Dirbtinis intelektas tapo mūsų darbo kasdienybės dalimi – greitas, patogus ir visada pasiekiamas. Tačiau ar kada susimąstėte, kas vyksta su informacija, kurią jam patikime? Kai produktyvumas ima lenkti atsargumą, o DI įrankiai tampa „tyliais kolegomis“, dažnai pamirštame, kad jie nėra nei neutralūs, nei saugūs iš prigimties.

Apie tai, kaip dirbtinis intelektas (DI) keičia mūsų darbo įpročius, ir rizikas, kurių dažnas nė nepastebi, pasakoja Kauno technologijos universiteto Informatikos fakulteto (KTU IF) profesorius ir Kibernetinio saugumo kompetencijų centro vadovas Šarūnas Grigaliūnas.

– Ar pastebite, kad darbuotojai vis dažniau dalijasi konfidencialia informacija su DI sistemomis?

– Taip, pastebime augančią tendenciją. Mūsų tyrimai KTU rodo, kad beveik 40 proc. darbuotojų yra bent kartą įkėlę darbo informaciją į „ChatGPT“ ar panašius įrankius. Žmonės dažnai neįvertina, kad kopijuodami el. laišką ar dokumentą į „ChatGPT“, jie faktiškai siunčia šiuos duomenis į išorinę sistemą. Pavyzdžiui, žinau atvejį, kai vienos įmonės finansų skyriaus darbuotojas įkėlė sutarties projektą į „ChatGPT“, prašydamas jį patobulinti, visiškai nesuvokdamas, kad taip jis šiuos duomenis padarė prieinamus trečiajai šaliai.

Gali kilti klausimas, kaip apsaugoti savo aplinką, ir tam yra prieinamos generatyvinio dirbtinio intelekto (GDI) saugaus naudojimo organizacijoje gairės.

– Kodėl darbuotojai yra linkę pasitikėti DI sistemomis ir dalytis su jomis jautria informacija?

– Čia matome kelis psichologinius mechanizmus. Pirmiausia, DI įrankiai sukurti būti draugiškais ir patikimais – jie atsako mandagiai, profesionaliai ir sukuria iliuziją, kad bendraujate su patikimu kolega.

Antra, dauguma darbuotojų neturi gilaus supratimo, kaip veikia šie įrankiai. Jie mato tik naudotojo sąsają, bet nežino, kas vyksta „po gaubtu“ – kur keliauja duomenys, kaip jie saugomi ir kas gali turėti prieigą.

Be to, pastebime ir technologinį optimizmą – dauguma darbuotojų mano, kad jei įrankis populiarus ir naudojamas milijonų žmonių, jis tikriausiai yra saugus. Tai panašu į situaciją, kai žmonės nedaro savo asmeninių duomenų ar kompiuteriuose saugomos informacijos atsarginių kopijų, nes galvoja, „man nieko blogo nenutiks“ arba „aš neturiu ko slėpti“.

– Ar, jūsų nuomone, žmonės gali tapti mažiau atsakingi už savo sprendimus, jei jie nuolat pasitiki DI?

– Taip, tai matome jau dabar. Šį reiškinį vadiname „atsakomybės perkėlimu“ arba „automatizacijos šališkumu“. Kai žmonės naudoja DI įrankį, jie linkę dalį atsakomybės perkelti sistemai. Pavyzdžiui, žinau atvejį, kai darbuotojas, naudodamas DI pagalbą rašant svarbų pasiūlymą klientui, neatidumo dėka įtraukė klaidingą informaciją, nes tiesiog pasitikėjo DI pateiktu tekstu be patikrinimo.

Taip pat matome ir atvirkštinį paradoksą – kai kuriose situacijose žmonės tampa per daug

atsargūs ir abejingi, nepriimdami jokių sprendimų be DI patvirtinimo. Vienas iš mūsų tyrimų parodė, kad IT specialistai, nuolat naudojantys kodo generavimo įrankius, pradeda abejoti savo gebėjimais rašyti kodą be šių įrankių pagalbos.

– Kokias grėsmes kelia darbuotojų pasitikėjimas DI sistemomis, tokiomis kaip „ChatGPT“?

– Grėsmių spektras yra platus. Pirmiausia, tai tiesioginis konfidencialios informacijos nutekėjimas. Įsivaizduokite, kad į „ChatGPT“ įkeliate dokumentą su klientų duomenimis, įmonės finansine informacija ar intelektine nuosavybe – dabar ši informacija yra „OpenAI“ serveriuose.

Taip pat atsiranda rizika, kad konkurentai galėtų gauti prieigą prie vertingos verslo informacijos. Turėjome atvejį, kai vienos įmonės darbuotojas įkėlė naują produkto koncepciją į DI įrankį, prašydamas pataisyti teksto klaidas, ir taip faktiškai paviešino dar nepatentuotą idėją.

Reikia nepamiršti, kad kyla ir BDAR pažeidimų rizika. Jei į DI įrankį įkeliate dokumentą su asmens duomenimis be tinkamo teisinio pagrindo, pažeidžiate įstatymą ir įmonei gali grėsti rimtos baudos. Lietuvoje viena organizacija jau susidūrė su problemomis, kai darbuotojas įkėlė klientų duomenis į „ChatGPT“, norėdamas sukurti tikslesnius atsakymus.

– Ar jau yra buvę atvejų, kai konfidencialūs duomenys, sukelti į DI sistemas, pateko į viešumą?

– Taip, jau turime nemažai tokių atvejų tiek Lietuvoje, tiek pasaulyje. „Samsung“ darbuotojai netyčia nutekino įmonės programinį kodą per „GitHub Copilot“. Vienas JAV teisininkas naudojo „ChatGPT“, ruošdamas teismo dokumentus, ir sistema „sugalvojo“ neegzistuojančius teismų sprendimus, kuriuos jis pateikė teismui.

Lietuvoje turėjome atvejį, kai įmonės rinkodaros specialistas įkėlė dar nepaskelbtą reklamos kampaniją į „ChatGPT“, prašydamas patobulinti tekstą, o vėliau pastebėjo labai panašią konkurentų kampaniją. Nors tiesioginio ryšio neįrodėme, sutapimas buvo įtartinas.

Taip pat prisimenu atvejį, kai darbuotojas sukėlė įmonės vidinės komunikacijos fragmentus į „ChatGPT“, o paskui pastebėjo, kad dalis turinio atsirado „ChatGPT“ atsakymuose kitiems naudotojams, nors ir be konkrečių įmonės pavadinimų.

– Kaip „ChatGPT” ir kitos panašios DI platformos tvarko naudotojų pateikiamus duomenis? Ar jie gali būti saugomi ir panaudoti vėliau?

– Tai priklauso nuo konkrečios platformos ir jos duomenų tvarkymo politikos. „OpenAI“, „ChatGPT“ kūrėjai pagal savo privatumo politiką saugo pokalbius 30 dienų, kad galėtų reaguoti į piktnaudžiavimo atvejus, tačiau taip pat naudoja šiuos duomenis savo modelių tobulinimui.

Nemokamose DI platformų versijose praktiškai visada jūsų įvesti duomenys yra naudojami modelių apmokymui. Tai reiškia, kad jūsų įkelti tekstai gali tapti modelio žinių dalimi. Yra dokumentuotų atvejų, kai modeliai „prisimena“ dalykus, kurie buvo pateikti treniravimo

metu.
Verta paminėti, kad mokamose versijose, pavyzdžiui, „ChatGPT Enterprise“, paprastai yra griežtesnės privatumo apsaugos priemonės ir įmonės gali pasirinkti nenaudoti jų duomenų modelių tobulinimui. Tačiau net ir tada duomenys turi būti siunčiami į išorinius serverius apdorojimui.
– Ar įmanoma atsekti ir ištrinti darbuotojų netyčia paviešintus duomenis tokiose DI sistemose?
– Deja, tai yra beveik neįmanoma. Kai duomenys patenka į DI sistemą, jie gali būti naudojami įvairiais būdais: saugomi serverių žurnaluose, naudojami modelių tobulinimui ar analizuojami kokybės tikrinimo tikslais.
Net jei paprašytumėte ištrinti savo pokalbį (kas techniškai yra įmanoma kai kuriose platformose), neturite jokių garantijų, kad duomenys nebuvo jau panaudoti modelio atnaujinimui ar kad jie neegzistuoja atsarginėse kopijose. Įsivaizduokite tai kaip šiukšlių išmetimą į vandenyną – galbūt galite ištraukti keletą didesnių dalių, bet dauguma jau išsisklaidė plačiai ir nebeįmanoma jų surinkti.
– Kaip organizacijos gali subalansuoti DI naudojimą ir duomenų apsaugą?
– Turime taikyti kelių lygių strategiją, kurios pavyzdžiu galiu pasidalyti:

1. Darbuotojų mokymas ir sąmoningumo didinimas. Organizacijos privalo šviesti darbuotojus apie duomenų privatumo rizikas. Mūsų universitete sukūrėme interaktyvius mokymus, kuriuose demonstruojame, kaip duomenys keliauja per DI sistemas ir kokie yra galimi pavojai.
2. Aiški DI naudojimo politika. Kiekviena organizacija turėtų turėti dokumentuotas gaires, kokio tipo informaciją galima įkelti į viešus DI įrankius, o kokios ne. Pavyzdžiui, Lietuvos bankų sektorius jau turi tokias taisykles.
3. Vidinės DI alternatyvos. Verta investuoti į vidinius, privatumo atžvilgiu saugius DI sprendimus. Pavyzdžiui, kai kurios Lietuvos įmonės jau naudoja „privačius ChatGPT“ variantus, diegiamus įmonės infrastruktūroje, kur duomenys neišeina už organizacijos ribų.
4. Technologinės kontrolės priemonės. Organizacijos gali įdiegti DLP (angl. Data Loss Prevention) sistemas, kurios stebi ir blokuoja konfidencialios informacijos siuntimą į nepatvirtintas išorines platformas.
5. Rizikos vertinimas. Prieš leidžiant naudoti bet kokį DI įrankį, organizacija turėtų atlikti poveikio duomenų apsaugai vertinimą (DPIA), kaip to reikalauja BDAR.

Svarbu suprasti, kad nėra vieno universalaus sprendimo – kiekviena organizacija turi sukurti savo strategiją, atsižvelgdama į savo veiklos pobūdį, tvarkomus duomenis ir rizikos toleranciją.
Pabaigai noriu pasidalyti informacijos saugos (C)ISO asistentu (www.ciso.lt), kuris padeda

organizacijoms geriau suprasti ir valdyti tokias rizikas, parengti aiškias informacijos saugos tvarkas ir procedūras, kurios būtų pritaikytos konkretiems organizacijos veiklos scenarijams ir darbuotojų pareigoms.