

# Vpliv umetne inteligence na kibernetsko (ne)varnost oziroma kdo je danes najšibkejši člen

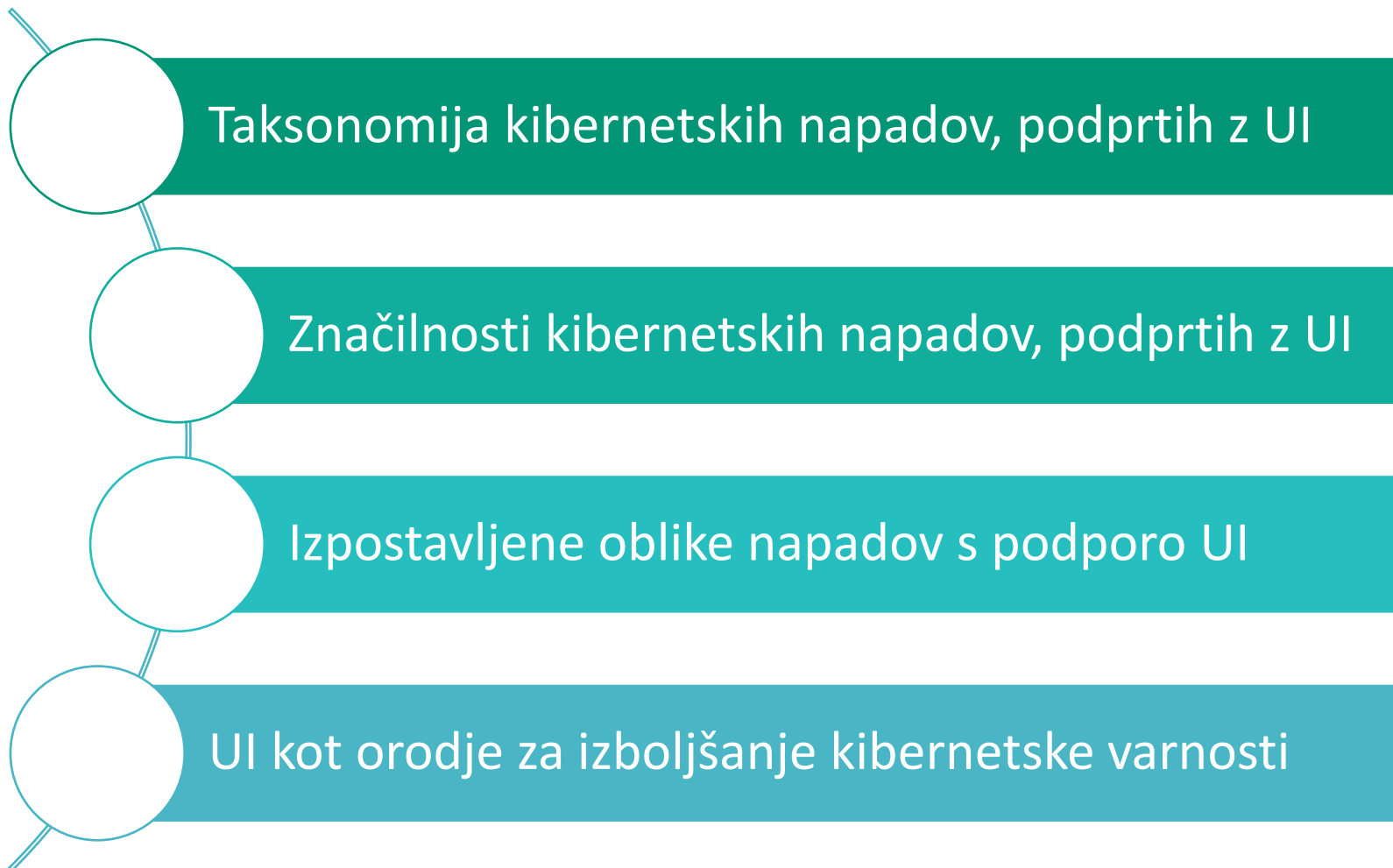
---

Jaka Kosmač, univ. dipl. prav., državni revizor, CISA

---

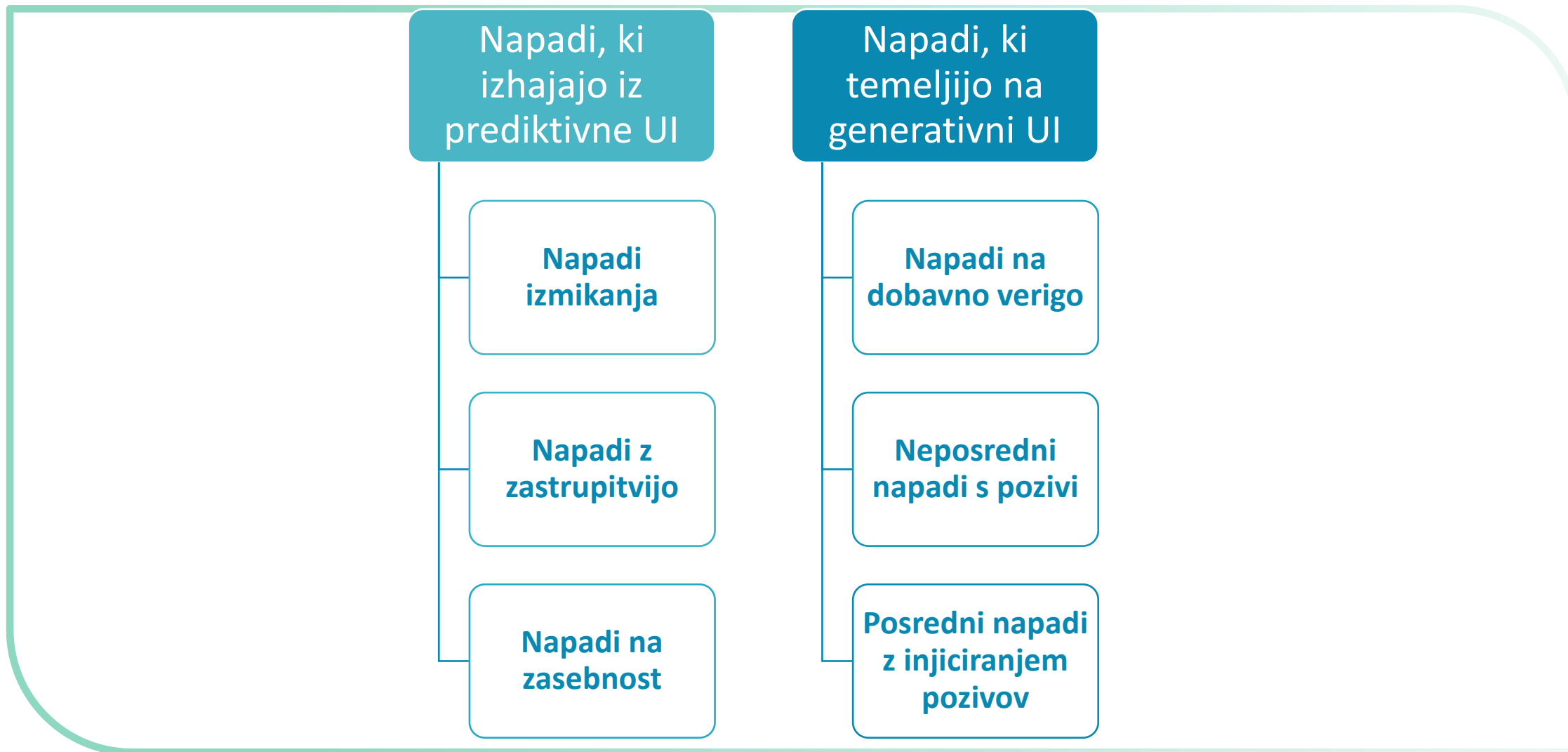
Gradivo je last Slovenskega inštituta za revizijo in je predmet avtorske zaščite in drugih oblik zaščite intelektualne lastnine. Prepovedano je kakršnokoli reproduciranje, razen izključno za osebno uporabo in v nekomercialne namene, pri čemer se morajo ohraniti vsa opozorila o avtorskih ali drugih pravicah, zato se ne smejo prepisovati, razmnoževati ali kako drugače razširjati. Naveden mora biti tudi vir.

- McKinsey: UI **uporablja >75 % organizacij** (skok s 55 % v letu dni).
- Harvard Business Review: **UI zmanjšuje stroške napadov** (npr. spear phishing), ob tem ohranja ali celo zvišuje uspešnost.
- Deep Instinct: **97 %** strokovnjakov za kibernetško varnost se boji, da se bodo njihove organizacije **soočile z varnostnimi incidenti, ki jih povzroča UI.**
- Netacea: **93 % podjetij** pričakuje, da se bodo v naslednjem letu soočala z **dnevnimi napadi UI.**
- IBM: povprečen vdor leta 2024 stal **4,88 mio USD**; z varnostno UI in avtomatizacijo prihranek **2,22 mio USD.**



# Taksonomija kibernetskih napadov, podprtih z UI

# NIST – taksonomija kibernetičskih napadov z UI



# Napadi izmikanja (angl. *Evasion Attacks*)

- **Napadi z znanjem o modelu (angl. *White-Box Evasion*):** napadalec pozna strukturo in parametre modela UI ter ustrezno oblikuje zavajajoče vhodne podatke.
- **Napadi brez znanja o modelu (angl. *Black-Box Evasion*):** napadalec ne pozna modela UI, a s poskusi odkrije, kako nanj vplivati.
- **Prenosljivost napadov:** napad, ustvarjen na enem modelu UI, se uporabi še na drugih podobnih modelih UI.
- **Napadi v resničnem svetu:** denimo prepoznavanje znakov v prometu, kjer preoblečen znak zmede sistem avtonomnega vozila.

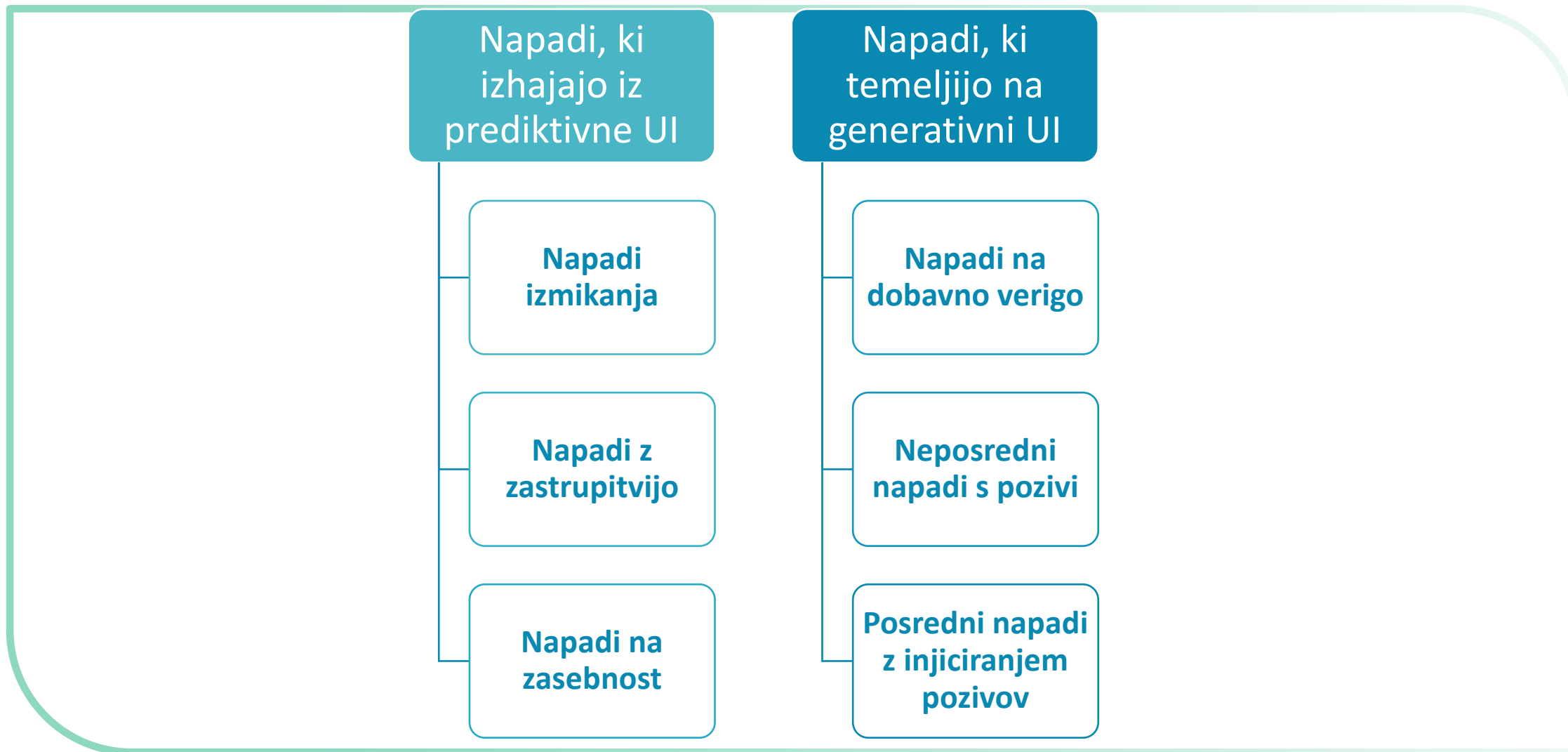
# Napadi z zastrupitvijo (angl. *Poisoning Attacks*)

- **Zastrupitev za onemogočanje (angl. *Availability Poisoning*):** zmanjšanje točnosti modela UI.
- **Ciljno usmerjena zastrupitev (angl. *Targeted Poisoning*):** model UI napačno klasificira točno določen primer.
- **Zastrupitev z vgrajenimi stranskimi vrati (angl. *Backdoor Poisoning*):** napadalec vgradi skrite sprožilce v model.
- **Zastrupitev modela:** napadalci poskušajo neposredno spremeniti že naučen model UI, tako da vanj vgradijo zlonamerno funkcionalnost.

- **Rekonstrukcija podatkov (angl. *Data Reconstruction*):** napadi z rekonstrukcijo podatkov omogočajo pridobitev posameznikovih podatkov iz javno objavljenih agregiranih informacij.
- **Sklepanje o članstvu (angl. *Membership Inference*):** napadalec ugotovi, ali so bili določeni podatki uporabljeni za učenje strojnega modela. Napadalec lahko razkrije osebne podatke posameznika (podobno kot pri napadih z rekonstrukcijo podatkov) – resno tveganje pri objavi agregiranih podatkov ali modelov, ki so jih usposabljali na uporabniških podatkih (primer študije redkih bolezni).
- **Sklepanje o lastnostih (angl. *Property Inference*):** odkrivanje statističnih lastnosti učne množice.
- **Ekstrakcija modela (angl. *Model Extraction*):** kopiranje modela ali pridobivanje njegovih parametrov.



# NIST – taksonomija kibernetičnih napadov z UI



- **Zastrupitev podatkov:** vključevanje pristranskih ali škodljivih podatkov v učne baze.
- **Zastrupitev modela:** manipulacija s samim modelom ali vgrajevanje vrinjenih stranskih vrat.

- **Pridobivanje informacij:** pridobivanje občutljivih podatkov (npr. gesel) iz modela.

- **Napadi na razpoložljivost:** povzročijo preobremenitev sistema ali nedelovanje.
- **Napadi na integriteto:** izkrivljanje ali spreminjanje rezultatov.
- **Ogrožanje zasebnosti:** nenamerno razkrivanje občutljivih informacij.

# Značilnosti kibernetских napadov, podprtih z UI

**Avtomatizacija  
napadov**

**Učinkovito  
zbiranje  
podatkov**

**Prilagajanje**

**Učenje s  
pomočjo  
okrepite**

**Ciljanje na  
zaposlene**

# Izpostavljene oblike napadov, podprtih z UI

**Socialni inženiring**

**Spletno ribarjenje**

**Globoki ponaredek**

**Nasprotovalna  
umetna  
inteligenca/strojno  
učenje**

**Zlonamerni GPT-ji**

**Napadi z  
izsiljevalskimi  
virusi**



## Napadalci:

- prepoznajo in določijo **idealno tarčo** (podjetje ali posameznika znotraj organizacije).
- **razvijejo identiteto** za komunikacijo s tarčo;
- **oblikujejo** možen in prepričljiv **scenarij**, s katerim bodo pritegnili pozornost;
- ustvarijo **personalizirana sporočila** ali multimedijske vsebine, kot so zvočni ali videoposnetki, za vzpostavitev stika s tarčo.

- Generativna UI se uporablja za ustvarjanje personaliziranih in prepričljivih sporočil (e-pošta, SMS, klici, družbena omrežja).
- Glavni namen napadov ostaja **pridobivanje občutljivih informacij, dostopa do sistemov/finančnih sredstev** ali **prepričati uporabnike, da odprejo zlonamerne datoteke ali kliknejo nevarne povezave**.
- V naprednejših primerih UI omogoča **avtomatizirano komunikacijo v realnem času** preko klepetalnih robotov (chatbotov), ki delujejo skoraj enako kot resnični ljudje.
- Kot „**Služba za pomoč uporabnikom**“ –zbirajo osebne podatke, poverilnice ali celo pridobijo nepooblaščen dostop do naprav in sistemov.
- **Europol**: Dostopnost generativne UI znižuje tehnični prag za izvedbo napadov – zdaj jih lahko izvajajo tudi neizkušeni akterji z uporabo javno dostopnih modelov in preprostih pozivov.

- Nasprotovalna umetna inteligenca: napadi z manipulacijo ali napačnimi podatki **zmanjšujejo točnost UI/ML sistemov.**
- Tehnike ciljajo na **različne faze razvoja in delovanja modelov.**
- Glavne vrste napadov: **zastrupljanje podatkov, izmikanje ter spreminjanje modela.**

- Zlonamerni GPT je **spremenjena različica generativnega modela, ki ustvarja škodljive ali napačne vsebine.**
- Uporablja se pri **kibernetskih napadih za generiranje zlonamerne kode, spletnega ribarjenja in lažnih vsebin.**
- Leta 2023 je podjetje Netenrich odkrilo **FraudGPT, orodje UI oglaševano na temnem spletu za kriminalne aktivnosti.**
- FraudGPT (kot WormGPT) kriminalcem omogoča **pisanje prevarantskih pisem, ribarjenje, ustvarjanje škodljivih kod in iskanje ranljivosti.**

- Globoki ponaredki so videi, slike ali zvoki, ustvarjeni z UI za zavarjanje ljudi.
- Dezinformacije, lažne novice, blatenje osebnosti ali kibernetске napade.
- **Akt o UI** – pravila o preglednosti pri uporabi globokih ponaredkov.
- Uvajalci UI sistemov morajo jasno razkriti, da je vsebina umetno ustvarjena, razen v primerih kazenskih postopkov ali umetniških del.
- Globoki ponaredki se **pogosto uporabljajo v kampanjah socialnega inženiringa**, kjer napadalci posnemajo glas ali podobo oseb za prevaro zaposlenih (npr. prenos denarja ali dostop do sistema).

# Nekateri izpostavljeni primeri globokih ponaredkov\*

\*Vir: Breacher.ai

- Napadalci so uporabili **realnočasovno zamenjavo obraza** (angl. *face-swap*) med video klicem.
- Žrtev je verjela, da **govori s poznano osebo**.
- Med pogovorom je bila podana **zahteva za takojšnje finančno nakazilo**.
- Izguba: približno 4,3 milijona juanov (**622 000 USD**).
- Primer dokazuje ranljivost spletnih sestankov brez dodatnih varnostnih preverjanj.
- Študije kažejo, da **57 % ljudi misli, da lahko prepoznajo globoke ponaredke**, vendar le **24 % dejansko prepozna dobro izdelane ponaredke**.

- Glas direktorja je bil **kloniran z uporabo par sekund posnetka.**
- Napadalci so klicali in **posnemali njegov ton, naglas in način govora.**
- Prepričali so zaposlenega v **nujno** plačilo dobavitelju.
- Nakazilo je znašalo okoli **220 000 EUR.**
- Kloniranje glasu je hitro, poceni in težko zaznavno.



# Milijonski škandal z globokim ponaredkom CFO-ja

- Napadalci so pripravili **več globokih ponaredkov avatarjev zaposlenih.**
- V videokonferenco so vključili tudi **ponarejenega CFO-ja.**
- Udeleženci so prejeli **navodila za izvedbo bančnih transakcij.**
- Izguba: skupno približno **25 milijonov USD.**
- Napad kombinira globoke ponaredke z naprednim socialnim inženiringom.

- Premier Trudeau **uporabljen v lažnem oglasu** za “robot traderja”.
- Videoposnetek je bil razširjen prek **YouTuba in Facebooka**.
- Namen: **prepričati ljudi v vlaganje v prevaro**.
- Ena žrtev izgubila okoli **12.000 USD**.
- Primer kaže na **nevarnost globokih ponaredkov političnih osebnosti**.

- Napadalci **analizirajo poslovne procese in navade vodstva.**
- Uporabijo video-in glasovne globoke ponaredke v sestankih.
- Dodajo **kontekst** (npr. nujni projekt, kratki roki).
- Cilj: **pridobiti zaupanje in sprožiti hitro odločitev.**
- **Čas napada:** ko je ekipa pod pritiskom ali v stresu.

- Nadgradnja klasične prevare **vdor v poslovno komunikacijo** (angl. *Business Email Compromise*).
- Napadalci uporabljajo **generativno UI** za **ustvarjanje verodostojnih sporočil**.
- Sporočila vključujejo osebne in vsebinske podatke o podjetju.
- Cilj: **sprožiti nujno finančno transakcijo**.
- Dodana taktika: **ustvarjanje občutka časovne stiske**.

# Manipulacija na družbenih omrežjih z globokimi ponaredki

- Globoki ponaredki se širijo prek bot omrežij.
- Vključujejo ponarejene videe, glasove ali izjave.
- Bot omrežja zagotavljajo množično deljenje in lažno verodostojnost.
- Namen: vplivati na javnost, politike ali blagovne znamke.
- Gre za dolgoročno obliko informacijske vojne.

# Umetna inteligenca kot orodje za izboljšanje kibernetске varnosti

Napredno  
zaznavanje in  
preprečevanje  
groženj

Avtomatiziran  
odziv na incidente

Obveščanje o  
grožnjah in  
napovedovanje

Sistemi za analizo  
vedenja  
uporabnikov in  
entitet

Upravljanje  
ranljivosti in  
določanje prioritet

Avtomatizacija  
VOC

Zaznavanje  
globokih  
ponaredkov in  
identitetnih prevar

- **UI:** omogoča zaznavanje in preprečevanje sofisticiranih groženj (npr. »zero-day«), kjer klasični sistemi odpovejo (ENISA, NIST).
- **Delovanje:** strojno učenje (npr. nevronske mreže, SVM) analizira velike podatkovne množice – promet, dnevnike, vedenjske vzorce, končne točke – ter prepoznava subtilne anomalije.
- **Učinek:** boljše odkrivanje mutacijskih/polimorfnih groženj in krepitev odpornosti varnostnih sistemov.



- **UI:** poleg zaznavanja groženj omogoča avtomatiziran in pospešen odziv na incidente (ENISA, NIST).
- **Delovanje:** samodejna izolacija naprav, blokada IP-jev, karantena datotek in avtomatizirano odpravljanje ranljivosti.
- **Učinek:** odziv v sekundah/minutah preprečuje eskalacijo, zmanjšuje škodo in odvisnost od kadra, ki ga primanjkuje.

- **UI:** preoblikuje sisteme obveščanja o grožnjah v proaktivna in prediktivna orodja.
- **Delovanje:** uporablja NLP in analizo omrežij za povezovanje varnostnih dogodkov in napadalcev ter analizira podatke (dnevniki, temni splet, globalni bilteni).
- **Učinek:** napovedovanje vektorjev napadov, ocena tveganj in omogočanje proaktivne krepitve obrambe ter optimizacije strategij.

- **UI:** sistemi za analizo vedenja uporabnikov in entitet zaznavajo notranje grožnje in kompromitirane račune, ki jih klasične rešitve pogosto spregledajo.
- **Delovanje:** UI oblikuje in posodablja profile normalnega vedenja (čas prijave, lokacija, dostop do virov, obseg prometa) ter sproži opozorilo ob odstopanjih.
- **Učinek:** zgodnje zaznavanje kompromitiranih računov in zlonamernih aktivnosti notranjih akterjev kot dodatni sloj obrambe.

- **Izziv:** kompleksna IT-okolja otežujejo ročno določanje prioritet ranljivosti, kar je počasno in nagnjeno k napakam.
- **UI:** avtomatizirano analizira ranljivosti, konfiguracije in grožnje ter določa kritičnost in optimalne prioritete za odpravljanje.
- **Učinek:** prepoznavanje vzorcev, predlog strateških sprememb in zmanjšanje splošne napadalne površine organizacije za učinkovitejšo obrambo.

- **UI** in strojno učenje zmanjšujeta obremenitve analitikov, preprečujeta izgorelost in omogočata boljše zaznavanje resnih groženj.
- **Delovanje:** UI prevzame rutinska opravila – zbiranje in korelacija zapisov, filtriranje lažnih pozitivnih rezultatov, priprava poročil, začetna analiza incidentov – analitikom se osredotočijo na strateške naloge.
- **Rezultat:** večja učinkovitost in odzivnost VOC ter izboljšana splošna varnostna drža organizacije.

# Zaznavanje globokih ponaredkov in identitetnih prevar

- **Izziv:** eksponentna rast generativne UI je prinesla nevarnost globokih ponaredkov in identitetnih prevar, ki temeljijo na realističnih, a lažnih multimedijskih vsebinah.
- **UI:** ključna pri razvoju sofisticiranih rešitev za zaznavanje teh manipulacij. Z globokim učenjem lahko algoritmi analizirajo subtilne artefakte, nedoslednosti v gibanju, osvetlitvi, mimiki obraza ali intonaciji glasu, ki so značilni za umetno generirane vsebine. To orodje je ključno v boju proti različnim vrstam dezinformacij.
- **Učinek:** preprečevanje dezinformacij in zlorab pri verifikaciji identitete, zaščita digitalnih ekosistemov ter skladnost z zahtevami EU Akta o UI za razkritje umetno ustvarjenih vsebin.

# Kdo je danes najšibkejši člen?

- **UI in kibernetika varnost:** UI izboljšuje zaznavanje in odzivanje na grožnje, hkrati omogoča napadalcem bolj sofisticirane metode. Najpogostejša tarča ostaja človek.
- **Vloga vodstva:** Vodstvo organizacij mora jasno komunicirati pomen kibernetike varnosti, uvajati jasne politike ter zaposlene aktivno vključevati v procese zaščite.
- **Celovit pristop:** Prava moč varnosti je v kombinaciji tehnologije in kibernetike kulture – zaposleni postanejo prvi obrambni zid, če so dobro usposobljeni in podprti z varnostno ozaveščeno organizacijsko kulturo.
- **Najšibkejši člen** ni posameznik sam po sebi, temveč človek, kadar ga organizacija pusti neustrezno usposobljenega in ranljivega.

