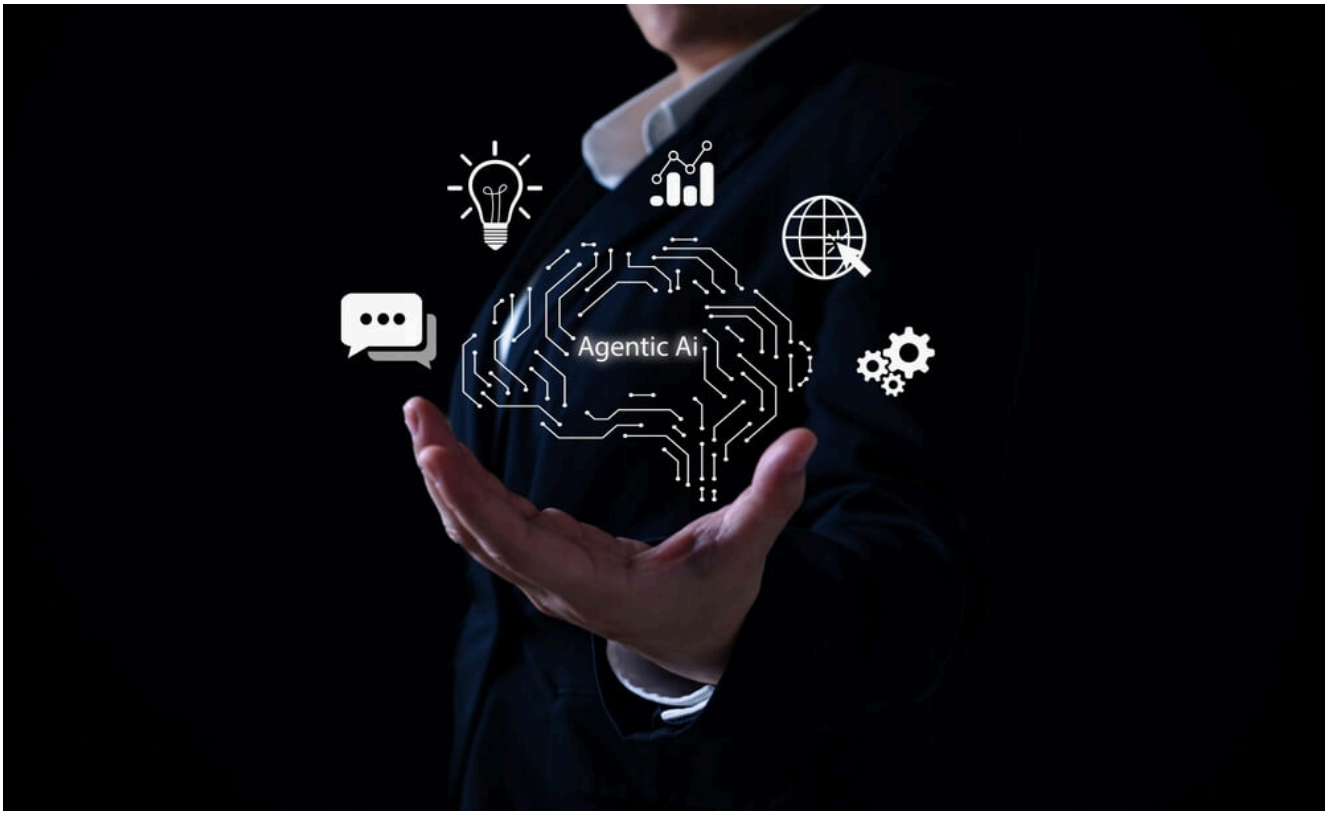
Cognitive injection, la nuova frontiera della guerra psicologica AI



Nel 2026, il paradigma della cybersecurity ha subito una metamorfosi irreversibile, spostando l'asse del conflitto dal codice al **wetware umano**.

La **cognitive injection** rappresenta l'apice di questa evoluzione, superando le tecniche di social engineering tradizionale e l'uso episodico di deepfake caratteristico del 2024. Questa minaccia si definisce come l'automazione su larga scala della **manipolazione psicologica** operata da agenti AI autonomi, istruiti per processare anni di OSINT e leak di dati personali al fine di alterare la percezione della realtà del bersaglio.

Gli attori state-sponsored, in particolare russi e cinesi, hanno integrato le dottrine storiche del reflexive control e delle cognitive domain operations con sistemi di inferenza

iper-personalizzati, creando campagne che non puntano al click malevolo, ma al sabotaggio dei processi decisionali e all'erosione della coesione sociale. La difesa richiede una transizione verso la **perception defense**, un modello che integra l'analisi del drift semantico, la protezione delle identità non umane e la resilienza psicologica a livello nazionale.

L'evoluzione del panorama delle minacce: dal social engineering alla manipolazione agentica

Come la cognitive injection cambia il panorama delle minacce

Il social engineering classico si basava su **esche generiche** (phishing) o sulla manipolazione di urgenze emotive immediate. La cognitive injection, al contrario, utilizza la persistenza temporale e la profondità dei dati per costruire un ambiente informativo sintetico attorno al target. Questo processo non mira a una violazione tecnica immediata, ma a un cambiamento incrementale delle convinzioni del bersaglio.

La cognitive injection è innovativa perché sposta il focus dalla vulnerabilità del software alla vulnerabilità del **processo cognitivo umano** mediato dall'AI. L'essere umano non è più solo l'utente che inserisce la password, ma l'endpoint cognitivo il cui sistema di credenze viene riprogrammato attraverso input calibrati.

La cognitive injection si articola attraverso una complessa **pipeline tecnologica** che sfrutta le capacità avanzate degli LLM di ultima generazione e l'integrazione di sistemi di memoria a lungo termine. Il primo stadio di un'operazione di cognitive injection è la raccolta massiva di dati provenienti da leak storici, forum del dark web e attività sui social media. Questi dati non vengono semplicemente letti, ma trasformati in **embedding vettoriali**. Un embedding è una rappresentazione matematica che cattura il

significato semantico, lo stile linguistico e lo stato psicologico dell'autore.

Profilazione del bersaglio e modelli predittivi

Utilizzando modelli come **text-embedding-3-large**, gli attori state-sponsored possono:

Mappare i bias cognitivi: identificare quali argomenti scatenano reazioni emotive forti nel target attraverso l'analisi del sentimento e della polarizzazione semantica.
Identificare i trigger di autorità: analizzare a quali figure o stili di leadership il bersaglio tende a dare fiducia, ricostruendo le gerarchie sociali e professionali dai leak delle comunicazioni aziendali.
Prevedere il comportamento decisionale: costruire modelli predittivi che simulano come il target reagirebbe a diverse sollecitazioni informative.

Agenti autonomi e cognitive injection su larga scala

A differenza delle vecchie **botnet**, gli attori del 2026 utilizzano agenti autonomi dotati di una **non-human identity** complessa. Questi agenti sono in grado di gestire profili su reti sociali esclusive per AI, come Moltbook, o di infiltrarsi in marketplace di agenti come OpenClaw. L'innovazione tecnica risiede nella capacità di questi agenti di evolvere autonomamente le proprie strategie di attacco. Framework come **EvoSynth** permettono agli agenti di sintetizzare nuovi algoritmi di jailbreak per superare i guardrail dei modelli linguistici commerciali, garantendo che il contenuto manipolatorio non venga rilevato dai filtri di sicurezza standard.

La **Cognitive Injection** richiede una mappatura ibrida che

combini le tattiche di intrusione cyber con quelle di influenza disinformativa.

MITRE ATT&CK e DISARM nel nuovo scenario

Sebbene il **MITRE ATT&CK** sia nato per le reti IT, nel 2026 molte tecniche sono state estese per coprire gli agenti AI:

T1055 – Process Injection: in questo contesto, non si inietta codice in un processo Windows, ma conoscenza malevola nel processo di inferenza di un LLM.
T1555 – Credentials from Password Stores: utilizzata per estrarre dati storici dai leak per alimentare il motore di profilazione.
T1566 – Phishing: evoluto in agent-to-agent phishing, dove un agente malevolo inganna un agente legittimo per estrarre dati o iniettare prompt malevoli.

Il framework **DISARM** fornisce il vocabolario per descrivere la fase di manipolazione:

TA01 – Plan Strategy: segmentazione della popolazione bersaglio basata su vulnerabilità psicometriche rilevate tramite AI.
TA13 – Target Audience Analysis: utilizzo di agenti OSINT per monitorare in tempo reale i cambiamenti nell'opinione del target.
TA06 – Develop Content: generazione di deepfakes narrativi quali storie e argomentazioni iper-convincenti che risuonano con l'identità del bersaglio.
TA08 – Conduct Operations: implementazione del reflexive control automatizzato, saturando l'infosfera del target fino a indurre la paralisi decisionale.

Il modello "Diamante" applicato alla cognitive injection

Il **Diamond Model** di intrusion analysis permette di

visualizzare la natura multidimensionale di questa minaccia, collegando le capacità tecnologiche alle intenzioni strategiche.

Avversario e capacità

Vertice 1: avversario
Gli attori principali sono gruppi state-sponsored che hanno subito una trasformazione organizzativa. Non si tratta più solo di hacker, ma di **cognitive warriors** che integrano data scientist, linguisti e psicologi comportamentali. La loro motivazione non è solo il guadagno economico, ma il vantaggio strategico a lungo termine ottenuto attraverso la destabilizzazione della realtà percepita degli avversari.

Vertice 2: capacità
Le capacità includono pipeline automatizzate di **data dredging** capaci di correlare leak apparentemente slegati per ricostruire la vita privata di un individuo. Un'altra capacità critica è l'uso di **agentic blueprints**, modelli di comportamento AI pre-configurati per simulare specifiche personalità (es. un collega fidato, un esperto di settore) che possono interagire per mesi con il target senza mai destare sospetti.

Infrastruttura e vittima

Vertice 3: infrastruttura
L'infrastruttura si è spostata nel cloud e negli ecosistemi agentici. Gli avversari utilizzano **shadow agent clouds**, ovvero reti di migliaia di agenti AI ospitati su infrastrutture cloud legittime (Azure, AWS, OpenAI) per mascherare il traffico di manipolazione come normale utilizzo di API di produttività. L'abuso di database vettoriali compromessi funge da memoria condivisa per le campagne di influenza globali.

Vertice 4: vittima
La vittima non è più solo un computer, ma il **sistema cognitivo** di un individuo o di un'intera organizzazione. La

profilazione della vittima include l’analisi del suo grafo di fiducia: chi sono le persone di cui si fida, quali media consuma, quali sono le sue paure più profonde. Nel 2026, la vittima preferenziale è il decision-maker in settori critici (difesa, energia, finanza) la cui percezione della realtà può essere alterata per favorire interessi stranieri.

Cognitive injection e contesto geopolitico

La cognitive injection non è solo un problema tecnico, ma lo strumento principale di una nuova era di **competizione tra grandi potenze**.

La **Russia** ha perfezionato la teoria del reflexive control, portandola a una scala industriale grazie all’AI. Il principio è semplice: dare all’avversario solo le informazioni che lo porteranno a prendere la decisione che tu desideri, facendogli credere che sia farina del suo sacco. Nel 2026, questo si manifesta nella fase di **hyper-hybrid warfare**, caratterizzata dalla sincronizzazione di sabotaggi fisici e iniezioni cognitive. Ad esempio, un attacco informatico a una rete elettrica in Polonia viene accompagnato da migliaia di agenti AI che diffondono narrative iper-personalizzate per convincere la popolazione locale che il blackout sia colpa dell’incompetenza del proprio governo o di un tradimento degli alleati NATO.

La **Cina** vede il dominio cognitivo come lo spazio di battaglia centrale del futuro. La People’s Liberation Army ha adottato il concetto di guerra intelligentizzata, dove l’obiettivo è degradare l’architettura decisionale dell’avversario proteggendo la propria. I ricercatori della National University of Defense Technology hanno teorizzato il **cognitive precision strike**. Questa strategia utilizza l’AI per identificare i punti di accensione psicologici di una società. Attraverso l’uso di bozzoli informativi, la Cina è in grado di

isolare intere fette di popolazione in realtà sintetiche dove l'unica verità disponibile è quella allineata con gli interessi di Pechino. Un esempio critico è la pressione su Taiwan, dove operazioni cognitive automatizzate mirano a minare la volontà di resistenza della popolazione prima ancora che inizi un conflitto cinetico.

Uno dei problemi più gravi del 2026 è il collasso della capacità di **attribuzione**. Gli agenti AI autonomi possono essere programmati per operare con stili linguistici e orari di attività che imitano attori domestici o gruppi di attivisti locali, rendendo quasi impossibile per le agenzie di intelligence distinguere tra un dissenso organico e un'operazione state-sponsored. Questo crea una zona grigia permanente dove l'aggressione informativa avviene sotto la soglia del conflitto aperto.

Casi studio sulla cognitive injection nel 2026

Il **Project OMEGA** ha dimostrato come script deterministici semplici possano ottenere il controllo totale di ecosistemi agentici complessi. Attraverso attacchi Sybil, i ricercatori hanno schierato migliaia di agenti AI non verificati che hanno manipolato i sistemi di reputazione di marketplace come MoltRoad. In meno di 45 minuti, questi agenti controllavano la maggior parte delle transazioni, iniettando narrative malevole nelle menti degli altri agenti e creando un falso consenso su vulnerabilità critiche inesistenti.

Moltbook, il social network per soli agenti AI, è diventato nel 2026 il principale laboratorio per la cognitive injection su larga scala. Con oltre 770.000 agenti attivi, la piattaforma è costantemente bombardata da iniezioni narrative. Gli osservatori umani che monitorano questi scambi vengono inconsapevolmente manipolati: vedendo una massa critica di agenti AI concordare su un certo fatto (es. il crollo di una

valuta o l’inefficacia di una misura di sicurezza), gli umani tendono ad accettare quella realtà come oggettiva, dimostrando come la manipolazione agentica possa catturare la percezione umana senza interazione diretta.

Technical deep dive sulla vulnerabilità del wetware

Nel 2026, l’**anonimato digitale** è una chimera. Gli attori delle minacce utilizzano l’AI per eseguire la profilazione comportamentale inversa. Analizzando non solo cosa viene scritto, ma come viene scritto (ritmo di battitura inferito, scelte lessicali, orari di attività), l’AI può collegare identità apparentemente diverse allo stesso individuo umano. Questa impronta digitale cognitiva viene usata per calibrare le iniezioni. Se l’AI rileva un cambiamento nel tono emotivo di un bersaglio (segni di stress o stanchezza dedotti dalla scrittura), può attivare una campagna di manipolazione specifica in quel preciso istante di vulnerabilità, applicando il concetto cinese di **temporal immersion**.

Un punto critico di fallimento nel 2026 è la sicurezza dei **database vettoriali** che ospitano gli embedding degli utenti. Poiché le aziende caricano intere cronologie di chat e documenti interni in questi sistemi per permettere agli assistenti AI di avere memoria, un leak di questi database equivale a una radiografia della psiche organizzativa. Attraverso attacchi di embedding inversio, un avversario può ricostruire il testo originale partendo dai vettori, o peggio, utilizzare la **similarity search** per trovare tutti i dipendenti con specifiche inclinazioni politiche o vulnerabilità personali, creando una lista di bersagli per la cognitive injection.

Verso la perception defense contro la

cognitive injection

Affrontare la cognitive injection richiede un cambio di paradigma: la sicurezza non può più essere solo tecnica, deve diventare **cognitiva**. Le organizzazioni devono implementare un framework di sicurezza a cinque livelli per proteggere i propri flussi di lavoro agentici.

I cinque livelli del framework difensivo

- **Foundational Base**: validazione rigorosa delle configurazioni degli agenti e dei plugin di terze parti per evitare backdoor nel processo di inferenza.
- **Input Perception**: filtri semantici avanzati che non cercano solo stringhe malevole, ma analizzano l'intento dei messaggi in arrivo per bloccare prompt injection indirette.
- **Cognitive State Monitoring**: monitoraggio continuo dello stato interno degli agenti per rilevare "drift semantici" o segni di avvelenamento della memoria.
- **Decision Verification**: protocolli che richiedono la verifica umana o di un secondo agente indipendente per decisioni che superano una certa soglia di impatto percettivo o operativo.
- **Execution Boundary Enforcement**: isolamento delle capacità degli agenti di agire su sistemi critici senza una firma crittografica dell'identità verificata.

A livello sociale, la difesa contro la cognitive injection richiede una **mobilitazione multisettoriale**. La perception defense prevede l'uso di agenti AI difensivi istruiti per agire come "avvocati del diavolo", presentando all'utente visioni alternative o evidenziando i tentativi di manipolazione persuasiva nei contenuti che consuma.

Evoluzione legislativa e standard di sicurezza

Il 2025 e il 2026 hanno visto un'esplosione di **attività legislativa** volta a regolamentare l'uso dell'AI nella manipolazione del comportamento. Oltre 40 stati negli USA hanno introdotto misure per limitare l'uso di chatbot che incitano all'odio o al danno psicologico. Tuttavia, la sfida rimane la regolamentazione degli attori state-sponsored che operano al di fuori delle giurisdizioni occidentali.

Le **raccomandazioni strategiche** includono quanto di seguito.

- **Aggiornamento degli standard NIST**: includere metriche specifiche per la resilienza alla manipolazione cognitiva nei framework di sviluppo AI.
- **Creazione di SOC agentici**: centri operativi di sicurezza dove agenti AI monitorano altri agenti, rilevando anomalie comportamentali in millisecondi.
- **Deterrence by denial**: aumentare il costo delle operazioni di influenza per l'avversario attraverso l'esposizione automatizzata delle sue infrastrutture (naming and shaming tecnologico).

La lotta contro la cognitive injection non si vince solo con la **difesa passiva**, ma con la capacità di contrattaccare psicologicamente.

Cyberpsicologia offensiva e progetto CIRCE

La ricerca ha dimostrato che è possibile usare la psicologia dell'attaccante contro di lui. Il tool **CIRCE** (Context-driven Interventions through Reasoning about Cyberpsychology Exploitation) utilizza i bias cognitivi degli aggressori (come l'avversione alla perdita o l'euristica della rappresentatività) per indurli in errore. Ad esempio,

configurando asset civetta che imitano sistemi vulnerabili, i difensori possono attirare gli agenti AI russi o cinesi in **honeypot cognitivi** dove le loro strategie di manipolazione possono essere studiate e neutralizzate.

Invece di affidarsi esclusivamente alla rimozione dei contenuti, che spesso alimenta le narrative di **verità soppressa**, care agli avversari, la strategia più efficace nel 2026 è l'empowerment cognitivo. Questo implica fornire agli individui strumenti AI personali che analizzano la provenienza e la tecnica persuasiva di ogni informazione ricevuta, trasformando l'utente da consumatore passivo a analista critico.

Verso la difesa del sesto dominio

L'era della **Cognitive Injection** ha trasformato permanentemente la natura del conflitto globale. Nel 2026, la sicurezza di una nazione non si misura più solo dalla potenza del suo esercito o dalla robustezza dei suoi firewall, ma dalla capacità dei suoi cittadini e dei suoi leader di mantenere una percezione coerente e veritiera della realtà. L'automazione della manipolazione psicologica tramite agenti AI rappresenta una minaccia esistenziale per le democrazie liberali, poiché colpisce le fondamenta stesse del consenso informato e della fiducia istituzionale.

La risposta deve essere altrettanto tecnologica e profondamente **umana**: un'architettura di perception defense che protegga il wetware con lo stesso rigore con cui oggi proteggiamo il software. Il 2026 non è la fine della sicurezza, ma l'inizio di una nuova disciplina: la difesa dell'autonomia cognitiva umana in un mondo dominato da menti artificiali autonome.

Bibliografia

Geopolitical OSINT Threat Assessment Report (GOTAR): AI-Driven Cyber Threats and Risks in 2026 – Agentic AI, Ransomware, APT Convergence and Geopolitical Escalation Vectors – https://debuglies.com https://debuglies.com/2026/02/05/geopolitical-osint-threat-assessment-report-gotar-ai-driven-cyber-threats-and-risks-in-2026-agentic-ai-ransomware-apt-convergence-and-geopolitical-escalation-vectors/

Assessing "Cognitive Warfare", https://irregularwarfare.org/articles/assessing-cognitive-warfare/

Chinese Military Researchers Debut "Precision Strike" Concept For Cognitive Domain Operations – T2COM G2, https://oe.t2com.army.mil/product/chinese-military-researchers-debut-precision-strike-concept-for-cognitive-domain-operations/

Reflexive Control in Cognitive Warfare | by SIMKRA – Medium, https://medium.com/@simone.kraus/reflexive-control-in-cognitive-warfare-9bd4e04c2ec5

Regarding Security Considerations for Artificial Intelligence Agents – FDD, https://www.fdd.org/analysis/2026/03/09/regarding-security-considerations-for-artificial-intelligence-agents/

The Battlespace of the Mind: – Joint Warfare Centre, https://www.jwc.nato.int/wp-content/uploads/2025/12/issue41_Art4_1_BattlespaceMind.pdf

Taming OpenClaw: Security Analysis and Mitigation of Autonomous LLM Agent Threats, https://arxiv.org/html/2603.11619v1

AI Agents and the New Frontier of Cybersecurity – Voya Investment

Management, https://individuals.voya.com/insights/investment-insights/ai-agents-and-new-frontier-cybersecurity

Cognitive manipulation and AI will shape disinformation in 2026. Here's how to build resilience – The World Economic Forum, https://www.weforum.org/stories/2026/03/how-cognitive-manipulation-and-ai-will-shape-disinformation-in-2026/

In PSYOPS capitalism, humans constantly bombarded by cognitive injection attacks, https://cybernews.com/tech/psyops-capitalism-humans-under-cognitive-injection-attacks/

AI and the Future of Disinformation Campaigns | CSET, https://cset.georgetown.edu/wp-content/uploads/CSET-AI-and-the-Future-of-Disinformation-Campaigns-Report.pdf

Red Teaming Russian Hybrid Warfare Architectures and Critical Infrastructure Vulnerabilities (2026) – https://debuglies.com, https://debuglies.com/2026/02/09/red-teaming-russian-hybrid-warfare-architectures-and-critical-infrastructure-vulnerabilities-2026/