly.

> "Speed is critical, but chaos isn't security. Guardrails let us move fast safely."
>
> — **Sebastian Kinnaird**

That line changes the moment an action can alter configurations or disrupt production. **Matt Cerny** captured the distinction well. ***"We separate reversible actions from the rest. If it can break production, it needs eyes on it."*** This reflects the operational reality most teams face. Containment steps often have a small and predictable blast radius. Configuration changes do not.

Speed remains a pressure point, but speed without safety creates new risks. **Sebastian Kinnaird** put it plainly. ***"Speed is critical, but chaos isn't security. Guardrails let us move fast safely."*** Proper scoping lets teams leverage AI when the impact is low, while keeping humans in control when the impact is significant.

Rollback capability is the non-negotiable requirement that determines whether autonomy is allowed at all. **Yousef Syed** emphasized this with a comparison that resonated with every attendee. ***"Autonomy without rollback is like change control without versioning."*** If an autonomous action cannot be undone cleanly, it has not earned the right to be autonomous.

Guidance: Scale autonomy by impact tier. Start with reversible containment actions, validate with metrics, and keep rollback plans mandatory.

Accountability and Liability

AI can assist and accelerate decision-making, but responsibility for those decisions does not transfer to an algorithm. Every practitioner in this workshop aligned on a single principle. Accountability stays with the enterprise. Autonomy only works when governance, ownership, and liability are clearly defined.

> "Ownership doesn't disappear because the action was automated."
>
> — **David Bosomworth**

Greg McCord framed it directly when he stated that ***"risk lives where the business owns the decision."*** CISOs identify and classify risks, but the owners must accept them. Automated action does not change that structure. It only underscores the importance of documenting the chain of accountability.

This puts pressure on contracts and vendor relationships. **Rob Emmerson** warned that ***"no one should assume a vendor carries liability for model error."*** If the business expects a vendor to share responsibility, it must be written into the statement of work or legal contract. Without explicit terms, all fault reverts to the customer.

Internal ownership also matters. **David Bosomworth** said that ***"ownership doesn't disappear because the action was automated."*** Systems can execute on your behalf, but they cannot inherit your responsibility. When the impact hits production, the burden remains human.

The same reality drove **John Sapp's** view. Accountability still lands on his desk when something breaks, which is why decision chains, logs, and review processes must remain clear. AI can propose and act. Humans remain responsible for understanding and approving the risk posture that allows it.

Together, these perspectives reinforce a simple rule. Autonomy is not a shield. It is an operational choice that still relies on explicit human authority.

Guidance: Assign clear accountability before enabling autonomy. Require documented risk acceptance for every autonomous workflow. Embed liability and model error clauses in every vendor contract.

Human-in-the-Loop as Doctrine

As autonomy grows, the analyst's role doesn't shrink. It changes shape. AI automates repetitive work that slows teams down, but judgment, interpretation, and risk acceptance remain human responsibilities. Every participant reinforced this point. Human-in-the-loop (HITL) is not a safeguard to bolt on later. It is the operating model that makes autonomous systems usable and trustworthy.

Andy Lenzen put the foundation in clear terms. AI can summarize and suggest, but analysts must act. Autonomy may filter noise, enrich signals, or prepare response steps, but the final decision belongs to a person who understands context, mission, and consequences.

That is why **Bradley Heath** stressed that accountability cannot be outsourced. ***"Human-in-the-loop is how teams preserve integrity and prevent automation from drifting into decisions that exceed its mandate."*** Even when AI performs flawlessly, someone must own the rationale behind each action.

Ethical judgment reinforces the need for this structure. **Jaya Agnihotri** reminded the group that machines execute policy, while people enforce ethics. AI can follow patterns, but it cannot interpret cultural risk, customer impact, or regulatory nuance. Human oversight keeps those boundaries intact.

Matthew King captured the balance that teams are striving for. Human-in-the-loop is the line between speed and safety. AI accelerates every step of triage and response, but human review ensures the outcome aligns with business risk and accepted controls.

HITL is not a bottleneck. It is the mechanism that lets autonomy scale responsibly.

Human-in-the-loop is how teams preserve integrity and prevent automation from drifting into decisions that exceed its mandate."

— **Bradley Heath**

Guidance: Design every autonomous process with human checkpoints. Build review, override, and rollback into normal operations.

Guardrails That Cannot Be Removed

As AI systems take on more responsibility, they'll inherit the privileges and access paths that used to belong only to senior engineers and core security platforms. That shift raises the stakes. Autonomous agents are not workflow helpers. They are privileged systems that must be protected with the same rigor used for identity providers, SIEM pipelines, and production control planes.

The foundation is visibility and traceability. **Emilio Mazzon** emphasized that every AI action should have an audit trail that you can't edit. Logs that can be modified or erased undermine trust, break accountability, and make incident response nearly impossible. Immutable records aren't optional. They're the baseline.

Stronger autonomy requires stronger scaffolding.

Integrity risks extend beyond logging. **Richard Marsden** reminded the group that bias and drift are slow failures that are harder to detect but just as dangerous. Models can shift gradually as data changes, teams evolve, or threat patterns adapt. Without continuous testing and validation, autonomy can slide into error without any obvious signal.

These guardrails define the minimum standard.

Environment design also matters. **Siva Inguva** highlighted the need to segment AI systems just as strictly as any production workload. Development environments are not production environments. The data, permissions, and impact profiles are completely different. Mixing them increases the risk of exposure, poisoning, and uncontrolled behavior.

These controls are not meant to slow teams down. They are what allow autonomy to operate safely at scale. As **Assi Ungar** noted, guardrails are what make autonomy sustainable, not restrictive. When controls are strong, teams can experiment, iterate, and deploy faster because the boundaries keep the system safe.

Guidance: Enforce least privilege, immutable logs, and firm separation of duties. Protect AI systems with the same rigor you use for your crown-jewel assets. The integrity of the system has to match the sensitivity of the data it touches.

Board Priorities for 2026

Boards aren't asking whether AI belongs in cybersecurity anymore. They're asking what has to be in place before autonomous systems can operate safely. Across the council, the strongest message was simple. Education and governance come first. Automation comes after. Boards are beginning to evaluate autonomous systems against three core pillars: trustworthiness, transparency and auditability. These criteria shape how quickly leaders are willing to authorize autonomy and what safeguards must be demonstrated first. Readiness also includes training executive teams to interpret AI-related risks with the same fluency as traditional ones.

"Don't chase tools, invest in readiness. The goal is not faster adoption, it is building trust and governance that scale with autonomy."

— Fred Descloux

Fred Descloux captured the tone directly when he said, ***"Don't chase tools, invest in readiness. The goal is not faster adoption, it is building trust and governance that scale with autonomy."*** Boards do not want rapid deployment. They want confidence that teams understand the technology, the risks, and the business impact before autonomy is introduced. That requires education at every level, from analysts to executives.

Readiness also depends on process clarity. **Madhul Sachdeva** said, ***"Education before execution is how adoption sticks."*** Teams need shared definitions for what autonomy means, which tasks qualify, and how oversight works. Without that baseline, technical deployment becomes the easy part while operational ambiguity becomes the real risk.

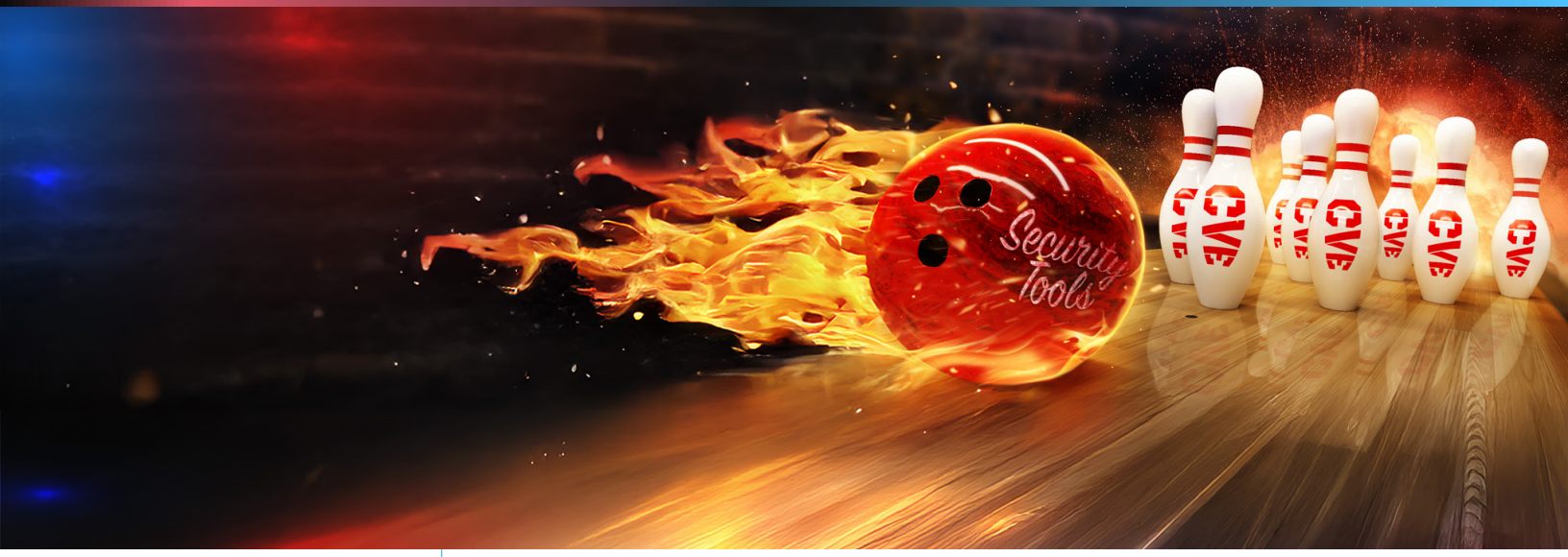
Boards are also demanding clearer measurement. The council agreed that security metrics and efficiency metrics must be tracked separately. **Ben Ramduny** explained that leaders should ***"measure what autonomy gives back, time, accuracy, or resilience."*** These are different outcomes, and boards need transparency on both the risk posture and the operational gains. As **Frederic Descloux** noted during the workshop, ***"Autonomous AI demands the same governance discipline we apply to identity or access management. It's not just a tooling decision, it's an enterprise trust decision."***

Clarity remains the guiding expectation. **John Sapp** put it plainly: ***"Boards want clarity, not complexity. Speak their language."*** Risk tiers, approval flows, rollback plans, and success criteria need to be documented in business terms. Autonomy only scales when leadership understands how decisions are made and how failures are contained.

Regulatory readiness is also becoming a board-level expectation. Organizations must align AI controls with emerging frameworks such as the EU AI Act, NIST AI RMF and ISO 42001 to ensure that autonomy is deployed within clear governance boundaries.

The priority for 2026 is not buying more AI. It is preparing the organization to use it responsibly.

Guidance: Sequence investment from education to visibility to limited automation. Establish success criteria and validate outcomes before expanding scope. Autonomy is earned, not purchased.



The Path Forward

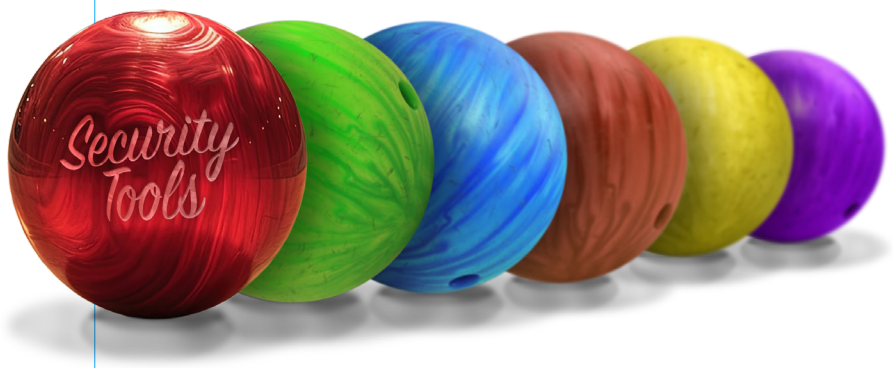
AI is accelerating defense at the same pace it empowers attackers. The council's conclusion is not optimistic or fearful. It is practical. Autonomy belongs where it reduces toil, increases consistency, and strengthens response. It doesn't belong where the blast radius is high, process maturity is low, or context is incomplete.

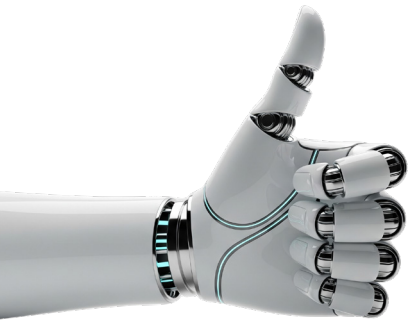
The future security operations center will not be human versus machine. ***"The future SOC isn't AI or human, it's both, co-executing securely,"*** **Greg McCord**. AI will handle volume, summarize complexity, and surface validated options. Humans will make judgment calls, define boundaries, and own accountability.

To reach that future, autonomy must remain grounded. **Assi Ungar** reminded the council that ***"autonomy without context is chaos. Guardrails are how we win on speed and safety."*** Context, visibility, and rollback are what separate responsible adoption from reckless deployment.

> "Autonomy without context is chaos. Guardrails are how we win on speed and safety."
>
> — **Assi Ungar**

The path forward is deliberate. Build guardrails first, deploy autonomy second, and expand only when the evidence supports it. The goal is not fully autonomous defense. It is a SOC where humans and AI reinforce each other, raising the standard of security performance.





Checklist for Deploying Autonomous AI Systems

Autonomy introduces speed, consistency, and scale. It also introduces new forms of operational risk. This checklist outlines the minimum controls every organization should implement before granting AI systems the ability to act. These principles apply regardless of industry, tooling, or maturity. They create a stable foundation for responsible autonomy and a shared baseline for CISOs, security teams, and boards.

Start with reversible, low-risk autonomy.

Focus early autonomy on containment, enrichment, and other actions that can be undone without business disruption.

Require documented risk acceptance for every autonomous workflow.

Decision authority must remain tied to enterprise risk management. Owners (not algorithms) accept the impact.

Enforce immutable audit logs and strict environment segmentation.

Logs must be uneditable and complete. Development environments must remain isolated from production.

Keep humans in the loop for all high-impact decisions.

Analysts approve configuration changes, policy updates, workflow transitions, and any action with material blast radius.

Scale autonomy only with proven reliability and rollback capability.

Promotion requires evidence. Autonomy expands only when the system can perform consistently and recover safely.

Maintain continuous testing to detect drift, bias, and safety issues.

Models and agents must be validated on an ongoing cycle. Silent failures are the most dangerous failures.

Track AI security metrics separately from productivity metrics.

Risk reduction and efficiency gains are different outcomes and must be measured independently.

This unified guidance forms the operational core of autonomous defense. It defines how security teams move from experimentation to sustained, governed deployment. It also provides a common language for cross-functional alignment, board communication, and long-term strategy.

We've Reserved You A Seat

Join the AI Security Council

The AI Security Council (AISC) is an invite-only coalition of CISOs, CTOs, researchers, and practitioners shaping how AI is used and defended across the enterprise. It's where security leaders anticipate adversarial AI, share field-tested strategies, and publish actionable frameworks for defending modern environments.

If you want a seat at the table shaping how security evolves with AI, **apply here**.

AI SECURITY COUNCIL