NACIONALINIS KIBERNETINIO SAUGUMO CENTRAS

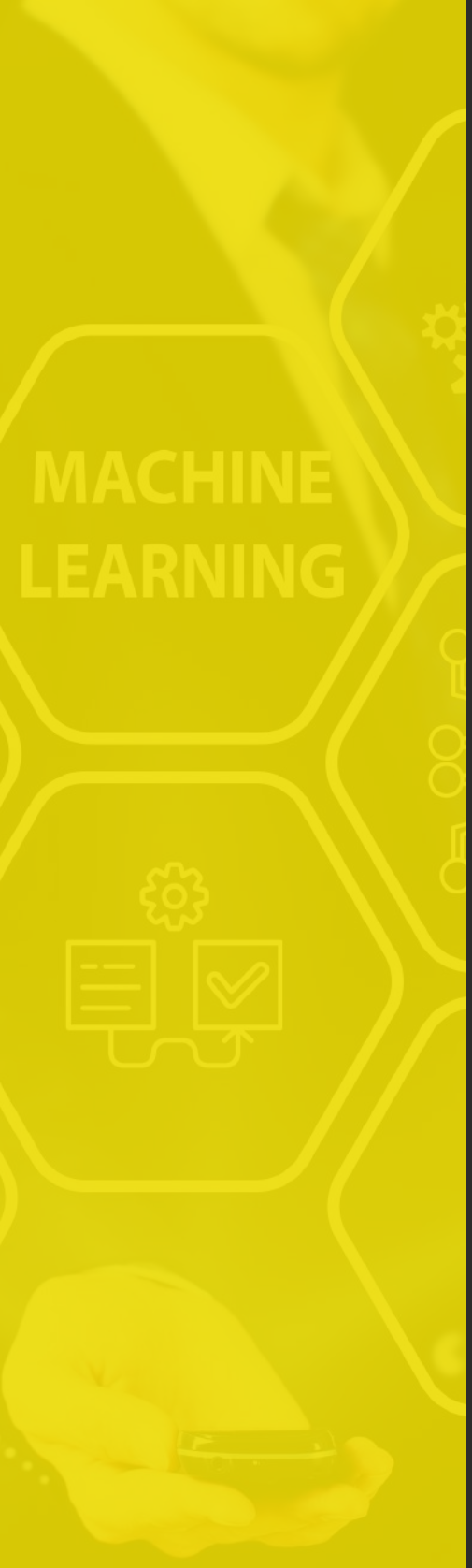
REKOMENDACIJOS SAUGIAM DIRBTINIO INTELEKTO (DI) SPRENDIMŲ NAUDOJIMUI ORGANIZACIJOJE

LAPKRITIS

2025

Bendrai finansuoja Europos Sąjunga

Bendrai finansuojama Europos Sąjungos lėšomis. Tačiau išsakytos nuomonės ir požiūriai yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos kibernetinio saugumo kompetencijų centro (ECCC) požiūrį. Europos Sąjunga ir dotaciją teikianti institucija nėra atsakingos už šią informaciją.



Turinys

NKSC rekomendacija

SAUGUS DIRBTINIO INTELEKTO TAIKYMAS ORGANIZACIJOJE

Rekomendacijų tikslas

Nors teorijoje dirbtinio intelekto (angl. *Artificial Intelligence* arba AI) (toliau – DI) pritaikymo galimybės egzistuoja jau ne vieną dešimtmetį, tačiau pastaraisiais metais reikšmingai išaugus duomenų apdorojimo greičiui, duomenų prieinamumui ir algoritmų pažangai praktinis DI pritaikymas įgavo naują pagreitį.

Šių **rekomendacijų tikslas** yra supažindinti organizacijas su esminiais DI technologijos principais, pateikti glaustą informaciją apie DI sistemų keliamas grėsmes ir joms valdyti rekomenduojamas taikyti saugumo kontrolės priemones.

Auditorija

Šios rekomendacijos yra orientuotos į kibernetinio saugumo subjektus ir kitas organizacijas, savo veikloje naudojančias DI sprendimus.

Populiarėjantys DI sprendimai ir kylantys iššūkiai

Per pastaruosius metus DI sprendimai reikšmingai pažengė į priekį, suteikdami galimybes kaip niekada anksčiau optimizuoti įvairių organizacijų veiklos procesus, tokius kaip: klientų aptarnavimą, logistiką, medicininę diagnostiką ar kitus. DI, ypač generatyvinio, sprendimai vis dažniau taikomi veikloms, kurias anksčiau atlikdavo išskirtinai žmonės, pavyzdžiui, didelių duomenų rinkinių sisteminimas, rutininių ar kūrybinių užduočių vykdymas, tačiau kaip ir bet kuri kita ankstyvosios stadijos technologija, taip ir DI, kelia iššūkius kibernetiniam saugumui.

Sparčiai populiarėjantys DI sprendimai tampa patraukliu taikiniu kibernetiniams nusikaltėliams. Piktavaliai siekia pasinaudoti DI sprendimų pažeidžiamumu, siekdami gauti konfidencialią organizacijų informaciją ar pasinaudoti DI sprendimais dezinformacijos sklaidai, kenkėjiškų kodų kūrimui ar automatizuotoms atakoms; taip pat nusikaltėliai dažnai bando pasisavinti privačius DI sprendimų modelius, taip siekdami išvengti modelio kūrimo ir apmokymo investicijų.

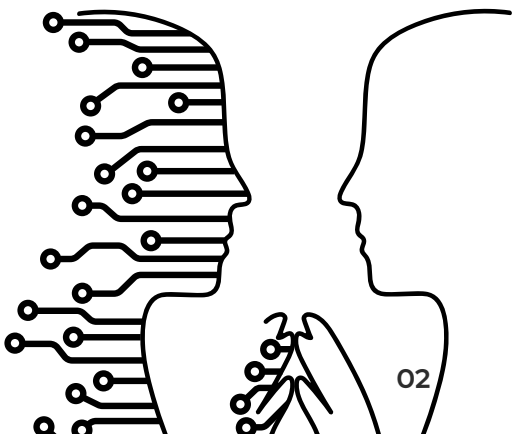
Nors tarptautiniu mastu kai kurios DI reguliavimo iniciatyvos dar tik įsigalioja, pavyzdžiui, 2024 m. Europos Sąjungos priimtas „Dirbtinio intelekto reglamentas“ (angl. *Artificial Intelligence Act)*, tačiau tokios tarptautinės organizacijos, kaip: International Organization for Standardization (ISO) ir International Electrotechnical Commission (IEC), Open Web Application Security Project (OWASP), National institute of technology (NIST), Nacionalinis kibernetinio saugumo centras prie Krašto apsaugos ministerijos (NKSC) bei Cybersecurity and infrastracture security agency (CISA) – jau keletą metų sėkmingai formuoja DI sprendimų saugumo, rizikos valdymo ir diegimo standartus bei gerąją praktiką, kuria remiamasi šiame leidinyje. Šiame leidinyje teikiamos rekomendacijos papildo jau 2024 metais NKSC išleistas „Generatyvinio dirbtinio intelekto (GenDI) saugaus naudojimo organizacijoje“ gaires.

DI sprendimų veikimo principai

Dirbtinis intelektas - plati sąvoka, apimanti keletą DI šakų ir liečianti visas informacinių technologijų sritis, kurios suteikia galimybę mašinoms atlikti užduotis, reikalaujančias žmogaus intelekto. Plačiai žinomi generatyvinio DI sprendimai dažniausiai yra kuriami naudojant mašininio mokymosi algoritmus (angl. M*achine Learning* arba ML).

> Esminis DI proveržis natūralios kalbos apdorojimo srityje įvyko dar 2017 m., kai buvo pristatytas VASWANI et al. mokslinis straipsnis pavadinimu „Attention is All You Need“, pasiūlęs transformerių architektūrą – pagrindą dabartiniams generatyvinio DI sprendimams.

DI populiarumą dar labiau paskatino viešoje erdvėje pasirodęs „ChatGPT“ produktas, kuris ir privatiems asmenis, ir įmonėms suteikė galimybę nemokamai naudotis jau apmokytais dideliais kalbos modeliais (angl. *Large Language Model* arba LLM) (toliau – LLM). Tai, kad daug resursų apmokymui pareikalavęs modelis tapo lengvai prieinamu ir nemokamu, ypač prisidėjo prie staigaus DI produktų ir sprendimų populiarumo augimo.



Pagrindinės DI šakos ir sritys:

Mašininis mokymas

(angl. Machine Learning)

Pagrindinė DI šaka, orientuota į algoritmų kūrimą, leidžianti DI sistemoms mokytis iš duomenų. Įprastai ML algoritmai apdoroja įvesties duomenis (pvz., tekstą, vaizdus, skaitines reikšmes) ir, taikydami tam tikrus duomenų apdorojimo ir vertinimo algoritmus, sukuria modelį, kuris geba prognozuoti norimą išvestį. Mokymo metu algoritmas analizuoja duomenų rinkinius, kuriuose kiekvienai įvesčiai yra priskirta teisinga išvestis (t. y. žinoma „teisinga“ atsakymo reikšmė). Toks ML mokymo būdas vadinamas prižiūrimu mokymu (angl. *Supervised learning*). Tokie duomenys leidžia sistemai „išmokti“ atpažinti sudėtingus ryšius, dėsningumus ar struktūras tarp įvesties ir išvesties duomenų. Ištreniruotas modelis vėliau gali būti taikomas naujiems, nematytiems duomenims prognozuoti ar vertinti. Mašininiu mokymu yra pagrįsta dauguma šių dienų DI sprendimų, tarp kurių populiariausias ML panaudojimas yra LLM modeliuose, tokiuose kaip GPT (angl. *Generative Pre-trained Transformer*);

Neuroniniai tinklai

(angl. NN arba Neural Network)

Tai matematinis ir programinis skaičiavimo modelis, sukurtas pagal biologinę žmogaus smegenų neuroninių tinklų struktūrą ir funkcijas. Būtent neuroniniai tinklai yra ypač dažnai naudojami vaizdo atpažinimo (angl. *Computer vision*), natūralios kalbos apdorojimo, generatyvinio DI gilaus mokymosi srityse;

Natūralios kalbos apdorojimas

(angl. NLP / Natural Language Processing)

DI sritis, kuri siekia suprasti, analizuoti ir generuoti žmogaus kalbą. NLP sistemomis apdorojama informacija iš teksto ar žodinės kalbos šaltinių, atliekama teksto analizė, klasifikavimas, vertimas, informacijos paieška bei vykdomos turinio generavimo užduotys;

Generatyvinis DI
(angl. GenAI arba GenDI)

Sistemos, kurios naudoja duomenų modelius generuojant naują teksto, paveikslėlių, garso įrašų, kodo ar kitokių duomenų turinį. GenDI taikomosios programos yra mokytos, naudojant didelius realaus pasaulio duomenų kiekius, ir gali prognozuoti žmogaus sukurtą turinį iš įvesties, net jei įvestis yra ne visa ar netiksli. Daugiau informacijos apie GenDI taip pat galima rasti „Generatyvinio dirbtinio intelekto (GenDI) saugaus naudojimo organizacijoje gairėse“;

Didieji Kalbos Modeliai
(angl. LLM arba Large Language Model)

Pažangūs neuroniniais tinklais paremti modeliai, gebantys analizuoti, suprasti ir generuoti tekstinį turinį. Jie mokomi pasitelkiant didžiulius tekstinių duomenų kiekius, todėl yra pajėgūs apdoroti kalbos struktūrą, semantiką bei kontekstą. Šie modeliai gali atlikti įvairias kalbos užduotis – versti, apibendrinti, atsakyti į klausimus, generuoti tekstą ar net programinį kodą. Vieni iš žinomiausių LLM pavyzdžių – ChatGPT (OpenAI), Mistral AI (Le Chat), Claude (Anthropic), Gemini (Google), LLaMA (Meta).

Rinkoje tarp populiariausių DI sistemų dominuoja LLM modeliai, kurie galėtų būti skirstomi į šias kategorijas – atvirojo kodo (angl. Open-source) ir licencijuotus.

Atvirojo kodo

Bendruomenės, tyrimų institucijų arba didelių kompanijų (pvz., Meta (LLaMA)) kuriami modeliai, kurių pirminis kodas ir mokymo architektūra yra viešai prieinami. Šiuos modelius galima modifikuoti, papildomai treniruoti bei pritaikyti specifiniams kiekvieno naudotojo poreikiams. Tokie modeliai dažnai yra pasitelkiami, kai organizacijai yra svarbu išlaikyti duomenų privatumą ir visiškai kontroliuoti modelio elgseną, svorius bei kitus parametrus.

Licencijuoti

Komerciniams tikslams sukurti ir gamintojų valdomi modeliai, pvz., Mistral AI (Le Chat), GPT (OpenAI) ar Claude Sonnet (Anthropic). Šie DI sprendimai dažniausiai pasiekiami per žiniatinklio sąsają, aplikacijos programavimo sąsają (angl. API arba *Application Programming Interface*) ar debesijos paslaugas. Dažnai šie LLM modeliai nereikalauja techninių išteklių, siūlo aukštą tikslumą bei užtikrina universalų ir nesudėtingą naudojimą, tačiau gali rinkti ir kaupti naudotojų įvedamus duomenis.

Grėsmės naudojant DI

Susipažinus su pagrindiniais DI sprendimų principais ir norint saugiai naudotis DI sprendimais, organizacijos privalo įvertinti visas aktualias grėsmes, kylančias jų organizacijos informaciniams ištekliams.

NAUDOJANT LICENCIJUOTUS DI SPRENDIMUS:

Neautorizuota prieiga per DI tiekimo grandinę
(angl. ML supply chain compromise)

Piktavaliai gali įgyti prieigą prie organizacijos tinklų ir informacinių sistemų (TIS), pažeisdami specifinius DI tiekimo grandinės komponentus, kurie gali apimti aparatinę įrangą, treniravimo duomenis, DI programinės įrangos paketus ar patį modelį;

NAUDOJANT POPULIARIUS SPRENDIMUS:

Šešėlinis DI
(angl. Shadow AI)

Darbuotojai gali nesankcionuotai naudoti viešai prieinamas DI paslaugas, naršyklės įskiepius ar trečiųjų šalių programas, taip iš dalies apeidami organizacijos taikomas TIS saugumo kontrolės priemones, bet kartu ir padidindami duomenų nutekėjimo bei kenkėjiškų atakų grėsmes;

Generatyvinio DI haliucinacijos
(angl. Hallucinations)

DI sugeneruoti atsakymai gali būti netikslūs arba visiškai išgalvoti ir kelti pavojų jais priimamų autonominių sprendimų pagrįstumui;

Neviešų duomenų atskleidimas

Į DI sprendimus įvedama nevieša informacija (pvz., asmens duomenys, organizacijos komercinės paslaptys) gali būti įrašyta, panaudota mokant kitą modelio versiją ir vėliau atsitiktinai atskleista trečiosioms šalims.

LOKALIUS ATVIROJO KODO AR ORGANIZACIJOS APMOKYTUS MODELIUS:

Duomenų užnuodijimas
(angl. Data poisoning)

Duomenų, naudojamų treniravimui, sąmoningas pakeitimas arba papildymas klaidingomis įrašų reikšmėmis, dėl ko modelis išmoksta neteisingas taisykles ir gali pateikti šališkus ar kenksmingus rezultatus;

Modelio vagystė
(angl. Model stealing)

Ataka, kai piktavalis teikia specialiai parengtas užklausas DI sprendimui ir pagal gautus atsakymus siekia sukurti DI sprendime taikomo modelio kopiją. Tokiu būdu yra siekiama pasisavinti intelektinę nuosavybę – DI sprendimo modelį, pavyzdžiui, konkurentas atkartoja unikalų draudimo bendrovės DI sprendime naudojamą rizikos vertinimo modelį ir taip siekia išvengti reikšmingų DI sprendimo modelio kūrimo ir mokymo investicijų.

Įvesties manipuliacija
(angl. Prompt injection)

Piktavaliai į DI sprendimus teikiamas užklausas įterpia kenkėjiškas komandas arba paslėptus nurodymus, siekdami apeiti saugumo filtrus ir išgauti įprastai nepasiekiamą informaciją;

Besivystant ir daugėjant DI naudojimo ir pritaikymo galimybėms verslo ir viešame sektoriuje, su DI susijusių grėsmių sąrašas nuolat plečiasi, todėl šis grėsmių sąrašas nėra baigtinis. Be tinkamos kontrolės, DI sprendimai gali susilpninti organizacijos kibernetinį atsparumą ir lemti kibernetinius incidentus, todėl yra svarbu, kad organizacijos, kurios naudoja arba svarsto galimybę naudoti DI sprendimus, įvertintų jų galimą poveikį organizacijos kibernetiniam saugumui. Norintiems apsisaugoti nuo šių grėsmių patariama vadovautis *„DI sprendimų naudojimui taikomos kibernetinio saugumo kontrolės priemonės“* skyriuje nurodyta informacija.

Pagrindiniai žingsniai integruojant DI sistemą

Kaip ir kiekvienas IT projektas, taip ir DI sistemų įtraukimas į organizacijos leistinų naudoti sistemų sąrašą, prasideda nuo organizacijos vadovybės vertinimo. Vadovai privalo pasverti siekiamų įgyvendinti DI sprendimų naudą ir grėsmes organizacijos vykdomai veiklai. Prieš nusprendžiant įtraukti DI sistemą į organizacijos veiklos procesus **rekomenduojame**:

1. Nustatyti aiškius ir išmatuojamus tikslus, suderinamus su organizacijos veiklos strategija ir informacinių išteklių apsauga. Plačiau apie strateginį planavimą, analizę ir tikslų nustatymą skaitykite NKSC „Generatyvinio dirbtinio intelekto (GenDI) saugaus naudojimo organizacijoje“ gairėse;

2. Apibūdinti aiškius DI sprendimų naudojimo scenarijus (angl. *Use cases*);

3 Jau ankstyvajame planavimo etape įtraukti numatomo diegti DI sprendimo paveikiamus organizacijos padalinių atstovus, įskaitant teisininkus, IT ir saugumo ekspertus;

4 Įvertinti kylančias DI sprendimų naudojimo rizikas organizacijos veiklos tęstinumui, ypač rizikas, susijusias su duomenų nutekinimu ir žala reputacijai.

Toliau apžvelgiami pagrindiniai DI sprendimų integravimo aspektai: DI naudojimo strategijos nustatymas, DI naudojimo politikos parengimas, kibernetinio saugumo kontrolės priemonių taikymas, naudojant DI sprendimus, bei jų tiekėjų pasirinkimo kriterijai.

DI sprendimų įgyvendinimo strategijos nustatymas

Įvertinę poreikį, galimas naudas ir DI sprendimo naudojimo keliamas grėsmes, pasirinkite tinkamiausią DI sprendimo įgyvendinimo strategiją:

- **Tiesioginė prieiga prie modelio** (angl. *Direct access*) – kai DI sprendimas yra naudojamas tiesiogiai per gamintojo sąsają (pvz., naršyklę ar mobiliąją programėlę);
- **Prieiga prie modelio per aplikacijos programavimo sąsają** (angl. *API access*) – kai DI sprendimas naudojamas per gamintojo ar trečiųjų šalių tiekėjų teikiamas API sąsajas;
- **Licencijuotas modelis** (angl. *Licensed model*) – kai DI sprendimas veikia įmonės nuomojamoje (dažniausiai debesijos) infrastruktūroje ir yra prieinamas tik įsigijus licencijas;
- **Patikrinto modelio papildomas patobulinimas** (angl. *Fine tune proven model*) – kai naudojami jau patikrinti modeliai, papildomai apmokomi su organizacijos duomenimis ar parametrais;
- **Iš anksto apmokytas modelis** (angl. *Pre-trained model*) – kai naudojamas bendros paskirties modelis, kuris gali būti pritaikomas specifiniams poreikiams;
- **Individualiai sukurtas modelis** (angl. *Custom model, local model*) – pagal specifinį įmonės poreikį kuriamas lokalus ir organizacijoje valdomas modelis.

Strategijos pasirinkimas privalo būti pagrįstas tikėtina DI sprendimų nauda, numatomų įvesti duomenų kritiškumu, kylančiomis grėsmėmis ir tikėtinu poveikiu organizacijos veiklai.

DI sprendimo pasirinkimo kriterijai

Prieš pasirenkant, kurio tiekėjo DI sistemą integruoti į organizacijos veiklos procesus, yra būtina įvertinti ne tik funkcinius reikalavimus, bet ir atsižvelgti į saugumo, atitikties, valdymo bei rizikos aspektus. DI sprendimo tiekėjo vertinimą **rekomenduotume atlikti vadovaujantis šiais minimaliais kriterijais**:

Sistemos saugumo priemonės. Įvertinti, ar DI sistema taiko būtinas kibernetinio saugumo kontrolės priemones, tokias kaip: šifravimas, API autentifikavimas, prieigos kontrolė (IAM, RBAC) s;

Teisinė ir reguliacinė atitiktis. Įvertinti, ar DI sprendimo tiekėjas užtikrina duomenų saugojimą Europos Sąjungoje, ar įgyvendina aktualius DI akte numatytus reikalavimus, BDAR, ar atitinka kitus kibernetinio saugumo teisės aktuose numatytus reikalavimus. Įvertinti, ar tiekėjas turi kibernetinio saugumo sertifikatus (ISO 27001, ISO 42001, SOC 2 ar kitus);

Duomenų srauto valdymas. Įvertinti, ar galima registruoti ir filtruoti naudotojų įvestis, išvestis ir užtikrinti duomenų srauto auditą;

Modelio tipo ir saugos valdymas. Jei pasirenkamas licencijuotas modelis, įvertinti, ar jį kuriant buvo taikyti MLSecOps (angl. *Machine Learning Security and Operations*) principai;

Rizikų valdymas. Įvertinti, ar buvo vykdytas modelio grėsmių vertinimas (pvz., remiantis „MITRE ATLAS“ ar „NIST AI RMF 1.0“ karkasais), ar buvo vertinami didelės žalos scenarijai. Įvertinti, ar sistemoje buvo vykdyti įsiskverbimo testavimai (angl. *Penetration testing* arba *Red team exercises*);

Tiekėjo patikimumas. Įvertinti, ar suteikiama pakankama DI sprendimo dokumentacija ir palaikymas, ar sudaromos SLA (*Service Level Agreement*) ir DPA (*Data Processing Agreement*) atsakomybės deklaracijos, ar vykdomi nepriklausomi auditai;

Integracijos galimybės. Įvertinti, ar sprendimas palaiko integraciją su kitomis sistemomis (DLP, IAM, SIEM ir kitoms), centralizuotą valdymą ir veiklos stebėseną.

DI technologijos pritaikymas vis dar išlieka itin dinaminis, o atsakingo taikymo bei naudojimo geroji praktika dar formuojama. Be technologinio DI sprendimų tinkamumo šių kriterijų taikymas leidžia įvertinti ir DI sprendimų tiekėjo patikimumą ilgalaikėje perspektyvoje.

DI naudojimo politikos parengimas

Pasirinkus DI sprendimą ir jo įgyvendinimo strategiją, DI sistemų saugumui užtikrinti yra būtina taikyti atitinkamas technines ir organizacines kontrolės priemones. Taikytinas kontrolės priemones rekomenduojame derinti su jau organizacijoje įgyvendintais reikalavimais (pvz., Kibernetinio saugumo įstatymu ir susijusiais teisės aktais) bei standartais (pvz., ISO/IEC 27001, SOC 2 ar kitais). Daugiau informacijos apie generatyvinio DI duomenų valdymo ir tvarkymo kontrolę galite rasti „Generatyvinio dirbtinio intelekto (GenDI) saugaus naudojimo organizacijoje“ gairių skyrelyje „Duomenų valdymas ir tvarkymas“. DI naudojimo politikoje siūlome apžvelgti bent šiuos aspektus:

- Kurios DI sistemos yra leistinos naudoti ir kuri informacija gali būti naudojama DI sistemose (pvz., tik vieši šaltiniai ir / ar vidiniai organizacijos duomenys);
- Naudotojų prieigos ir teisių valdymas – kas valdo teises prieiti prie DI sistemų ir kurios darbuotojų rolės turi teisę naudotis DI sistema;
- Kontrolės priemonių reikalavimai, siekiant užtikrinti tinkamą DI naudojimą – įvesties ir išvesties stebėjimas bei filtravimas, modelio tinkamumo ir etiškumo rizikos vertinimas ir kita.

Siekiant geriau suprasti visus DI sprendimų valdymo aspektus, rekomenduojame vadovautis ISO/IEC 42001 Informacijos technologijų – dirbtinio intelekto – valdymo sistemos standartu.

DI sprendimų naudojimui taikomos kibernetinio saugumo kontrolės priemonės

Saugus DI sistemų integravimas į organizacijos aplinką reikalauja pasirengimo, konfigūravimo ir duomenų srautų valdymo, kurie priklauso nuo DI sprendimo sudėtingumo. Tiek DI sistemoms, tiek IT aplinkoms, kuriose organizacija jas diegia rekomenduojame papildomai taikyti žemiau išvardintas saugumo kontrolės priemones.

Siekiant apsaugoti DI sistemą viso jos veikimo laikotarpiu, **rekomenduojame**:

- taikyti griežtą prieigos kontrolę ir įjungti kelių veiksnių autentifikaciją (angl. *Multi Factor Authentication* arba MFA);
- užtikrinti, kad visos DI sistemos programavimo sąsajos (angl. API) būtų žinomos, valdomos ir apsaugotos;
- naudoti duomenų nutekėjimo prevencijos (angl. *Data Loss Prevention* arba DLP) priemones, kurios geba identifikuoti į DI sistemą keliamą informaciją, įskaitant įvedamo turinio filtravimo ir auditavimo mechanizmus;
- išjungti modelio tobulinimo jūsų organizacijos įvedamais duomenimis funkciją;
- esant galimybei DI sistemą įtraukti į reguliarius saugumo auditus;
- užtikrinti reguliarų įdiegtų DI sistemų atnaujinimą ir saugumo pataisų diegimą;
- numatyti saugias DI sistemos pašalinimo galimybes;
- įsitikinti, kad duomenys yra šifruojami, tiek saugojimo, tiek perdavimo į DI sistemą metu (pvz., naudojant TLS 1.3 protokolą).

DI sprendimų saugumas yra sparčiai besivystanti ir tyrinėjama sritis. Tyrėjams atrandant naujas DI sprendimų silpnąsias vietas ir būdus joms išnaudoti, organizacijos, siekdamos apsaugoti naudojamus DI sprendimus, privalo nuolatos sekti, prisitaikyti bei laiku reaguoti į kintančias grėsmes.

Duomenų srauto registravimas ir stebėjimas

DI sistemos grąžinti atsakymai gali daryti įtaką priimamiems sprendimams ir priklausomai nuo sistemos pritaikymo gali patys priimti autonominius sprendimus, todėl yra svarbus pakankamas duomenų įvesties ir išvesties atsekamumas ir skaidrumas. DI sistemai stebėti rekomenduotume įdiegti sprendimus, kurie leistų:

- rinkti ir registruoti įvesties ir išvesties audito žurnalus;
- analizuoti ir filtruoti įvedamos informacijos srautus;
- identifikuoti anomalijas ar saugumo pažeidimus;
- atsekti klaidų ar įtartinų veiksmų šaltinį.

DI sistemų įvesties ir išvesties valdymo ir stebėjimo įrankių rinkoje pradeda atsirasti, tačiau vertinama, jog jų pritaikymas ir galimybės vis dar yra ankstyvoje stadijoje, todėl organizacijoms rekomenduojame apsvarstyti ir alternatyvias stebėjimo priemones, pavyzdžiui, pasitelkti vidinio tinklo SSL/TLS srauto dešifravimo sprendimus, leidžiančius stebėti ir filtruoti organizacijos naudotojų į DI sistemą įvedamą ir gaunamą informaciją.

Svarbu paminėti, jog SSL/TLS dešifravimo priemonės turėtų būti taikomos išskirtinai DI sistemų naudojimui, laikantis griežtų teisinių, techninių ir organizacinių reikalavimų, užtikrinant naudotojų duomenų konfidencialumą.

Darbuotojų mokymai

Siekdami sėkmingo DI sistemų integravimo ir bendros naudojimo saugumo higienos organizacijoje, didinkite darbuotojų sąmoningumą apie DI ir bent kartą per metus vykdykite periodinius darbuotojų mokymus. Siūloma, jog mokymai apimtų:

DI naudojimo principus

Užtikrinkite, kad visi darbuotojai būtų supažindinti su organizacijos DI naudojimo politika, galimomis rizikomis ir atsakomybėmis.

Specializuotas temas

IT, saugumo, teisės bei kokybės specialistams būtina organizuoti mokymus apie DI modelių tipus, duomenų apsaugą, rizikos vertinimą ir reguliacinių reikalavimų laikymąsi.

Simuliacijas ir praktines užduotis

Į praktinius kibernetinio saugumo mokymus, apimančius incidentų valdymą ir veiklos tęstinumą, įtraukite duomenų nutekėjimo įvykius ir DI sistemų naudojimo pažeidimus.

Etikos ir atsakomybės ribas

Organizacijoje, diegiant lokalius LLM modelius, nustatykite DI naudojimo etikos ribas, šališkumo riziką bei atsakomybę už išvesties rezultatų testavimą.

DI technologija sparčiai kinta, todėl rekomenduojama reguliariai atnaujinti žinias apie DI naujoves, saugumo praktikas ir aktualų teisinį reglamentavimą. Mokymų turinio, etikos aspektų ir komunikacijos svarbos apžvalga detaliau išdėstyta generatyvinio DI gairėse.

NACIONALINIS KIBERNETINIO SAUGUMO CENTRAS

SAUGUS DIRBTINIO INTELEKTO TAIKYMAS ORGANIZACIJOJE

Dirbtinis intelektas (DI) sparčiai tobulėja dėl pažangių algoritmų ir greitesnio duomenų apdorojimo. Ši rekomemdacija padės organizacijoms suprasti DI principus, galimas grėsmes ir siūlomas saugumo priemones joms valdyti.

Grėsmės naudojant DI

DI sprendimai vis dažniau tampa kibernetinių nusikaltėlių taikiniu. Piktavaliai išnaudoja jų pažeidžiamumus siekdami pasisavinti duomenis, skleisti dezinformaciją, kurti kenkėjiškus kodus ar pasisavinti privačius modelius.

Naudojant atvirojo kodo DI

- Duomenų užnuodijimas
- Modelio vagystė
- Įvesties manipuliacija

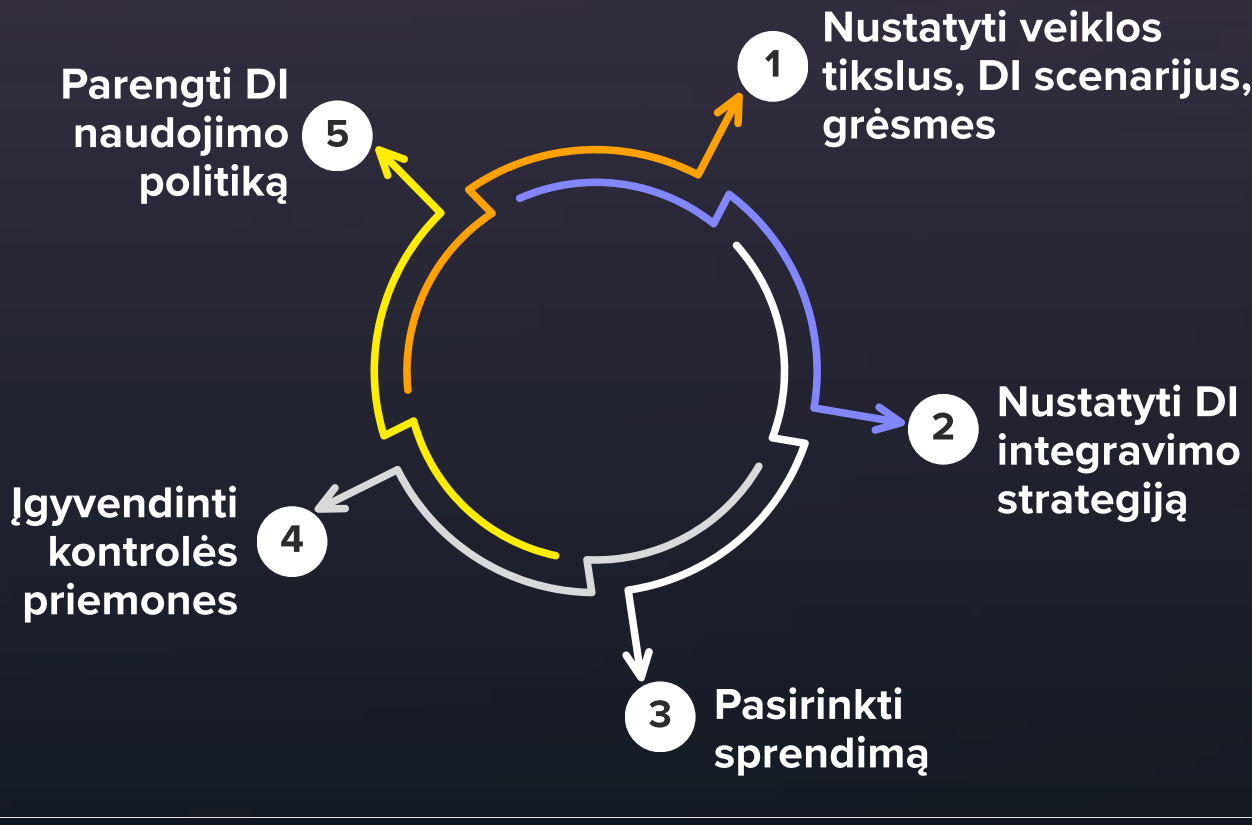
Naudojant populiarius DI

- Generatyvinio DI haliucinacijos
- Neviešų duomenų atskleidimas
- Šešėlinis DI

Naudojant licencijuotus DI

- Neautorizuota prieiga per DI tiekimo grandinę

Pagrindiniai žingsniai integruojant DI sistemą



DI sprendimų naudojimui taikomos kibernetinio saugumo kontrolės priemonės

- Prieigos kontrolė
- Duomenų nutekėjimo prevencija
- Tinkamas DI sistemos saugumo konfigūravimas
- Duomenų srauto registravimas ir stebėjimas
- Darbuotojų mokymai

DI iššūkiai

- DI technologijų pritaikymas išlieka itin dinamiškas, o atsakingo taikymo bei naudojimo geroji praktika dar tik formuojasi.
- DI sprendimų saugumas nuolat kinta - naujai atrandamos silpnybės reikalauja organizacijų budrumo ir gebėjimo greitai prisitaikyti prie kintančių grėsmių.
- DI technologijos sparčiai tobulėja, todėl būtina reguliariai atnaujinti žinias apie naujoves, saugumo priemones ir teisinį reglamentavimą.

Bendrai finansuoja Europos Sąjunga

Bendrai finansuojama Europos Sąjungos lėšomis. Tačiau išsakytos nuomonės ir požiūriai yra tik autoriaus (-ių) ir nebūtinai atspindi Europos Sąjungos ar Europos kibernetinio saugumo kompetencijų centro (ECCC) požiūrį. Europos Sąjunga ir dotaciją teikianti institucija nėra atsakingos už šią informaciją.