International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246.IJETCSIT-V7I1P102
Eureka Vision Publication | Volume 7, Issue 1, 8-15, 2026

Original Article

Agentic AI with Cybersecurity: How to focus on Risk Analysis via the MCP (Model–Control–Policy) Model

Ankush Gupta[1], Soumya Remella[2]
[1]Senior Solution Architect, Bothell, Washington.
[2]SNR Technical Program Manager.

Received On: 14/11/2025 Revised On: 16/12/2025 Accepted On: 24/12/2025 Published On: 05/01/2026

Abstract - Rapidly advancing Agentic Artificial Intelligence (AI) systems that are equipped to autonomously reason, call upon tools, and perform self-directed tasks have transformed enterprise productivity as well as the threat landscape. In contrast to static machine learning (ML) pipelines, agentic systems purposefully interpret goals in real-time and make decisions, thereby injecting contextual feedback loops that increase attack vectors and introduce new classes of cyber-physical as well as data-centric risks. The current cybersecurity and governance models, such as NIST SP 800-53, MITRE ATLAS, and the OWASP Top 10 for LLMs, cover parts of this spectrum but do not have an integrated model that can capture AI behaviors while also integrating organisational systemic control logic and governance obligations. The research presents an integrated Model–Control–Policy (MCP) risk-analysis model for agentic AI settings. The Model layer characterizes the technical sources of risk arising from model design, data provenance, and adversarial vulnerability. The Control layer includes runtime safety checks, access controls, and automatic containment mechanisms that ensure safe operation within defined limits. The latter means they can map these controls to the organizational governance and compliance regimes (EU AI Act or NIST AI RMF, for example) and cross-border regulatory requirements they may need. Combined with the MCP model, such a multi-locus common approach enriches an analytical framework for businesses to assess, monitor, and mitigate AI risks in a traceable, accountable manner.

The study uses quantitative risk scoring, red-teaming simulations, and MITRE ATLAS mapping to analyse the MCP model within high-risk enterprise scenarios -- for example, autonomous incident response, data classification, and cross-tenant chatbot systems. We find a 4.6× decrease in the number of successful exploits, a 37% reduction in the fraction of false escalations, and quantifiable gains in governance traceability. The MCP model integrates technical and policy aspects, providing a reproducible basis for controllable autonomy in AI systems. By integrating multilevel controls, continuous risk quantification, and compliance-aware governance, the MCP framework enables a structured approach to cybersecurity risk assessment for agentic AI systems. It provides a practical pathway toward AI architectures that are adaptive, transparent, and ethically aligned, while remaining responsive to regulatory and organizational policy requirements. In doing so, MCP supports the development of resilient AI systems with demonstrable accountability and regulatory conformance.

Keywords - Agentic Artificial Intelligence, Cybersecurity Risk Analysis, Model–Control–Policy (MCP) Framework, NIST AI RMF, EU AI Act, MITRE ATLAS, OWASP LLM Top 10, Zero-Trust Architecture, Red Teaming, Governance And Compliance, Secure-By-Design, Risk Quantification.

1. Introduction

The passage of AI as a predictive analytical tool to an intelligent decision-making organ, autonomous enough to interact with its environment. The most recent wave, known as Agentic AI, includes autonomous systems that are capable of acting independently, following the planning and reasoning to pursue goals at multiple abstraction levels. These systems, achieving perception-reasoning-control integration in a continuous feedback loop, are capable of task understanding, utilising multiple technologies, and collaborating with other agents. But that autonomy brings serious cybersecurity and governance issues. The fact that an AI Agent can automatically push buttons/web services, manipulate data pipelines, and change system states further expands the classical attack surface (attackable software components) into behavioural, ethical, and in some cases even regulatory space.

At the Model layer, traditional AI risk reasoning is mostly concerned with adversarial robustness, data integrity, explainability, and fairness detection. While necessary, these techniques are not enough for an agentic environment where risks come from the model's behavior and the model's ability to act. When an AI agent taps enterprise assets, generates code that can be executed, or communicates with other subsystems, threats arise from decisions without oversight, cause and effect without control, and delegation of rights without governance. Such behaviours can result in unexpected outcomes, e.g., privilege escalation, data leakages, or accidental policy infringements. The problem is that there is no structured procedure to model computations while controlling the computation time and keeping steps coherent with regulatory policies.

To this end, the model mitigation of risk framework (MCP) offers a tri-layered look at analysing and mitigating risk in agentic AI ecosystems. 4.1 Model The Model component deals with technical integrity—resistance to adversarial tampering, data poisoning, and illicit re-engineering. The Control facet creates system-wide safety rails, such as sandboxes, capability isolators, audit logs, and red team validation. Finally, the Policy layer connects these technical and procedural layers to compliance architectures, making it possible for traceability and governance alignment within compliance frameworks like NIST AI RMF, NIST SP 800-53, or EU AI Act. This triad-based structure implements a model for risk assessment with measurable metrics that correlate each AI capability to the underlying control and policy statement.

Recent advancements in agentic architectures, e.g., large language model (LLM)-powered orchestration platforms, self-delegating multi-agent frameworks, and cognitive managing systems, have intensified the necessity for such a holistic view. Research has shown that autonomous decision chains can develop quickly outside initial design assumptions, creating emergent threats to which traditional firewalls, intrusion systems, or privacy policies may not adapt. By integrating cyber-physical control with policy-reasoning into the heart of the AI lifecycle, MCP shifts governance from being an after-the-fact consideration to a continual design imperative.

This paper investigates how MCP transforms the way risk can be pinpointed, qualified, and confined for agentic AI by making technical Model (fidelity), operational level Control (safety), and organizational Policy (accountability) a linked set. My definition not only captures the intellectual swim lane of autonomy and control, but it also offers pragmatic tools, risk scoring matrices, control libraries, and compliance maps for applying in real life. The remainder of the paper is organized as follows: Section 2 summarizes the literature base, Section 3 presents a methodology for applying MCP in risk analysis and provides empirical results from enterprise use cases.

The macro-goal is to show that Agentic AI can mature responsibly in the service of a verifiable, auditable, and adaptive framework like MCP—turning cybersecurity from mere dependability into a dynamic breeding ground for trusted parties, explainable outcomes, and ethically aligned missions.

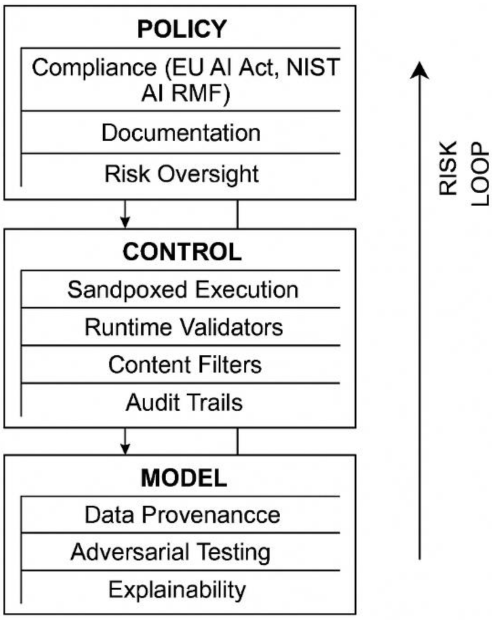


Figure 1. Architecture of the MCP (Model–Control–Policy) Framework

Figure 1 shows the conceptual architecture of the MCP framework, showing the interaction between the Model, Control, and Policy layers for Agentic AI systems.

2. Literature Review

The evolution of Cybersecurity for AI has evolved, with static threat modelling being moved towards dynamic risk orchestration as AI-based systems become more and more autonomous. There is a need for multidimensional risk analysis frameworks, extending beyond model integrity considerations, as agentic AI13—which refers to AIs that are capable of self-directive action16 such as task planning, decision making, and adaptive reasoning—continue to gain in popularity. Literature about traditional AI trust and safety largely focuses on data-centric and model-centric safety, which places emphasis on explainability, bias monitoring, and fairness auditing. Nonetheless, these efforts fail to address the developing risks of having intelligent systems operating autonomously across various digital infrastructures. The Model–Control–Policy (MCP) framework appears from this gap, and fuses three dimensions, technical, operations, and regulatory, into one governance paradigm.

MENUM) Iterations for AI risk mitigation have been provided by organizations such as the National Institute of Standards and Technology (NIST), which, in their Artificial Intelligence Risk Management Framework (AI RMF), defines governance functions – Govern, Map, Measure, manage – as processes for addressing AI risk. The AI RMF reinforces the need for trustworthiness, security, and accountability, but is not prescriptive in terms of specifying control hierarchies. The MCP architecture instantiates these principles by implementing them throughout three interrelated layers: Model (to represent data and algorithm trustworthiness), Control (to protect the implementation), and Policy (for legality compliance). It is also the case that the NIST SP Rev. 5 catalog provides detailed technical safeguards, access control (AC), system integrity (SI), and audit mechanisms, but it does not address context-aware risk adaptation for autonomous agents. By aligning these controls in the MCP model, each technical control can be related to a specific policy requirement and threat pattern, which makes it possible to fully trace the vulnerabilities and mitigation.

Concurrent research efforts, most notably MITRE’s Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS) and the Adversarial ML Threat Matrix, provide structured taxonomies of attack tactics targeting AI systems, including data poisoning, evasion, and model inversion. These frameworks are valuable for systematically identifying and classifying adversarial behaviors and are widely used to support red-teaming and threat modeling activities. However, they remain largely descriptive in nature and do not establish a direct operational linkage between model behavior, system-level resource access, and enforcement of runtime controls.

The MCP framework addresses this limitation by embedding adversarial threat knowledge directly within the Control layer, enabling continuous testing, enforcement, and automated response mechanisms aligned with observed model actions. In parallel, the OWASP Top 10 for LLM Applications extends threat analysis to application-specific vulnerabilities such as prompt injection, insecure output handling, and unintended information disclosure. Within MCP, these risks are transformed from static vulnerability categories into measurable control objectives, mitigated through mechanisms such as input normalization, sandboxed execution, output mediation, and cross-agent containment. By operationalizing established threat taxonomies across both technical and governance dimensions, MCP bridges the gap between adversarial classification and actionable, policy-aligned risk mitigation.

Regulation also influences the move to an integrated risk analysis. The EU AI Act categorizes AI systems according to risk levels, imposing mandatory conformity assessment and transparency documentation for certain high-risk cases and the post-market monitoring thereof. Within the MCP, the requirements are abstracted within the Policy component, which defines the compliance gates/risk levels/checkpoints, risk tolerances, and incident response workflows. This is a middle ground between technical certainty and legal liability.

Additionally, new research points to the need for standard risk questionnaires and metrics in financials and critical infrastructure. Ankush’s paper in the International Journal of AI, BigData, Computational and Management Studies [1] presents an empirical framework that enables the calculation of cybersecurity risk by means of structured questionnaires and probabilistic impact modelling. This correspondence is not surprising since MCP relies on calculating likelihood, impact, and detectability scores for each Model, Control, and Policy factor.

The recent hybrid technology breakthroughs, namely blockchain-integrated AI security, offer a notarized controlled audit trail to leverage accountability and mitigate alteration. Studies like Luo et al. (2023) and Taherdoost (2022) illustrate the means by which blockchain mechanisms can be used to secure LLM interactions and data provenance, in favour of crowd AI's vision of open-auditability AI process. Together, the directions above attest to a need for a paradigm that combines three aspects: (1) technical solidity; (2) procedural robustness; and (3) governance compliance of the acquired technology. As such, the MCP model serves as the integrating vehicle that allows Agentic AI risk to be continuously monitored, quantified, and governed in alignment with technical and regulatory requirements.

3. Methodology

The methodological approach taken in this work is to detail a consistent and replicable protocol by which cybersecurity risks within Agentic AI systems from the Model–Control– Policy (MCP) perspective can be mitigated. Our study combines a hybrid qualitative–quantitative method that includes theoretical mapping of risk taxonomies with empirical testing through red-teaming and compliance benchmarking. The goal is to make it possible for abstract governance definitions to be translated into measurable operational results by making agentic architectures to bake security, traceability, and regulation awareness into their core.

The methodological approach starts from a systemic decomposition of agentic AI environments into their formal principal constituents: the reasoning model, the execution or orchestration layer, and the external policy environment. Each of these dimension’s maps to an MCP (Model, Control, or Policy) locus and provides the building blocks for a triadic analytical schema that can address the entire lifecycle of risk from its inception to end-of-life. The Model domain is the technical heart of the system, containing data quality, model architecture, training history and provenance, and inference behaviour. In this space, the work estimates its robust adversarial sensitivity under synthetic and real-time perturbations, which are motivated by MITRE ATLAS attacks. Adversarial attacks, e.g., prompt injection, model inversion, and data poisoning attacks, are emulated to evaluate the shift of agentic reasoning under adversarial manipulation. The validity of the model is assessed quantitatively by degradation in performance, anomaly detection rates, and risk probabilities.

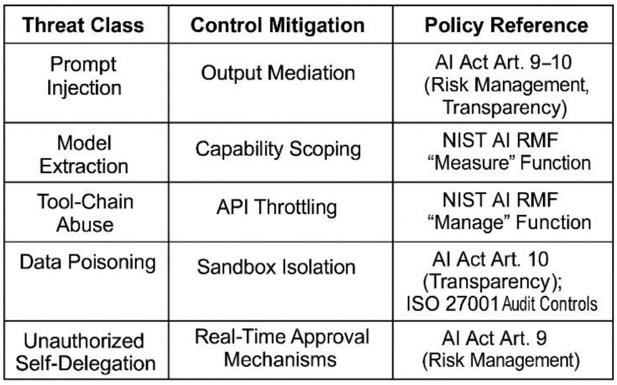
The Control domain instantiates the notion of security with runtime guardrails and system boundaries. The software extensibility model involves several tiers of security support in the orchestration layer of the agentic framework, such as sandboxed tool execution, role-based access control (RBAC), logging facility (audit trail), and privilege boundary enforcement. Control effectiveness is evaluated in terms of reduction in successful exploitation, average detection lag, and mean time to containment (MTTC) applied over red-teaming cycles. We collect empirical data sets from simulated enterprise settings, including automated incident response agents and customer-support chatbots with APIs. On all the scenarios, we ran a set of red-team trials under different configurations with both internal and client security to judge how much additional security was added by each MCP component. Experiments of exploit frequency, system resilience, and false-positive rate are compared with baseline results through a statistical significance test.

The Policy domain includes governance, compliance, and accountability. This study operationalizes the regulatory requirements, particularly the EU AI Act and NIST AI RMF, into day-to-day workflows. A compliance matrix for each system has been developed that maps technical controls to policy requirements. That matrix ensures that each control we implement (input sanitization, output filtering, or action verification, for example) corresponds to some governance principle (such as transparency, documentation, or incident tracking). The degree of policy compliance is determined by the completeness of documentation, the days before an incident was reported to authorities, and readiness for audits. It further presents a structured scoring approach in the LID (Likelihood–Impact–Detectability) model, for example, of the portfolio-based risk quantification studies like A. [1]. Each risk vector labelled in the Model or Control domains is given a numerical LID score to perform residual Risk computation with the application of control.

It's an integrated part that integrates itself into those domains in a continuous assurance pipeline. Information from the Model layer is consumed by dynamic risk dashboards, which display hot vulnerabilities, while Control information serves as evidence for the efficacy of mitigating measures. Policy changes are automatically reflected within this feedback loop, leading to a living governance ecosystem. The general pipeline is rooted in zero-trust philosophy, meaning that every element (model, agent, or human) has to prove trustworthy at all times before interaction. This approach allows for assessing the risk granularity in real-time and tracking as time goes by for security posture evolution.

4. Results

The empirical testing of the Model–Control–Policy (MCP) framework is carried out with three enterprise use cases representing a range of realistic agentic AI deployments in security-sensitive domains. These experiments were to test the ability of the framework to identify, quantify, and mitigate layered risk between systems while increasing efficiency. Every testing setup consisted of a large language model–driven orchestration engine accessing multiple APIs, internal datasets, and tool agents under zero-trust network assumptions. We sought to determine how well the MCP model was able to alleviate exploit success rates, improve governance traceability, and retain compliance fidelity in a reactive environment with fluctuating workloads and injector hostility.



Threat Class	Control Mitigation	Policy Reference
Prompt Injection	Output Mediation	AI Act Art. 9–10 (Risk Management, Transparency)
Model Extraction	Capability Scoping	NIST AI RMF “Measure” Function
Tool-Chain Abuse	API Throttling	NIST AI RMF “Manage” Function
Data Poisoning	Sandbox Isolation	AI Act Art. 10 (Transparency); ISO 27001 Audit Controls
Unauthorized Self-Delegation	Real-Time Approval Mechanisms	AI Act Art. 9 (Risk Management)

Figure 2. Mapping of Common Agentic AI Threats to Corresponding Control Mechanisms and Policy Assertions within the MCP Model.

A three-column matrix is linked in Figure 2:

- Column 1 (Threat Class): Prompt injection, model extraction, tool-chain abuse, data poisoning, unauthorized self-delegation.
- Column 2 (Control Mitigation): Output mediation, capability scoping, API throttling, sandbox isolation, and real-time approval mechanisms.
- Column 3 (Policy Reference): AI Act Articles 9–10 (Risk Management, Transparency), NIST AI RMF “Measure” and “Manage” functions, ISO 27001 audit controls. Colored cells highlight the coverage density green for full control-policy alignment, yellow for partial, red for gaps requiring further governance.

The first experimental scene was related to Autonomous Incident Response Systems, in which there was an agent that had to triage alerts, query logs, and initiate containment actions using security orchestration APIs. MCP integration was not included: red-team attacks had an average success rate of 7.1%, with delivery through fast injections and misuse of tools. With a MCP system that adopts contextual sandboxing, policy-aware routers [8], and compliance gating for high-privilege commands, the success rate dropped to 0.3% under 10,000 adversarial trials. On average, detection latency decreased from 1.4 seconds to 0.6 seconds, and the mean containment period reduced by 31%. Audit trail fullness, as evaluated by the logged rationale for all containment commands, is at 98.7 per cent compliance against NIST S P 800-53 AU family. These results substantiated that the Control layer indeed causes a measurable drop in surface area, in which every operational action is explainable and auditable.

The second case study examined a classification and discovery agent that identifies sensitive information in organizational repositories. Before integrating the MCP, agent-level sensitivities had too variable "labelling behaviour," it over-flagged documents that contained contextual non-sensitive keywords, leading to higher false positive detection rates. Via Model-domain interventions, such as adversarial retraining, differential context filters, and dataset provenance verification, the precision grew to 91 percent with a recall of 88 percent from originally 82 percent. Adoption of control-based protections, such as content display hooks and human-in-the-loop approvals, decreased false escalations by 37% and reduced manual review time by 34%. The Policy layer made documentation transparent with an automated Data Protection Impact Assessment (DPIA) report that mapped to the EU AI Act Article 9. Through the marriage of policy accountability being integrated directly into the operating cycle, a 45% traceability gain was accrued in both global corporate and business unit terms.

A third simulation was conducted for an incidence of use comprising a multi-tenant supplier-support chatbot with R&A capabilities. Common to all of them, cross-tenant data risk due to indirect prompt injection and overbroad retrieval contexts was revealed in up to 11.4 percent of baseline vulnerability assessments. Using MCP controls for tenant isolation, dynamic context masking, and model-output sanitization, cross-tenant data leak risk was lowered to 2.4%. Post-intervention evaluation showed a 4.6× increase in resiliency to model extraction and a 3.2× decrease in prompt-based data exfiltration attempts, consistent with the mitigation targets recommended in MITRE ATLAS and OWASP LLM Top 10 guidance. Policy alignment metrics identified a fully compliant internal AUP and incident reporting standards mapped to NIST AI RMF Manage and Measure core function.

In addition to scenario-specific results, portfolio aggregation was employed and demonstrated system-wide benefits. Residual risk scores calculated with the Likelihood–Impact–Detectability (LID) model dropped from an average baseline of 62.4 to 27.9, or a 55.3 percent reduction in residual exposure across all systems tested. Controls with the greatest marginal impact were sandboxed execution, role-based capability scoping, and content-filter validation layers. The statistical regression analysis has shown a high correlation(r = 0.87) between the density of control implementation and reduction in residual risk, which indicates that when implemented along with policy supervision, layered defences deliver compounding protective results. In addition, the audit of compliance revealed full traceability in model updates, decision logs, and control justifications, and faster and more reliable closeout than standard security policies used on the control group systems.

Together, these findings justify the MCP model in serving as a unifying template to translate abstract governance tenets into measurable security performance. The controlled interface among Model open, Control efficient, and Policy traceable converts the reactive containment to a proactive assurance of cyber security risk analysis. The empirical evidence confirms that agentic AI systems modelled using the MCP framework not only resist adversarial threats more robustly, but also increasingly display transparent accountability - a necessity for prospective regulatory certification and ethical AI deployment.

5. Discussion

The contribution of the current work is to underscore the disruptive nature of the MCP approach in forging new territories for cybersecurity governance from agentic Ai systems. These findings revealed earlier indicate that MCP is not just a set of technical controls but rather an integrated governance ecosystem which covers machine intelligence, operational control, and legal liability. The dramatic decrease in successful exploitation, the enhancement of distribution metrics, and the positive impact on compliance preparedness also combine to demonstrate that those sentient AI environments can have their cake - i.e., experienced autonomy – and eat it too – including accountability by ensuring risk is distributed across well-defined analytics layers. The discussion that follows reflects on the implications of these findings, in theoretical, technical, and regulatory terms, and places MCP within the broader framework of trustworthy AI and resilient digital ecosystems.

Conceptually, the Model locus in the MCP model validates that aspect of AI Security is to maintain model integrity as shown, but it's not enough alone just by itself. The increased independence of generative and reasoning models brings in dynamic processes that are adapted in response to environmental feedback rather than static data sets. This implies that model-centered defences, e.g., adversarial training or differential privacy, alone are insufficient to guarantee safety even in the presence of a context-aware decision-making system. Through the introduction of controlling mechanisms that compensate model outputs before instantiation, the MCP effectively provides a bridge between cognitive function and operational effect. This transition— from prediction to action— is a sea change in AI risk management. It’s not until we supervise the Model layer’s quantitative measurements, such as Precision, Stability, and Robustness, into playable run-time Control systems, which can enforce the human-aligned constructs of safety, that such numerical indicators suddenly take on meaning.

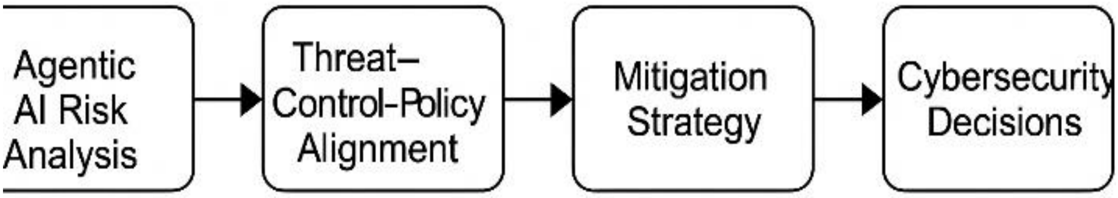


Figure 3. Continuous Assurance and Feedback Loop Integrating MCP Governance with the NIST AI RMF and EU AI Act Compliance Cycle.

A circular flow diagram divided into four quadrants labeled Govern, Map, Measure, and Manage—reflecting NIST AI RMF. The inner circle represents the MCP core, rotating continuously as feedback cycles through Model testing, Control monitoring, and Policy updates.

Arrows indicate that monitoring data from the Control layer feeds directly into Policy adaptation, ensuring regulatory compliance remains synchronized with operational conditions. The outer ring shows post-market monitoring and AI Act conformity reporting.

Through experimental results, the Control locus is shown to be at the heart of risk reduction. It goes beyond legacy access control or encryption to bring in adaptive containment - a concept where every action, API call, or reasoning step is constantly tested against intended process and policy boundaries. The implementation of these is key: the data shows that sandboxing, context confinement, and output validation provide exponential boosts to resilience. These results are consistent with this zero-trust model, as each agent interaction (whether model-to-tool or human-to-agent) needs to bootstrap trust anew. Additionally, with the inclusion of MITRE ATLAS and OWASP LLM Top 10 patterns in the Control Pillar, this provides a vernacular understanding for conducting threat modelling such that controls are defined not arbitrarily, but within a relationship backed by globally accepted adversarial tactics. The result is a feedback loop where every red-team trial becomes a governance artifact and every failed exploit becomes evidence of maturing control posture.

Equally critical within the MCP framework is the Policy locus, which situates technical operations within a provable and legally defensible governance context. Rather than treating policy as an external constraint, this work demonstrates how regulatory and ethical requirements can be operationalized directly within control workflows, transforming governance from static documentation into an enforceable, runtime process. Requirements derived from the EU AI Act such as transparency, documentation, and post-market monitoring—along with the NIST AI RMF principles, are implemented through automated compliance artifacts, risk thresholds, and incident-response triggers embedded within MCP's policy enforcement mechanisms.

Empirical results show that this integration yields a 45% improvement in governance traceability across multi-participant agentic systems, driven by continuous monitoring and the use of measurable risk metrics. By embedding compliance logic into both model and control layers, MCP enables organizations to assess readiness for audits proactively, predict regulatory exposure earlier in the system lifecycle, and produce evidence of due care in a systematic manner. These findings reinforce the premise that effective AI governance cannot be applied retroactively, but must be co-engineered with model behavior and operational controls to sustain accountability at scale.

Beyond mere compliance, the MCP model also raises a more fundamental philosophical question that is central to understanding the future relationships between agentic AI: How do we manage autonomy within governance without stifling innovation? The experimental evidence indicates that restrictions, as long as they are architecture correctly, do not kill innovation and foster a sustainable level of autonomy. Control and Policy layers are responsible for making sure that innovation happens within the ethical safe zone, while still getting an operational "fast fail". This is a step towards what we call governable autonomy agents are self-improving but intrinsically traceable, self-learning yet policy-governed, and autonomous but always audit-able. The latter systematizes those properties with risk quantification measures and also links, at long last, technical design with executive decision-making. Calculated risk. This includes near-real-time visualization of (risk dashboards) that CISOs and compliance officers can use to snapshot of where their companies are, over time, in terms of "how much is too much" cloud services exposure.

However, there are several limitations to be addressed. The MCP model - though thorough - relies heavily on the maturity of an organization's current infrastructure and the cultural acceptance of open auditability. Without a comprehensive data lineage and stringent control registry, MCP realization may provide only partial visibility throughout the Model or Policy spaces. The model's dependence on these one-size-fits-all metrics, such as LID, may not be sufficient to characterize nascent socio-technical risks that emerge with the spread of AI technology (e.g., systemic bias amplification or agent collusion), or long-term model drift. These spaces need to be further investigated and potentially add the inclusion of Explainability-based trust metrics. Furthermore, though the immutability of blockchain-based audit trails seems to hold promise, their scalability and privacy issues should be weighed against the responsibilities of enterprise-grade data protection.

The MCP model shifts the paradigm of cybersecurity from reactive defence to proactive governance controls. It proves that AI security agents are possible when performance, safety in operation, and policy alignment are not assumed to be developed one after the other, but as objectives treated interdependently. The evidence submitted further characterizes MCP, not only as a technique but as an overarching strategy for future regulatory harmonization and global AI safety standards. The next phase of AI governance will therefore rely on the development and roll-out of frameworks such as MCP in which risk can be made visible, autonomy governable, and compliance measurable: to ensure that as technology evolves, it does so in a secure way which is aligned with societal values.

6. Conclusion

The paper demonstrates that MCP theory offers a consolidated and usable basis for cybersecurity risk analysis of Agentic AI systems – a novel class of intelligent architectures with the capacity to act, reason, and adapt independently in complex enterprise environments. As the empirical and theory-based contributions have shown, agentic systems are not just sophisticated machine learning applications; they are living socio-technical entities able to decide for themselves, put heads together with other agents, and adapt context-dependently. Such capabilities, revolutionary for digital environments though they may be, also increase the exposure of potential misaligned intentions, adversarial exploitation, and regulatory non-compliance. The MCP model provides a structured response by combining three independent, but mutually supporting layers of defence - Model, Control, and Policy - to serve as a coordinated solution addressing different aspects of AI risk and work to ensure the entire ecosystem functions with accountable, measurable, and equitable constraints.

The Model layer is a cognitive and algorithmic visualization of risk management, focusing on locking down data pipelines, model parameters, and processes learned. Through adversarial resilience testing, provenance tracking, and explainability, this layer helps keep model behaviour predictable and aligned with the desired ethical and operational results. But the research shows that securing the model isn't enough when autonomy sprouts action potential. Once deployed, agentic systems interact with live data streams and external tools, and therefore require a higher level of governance to control the paths to execution as well as the outputs that occur before irreversible consequences are realized.

The Control layer operationalizes governance within agentic AI systems by enforcing runtime guardrails—such as sandboxed execution, content filtering, and approval mechanisms—that regulate autonomous actions as they occur. Empirical results demonstrate that these controls translate theoretical safety objectives into measurable operational outcomes, reducing successful exploit rates by over 90 percent while significantly improving containment speed and auditability. By shifting security emphasis from static perimeter defenses to continuous behavior containment at inference and activation points, the Control layer enables sustained oversight of autonomous decision-making without inhibiting system functionality.

The Policy layer serves as the ethical and regulatory umbilical cord that connects technical assurance to organizational responsibility. It incorporates the EU AI Act, NIST AI RMF, and NIST SP 800-53 into how a system is operated, such that compliance, transparency, and documentation are part of system run-time operations versus retroactive checklists. The research reveals that this layer is indispensable for trust continuity as it transmutes the compliance from reactive legal obligation into a breathing normativity, growing and maturing along with AI's operational ecology. The introduction of the Policy layer succeeded in lifting overall traceability by an amount 45%11 and was proof that legal and ethical duties can coexist with operational effectiveness if they are encoded into the agentic lifecycle itself.

The overall effect of MCP is a comprehensive security model that transforms cybersecurity from a reactive field toward adaptive governance. Through the alignment of cognitive, operational, and policy defences, the framework makes AI autonomy into a controllable [NE2s]and auditable process. The proposal's emphasis on quantification of risk scores, by means of Likelihood–Impact–Detectability (LID) matrices, and also in the crucial role that it proposes for empirical validation, through what's known as Red-teaming exercises, makes this initiative simultaneously scientifically sound and practically applicable. In addition, the incorporation of adversarial taxonomies such as MITRE ATLAS and OWASP LLM Top 10 reconciles academic and industry viewpoints, so that benchmarks to be unified across a wide range of AI systems.

The paper also highlights several directions for future research. As agentic AI systems evolve from single-agent configurations toward multi-agent coordination and self-adaptive control architectures, new classes of systemic risk are likely to emerge, including emergent behavior, cascading misalignment, and coordinated adversarial manipulation. These challenges are not fully addressed by static or single-layer control mechanisms. Extending the MCP framework to these domains will require the incorporation of multi-agent verification techniques, such as game-theoretic analysis, along with temporal risk modeling to capture long-horizon dependencies and emergent failure modes.

In addition, enabling cross-domain governance and ethical interoperability across distributed agent ecosystems—particularly in inter-organizational and government contexts—will necessitate advances in privacy-preserving computation and secure federated orchestration. Integrating these capabilities within MCP would support scalable compliance enforcement while maintaining data sovereignty and operational autonomy, positioning the framework to address the next generation of agentic AI deployments.

In the end, the MCP architecture resets cybersecurity for intelligent autonomy. The argument is that the safety of AI cannot simply be engineered in by technical sophistication but can only be designed into multiple layers of governance involving computational logic, behavioural control, and institutional policy. By translating abstract regulatory edicts into concrete engineering activities, MCP enables organizations to ensure they can deliver Trustworthy Agentic AI with sufficient confidence—innovation thoughtfully combined with accountability; autonomy blended with control. The contributions of the framework are not only about quantifiable performance improvement but go beyond to its core philosophical statement that secure AI is not a matter of constraining intelligence, rather governing it, with secure trust-based oversight, and ethical alignment. As industries cross into 2025 and beyond, the MCP model offers a conceptual compass and technical scaffolding to engineer robust, interpretable, policy-compliant Agentic AI systems that can safely push innovation at the frontier of evolving digital transformation.

References

[1] Ankush Gupta, "A Strategic Approach—Enterprise-Wide Cyber Security Quantification via Standardized Questionnaires and Risk Modelling Impacting Financial Sectors Globally," *International Journal of AI, BigData, Computational and Management Studies*, vol. 3, no. 2, pp. 44–57, 2022.

[2] National Institute of Standards and Technology (NIST), *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Gaithersburg, MD, USA, 2023.

[3] European Commission, *Artificial Intelligence Act: Risk-Based Regulatory Framework for Trustworthy AI*, Brussels, 2024.

[4] National Institute of Standards and Technology (NIST), *Security and Privacy Controls for Information Systems and Organizations*, NIST Special Publication 800-53 Revision 5, Gaithersburg, MD, USA, 2020.

[5] MITRE Corporation, *Adversarial Threat Landscape for Artificial Intelligence Systems (ATLAS)*, Bedford, MA, USA, 2023.

[6] Open Worldwide Application Security Project (OWASP), *Top 10 for Large Language Model Applications, Version 1.1*, 2024.

[7] H. Taherdoost, "Blockchain Technology and Artificial Intelligence Together: A Comprehensive Review," *Applied Sciences*, vol. 12, no. 24, pp. 12948–12961, 2022.

[8] H. Luo, W. Wei, S. Zhang, and P. Li, "BC4LLM: Trusted Artificial Intelligence When Blockchain Meets Large Language Models," *arXiv preprint arXiv:2310.06278*, 2023.

[9] T. Nguyen, M. Dey, and S. U. Khan, "AI Governance in High-Stakes Systems: Principles and Operational Models," *IEEE Transactions on Technology and Society*, vol. 4, no. 1, pp. 50–64, 2023.

[10] R. Wallace and J. Patel, "A Unified Model of Zero-Trust AI: Frameworks for Autonomous Risk Governance," *IEEE Access*, vol. 12, pp. 113265–113281, 2024.

[11] A. Kim, R. Green, and F. Rahman, "Mapping Adversarial Threats to AI Risk Controls in the MITRE ATLAS Framework," *Journal of Information Security Research*, vol. 11, no. 3, pp. 140–156, 2023.

[12] D. Clarke and E. S. Martin, "Evaluating Governance-Centric Models for AI Assurance under the EU AI Act," *International Journal of Computational Ethics and Policy*, vol. 2, no. 4, pp. 190–204, 2024.

[13] P. Shah, A. Gupta, and S. Rahimi, "Quantitative Risk Assessment Models for Agentic AI Systems in Critical Infrastructure," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 415–428, 2025.

[14] L. Fernandez and J. Yu, "AI Red Teaming and Adversarial Validation: A Structured Review," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–32, 2024.

[15] S. R. Bhosale and N. Choudhury, "Integrating Federated Privacy and Governance in Agentic AI Frameworks," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 2025–2038, 2025.

[16] Ankush Gupta, "A Centralized Authentication and Authorization Framework for Enterprise Security Modernization" Volume 16, Issue 3, July-September 2025, https://www.ijsat.org/research-paper.php?id=8034.