Governare l'autonomia degli agenti nella Cyber Security

Maggio 2026 · Cyber Resilience Forum · Richmond Events

Federico Cerutti
federico.cerutti@unibs.it

Università di Brescia

IMPERIAL
University of Southampton

CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

Executive Summary

L'Agentic AI sta ridefinendo il perimetro della cyber security con sistemi capaci di osservare il contesto, prendere decisioni intermedie, coordinarsi con strumenti e soggetti diversi e agire sull'ambiente operativo. Il problema centrale diventa il governo della delega operativa: ogni incremento di autonomia richiede infatti un mandato, limiti di azione, evidenze, supervisione e un modello di responsabilità coerente con il rischio dell'organizzazione.

Il valore degli agenti in cyber security è reale e si estende dalla gestione degli incidenti alla cyber threat intelligence fino al supporto alle crisi ransomware. Tuttavia, quando un sistema non si limita a suggerire ma può agire, la governance diventa il tema dominante. La discussione deve quindi spostarsi da *quali agenti possiamo adottare* a *quale autonomia siamo disposti ad ammettere e a quali condizioni*.

L'autonomia è governabile solo se è governata la conoscenza che la alimenta. Fonti, memoria, provenienza, rappresentazione del contesto, policy e qualità dello stato informativo sono parte integrante del problema di governo dell'autonomia, non una questione separata di implementazione.

Dall'automazione alla delega operativa

L'Agentic AI è la risposta ad a una pressione di scala che il modello interamente umano fatica sempre più a sostenere. La cyber security contemporanea è caratterizzata da una crescente complessità operativa: ambienti ibridi, filiere interdipendenti, segnali eterogenei, superfici d'attacco in continua espansione e minacce che evolvono più rapidamente della capacità umana di assorbirle e coordinarne la risposta. In parallelo, gli agenti stanno uscendo dalla fase puramente dimostrativa. Anche in ambito security il mercato si sta muovendo in questa direzione, usando agenti per sia per investigazioni che per identity e data security, e triage operativo.[1]

L'agency dipende sia dal modello linguistico che dall'architettura complessiva che lo rende capace di percepire il contesto, mantenere stato e produrre effetti nel mondo operativo.[2, 3] Un sistema di automazione esegue istruzioni definite; un copilota suggerisce o sintetizza; un agente combina invece memoria, recupero di contesto, uso di strumenti, permessi e decisioni intermedie in funzione di un obiettivo.

Questo cambia la natura del rischio. Quando un sistema può agire, i suoi errori, le sue omissioni o le sue interpretazioni distorte possono trasformarsi in eventi operativi con impatto su asset, persone, processi e reputazione.

Il valore degli agenti nella cyber security

Gli agenti estendono il modello di difesa verso domini nei quali la sicurezza interagisce con IT operations, risk management, legal/privacy, business continuity e comunicazione di crisi. Più cresce l'autonomia, più la cyber security si mostra come funzione organizzativa distribuita, e non come semplice sommatoria di controlli tecnici.[4]

Tre casi rendono visibile questo cambio di scala. Il primo è quello delle investigations e dell'incident response. Il secondo è la cyber threat intelligence. Il terzo è il supporto alle fasi più critiche di una crisi ransomware.

Incident response. Il valore degli agenti nell'incident response risiede soprattutto nella capacità di costruire un caso: correlare segnali eterogenei, ricostruire timeline, collegare identità, endpoint, asset e priorità operative, proporre ipotesi plausibili e suggerire next step coerenti con il contesto. In questo senso, gli agenti possono aumentare la qualità della comprensione operativa prima ancora della velocità di esecuzione.[5] Il fine non è sostituire il giudizio dell'analista senior, ma aumentare la continuità e la coerenza del processo investigativo, mantenendo al contempo la possibilità di escalation quando il caso entra in aree ambigue, ad alto impatto o a elevata incertezza.

Cyber Threat Intelligence. In questo dominio gli agenti hanno valore perché possono aumentare la capacità di lettura e di sintesi su dati e conoscenza strutturata e semi-strutturata.[6] Possono quindi trasformare fonti eterogenee in significato operativo: collegare segnali deboli, interpretare pattern emergenti, allineare eventi a tattiche note, prioritizzare scenari e supportare valutazioni strategiche.

Crisi ransomware. Qui il vantaggio potenziale degli agenti riguarda principalmente la capacità di mantenere coerenza informativa e decisionale in un momento in cui più funzioni devono allinearsi rapidamente: security, IT, legale, compliance, comunicazione, top management e talvolta terze parti. In queste condizioni, il supporto agentico può aiutare nella valutazione delle opzioni, nella gestione delle comunicazioni, nella tenuta del quadro informativo e nella tracciabilità delle decisioni, pur lasciando agli esseri umani la titolarità delle scelte critiche.

La conoscenza come condizione di governabilità

L'aumento di autonomia — con la possibilità di agire nel mondo cyber o reale — introduce nuove categorie di rischio o amplifica categorie già note: compromissione del-

l'agente, manipolazione della memoria, abuso delle autorizzazioni, perdita di controllo sul perimetro d'azione, opacità del flusso decisionale e mancato coordinamento tra agenti e strumenti.[3, 7] L'impatto può essere che l'agente *funzioni come progettato* e tuttavia produca esiti non accettabili dal punto di vista organizzativo. Questo accade quando il mandato è ambiguo, i permessi sono troppo larghi, la memoria conserva informazioni non affidabili, o il sistema non lascia evidenze sufficienti per comprendere perché abbia agito in un certo modo.

Governare l'autonomia significa allora progettare le condizioni entro cui la delega resta compatibile con accountability, resilienza e fiducia organizzativa.[8, 9, 10, 11]

L'autonomia è governabile solo se si governa la conoscenza che la alimenta. Fonti, memoria, provenienza, rappresentazione del contesto, policy, qualità dei metadati e allineamento semantico non sono dettagli implementativi; sono parte della stessa architettura di controllo.[12]

In cyber security questa osservazione è particolarmente forte, perché il dominio è già organizzato intorno a conoscenza formalizzata. STIX esprime informazioni di minacce in forma machine-readable; ATT&CK è una knowledge base di tattiche e tecniche; ATLAS estende questa logica ai sistemi AI-enabled.[13, 14, 15] Inoltre, i cybersecurity knowledge graph e la letteratura su ontologie e semantic web tools in CTI mostrano che la qualità dell'azione dipende in larga parte dalla qualità della rappresentazione delle entità, delle relazioni e delle fonti.[6, 16]

Questa non è una tesi alternativa a quella della governance. È la sua infrastruttura nascosta. In pratica, non si governa davvero l'azione di un agente se non si governa la conoscenza che la produce.

Una delega controllata

È quindi necessario che l'azienda — ad esempio, nella persona del CISO — progetti le condizioni di sicurezza, evidenza e responsabilità entro cui l'organizzazione può delegare azione a sistemi agentici. Il CISO, in altri termini, non governa un tool; disegna una delega controllata con l'obiettivo di raccordare l'intera architettura di sicurezza e i rapporti con le altre funzioni aziendali.

Una sintesi utile può essere articolata in sei domande che offrono una euristica compatta per distinguere la sperimentazione curiosa da un progetto che sta entrando in un perimetro di rischio reale.

Mandato: che cosa stiamo delegando davvero? La delega va descritta in termini operativi, non in formule generiche di supporto intelligente.

Fonti: su quali basi informative l'agente costruisce il proprio quadro del mondo? La questione non è solo quali dati usa, ma quali fonti siano considerate autorevoli.

Memoria: che cosa conserva, aggiorna o dimentica? Ogni memoria persistente è una forma di base di conoscenza e richiede regole di scrittura, revisione e protezione.

Permessi: dove può leggere, scrivere ed eseguire? Questi permessi sono la traduzione concreta del mandato.

Evidenze: quali tracce lascia? Senza provenance, motivazione operativa e audit trail, l'autonomia non è davvero controllabile.

Escalation: quando si ferma e chi decide? Ogni delega credibile include condizioni di arresto, revisione e revoca.

Conclusione

Il vantaggio degli agenti nella cyber security non dipende dal loro numero, ma dal loro governo. L'ingegneria della conoscenza ne è la chiave strutturale, perché rende visibile il legame fra ciò che l'agente sa, ciò che può fare e ciò di cui l'organizzazione resta responsabile.

Acronimi

AI Artificial Intelligence
CISO Chief Information Security Officer
CTI Cyber Threat Intelligence
IT Information Technology
SOC Security Operations Center

Riferimenti

[1] Microsoft. *Microsoft Security Copilot agents overview*. Accessed 2026-04-27. 2026. URL: https://learn.microsoft.com/en-us/copilot/security/agents-overview.
[2] World Economic Forum and Capgemini. *AI Agents in Action: Foundations for Evaluation and Governance*. 2025. URL: https://reports.weforum.org/docs/WEF_AI_Agents_in_Action_Foundations_for_Evaluation_and_Governance_2025.pdf.
[3] Microsoft AI Red Team. *Taxonomy of Failure Mode in Agentic AI Systems*. White paper. 2025. URL: https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf.
[4] World Economic Forum. *Elevating Cybersecurity: Ensuring Strategic and Sustainable Impact for CISOs*. 2025. URL: https://reports.weforum.org/docs/WEF_Elevating_Cybersecurity_2025.pdf.
[5] Asif Shahriar, Md Nafiu Rahman, Sadif Ahmed, Farig Sadeque e Md Rizwan Parvez. "A Survey on Agentic Security: Applications, Threats and Defenses". In: *arXiv preprint arXiv:2510.06445* (2025). URL: https://arxiv.org/abs/2510.06445.
[6] Luca Cotti, Idilio Drago, Anisa Rula, Devis Bianchini e Federico Cerutti. "OntoLogX: Ontology-Guided Knowledge Graph Extraction From Cybersecurity Logs With Large Language Models". In: *Advanced Intelligent Systems*. DOI: https://doi.org/10.1002/aisy.202501381.
[7] OWASP Gen AI Security Project. *Agentic AI Threats and Mitigations*. Accessed 2026-04-27. 2026. URL: https://genai.owasp.org/resource/agentic-ai-threats-and-mitigations/.
[8] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. 2023. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.
[9] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. 2024. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf.
[10] National Institute of Standards and Technology. *The Cybersecurity Framework (CSF) 2.0*. NIST CSWP 29. 2024. URL: https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf.
[11] National Institute of Standards and Technology. *Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile*. Rapp. tecn. NIST IR 8596 (Initial Preliminary Draft). Initial Preliminary Draft. National Institute of Standards e Technology (NIST), dic. 2025. URL: https://csrc.nist.gov/pubs/ir/8596/iprd.
[12] Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang e Xindong Wu. "Unifying Large Language Models and Knowledge Graphs: A Roadmap". In: *IEEE Transactions on Knowledge and Data Engineering* (2024). DOI: 10.1109/TKDE.2024.3352100. URL: https://arxiv.org/abs/2306.08302.
[13] Bret Jordan, Rich Piazza e Trey Darley. *STIX Version 2.1*. OASIS Standard. 2021. URL: https://docs.oasis-open.org/cti/stix/v2.1/os/stix-v2.1-os.html.
[14] MITRE. *MITRE ATT&CK*. Accessed 2026-04-27. 2026. URL: https://attack.mitre.org/.
[15] MITRE. *MITRE ATLAS*. Accessed 2026-04-27. 2026. URL: https://atlas.mitre.org/.
[16] Leslie F. Sikos. "Cybersecurity knowledge graphs". In: *Knowledge and Information Systems* 65.9 (2023), pp. 3511–3531. DOI: 10.1007/s10115-023-01860-3.