MIDA PEAKS TEADMA AI-OHTUDEST?

Tehisarul on paljude hüvede kõrval ka mitmepalgelised ohud.

Andmehalduse ja küberturvalisuse reeglites peab olema määratud AI-lahenduste kasutamise poliitika.

AI-lahendustega teavet jagavad organisatsioonid peavad oma töötajaid koolitama, et nad oleksid võimelised tehisintellektiga kaasnevaid ohte maandama.

Tehisaru (AI) annab avalikule sektorile, eraettevõtetele ja üksikisikutele suurepärase võimaluse luua lisaväärtust, parandada tööprotsesse ja kannustada innovatsiooni. Kuid võimaluste kõrval seisavad alati ka olulised riskid, mis võivad vale haldamise või hooletuse korral tuua kaasa märkimisväärset kahju isikutele, asutustele ning ühiskonnale laiemalt.

Tehisaruga kaasnevad ohud nõuavad oma teavet AI-lahendustega töötlevatelt organisatsioonidelt süsteemset mõtestamist, tugevat riskijuhtimist ning selgete poliitikate ja turbemeetmete rakendamist. Seetõttu tuleb AI kasutamisel riske maandada nii riigi kui ka organisatsiooni tasemel, rakendades selleks puhuks AI kasutamise strateegiat ning selle alusel loodud organisatsioonilisi, eetilisi, juriidilisi ja ka tehnilisi meetmeid.

Suurimad ohud AI kasutamisel jagunevad viie valdkonna vahel: andmelekked, vale treening, küberründed, valeinfo levitamine ja manipulatsioon.

AI kasutamine kõrge riskiga, tundliku, eriti aga just salastatud teabe töötlemisel peab põhinema teadlikul ja kaalutletud otsusel.

Andmelekked kujutavad endast otsest ohtu salastatud või muule tundlikule teabele, näiteks ärisaladustele. Tehisaru vajab keelemudeli treenimiseks suurtes kogustes andmeid, mille seas võivad olla tundlikud, isiku identifitseerimist võimaldavad või strateegiliselt tähtsad andmed. Kuigi seadusandlikud või lepingulised kohustused ei luba selliseid andmeid otse avaldada, suudab AI-lahendus sellesse korra juba sisestatud teavet ikkagi kasutada ning oskusliku küsimise korral ka kolmandatele osapooltele välja anda. 2023. aastal raputas tehnoloogiamaailma juhtum, kus Samsungi töötajad sisestasid ärisaladusi juturobotivestlusesse, mis muutis need potentsiaalselt kättesaadavaks teistele kasutajatele. Teada on ka juhtum, kus patsiendi andmed jäeti kaitsmata pilveserverisse, kus kõrvalised isikud neile ligi pääsesid.

Selliste lekete vältimiseks tuleb lähtuda teabe tervikliku infoturbejuhtimise printsiipidest ning rakendada andmekaitse- ja privaatsusmeetmeid, nagu pseudonüümimine, anonüümimine ja andmeminimeerimine.

Pseudonüümimine	Anonüümimine	Andmeminimeerimine
Reaalsete isikute andmed asendatakse võltsnimedega või koodidega, mis ei ole otseselt seotud isiku tegeliku identiteediga	Pöördumatu tegevus: isikuandmeid muudetakse või need kustutatakse nii, et neid ei saa enam seostada reaalse isikuga	Isikuandmete töötlemist vähendatakse, piirdudes ainult nende andmetega, mille töötlemiseks on tegelik vajadus

Andmelekete ennetamiseks ja tuvastamiseks on vaja organisatsioonis kehtestada infohalduse ja küberturbe poliitikate osana AI-lahenduste kasutamise poliitika ja ligipääsupiirangud AI-lahendusele ning rakendada logimist, monitooringut ja anomaaliate tuvastamist. Hea tava on järgida andmekaitse aluspõhimõtteid juba enne lahenduste kasutusele võtmist.

Vale treening on üks suurimaid riske suurte keelemudelite arendamisel. Kui mudelit treenitakse enne kasutuselevõttu kallutatud, ebatäpsete või tundlike andmetega, võivad selle väljundid peegeldada samu vigu. Selline viga ei ole pelgalt tehniline, vaid võib mõjutada otseselt inimeste õigusi ja usaldust süsteemi vastu.

Keelemudelite treenimine ja rakendamine peab toimuma eetiliselt ja järelevalve all, et vältida salastatud või muu tundliku teabe lekkimist ning ühiskondlikku kahju.

Näiteks muutus Microsofti eksperimentaalne Twitteri juturobot Tay vähem kui 24 tunniga rassistlikuks ja seksistlikuks, kuna õppis sotsiaalmeedia kasutajate pahatahtlikest sisenditest. Veebirobot tuli seetõttu kiiresti võrguühendusest eemaldada. Näiteid saab tuua ka sellest, kui AI-lahendused annavad faktide asemel enda väljamõeldud vastuse, kuna nad on õpetatud andma positiivseid vastuseid, et mitte tunnistada vastuse andmiseks vajalike andmete puudumist.

Valest treeningust tuleneva ohu vähendamiseks on oluline treenimisel kasutada kvaliteetseid, mitmekesiseid ja esinduslikke andmekogumeid. Seejuures tuleb kontrollida andmete puhul nende kvaliteeti ja päritolu, lahenduse enda komponentide puhul aga lisaks ka võimalikku kallutatust analüüsimisel. Hea mõte on mudeleid enne kasutusele võtmist valideerida ja testida erinevais stsenaariumeis. Kõrge riskiga süsteemides tuleb tagada püsiv inimjärelevalve.

Küberrünnete oht kasvab kiiresti koos AI arenguga. AI-d saab kasutada pahatahtlikel eesmärkidel, näiteks turvaaukude leidmiseks või isegi pahavara loomiseks. On dokumenteeritud juhtumeid, kus AI genereeritud kood suudab vältida traditsioonilisi viirusetõrjeprogramme. Mitmed raportid on juhtinud tähelepanu, et AI-d kasutatakse kõigis küberrünnakute faasides alates planeerimisest teostuseni.

Häkkimisohu vähendamiseks tuleb rakendada tehnilisi turvameetmeid, nagu tulemüürid, krüpteerimine ja tuvastussüsteemid. Regulaarsed turvaauditid, haavatavuste testimised ja ründeanalüüs on samuti tõhusad meetmed. Sarnaselt mistahes muu süsteemiga on oluline turvapaikade õigeaegne rakendamine. Hea on rakendada üldist küberturvalisuse riskihalduse metoodikat ning pidevat järelevalvet.

Valeinfo levitamine on AI üks keerukamaid ja ühiskondlikult ohtlikumaid aspekte. AI-mudelid suudavad luua veenvalt kõlavaid tekste, helisid ja videoid, mis võivad sisaldada faktivigu või olla sihilikult eksitavad. Äärmuslikul juhul võidakse kehastuda ka autoriteediga isikuks, kellena edastatakse eksitava infoga korraldusi. Valeinfo oht on eriti suur riikidevaheliste konfliktide, valimiste või kriisisituatsioonide ajal. Näiteks

on AI abil loodud valeuudised mõjutanud valimiskampaaniaid nii Moldovas kui ka Iirimaal, kus on tekitatud paanikat seoses väljamõeldud sündmustega. Harvad pole ka näited juhtumitest, kus AI abil on suunatud isikuid tegema ebaratsionaalseid otsuseid, näiteks kandma raha kurjategijate kontole.

Valeinfo levitamisest tuleneva ohu vähendamiseks tuleb AI lahendusi kasutaval organisatsioonil rakendada faktikontrolli mehhanisme ning sisukontrolli ja süvavõltsingute (*deepfake*) tuvastamist. Lisaks on oluline valeinfot levitavate kontode kiire tuvastamine ja eemaldamine (nt sotsiaalmeediaplatvormidel). Lisaks on hea märgistada AI abil loodud sisu ning rakendada inimjärelevalvet loodud info adekvaatsuse hindamiseks. Kindlasti on vaja suurendada inimeste teadlikkust ja kriitilist mõtlemist, et teha vahet pärisinfol ja peidetud eesmärki kandval valeinfol.

Manipulatsioon viitab AI võimele mõjutada inimeste otsuseid, emotsioone ja käitumist. Cambridge Analytica juhtum näitas, kuidas algoritme saab kasutada valijate psühholoogilise profiili loomiseks ja nende sihitud mõjutamiseks. AI suudab luua isikustatud sisu, mis võib olla emotsionaalselt laetud ja kallutatud.

Selliste ohtude leevendamiseks on esmatähtis läbipaistvuse nõue: igal kasutajal peab olema õigus ja võimalus aru saada, kas talle suunatud materjal, suhtlus või otsus on AI tehtud ning milliste põhimõtete alusel. Selleks peab tehisaru kasutav organisatsioon teadlikkust tõstma ning korraldama vajadusel koolituse, et AI kasutajad oskaksid manipuleerimist tuvastada.

OHT	VASTUMEETMED
Vale treening	• Kasuta kontrollitud ja mitmekesiseid andmestikke • Rakenda andmefiltreerimist ja eelhindamist • Väldi tundlikku või kallutatud sisu
Andmelekked	• Eemalda treeningandmetest salastatud info • Kasuta *sandbox*-keskkondi ja piiratud ligipääsu • Logi ja jälgi mudeli väljundeid
Küberründed	• Piira mudeli ligipääsu tundlikule kübertaristule • Kasuta rakenduse tasemel autentimist • Vii läbi regulaarseid turvaauditeid ja testimisi
Valeinfo levitamine	• Rakenda faktikontrolli mehhanisme • Kasuta sisufiltreid ja modereerimist • Koolita kasutajaid AI väljunditesse kriitiliselt suhtuma
Manipulatsioon	• Märgista AI loodud sisu läbipaistvuse huvides • Piira personaliseeritud sisusoovitusi tundlikes valdkondades • Rakenda eetilisi sisuloome juhiseid

AI lahenduste üha laiema kasutuselevõtmisega muutub küberturbemeetmete rakendamine igal tasandil senisest veelgi olulisemaks. Tehisaru kasutamine nõuab teadlikkust, vastutust ja läbimõeldud turvameetmeid. Ainult läbipaistva ja turvalise lähenemise kaudu saab AI ühiskonda teenida, mitte seda ohustada.