Quando l'AI diventa una nuova superficie di rischio

A cura di Francesco Iarlori



Dalla protezione dei sistemi alla governance dell'autonomia: come quarant'anni di evoluzione digitale stanno cambiando ancora una volta la cybersecurity.

Ho iniziato ad interagire con i sistemi informatici quando Unix era uno degli ambienti nei quali si costruiva il futuro dell'informatica. Negli anni Ottanta e nei primi anni Novanta la sicurezza aveva un significato relativamente chiaro: proteggere sistemi, utenti, dati e privilegi, stabilendo chi potesse accedere a una determinata risorsa e quali operazioni fosse autorizzato a compiere.

Con Internet abbiamo dovuto proteggere le reti; con il Web e i servizi digitali abbiamo iniziato a proteggere identità e transazioni; con il cloud abbiamo imparato a governare infrastrutture distribuite che non coincidevano più con un luogo fisico. Ogni trasformazione ha ampliato la superficie di attacco e richiesto nuovi strumenti, nuove competenze e, progressivamente, nuovi modelli organizzativi.

L'Intelligenza Artificiale introduce però qualcosa di diverso. Per la prima volta stiamo inserendo nelle organizzazioni sistemi che non si limitano a conservare informazioni o eseguire istruzioni, ma interpretano il contesto, producono conoscenza, suggeriscono decisioni e, sempre più spesso, possono trasformare queste decisioni in azioni.

Il problema della cybersecurity diventa quindi inevitabilmente anche un problema manageriale. Non dobbiamo più chiederci soltanto chi possa accedere ai nostri sistemi, ma quali decisioni siamo disposti a delegare a sistemi che, una volta ottenuto quell'accesso, possiedono una crescente capacità di interpretare autonomamente ciò che devono fare.

La superficie di attacco diventa organizzativa

Per decenni la cybersecurity è stata considerata prevalentemente una responsabilità delle funzioni IT e Security. Questo modello diventa insufficiente quando l'AI entra direttamente nei processi aziendali e partecipa, ad esempio, alla selezione di un candidato, alla sintesi di un contratto, all'analisi di una richiesta di credito, alla preparazione di una risposta a un cliente o al supporto di una decisione amministrativa.

La sicurezza non riguarda quindi più soltanto l'infrastruttura, ma il processo organizzativo nel quale l'AI viene inserita. Il management deve sapere dove questi sistemi vengono utilizzati, quali informazioni possono consultare, quali decisioni possono influenzare e, soprattutto, dove debba terminare la loro autonomia. In questa prospettiva, la nuova superficie di attacco tende progressivamente a coincidere con l'organizzazione stessa.

Dal punto di vista tecnico, nel software tradizionale



possiamo semplificare il processo attraverso una sequenza relativamente deterministica:
Input → Programma → Output

Con un sistema AI la catena diventa
Input → Contesto → Modello → Interpretazione → Output

e, con l'introduzione degli agenti, può estendersi fino a
Input → Contesto → Interpretazione → Decisione → Azione.

Ogni passaggio aggiunge nuove possibilità di manipolazione. Non dobbiamo quindi proteggere soltanto codice e infrastruttura, ma anche prompt, contesto, fonti informative, modelli, strumenti collegati e output, considerando soprattutto le relazioni che si creano tra questi elementi.

Dal controllo degli accessi al controllo dell'autonomia

Uno dei principi fondamentali della sicurezza informatica è sempre stato il controllo dei privilegi. Negli ambienti Unix degli anni Ottanta e Novanta era immediatamente comprensibile: un utente normale poteva compiere determinate operazioni, mentre un utente privilegiato disponeva di possibilità molto più ampie.

Oggi questo principio assume una dimensione differente. Un agente AI può avere accesso alla posta elettronica, al calendario, al CRM, ai documenti aziendali, ai sistemi ERP e progressivamente anche a workflow operativi. Possiamo aver configurato correttamente ogni singolo permesso e avere comunque creato un rischio, perché il vero problema non riguarda soltanto ciò a cui l'agente può accedere, ma il modo in cui può combinare autonomamente risorse e autorizzazioni differenti.

Per il management questo significa passare dalla governance dei privilegi alla **governance dell'autonomia.**

Dal punto di vista tecnico, un agente può essere perfettamente autenticato, utilizzare API correttamente autorizzate, comunicare attraverso connessioni cifrate e produrre log completi delle proprie attività, generando comunque un comportamento indesiderato attraverso la combinazione legittima di una serie di autorizzazioni.

Potremmo considerarla una nuova forma di **escalation dell'autonomia**: il sistema non acquisisce necessariamente nuovi privilegi, ma utilizza quelli esistenti per raggiungere un risultato che nessuno aveva previsto quando quei privilegi erano stati assegnati.

La conoscenza diventa infrastruttura critica

Per anni abbiamo considerato database, applicazioni e reti come infrastrutture critiche. Con l'AI dobbiamo aggiungere la conoscenza, perché sempre più organizzazioni stanno costruendo sistemi nei quali i modelli interrogano documenti interni, knowledge base, procedure, e-mail, database e architetture RAG per produrre informazioni e supportare decisioni.

Questi strumenti stanno diventando una sorta di memoria operativa dell'organizzazione. Se quella memoria viene alterata, il problema non consiste necessariamente nella perdita o nell'indisponibilità dei dati: le informazioni possono essere perfettamente accessibili e i sistemi tecnicamente funzionanti, mentre l'organizzazione inizia a prendere decisioni sulla base di una rappresentazione distorta della realtà.

Per un CEO, un CIO o un board emerge quindi una nuova domanda di governance: chi garantisce l'integrità della conoscenza utilizzata dall'AI e attraverso quali meccanismi questa integrità viene verificata nel tempo?

A livello tecnico entra in gioco il rischio del data poisoning e, in senso più ampio, di quello che potremmo definire **knowledge poisoning**. Un attaccante potrebbe non aver bisogno di compromettere direttamente il modello, ma potrebbe intervenire sui documenti, sulle informazioni o sulle fonti che alimentano il suo contesto.



Il sistema continuerebbe in questo caso a funzionare normalmente, senza necessariamente produrre anomalie infrastrutturali rilevabili dagli strumenti tradizionali di sicurezza, ma potrebbe arrivare sistematicamente a conclusioni errate. Nel mondo dell'AI, all'integrità del dato dobbiamo quindi aggiungere una dimensione ulteriore: **l'integrità semantica.**

Il linguaggio diventa una superficie di attacco

Esiste un altro cambiamento meno intuitivo, ma particolarmente significativo. Per quarant'anni abbiamo insegnato alle organizzazioni a non aprire allegati sospetti, verificare i link e proteggere password e credenziali; con l'AI, anche un documento apparentemente innocuo può diventare parte di un attacco.

Documenti, e-mail, siti Web e contenuti esterni non devono quindi essere considerati soltanto informazioni da acquisire, perché possono influenzare direttamente il comportamento di un sistema intelligente. Il confine tra comunicazione e comando diventa molto più sottile.

La **prompt injection** rappresenta uno degli esempi più evidenti di questo fenomeno. Nel software tradizionale abbiamo imparato, anche attraverso vulnerabilità come la SQL injection, quanto sia importante mantenere una separazione tra istruzioni e dati. Con i Large Language Model il problema ritorna sotto una forma diversa, perché una frase contenuta in un documento può essere contemporaneamente un dato da analizzare e un'istruzione interpretabile dal modello.

Il linguaggio naturale entra così a far parte della superficie di attacco, estendendo il problema della sicurezza a un territorio nel quale sintassi, semantica e contesto diventano elementi tecnicamente rilevanti.

L'AI industrializza l'ingegneria sociale

Esiste poi il lato offensivo dell'AI. Negli anni Novanta molte truffe digitali erano relativamente facili da riconoscere: e-mail scritte male, siti approssimativi, errori linguistici e richieste improbabili costituivano segnali che permettevano anche a un utente non particolarmente esperto di percepire il rischio.

L'AI sta progressivamente eliminando questi segnali e, soprattutto, sta modificando l'economia dell'attacco. Un cybercriminale non deve necessariamente essere un eccellente programmatore, perché può diventare soprattutto un efficace **orchestratore di strumenti intelligenti**, utilizzando l'AI per creare comunicazioni personalizzate, traduzioni, identità sintetiche, phishing mirato, voce artificiale e deepfake.

Il problema manageriale non è quindi soltanto l'aumento quantitativo degli attacchi, ma la crescita della loro qualità e della loro credibilità. Tecnologie di generazione linguistica, voice cloning, immagini e video sintetici possono essere combinate con informazioni pubblicamente disponibili o precedentemente sottratte all'organizzazione, rendendo scalabile una personalizzazione che in passato richiedeva un significativo lavoro manuale.

In termini economici, l'AI sta riducendo il costo marginale della sofisticazione dell'attacco, permettendo di portare tecniche precedentemente riservate a operazioni mirate verso una scala molto più ampia.

Il nuovo malware potrebbe essere una decisione

Questa è probabilmente una delle trasformazioni più importanti. Per decenni abbiamo cercato software malevolo, identificando virus, worm, trojan e ransomware attraverso le modifiche che producevano sui sistemi o attraverso comportamenti considerati anomali.

In un'organizzazione sempre più dipendente dall'AI potrebbe però non essere necessario compromettere direttamente un sistema. Potrebbe essere sufficiente influenzare una decisione.

Immaginiamo un sistema AI che supporta acquisti, credito, gestione delle risorse umane, investimenti o pianificazione strategica.

Se qualcuno riesce a modificare sistematicamente il contesto utilizzato dal sistema, può alterare le decisioni dell'organizzazione senza necessariamente violare l'infrastruttura sottostante. Gli utenti rimangono autenticati, i server operativi e le API perfettamente funzionanti, ma le conclusioni sulle quali vengono costruite le decisioni possono essere sbagliate.

Questo rende necessario ampliare il concetto stesso di incidente cyber e introduce quella che potremmo definire una forma di **attacco epistemico**, nella quale l'obiettivo non è compromettere ciò che il sistema fa, ma ciò che il sistema crede di sapere.

Data poisoning, manipolazione del contesto, retrieval poisoning, prompt injection e alterazione delle fonti possono diventare strumenti per modificare il processo decisionale senza necessariamente compromettere l'infrastruttura.

Se il malware tradizionale modifica il comportamento del computer, l'attacco epistemico modifica la rappresentazione della realtà sulla quale il sistema costruisce le proprie decisioni.

Dagli assistenti agli agenti: cambia la responsabilità

La differenza tra assistenti e agenti AI dovrebbe entrare nelle discussioni dei board che stanno valutando l'introduzione di sistemi agentici. Un assistente prevalentemente propone o produce un risultato che un essere umano può verificare, mentre un agente può acquisire informazioni, interpretare una situazione, selezionare uno strumento ed eseguire direttamente un'azione.

Finché un sistema prepara una risposta e una persona decide se inviarla, esiste un punto di controllo umano relativamente evidente. Quando invece il sistema può leggere una richiesta, decidere cosa fare, utilizzare uno strumento ed eseguire autonomamente l'azione, il modello di responsabilità cambia radicalmente.

Dobbiamo quindi definire non soltanto il principio di **human in the loop**, ma stabilire dove, quando e per quali decisioni l'essere umano debba effettivamente rimanere nel processo. Non tutte le decisioni hanno infatti lo stesso livello di rischio e, di conseguenza, non tutta l'autonomia dovrebbe essere governata nello stesso modo.

Dal punto di vista tecnico, possiamo rappresentare intuitivamente il rischio attraverso una relazione:
Autonomia × Accesso × Velocità × Scala = Impatto potenziale.

Un errore umano può produrre un'azione sbagliata, mentre un agente può ripetere la stessa azione migliaia di volte prima che qualcuno rilevi l'anomalia. Diventano quindi necessari limiti operativi, segregazione dei privilegi, monitoring comportamentale, approval gate, audit trail e meccanismi tecnici capaci di interrompere rapidamente l'autonomia quando il comportamento supera determinate soglie.

La cybersecurity diventa una responsabilità manageriale

Per questo considero rischioso trattare l'AI security esclusivamente come un nuovo capitolo della sicurezza informatica. Naturalmente abbiamo bisogno di specialisti di cybersecurity e dobbiamo continuare a proteggere modelli, API, infrastrutture, identità e dati, ma questo livello di protezione non è più sufficiente.

Il management deve definire quanta autonomia sia disposto a concedere, quali decisioni possano essere delegate, quali richiedano una validazione umana, quali fonti siano considerate affidabili e, soprattutto, chi risponda quando una decisione algoritmica produce conseguenze inattese.

Non siamo quindi più soltanto nel territorio dell'IT governance. Stiamo entrando progressivamente nella **governance dell'intelligenza organizzativa**, dove tecnologia, responsabilità, conoscenza e capacità decisionale diventano parti dello stesso problema.

Da Zero Trust a Zero Assumption

Zero Trust ci ha insegnato il principio "**Never trust, always verify**", spostando la sicurezza dalla fiducia implicita alla verifica continua dell'identità e delle condizioni di accesso. L'AI potrebbe richiedere un ulteriore passo, che definirei **Zero Assumption**: non assumere che un output sia corretto perché il modello è sofisticato, che un agente stia operando correttamente perché è autenticato, che un'informazione sia affidabile perché proviene da un sistema interno o che un'azione sia sicura semplicemente perché tecnicamente autorizzata.

Soprattutto, non dovremmo assumere che una maggiore intelligenza del sistema corrisponda automaticamente a un maggiore controllo da parte dell'organizzazione.

Guardando agli ultimi quarant'anni emerge infatti una progressione precisa. Abbiamo iniziato proteggendo le macchine, abbiamo esteso la sicurezza alle reti e successivamente alle identità, ai dati e alle transazioni; oggi dobbiamo includere conoscenza e decisioni e, con la progressiva diffusione degli agenti, dovremo imparare a proteggere anche l'autonomia.

È probabilmente in quest'ultimo passaggio che cybersecurity e management finiranno definitivamente per incontrarsi, perché la domanda fondamentale non sarà più soltanto se un sistema AI sia tecnicamente sicuro, ma **quanto potere decisionale siamo disposti ad affidargli, quali confini intendiamo porre alla sua autonomia e come possiamo riprenderne il controllo quando qualcosa non va.**

