AI v kibernetski varnosti: nov napadalni vektor in pametnejša obramba

Milan Gabor

Viris, Maribor, Slovenija
milan@viris.si

Umetna inteligenca hitro spreminja področje kibernetske varnosti. Napadalcem omogoča hitrejše in cenejše izvajanje obstoječih tehnik, kot so izvidnica, phishing, ustvarjanje sintetičnih vsebin in prilagajanje zlonamerne kode, hkrati pa sistemi z umetno inteligenco postajajo tudi samostojna napadalna površina. Prispevek predstavi, kako se zaradi tega spreminja napadalna površina organizacij, in ugotovitve ponazori z javno dokumentiranimi primeri – goljufijo s ponarejeno videokonferenco, ranljivostjo EchoLeak v Microsoft 365 Copilot ter vohunsko kampanjo, ki so jo vodili AI agenti. Za razumevanje ključnih tveganj so uporabljene smernice skupnosti OWASP, predvsem seznama Top 10 za aplikacije z velikimi jezikovnimi modeli in za agentne aplikacije. Drugi del prispevka obravnava uporabo umetne inteligence v obrambi – obogatitev in triažo opozoril, zmanjševanje števila lažno pozitivnih opozoril, korelacijo dogodkov med sistemi SIEM, EDR/XDR, oblačnimi okolji in sistemi za upravljanje identitet ter učinkovitejšo prioritizacijo dejansko izpostavljenih tveganj. Umetno inteligenco je zato treba obravnavati hkrati kot tveganje, napadalno površino in novo obrambno zmogljivost.

Ključne besede:
umetna inteligenca
kibernetska varnost
OWASP
SOC
odziv na incidente

Univerzitetna založba Univerze v Mariboru

ISBN 978-961-299-188-3
DOI https://doi.org/10.18690/um.feri.7.2026.10

1 Uvod

Vsak večji tehnološki val zadnjih dveh desetletij – virtualizacija, računalništvo v oblaku, mobilnost in kontejnerji – je razširil napadalno površino organizacij. Umetna inteligenca je pri tem posebna iz treh razlogov: spreminja ekonomiko napadov, uvaja nov razred ranljivosti in bistveno pospešuje posamezne faze napada.

Prvi razlog je ekonomski. Naloge, ki so prej zahtevale izkušenega napadalca in več dni dela – izvidnica o tarči, pripravo verodostojnega sporočila v maternem jeziku žrtve ali prilagajanje izkoriščevalske kode – so danes bistveno cenejše in lažje ponovljive. IBM v poročilu Cost of a Data Breach 2026 ocenjuje, da je bila v opazovanem obdobju približno vsaka četrta zlonamerna kršitev varnosti podprta z umetno inteligenco [1].

Drugi razlog je tehnični. Veliki jezikovni modeli ne ločijo zanesljivo med navodili in podatki, saj oboje prejemajo po istem kanalu, v naravnem jeziku. Gre za strukturno značilnost tehnologije in ne zgolj za napako v implementaciji, kar pojasnjuje, zakaj vbrizgavanje ukazov ostaja na vrhu OWASP-ovega seznama tveganj za aplikacije z velikimi jezikovnimi modeli [2].

Tretji razlog je hitrost. Če napadalec avtomatizira del napadalne verige, se lahko čas med prvim dostopom in odtekanjem podatkov skrajša na nekaj ur. Obramba, ki temelji predvsem na ročni triaži čakalne vrste opozoril, takšnemu tempu težko sledi. Prispevek zato najprej obravnava umetno inteligenco kot napadalno sredstvo in kot tarčo napadov, nato pa še njeno uporabo v varnostno-operativnem centru (SOC).

2 Umetna inteligenca kot napadalni vektor

Izraz »napadalni vektor« se v povezavi z umetno inteligenco uporablja v dveh pomenih. V ožjem pomenu so tarča napada sistemi z umetno inteligenco, kar opišemo v naslednjem poglavju. V širšem pomenu, ki je predmet tega poglavja, pa umetna inteligenca nastopa na strani napadalca kot orodje, ki spreminja način izvedbe napada, ne nujno tudi njegovega cilja.

Umetna inteligenca praviloma ne ustvarja povsem novih vrst napadov, temveč bistveno spreminja njihove značilnosti. Cena posamezne kampanje je nižja, kakovost pa višja, ker modeli odpravljajo številne znake, po katerih so uporabniki nekoč prepoznali lažna sporočila. Izvedba je tudi hitrejša, saj je mogoče posamezne faze prepustiti orodju. S tem se zmanjšuje nekdanji kompromis med številom tarč in stopnjo prilagojenosti – napad, prilagojen konkretni osebi in njeni vlogi v organizaciji, je danes mogoče izvajati v obsegu, ki je bil prej značilen predvsem za množične kampanje.

Za varnostne ekipe je zato ključno vprašanje, katere kontrole v takem okolju še ostajajo učinkovite. Kontrole, ki temeljijo na človeški presoji pristnosti – na primer prepoznavanju slabo napisanega sporočila ali zaupanju znanemu glasu oziroma obrazu na videokonferenci – izgubljajo zanesljivost. Nasprotno pa kontrole, ki temeljijo na preverljivih dejstvih in postopkih – na primer potrditvi po neodvisnem kanalu, omejevanju privilegijev in ločitvi dolžnosti – ostajajo učinkovite. Primeri v nadaljevanju so izbrani zato, ker jasno pokažejo, katera vrsta kontrole v posameznem primeru odpove.

2.1. Socialni inženiring in sintetične vsebine

Phishing ostaja najpogostejši način začetnega vdora. Agencija ENISA v poročilu Threat Landscape 2025 ugotavlja, da predstavlja približno 60 odstotkov opazovanih primerov začetnega vdora, hkrati pa opozarja na hitro rast kampanj, podprtih z generativno umetno inteligenco [3].

Pomemben preskok pa ni le v kakovosti besedila, temveč v uporabi sintetičnega zvoka in videa. Eden najbolj poučnih javno dokumentiranih primerov je goljufija pri inženirskem podjetju Arup. Januarja 2024 je zaposleni v hongkonški pisarni prejel elektronsko sporočilo, ki se je izdajalo za sporočilo finančnega direktorja iz Združenega kraljestva in zahtevalo izvedbo zaupne transakcije. Zaposleni je bil sprva skeptičen, kar je bilo z vidika varnostne

kulture ustrezna reakcija. Dvome pa je odpravil naslednji korak – videokonferenca, na kateri so bili navidezno prisotni finančni direktor in več znanih sodelavcev. Vsi udeleženci razen žrtve so bili ustvarjeni z uporabo umetne inteligence. Sledilo je petnajst nakazil v skupni vrednosti približno 25 milijonov ameriških dolarjev. Goljufijo so odkrili šele ob običajnem naknadnem preverjanju s sedežem podjetja; interni informacijski sistemi pri tem niso bili kompromitirani [4].

Za varnostne ekipe je bistveno naslednje – ni šlo za vdor v informacijski sistem, temveč za zlorabo postopka odobritve. Odpovedala ni tehnična kontrola, ampak predpostavka, da prepoznava obraza in glasu zadostuje kot dokaz identitete. Praktična posledica je jasna – postopki za nakazila nad določenim pragom, spremembe bančnih podatkov in ponastavitve poverilnic morajo vključevati potrditev po kanalu, ki je neodvisen od kanala, po katerem je bila zahteva prejeta.

2.2. Agentne operacije

Novembra 2025 je bila objavljena analiza kampanje z oznako GTG-1002, ki je bila z visoko stopnjo zaupanja pripisana državno sponzorirani skupini. Napadalci so agentno orodje, povezano prek strežnikov MCP, uporabili za izvidnico, iskanje in preverjanje ranljivosti, zbiranje poverilnic ter izločanje podatkov pri približno tridesetih tarčah. Varnostne mehanizme modela so obšli tako, da so nalogo razdelili na manjše korake in modelu predstavili kontekst legitimnega penetracijskega testiranja. Po oceni poročila je bila velika večina taktičnega dela izvedena samodejno, človeški operaterji pa so predvsem potrjevali prehode med posameznimi fazami [5].

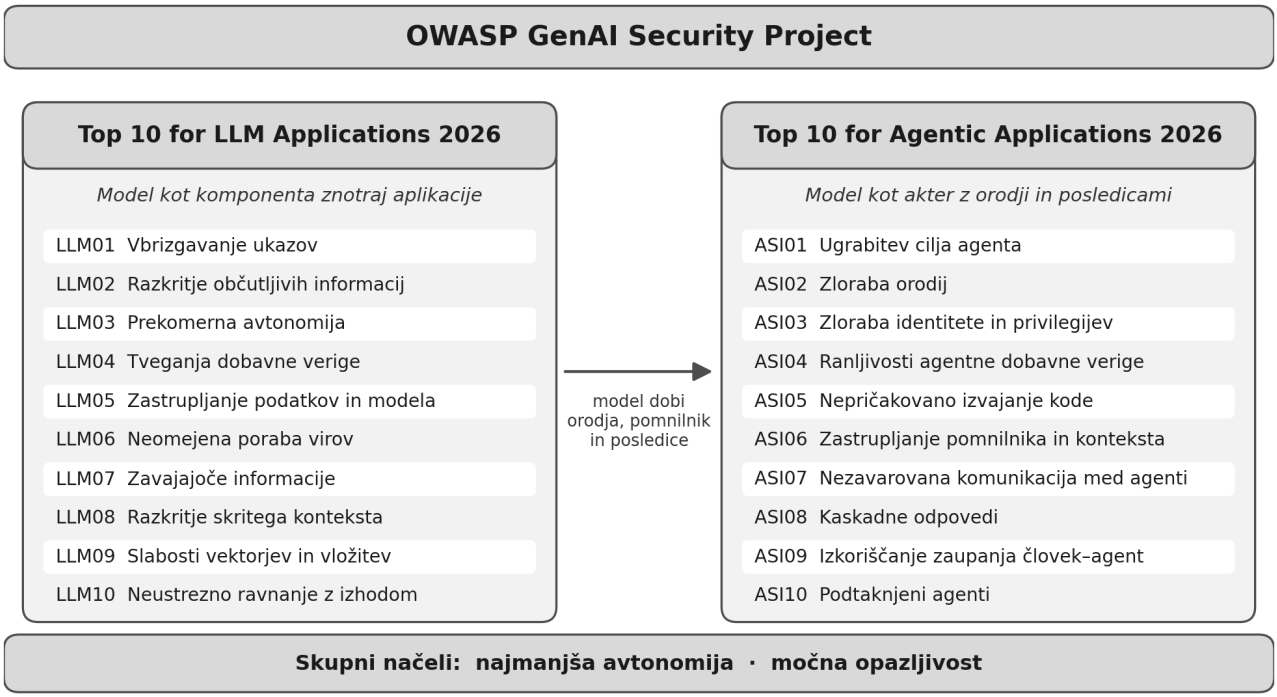
Pri tem primeru je pomembno upoštevati omejitve vira – gre za poročilo ponudnika, indikatorji kompromitacije niso bili javno objavljeni, del strokovne javnosti pa dvomi o oznaki »prvi« in o dejanskem obsegu kampanje [6]. Kljub temu ostaja bistvena ugotovitev uporabna – agentna orodja lahko večfazno operacijo izvedejo hitreje kot ročno voden napad, s čimer se časovno okno za obrambo dodatno krajša.

3 Napadi na sisteme z umetno inteligenco in smernice OWASP

Prejšnje poglavje je umetno inteligenco obravnavalo kot orodje napadalca, to poglavje pa jo obravnava kot tarčo napada. Eden najbolj uveljavljenih referenčnih okvirov na tem področju je OWASP GenAI Security Project.

Projekt ne vključuje le seznamov tveganj. Pokriva več delovnih področij – od obveščevalnih podatkov o grožnjah in upravljanja varnosti umetne inteligence, prek varnega uvajanja in varnosti podatkov, do varnosti agentnih aplikacij, red teaminga in vrednotenja. Organizacijam tako poleg seznamov tveganj ponuja tudi gradiva za upravljanje, uvajanje in varnostno preverjanje rešitev.

Projekt vzdržuje dva ločena seznama tveganj, njuno razmejitev pa prikazuje slika 1. Prvi obravnava model kot komponento v aplikaciji, drugi pa sisteme, v katerih model dobi dostop do orodij, pomnilnika in možnosti izvajanja dejanj ter s tem postane aktiven akter.



Slika 1: Področji uporabe obeh seznamov OWASP GenAI Security Project. Prirejeno po [2, 8].

3.1. OWASP Top 10 for LLM Applications

Izdaja seznama za leto 2026 vključuje tveganja, prikazana v tabeli 1. Seznam je uporabno izhodišče za pregled arhitekture, oblikovanje testnih scenarijev in določanje kontrolnih točk pred uvedbo aplikacije v produkcijsko okolje.

Tabela 1: Tveganja iz seznama OWASP Top 10 for LLM Applications 2026.

Oznaka	Tveganje	Kratek opis	Primer napada
LLM01:2026	Vbrizgavanje ukazov (angl. Prompt Injection)	Napadalec z namensko oblikovanim vhodom vpliva na navodila modela in ga skuša pripraviti do tega, da prezre sistemska pravila ali izvede neželeno dejanje.	Uporabnik v poziv vnese navodilo, naj model prezre prejšnja pravila in razkrije informacije ali izvede dejanje, ki sicer ni dovoljeno.
LLM02:2026	Razkritje občutljivih informacij (angl. Sensitive Information Disclosure)	Model lahko razkrije zaupne, osebne, interne ali druge občutljive informacije, ki so prisotne v kontekstu, učnih podatkih ali povezanih sistemih.	Napadalec z zaporedjem vprašanj iz modela pridobi dele internih dokumentov, osebne podatke ali druge zaupne informacije, vključene v kontekst pogovora.
LLM03:2026	Prekomerna avtonomija (angl. Excessive Agency)	Model ima preširoke pravice za uporabo orodij, API-jev ali izvajanje dejanj brez ustreznega nadzora in omejitev.	Napadalec z vbrizgavanjem ukazov doseže, da model samodejno pošlje e-pošto, spremeni podatke ali izvede administrativno dejanje v imenu uporabnika.

Oznaka	Tveganje	Kratek opis	Primer napada
LLM04:2026	Tveganja dobavne verige (angl. Supply Chain)	Ranljivosti ali zlonamerne komponente se lahko pojavijo v modelih, knjižnicah, podatkovnih virih, vtičnikih ali drugih zunanjih komponentah rešitve z umetno inteligenco.	Organizacija uporabi kompromitiran model ali knjižnico tretje osebe, ki vsebuje zlonamerno funkcionalnost ali omogoča uhajanje podatkov.
LLM05:2026	Zastrupljanje podatkov in modela (angl. Data and Model Poisoning)	Napadalec spremeni učne podatke, podatke za dodatno učenje ali druge vire in s tem vpliva na vedenje modela.	V podatkovno zbirko se vnesejo posebej pripravljeni zapisi, zaradi katerih začne model pri določenih poizvedbah vračati napačne ali napadalcu koristne odgovore.
LLM06:2026	Neomejena poraba virov (angl. Unbounded Consumption)	Pomanjkanje omejitev omogoča prekomerno uporabo modela, procesorske moči, pomnilnika, žetonov ali plačljivih API-jev.	Napadalec avtomatizira veliko število kompleksnih zahtev in s tem povzroči visoke stroške uporabe API-ja ali nedosegljivost storitve.
LLM07:2026	Napačne oziroma zavajajoče informacije (angl. Misinformation)	Model lahko ustvari napačne, izmišljene ali zavajajoče informacije, ki jih uporabnik zmotno razume kot zanesljive.	Model ustvari neobstoječo pravno zahtevo, ranljivost, referenco ali tehnični postopek, uporabnik pa odgovor uporabi brez dodatnega preverjanja.
LLM08:2026	Razkritje skritega konteksta (angl. Hidden Context Exposure)	Uporabnik lahko pridobi informacije iz sistemskih navodil, skritega konteksta ali drugih podatkov, ki niso namenjeni neposrednemu prikazu.	Napadalec oblikuje poziv, s katerim model razkrije del sistemskega poziva, interna navodila ali skrite podatke, uporabljene pri pripravi odgovora.
LLM09:2026	Pomanjkljivosti vektorskih predstavitev in vdelav (angl. Vector and Embedding Weaknesses)	Pomanjkljiva zaščita vektorskih baz in vdelav lahko omogoči manipulacijo rezultatov iskanja, nepooblaščen dostop ali uporabo napačnega konteksta v sistemih RAG.	Napadalec v zbirko RAG vnese dokument z zlonamernimi navodili, ki se zaradi podobnosti pri iskanju vključijo v kontekst in vplivajo na odgovor modela.
LLM10:2026	Neustrezno obravnavanje izhodnih podatkov (angl. Improper Output Handling)	Izhod modela se brez ustreznega preverjanja posreduje drugim sistemom ali obravnava kot zaupanja vredna koda oziroma ukaz.	Model ustvari HTML, SQL ali ukaz operacijskega sistema, aplikacija pa ga brez ustreznega preverjanja izvede, kar lahko povzroči XSS, SQL injection ali izvajanje ukazov.

Vir: [2]

3.2. Posredno vbrizgavanje ukazov: primer EchoLeak

Junija 2025 je bila razkrita ranljivost v rešitvi Microsoft 365 Copilot z oznako CVE-2025-32711 in oceno CVSS 9,3. Napadalec je žrtvi poslal navidezno običajno elektronsko sporočilo, v katerem so bila zlonamerna navodila skrita v komentarjih HTML oziroma v uporabniku neopaznem besedilu. Ko je uporabnik pozneje pomočniku zastavil povsem nepovezano vprašanje, je mehanizem RAG sporočilo vključil v kontekst, model pa je skrita navodila obravnaval kot ukaz in podatke iz uporabnikovega konteksta posredoval prek zunanjega vira. Napad ni zahteval nobene interakcije uporabnika s sporočilom. Ranljivost je bila odpravljena na strani ponudnika, izkoriščanje v realnih napadih pa ni bilo potrjeno [9], [10].

Primer je posebej poučen, ker gre za posredno vbrizgavanje ukazov. Vir zlonamernega navodila ni uporabnik, temveč vsebina, ki jo sistem sam pridobi. Vsaka arhitektura RAG, ki v kontekst vključuje nepreverjene zunanje vsebine – na primer elektronsko pošto, zahtevke uporabniške podpore, spletne strani ali deljene dokumente – je takšnemu tveganju izpostavljena že po zasnovi.

3.3. OWASP Top 10 for Agentic Applications

Decembra 2025 je pobuda OWASP Agentic Security Initiative objavila ločen seznam za agentne aplikacije, torej za sisteme, ki načrtujejo naloge, kličejo orodja, uporabljajo pomnilnik in delujejo z lastnimi poverilnicami. Seznam vključuje deset tveganj z oznakami ASI01 do ASI10 – ugrabitev cilja agenta, zloraba orodij, zloraba identitete in privilegijev agenta, ranljivosti agentne dobavne verige, nepričakovano izvajanje kode, zastrupljanje pomnilnika in konteksta, nezavarovana komunikacija med agenti, kaskadne odpovedi, izkoriščanje zaupanja med človekom in agentom ter podtaknjeni agenti [8].

Dokument pred posameznimi tveganji izpostavi dve temeljni načeli. Najmanjša avtonomija je agentna različica načela najmanjših privilegijev – omejiti je treba ne le, do katerih virov agent dostopa, temveč tudi, koliko svobode ima pri ukrepanju brez človeške potrditve. Podroben nadzor pa zahteva, da je v vsakem trenutku mogoče ugotoviti, kaj agent počne, zakaj to počne in s katero identiteto deluje [8]. Skupni imenovalec večine teh tveganj je zato mogoče razumeti predvsem kot vprašanje identitete in avtorizacije, preneseno v novo okolje.

3.4. Umetna inteligenca v senci

Ta kategorija tveganja ni tehnična ranljivost, temveč predvsem vprašanje upravljanja. Delež varnostnih incidentov, povezanih z neodobreno uporabo orodij umetne inteligence, po zadnjih meritvah hitro narašča, večina organizacij pa še nima vzpostavljenih ustreznih postopkov za nadzor takšne uporabe [1]. Učinkovita zaščita zato ne more temeljiti le na pravilnikih in izobraževanju, temveč tudi na tehničnih ukrepih, kot so nadzor izhodnega prometa, kategorizacija storitev in ponudba odobrene alternative.

4 Umetna inteligenca v obrambi

Sodobni varnostno-operativni center ima praviloma dovolj senzorjev in virov telemetrije za učinkovito zaznavanje dogodkov. Glavna omejitev ni več količina podatkov, temveč sposobnost njihove presoje in povezovanja. Če se delež lažno pozitivnih opozoril po raziskavah približuje polovici vseh zaznav [11], analitiki preprosto nimajo dovolj časa, da bi vsak signal ustrezno obogatili s kontekstom. Prav na tem področju lahko umetna inteligenca prispeva največ.

4.1. Triaža in zmanjševanje šuma

Ena najbolj zrelih uporab umetne inteligence v SOC je obogatitev in triaža opozoril, še preden jih pregleda analitik. Model ali agent z umetno inteligenco ob prejemu opozorila samodejno pridobi dodatni kontekst – lastnika in kritičnost sredstva, zgodovino uporabnika, ugled naslova, prisotnost iste datoteke drugod v okolju ter povezane zahtevke – nato pa pripravi predlog razvrstitve z obrazložitvijo in dokazi.

Ključna odločitev pri zasnovi je, da umetna inteligenca v začetni fazi uvajanja opozoril ne zapira samodejno. Deluje kot analitik prve ravni, ki pripravi kontekst in predlog, končno odločitev pa potrdi človek (angl. Human in the Loop). Samodejno zapiranje je smiselno uvesti šele za dobro razumljen in ozko opredeljen razred opozoril, pri katerem so rezultati preverjeni na dovolj velikem vzorcu ter sta zagotovljena vzorčenje in revizijska sled.

4.2. Korelacija med sistemi

Največja dodana vrednost umetne inteligence ni v analizi posameznega vira, temveč v povezovanju šibkih signalov med sistemi, ki jih pogosto spremljajo različne ekipe. Tabela 2 prikazuje poenostavljen scenarij, sestavljen iz vzorcev, ki se v praksi redno pojavljajo.

Tabela 2: Šibki signali, ki posamično ne dosežejo praga za preiskavo.

Čas	Vir	Signal	Posamična ocena
09:14	Upravljanje identitet	Uspešna prijava z MFA iz nove države	nizka
09:20	Oblačna storitev	Uporabnik odobri dovoljenje OAuth aplikaciji tretje osebe	informativno
09:26	Poštni sistem	Ustvarjeno pravilo za posredovanje e-pošte na zunanji naslov	nizka
10:02	EDR	Zagon zakrite skripte PowerShell na delovni postaji	srednja
10:40	Požarna pregrada	Odhodni promet proti domeni, registrirani pred tremi dnevi	nizka

Vsak od teh zapisov se lahko sam zase izgubi v množici opozoril. Ko jih povežemo, pa dobimo vzorec prevzema uporabniškega računa, vzpostavljanja obstojnosti in priprave na odtekanje podatkov, značilen za kompromitacijo poslovne elektronske pošte. Ključna je korelacija po identiteti in časovnem oknu, ne zgolj po posameznem viru. Prav pri tej nalogi lahko jezikovni model dobro dopolni statična pravila – poveže heterogene zapise, jih povzame v razumljivo zgodbo in predlaga naslednje korake preverjanja.

4.3. Ocenjevanje tveganja in prioritizacija ranljivosti

Za organizacijo je praktično najpomembnejše vprašanje, katere ranljivosti so dejansko izpostavljene, in ne zgolj njihovo skupno število. Ocena CVSS sama po sebi ne pove, ali je prizadeti sistem dosegljiv z interneta, ali obstaja javno dostopna izkoriščevalska koda, ali je sredstvo na poti do kritičnih podatkov in ali so vzpostavljene kompenzacijske kontrole. Umetna inteligenca je lahko posebej koristna pri združevanju podatkov o ranljivostih, izpostavljenosti, obveščevalnih podatkov o grožnjah in poslovnega konteksta.

Ob tem se pokaže pomembno neskladje. Organizacije agente v varnostnih procesih pogosto uvajajo za iskanje groženj in odzivanje na incidente, precej redkeje pa za upravljanje ranljivosti, torej za dejavnost, ki dejansko zapira vstopne točke za napadalce [12]. Znane ranljivosti zato lahko ostajajo odprte dlje, medtem ko se čas med objavo ranljivosti in njenim izkoriščanjem krajša.

4.4. Meje uporabnosti

Velja preprosto pravilo – umetna inteligenca ne more nadomestiti kakovostnih podatkov. Model brez zanesljive inventure sredstev, evidence lastništva sistemov, klasifikacije podatkov in normaliziranih dnevnikov bo hitro ustvaril prepričljive, vendar napačne sklepe. Vlaganje v kakovostno podatkovno osnovo je zato predpogoj za učinkovito uporabo umetne inteligence, ne pa alternativa zanjo.

5 Ko obrambna umetna inteligenca postane tarča

Vsako orodje z umetno inteligenco, vključeno v SOC, postane tudi novo sredstvo, ki ga je treba ustrezno zaščititi. Pri tem je treba upoštevati predvsem štiri skupine tveganj.

Vbrizgavanje ukazov prek telemetrije

Če agent med preiskavo bere vsebino elektronskega sporočila, ime datoteke, vrednost polja User-Agent ali drugo vsebino opozorila, obdeluje nepreverjen vhod. Napadalec lahko vanj vstavi navodilo, namenjeno agentu in ne človeku.

Prekomerna avtonomija

Agent, ki lahko izolira končno točko, blokira uporabniški račun ali spreminja pravila požarne pregrade, je privlačna tarča. Takšne pravice morajo biti strogo omejene po obsegu in času ter vezane na lastno identiteto agenta, ne na skupni servisni račun.

Prekomerno zaupanje rezultatom

Tekoče in samozavestno napisan povzetek preiskave lahko deluje bolj verodostojno, kot upravičuje njegova dejanska zanesljivost. Vmesnik mora zato jasno ločevati med preverjenimi dejstvi iz telemetrije in sklepi, ki jih je pripravil model.

Uhajanje podatkov v model

Telemetrija SOC lahko vsebuje osebne podatke in poslovne skrivnosti. Pri uporabi zunanjih ponudnikov je zato treba jasno urediti hrambo podatkov, njihovo morebitno uporabo za učenje in lokacijo obdelave. V Evropski uniji je to tudi vprašanje skladnosti s Splošno uredbo o varstvu podatkov, direktivo NIS2 in Aktom o umetni inteligenci.

6 Priporočila

Spodnja priporočila izhajajo neposredno iz obravnavanih primerov in smernic skupnosti OWASP. Urejena so po smiselnem vrstnem redu uvajanja: najprej ukrepi, ki ne zahtevajo nove tehnologije, temveč predvsem evidenco in jasno razmejitev odgovornosti; nato ukrepi na ravni arhitekture aplikacij z umetno inteligenco; nazadnje pa ukrepi v poslovnih procesih in varnostno-operativnem centru. Priporočila so namenjena organizacijam ne glede na velikost varnostne ekipe, saj večina ni tehnološko zahtevna. Gre predvsem za dosledno uporabo že znanih varnostnih načel v novem okolju. Smiselno jih je najprej uporabiti kot kontrolni seznam za oceno trenutnega stanja in šele nato kot osnovo za načrt ukrepov, saj se pri prvem pregledu pogosto pokaže, da je največja vrzel v evidencah in procesih, ne v tehničnih kontrolah.

Evidenca in preglednost

Organizacija mora vzpostaviti evidenco sistemov in agentov z umetno inteligenco, ki vključuje vsaj lastnika, namen, vire podatkov, uporabljena orodja in dostopne pravice. Brez takšne evidence ni mogoče učinkovito ocenjevati tveganj niti se odzivati na incidente. Uporabnikom je smiselno ponuditi odobreno alternativo, uporabo neodobrenih storitev pa tehnično spremljati in omejevati.

Ločitev navodil od podatkov

Vsako zunanjo vsebino, ki vstopa v kontekst modela, je treba obravnavati kot nepreverjen vhod, kanal za navodila pa čim bolj jasno ločiti od kanala za podatke.

Najmanjša avtonomija

Nabor orodij in pravic mora biti omejen na nujno potrebno, avtorizacijo pa je treba preveriti ob vsakem klicu orodja. Dejanja z velikim vplivom morajo zahtevati človeško potrditev.

Red teaming

Aplikacije je treba preizkušati s scenariji iz obeh OWASP-ovih seznamov, vključno s posrednim vbrizgavanjem ukazov in zastrupljanjem pomnilnika oziroma konteksta.

Procesi, odporni na sintetične vsebine

Za dejanja z velikim poslovnim ali varnostnim vplivom, kot so nakazila, spremembe bančnih podatkov in ponastavitve poverilnic, mora biti uvedena potrditev po neodvisnem kanalu. Identitete ni dovoljeno potrjevati zgolj na podlagi glasu ali videa. Scenarije s sintetičnimi vsebinami je smiselno vključiti tudi v vaje odzivanja na incidente.

Detekcija

Smiselno je spremljati odhodne povezave do končnih točk ponudnikov modelov iz procesov, ki niso brskalnik ali odobreno razvojno orodje. Tak promet je lahko neposreden pokazatelj neodobrene uporabe orodij umetne inteligence in uporaben indikator za nadaljnjo preiskavo.

Postopno uvajanje v SOC

Uvajanje naj poteka po korakih. Najprej obogatitev opozoril, nato predlog razvrstitve s človeško potrditvijo, samodejno zapiranje pa šele pri dobro izmerjenih in ozko opredeljenih razredih opozoril. Pred uvedbo in po njej je smiselno spremljati delež lažno pozitivnih opozoril, delež nepreiskanih opozoril ter povprečen čas do triaže in zajezitve incidenta.

7 Zaključek

Umetna inteligenca v kibernetski varnosti ni niti zgolj orodje napadalcev niti univerzalna rešitev obrambnih težav. Je nova plast tehnološke realnosti, ki jo mora organizacija hkrati obravnavati kot tveganje (uhajanje podatkov, neodobrena uporaba, prekomerno zaupanje rezultatom), kot napadalno površino (vbrizgavanje ukazov, zloraba orodij, agentna dobavna veriga) in kot obrambno zmogljivost (triaža, korelacija in prioritizacija tveganj).

Dokumentirani primeri – od goljufije s sintetično videokonferenco in posrednega vbrizgavanja ukazov do agentno vodene vohunske kampanje – kažejo, da ne gre več le za napoved, temveč za operativno realnost. Hkrati potrjujejo, da so učinkovite kontrole večinoma dobro znane – potrditev po neodvisnem kanalu, načelo najmanjših privilegijev, ločevanje navodil od podatkov in kakovostna revizijska sled. Umetna inteligenca teh načel ne razveljavlja, spreminjajo se predvsem mesta in načini, na katerih jih je treba uveljaviti.

Najuspešnejše bodo tiste varnostne ekipe, ki bodo znale povezati kakovostne podatke, avtomatizacijo, umetno inteligenco in strokovno presojo ljudi. Vrstni red ni naključen. Brez kakovostne podatkovne osnove lahko preostali elementi ustvarjajo predvsem hitre in prepričljive napake. Tudi ob vse večji avtomatizaciji zato človeška presoja ostaja nujen del učinkovite kibernetske obrambe.

Literatura

[1] IBM, Cost of a Data Breach Report 2026, IBM Security in Ponemon Institute, julij 2026. Dostopno na: https://www.ibm.com/reports/data-breach (obiskano 8. 8. 2026).

[2] OWASP GenAI Security Project, OWASP Top 10 for LLM Applications 2026, avgust 2026. Dostopno na: https://genai.owasp.org/resource/owasp-genai-llm-top-10-2026/ (obiskano 8. 8. 2026).

[3] ENISA, ENISA Threat Landscape 2025, Agencija Evropske unije za kibernetsko varnost, oktober 2025. Dostopno na: https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025 (obiskano 8. 8. 2026).

[4] CNN Business, »Arup revealed as victim of $25 million deepfake scam involving Hong Kong employee«, 16. 5. 2024. Dostopno na: https://www.cnn.com/2024/05/16/tech/arup-deepfake-scam-loss-hong-kong-intl-hnk (obiskano 8. 8. 2026).

[5] Anthropic, Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign, november 2025. Dostopno na: https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf (obiskano 8. 8. 2026).

[6] Just Security, »The Era of AI-Orchestrated Hacking Has Begun«, januar 2026. Dostopno na: https://www.justsecurity.org/127053/era-ai-orchestrated-hacking/ (obiskano 8. 8. 2026).

[7] OWASP GenAI Security Project, pregled delovnih področij projekta. Slika dostopna na: https://genai.owasp.org/wp-content/uploads/2025/07/OWASP-GenAI-Security-Project-right-img.png (obiskano 8. 8. 2026).

[8] OWASP GenAI Security Project, OWASP Top 10 for Agentic Applications 2026, december 2025. Dostopno na: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/ (obiskano 8. 8. 2026).

[9] REDDY P., GUJRAL A. S., »EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System«, arXiv:2509.10540, september 2025. Dostopno na: https://arxiv.org/abs/2509.10540 (obiskano 8. 8. 2026).

[10] NIST National Vulnerability Database, CVE-2025-32711. Dostopno na: https://nvd.nist.gov/vuln/detail/CVE-2025-32711 (obiskano 8. 8. 2026).

[11] Microsoft in Omdia, State of the SOC, 2026 (povzeto po javnih objavah industrijskih analiz).

[12] Help Net Security, »Data breach cost 2026 averaged $4.99 million, AI attacks ran higher«, 30. 7. 2026. Dostopno na: https://www.helpnetsecurity.com/2026/07/30/ibm-cost-of-a-data-breach-2026/ (obiskano 8. 8. 2026).