TARTU ÜLIKOOL

Sotsiaalteaduste valdkond

Johan Skytte poliitikauuringute instituut

Jana Kotšnova

RAKENDAMISLÕHE: EESTI AVALIKU SEKTORI TÖÖTAJATE TEHISINTELLEKTI KASUTAMISE ANALÜÜS

Bakalaureusetöö

Juhendaja: Heiko Pääbo, PhD

Tartu 2026

AUTORLUSE DEKLARATSIOON

Olen koostanud töö iseseisvalt. Kõik töö koostamisel kasutatud teiste autorite seisukohad, ning kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

Jana Kotšnova

18.05.2026

Töö sõnade arv: 12 054

Töö koostamisel kasutati tehisintellekti tööriista *Claude Sonnet 4.6* kolmel otstarbel. Esiteks kasutati seda inglisekeelsete erialaterminite tõlkimiseks eesti keelde, näiteks terminite "*prompt injection*", "*automation complacency*" või "*chain-of-thought*" puhul. Näide päringust: "*How would you translate "prompt injection" into Estonian?*". Teiseks tõlgiti mudeli abil pikemaid ingliskeelseid lõike eesti keelde, kusjuures originaaltekstid olid autori enda kirjutatud. Näide päringust: "*Translate this text to Estonian. Keep the style of writing as close to the original as possible*". Küsiti nõu ka iseseisvalt tehtud tõlgete kohta: "*Is this text grammatically correct in Estonian? Bring out possible mistakes and typos as separate bullet points*". Kolmandaks küsiti mudelilt soovitusi töö struktuuri puudutavates küsimustes, näiteks lõikude järjestuse osas. Näide päringust: "*The structure of this chapter feels off. In what order could I make my points for it to be more clear?*". Küsiti ka üldisemaid soovitusi töö kirjutamise kohta. Näide päringust: "*Does this introduction contain all the elements listed in the requirements*".

ANNOTATSIOON

Bakalaureusetöö uurib, kuidas Eesti avaliku sektori töötajate tajutud tehisintellekti (TI) riskid erinevad akadeemilises kirjanduses kaardistatud riskidest ning kuidas need riskitajud mõjutavad töötajate käitumismustreid TI kasutamisel. Uurimuse fookusasutuseks on Majandus- ja Kommunikatsiooniministeerium, kus viidi läbi 11 poolstruktureeritud intervjuud. Intervjuude analüüs tugineb Slovici (1987) riskitaju psühhomeetrilisele paradigmale, mille raames hinnati osalejate riskitaju positsiooni hirmu ja tundmatuse dimensioonides, ning inimese ja automatiseeritud süsteemi koostoime raamistikule (Sheridan, 1992; Parasuraman & Riley, 1997; Parasuraman et al., 2000), mille abil klassifitseeriti osalejate käitumine automatiseerimise tasemete järgi.

Tulemused näitavad, et osalejate riskiteadlikkus koondus peamiselt nähtavate riskide ümber: andmelekke ja hallutsineerimise risk olid laialdaselt teadvustatud, samas kui algoritmilise kallutatuse, läbipaistvuse puudumise ja küberrünnakute riskid jäid enamiku osalejate riskitajust välja. Lisaks kirjanduses kaardistatud riskidele tõid osalejad esile kaks täiendavat riskikategooriat: inimliku vea riski ning valdkonnateadmiste puudumisest tuleneva riski. Analüüs näitas ka, et riskitaju ja käitumine ei ole lineaarses seoses: keskmise teadlikkuse ja suurema kasutamiskogemusega osalejatel ilmnes tendents kontrolli vähendada, samas kui madalama või sügavama teadlikkusega osalejate järelevalve oli kõrgem. Enamik osalejatest ei poolda tehisintellekti kasutamist piiravaid meetmeid, kuid toetab kasutuspraktikate ja tööriistade ühtlustamist riigi tasandil.

SISUKORD

Joonised:

SISSEJUHATUS

Tehisintellekti (TI) kasutamine organisatsioonides on kiiresti kasvanud, sealhulgas abistamiseks ülesannete puhul, mis hõlmavad organisatsioonide siseandmeid (Balashov et al., 2025, lk 1). Käesolev uurimistöö keskendub kitsamalt TI rakendamisele avalikus sektoris, kus tehnoloogial on potentsiaal muuta poliitikakujundamise protsesse tõenduspõhisemaks, tõhusamaks ja läbipaistvamaks (Upreti et al., 2023, lk 231). Tehnoloogia kasutuselevõtu kiirus on ületanud seda reguleerivate raamistike arengut (Kim et al., 2025, lk 1), mistõttu toob tehnoloogia lisaks uutele võimalustele ka mitmeid uusi riske.

Keskseim tehisintellekti reguleeriv raamistik Euroopa Liidus, sealhulgas Eestis, on Tehisintellekti määrus (*AI Act*), mis jõustus 2024. aasta augustis (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024). Määrus toob esimest korda otsese regulatsiooni alla ka üldotstarbelise tehisintellekti mudelid (*General-Purpose Artificial Intelligence*, *GPAI*), sealhulgas suured keelemudelid. Neile kehtestatakse dokumentatsioonikohustused ning kõige võimekamatele mudelitele ka rangemad süsteemse riski hindamise nõuded (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, ptk V). Määruse praktiline mõju on aga hetkel piiratud mitmel põhjusel. Esiteks kohalduvad rangemad reeglid võimekamatele mudelitele alles 2027. aastast, kuna enne 2025. aasta augustit turule toodud mudelitele on ette nähtud üleminekuperiood (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, art 111). Teiseks on paljud tehnilised standardid alles väljatöötamisel ning kontrollimehhanismid ei ole täielikult toimivad. Lõpuks on enamik tänaseid tippmudeleid treenitud enne lõplike reeglite vastuvõtmist interneti ulatusliku kraapimise (*web scraping*) teel kogutud andmetel, mille vastavust autoriõiguse ja isikuandmete kaitse nõuetele on tagantjärele võimatu kontrollida. Lisaks on määruse kohustused suunatud mudelite pakkujatele ja arendajatele, mitte rakendajatele ega lõppkasutajatele. See tähendab, et Euroopa Liidu tasandil puuduvad hetkel ühtsed reeglid selle kohta, kuidas avalik sektor peaks tehisintellekti oma tööprotsessides rakendama.

Eestis on välja töötatud strateegilised dokumendid, mis kirjeldavad hetkeolukorda ning kaardistavad üldisi põhimõtteid tehisintellekti rakendamiseks avalikus sektoris: peamised neist on "Andmete ja tehisintellekti valge raamat 2024–2030" ja "Tehisintellekti tegevuskava 2024–2026" (MKM et al., 2024a; MKM et al., 2024b). Avaliku sektori asutustele ei ole aga kehtestatud siduvaid nõudeid selle kohta, kuidas hinnata või maandada riske, mis kaasnevad üldotstarbeliste tehisintellektitööriistade kasutuselevõtuga. Selle tulemusena sõltub TI

vastutustundlik kasutamine asutustes suuresti töötajate individuaalsetest teadmistest, hoiakutest ja riskitajust.

Uurimislünk seisneb selles, et Eestis ei ole seni põhjalikult analüüsitud, kuidas avaliku sektori töötajate tegelikud tajud ja kogemused erinevad akadeemilises kirjanduses kaardistatud riskidest, mis kaasnevad tehisintellekti kasutuselevõtuga. Küsimused andmete turvalisuse, algoritmilise kallutatuse, hallutsineerimise ja läbipaistvuse puudumise kohta on kirjanduses käsitletud, kuid see, kuidas avaliku sektori töötajad neid riske tajuvad ja kas nende arusaam vastab dokumenteeritule, on uurimata. Kui töötajate riskitaju erineb oluliselt sellest, mida kirjandus dokumenteerib, tekib rakendamislõhe, kus tehnoloogia võetakse kasutusele ilma, et selle ohte piisavalt teadvustataks. Uuringu eesmärk on kaardistada lõhe Majandus- ja Kommunikatsiooniministeeriumi näitel. Uurimisküsimus on järgmine: Kuidas erinevad Eesti avaliku sektori töötajate tajutud riskid TI kasutamisel akadeemilises kirjanduses kaardistatud riskidest? Sellest tuleneb alamküsimus: Kuidas võivad need riskitajud mõjutada TI-põhiste tööriistade praktilist rakendamist poliitikakujundamises?

Uurimisküsimusele vastamiseks tuginetakse kahele teoreetilisele raamistikule: Slovici (1987) riskitaju psühhomeetrilisele paradigmale ning inimese ja automatiseeritud süsteemi koostoime raamistikule (*Human-Automation Interaction, HAI*) (Sheridan, 1992; Parasuraman & Riley, 1997; Parasuraman jt, 2000). Riskitaju teooria aitab kategoriseerida ja mõõta tehisintellektiga seotud riskihinnanguid. Inimese ja automatiseeritud süsteemi koostoime raamistik aitab omakorda mõista, kuidas riskitaju avaldub ametnike käitumises TI-süsteemidega töötamisel. Teoreetilisele raamistikule järgneb uurimistöö tausta ülevaade, milles kaardistatakse akadeemilises kirjanduses dokumenteeritud tehisintellekti riskikategooriad, eesmärgiga luua võrdlusbaas, mille suhtes osalejate tajutud riske hinnata. Teoreetilisest lähtepunktist eeldatakse, et kirjanduses kaardistatud riskide ja töötajate riskitaju vahel esineb rakendamislõhe: töötajad ei pruugi kõiki dokumenteeritud riske teadvustada või võivad neid hinnata teisiti, kui kirjandus ette näeb. Inimese ja automatiseeritud süsteemi koostoime raamistiku abil uuritakse lähemalt, kuidas riskitaju on seotud käitumisega, kuna riskid avalduvad selles, kuidas töötajad TI väljundit kontrollivad, millisel määral nad automatiseeritud süsteemidele toetuvad ning kuidas nende riskitaju kujundab igapäevaseid otsuseid tööriistade kasutamisel. Metoodika peatükis kirjeldatakse uurimisstrateegiat, valimi moodustamise põhimõtteid ning poolstruktureeritud intervjuude läbiviimise ja temaatilise analüüsi protsessi. Tulemuste peatükis esitatakse taustapeatükis käsitletud riskide kaupa, kuidas intervjueeritavad neid riske tajuvad, ning tuuakse välja ka riskid, mida kirjanduses ei käsitletud, kuid mida osalejad intervjuudes

mainisid. Lisaks näidatakse, kuidas siduvate juhiste puudumisel tuginevad töötajad tehisintellekti kasutamisega seotud otsustes oma tunnetusele ja individuaalsetele arusaamadele kaasnevatest riskidest. Töös ilmneb, et riigiasutused töötavad välja sisemisi põhimõtteid, mis ei pruugi teineteisega kooskõlas olla. Ka tööriistade valik sõltub asutusest, kuigi Riigi IT Keskus (RIT) on TI-tööriistade riskianalüüsi läbi viinud. Tulemuseks on kasutuspraktikate ebajärjepidevus nii asutuste vahel kui ka asutuste sees, mis võimendab tehnoloogia kasutuselevõtuga kaasnevaid riske.

1. TEOREETILINE RAAMISTIK

Selles peatükis esitatakse teoreetilise raamistiku, mis aitab mõista, miks avaliku sektori töötajate suhtumist TI-tööriistadesse on oluline uurida ning miks selleks ei piisa riskide mõistmist pelgalt tehnilise poole pealt. Esmalt tutvustatakse Paul Slovici (1987) riskitaju teooriat, mis selgitab, kuidas inimesed tajuvad ja hindavad erinevaid riske ning millised tegurid mõjutavad nende hinnanguid ja hoiakuid. Paul Slovici lähenemine aitab selgitada, miks samad tehisintellektiga seotud riskid võivad erinevate inimeste jaoks tunduda erinevalt olulised. Järgmisena tutvustatakse inimese ja automatiseeritud süsteemi koostoime raamistikku (Sheridan, 1992; Parasuraman & Riley, 1997; Parasuraman jt, 2000), mis käsitleb inimese ja tehisintellekti interaktsiooni erinevaid tüüpe ja tasandeid. Nimetatud raamistik lisab uue tehnoloogia tajumisele käitumusliku perspektiivi ning aitab kirjeldada, mis ulatuses ja millistel tingimustel töötaja tehisintellekti väljundeid kontrollib. See on töö kontekstis oluline küsimus, kuna käsitletud riskid avalduvad praktikas just nendel hetkedel, kus inimlik kontroll on nõrgenenud või puudub.

Integreeritud mudel aitab selgitada nähtust, mida töö empiiriline osa uurib: rakendamislõhet selle vahel, mida teatakse tehisintellekti riskide kohta, ning selle vahel, kuidas neid tegelikkuses tajutakse ja kuidas nendele reageeritakse. Olemasolevas rakendusuuringute (*implementation research*) kirjanduses defineeritakse rakendamislõhet (*implementation gap*) tavaliselt kui erinevust poliitikakujundamisel seatud eesmärkide ja tegelikult saavutatud tulemuste vahel (Leyer et al., 2026, lk 3). Definitsioon ei kata aga selle töö konteksti täies mahus: kuna Eesti avalikus sektoris puuduvad tänapäeval kindlad eeskirjad TI kasutamise kohta, ei saa rääkida olukorrast, kus poliitika elluviimine kaldub algsest plaanist kõrvale. Selle tõttu, kirjeldan selle töö kontekstis rakendamislõhe täpsemalt kui olukorda, kus teaduskirjanduses kirjeldatud riskid ei kajastu praktikas: töötajate tajudes ja käitumises. Sarnaste olukordade kirjeldamiseks leidub inglisekeelses kirjanduses mõiste "*research–practice gap*", mis kirjeldab ajavahemikut uute tõenduspõhiste praktikate tuvastamise ja nende praktilise kasutuselevõtu vahel erinevate valdkondade spetsialistide poolt (Ungvarsky, 2025). Selle töö raames kasutatakse mõistet "rakendamislõhe", kuna see on piisavalt ülevaatlik, et laiemas tähenduses käsitleda.

1.1 Riskitaju teooria

Riskitaju kui uurimisvaldkond käsitleb erinevaid viise, kuidas inimesed hindavad, tõlgendavad ja kirjeldavad tegevusi ja tehnoloogiaid, mis võivad kaasa tuua potentsiaalseid riske (Slovic,

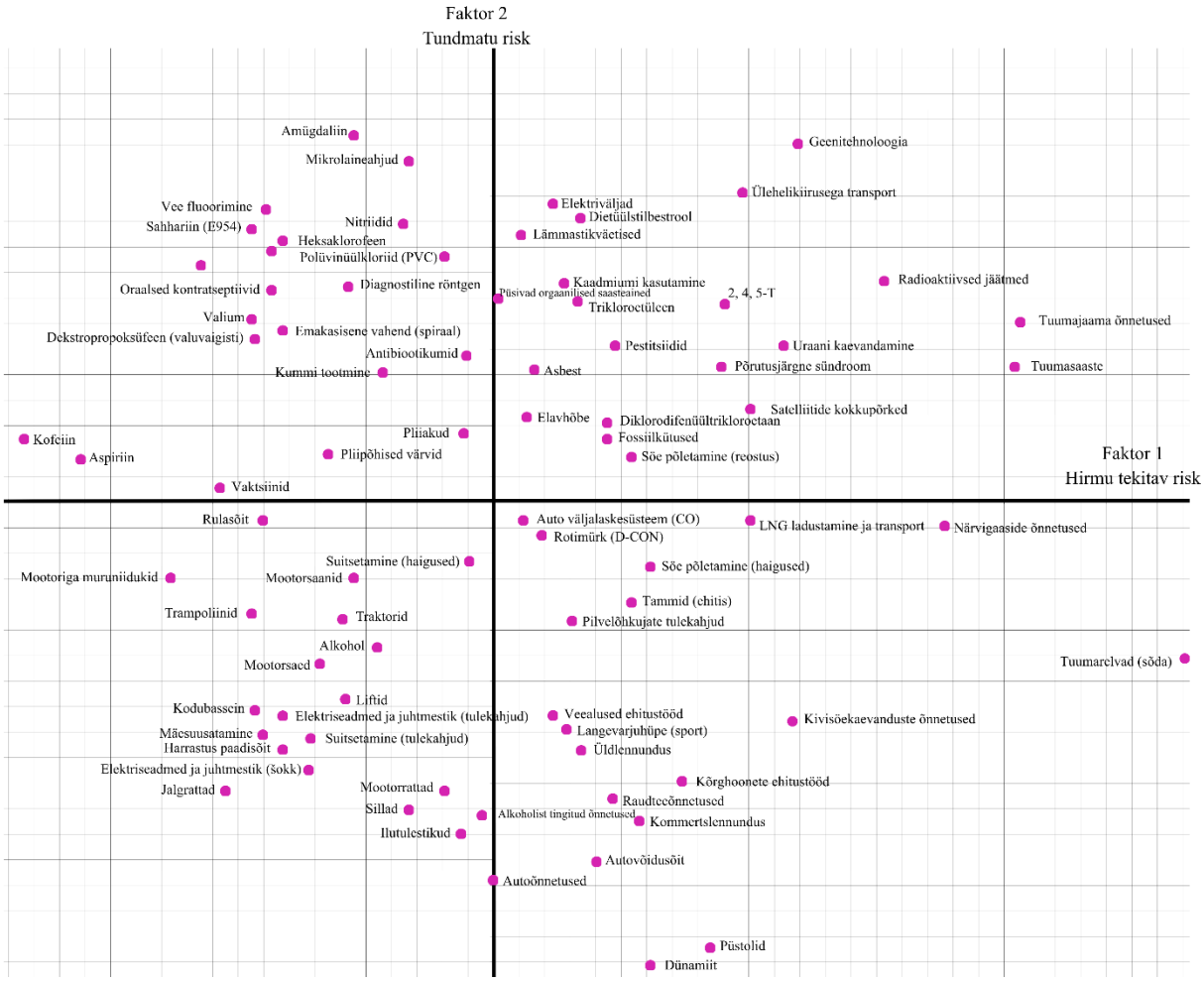
1987, lk 280). Tänapäeva kiire tehnoloogiline areng teeb riskitaju uurimise eriti aktuaalseks, kuna selle tulemusi saab kasutada, et hinnata, kuidas inimesed uute tehnoloogiate puhul riske tajuvad ja tehnoloogiaid omaks võtavad. 1987. aastal töötas Paul Slovic välja riskitaju teooria, mis pani aluse tänapäevasele riskitaju uurimisele, näidates, et avalikkuse riskitaju on võimalik mõõta ja prognoosida. Autor uuris riskitaju psühhomeetrilise lähenemise kaudu, kasutades skaleerimist ja mitmemõõtmelist analüüsi, mille abil sai riskitajusid kvantitatiivselt esitada niinimetatud "kognitiivsete kaartidena" ("*cognitive maps*") (Slovic, 1987, lk 281).

Oma uurimusega toetas Slovic poliitikakujundamist pakkudes raamistiku, mis aitab ennustada, kuidas avalikkus tajub ja reageerib ohtlikele tegevustele ja tehnoloogiatele (Slovic, 1987, lk 280). Slovici sõnul aitab riskitajude hindamine ja ennustamine parandada riskiteabe suhtlust avalikkuse, tehniliste ekspertide ja poliitikakujundajate vahel (Slovic, 1987, lk 280). Autori uurimuse aluseks oli arusaam, et ilma arusaamata sellest, kuidas inimesed riskidest mõtlevad ja neile reageerivad, võivad ka hästi kavandatud poliitikad jääda ebaefektiivseks (Slovic, 1987, lk 280). Oluline erinevus varasematest riskitaju uurimistest seisneb selles, et Slovic seadis kahtluse alla eelduse, et riskitaju põhineb suuremas osas kättesaadaval statistikal. Slovici sõnul "riskantsus tähendab inimeste jaoks enamat kui eeldatav surmajuhtumite arv" *(""riskiness" means more to people than "expected number of fatalities"*") (Slovic, 1987, lk 285). Autor näitas, et riskitaju kujuneb hoopis erinevate kvalitatiivsete tegurite koosmõjul.

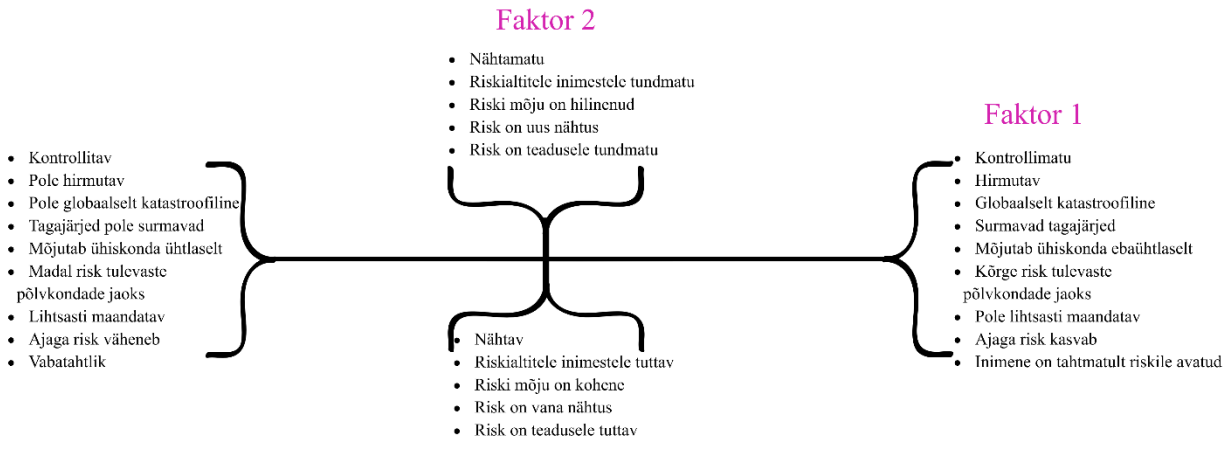
Oma empiirilises uuringus palus Slovic erinevatel ühiskonnagruppidel hinnata 81 erinevat ohuallikat (*hazards*). Osalejad hindasid iga ohuallikat 18 erineva kvalitatiivse omaduse alusel (Slovic, 1987, lk 282). Faktoranalüüsi abil eristas Paul Slovic riskitaju mõjutavaid tegureid kaheks peamiseks dimensiooniks. Esimene faktor (*Factor 1*) "hirmu tekitav risk" (*Dread risk*), iseloomustab riske, mida tajutakse kontrollimatute, hirmutavate, katastroofiliste ja ebaõiglaste tagajärgedega ohtudena (Slovic, 1987, lk 282-283). Teine faktor (*Factor 2*) "tundmatu risk" (*Unknown risk*), hõlmab riske, mida peetakse varjatuks, tundmatuks, uueks ja viivitatud tagajärgedega (Slovic, 1987, lk 282-283). Täpsemad tegurid on esitatud Joonisel 2. Oluline järeldus oli see, et hirmu faktor (*Factor 1*) on tugevaim ennustaja selle kohta, kuivõrd kõrgeks inimesed konkreetset riski hindavad ja kui suurel määral seda soovitakse reguleerida (Slovic, 1987, lk 283).

Slovic paigutas hirmu tekitava riski tegureid x-teljele ja tundmatu riski tegureid y-teljele. Selle tulemusena kujunes välja niinimetatud faktorruum (*factor space*), mis on esitatud Joonisel 1. Faktorruum on oma olemusest kahemõõtmeline tasand, kus iga ohuallikas on esitatud punktina.

Autori edasine analüüs kinnitas, et inimese riskitaju ja -hoiakud on tihedalt seotud sellega, millisel positsioonil (koordinaatidel) teatud ohuallikas faktorruumis paikneb (Slovic, 1987, lk 283). Faktorruum aitab selgitada inimeste äärmist vastumeelsust mõnede ohtude suhtes, nende ükskõiksust teiste riskide suhtes ning lahknevusi nende reaktsioonide ja ekspertide arvamuste vahel.



Joonis 1. Autori töö. "81 erineva ohuallika faktorruumis paiknemine riskiteguri vaheliste seoste analüüsi põhjal". Slovic, 1987. lk 282.



Joonis 2. Autori töö. "Faktor 1 ja Faktor 2 moodustavad omadused". Slovic, 1987. lk 282.

Lisaks mainib autor, et inimesed tõlgendavad uut informatsiooni teatud riski kohta erinevalt: hinnangu kujunemine sõltub sellest, kas inimesel on eelnevalt olnud kindel seisukoht teatud

riski kohta (Slovic, 1987, lk 281). Sel juhul keskendub inimene informatsioonile, mis langeb kokku tema uskumustega, ja suhtub skeptiliselt tulemustesse, mis sellest erinevad (Slovic, 1987, lk 281). Kui aga kindlad eelnevad arvamused puuduvad, sõltub inimese arusaam suuresti sellest, kuidas probleem on temale esitatud. Sama riskiteabe esitamine eri vormides võib mõjutada inimeste edasist tajumist (Tversky & Kahneman, 1981, lk 456).

Kuigi Slovici mudel on algselt välja töötatud avalikkuse riskitaju mõõtmiseks, saab seda kasutada ka poliitikakujundajate tajude analüüsimiseks. Autor ise kinnitab, et ekspertide hinnangud on sageli mõjutatud samadest kallutatustest nagu üldsuse omad, eriti kui eksperdid peavad tegutsema ilma piisavate andmeteta ja toetuma suurel määral intuitsioonile (Kahneman et al., 1982, lk 1130). Nagu eelnevalt mainitud, tegutsevad hetkel avaliku sektori töötajad "regulatiivses vaakumis" (*regulatory vacuum*), kasutades TI-põhiseid tööriistu ilma siduvate juhisteta, mistõttu võib eeldada, et ka nemad kalduvad riskide hindamisel oma intuitsioonile toetuma.

Samuti, kuigi Slovic ei käsitle oma teooria raames tehisintellekti, sobib see mudelisse kui uus, potentsiaalseid riske kaasatoov tehnoloogia. Slovici mudelit kasutatakse näiteks 2023. aasta uuringus, mille eesmärk oli välja selgitada, kuidas inimeste ebamäärane tajumine tehisintellektist mõjutab nende valmisolekut kasutada TI-põhiseid tööriistu (Schwesig et al., 2023, lk 1054). Osalejad lugesid stsenaariume, mis kirjeldasid erinevaid tehisintellekti rakendusi (meditsiin, transport, meedia, psühholoogia), ja seejärel hindasid tehisintellekti riskitaju mõõdikute abil (Schwesig et al., 2023, lk 1058). Käesolev töö keskendub aga tehisintellektiga seotud riskitaju analüüsimisele kitsamalt, uurides seda eelkõige avaliku sektori töötajate seas.

1.2 Inimese ja automatiseeritud süsteemi koostoime raamistik

Inimese ja automatiseeritud süsteemi koostoime raamistik (*Human-Automation Interaction Framework, HAI*) kirjeldab, kuidas inimesed ja TI-süsteemid koos töötavad ning millisel määral inimene säilitab kontrolli otsustusprotsessi üle. Raamistik täiendab riskitaju teooriat, kuna eelduseks on, et riski tajumine ja edasine käitumine teatud tehnoloogia suhtes on omavahel seoses: töötajad tõlgendavad tajutud riske käitumismustriteks, mis omakorda mõjutavad nende edasist tegevust ja seda, kuidas nende riskitaju edasi kujuneb. Raamistiku algsed põhimõtted pärinevad Thomas B. Sheridani (1992) automatiseerimise tasemete uurimustest, mis tõi esile, et inimtegevuste automatiseerimisel on erinevaid tüüpe ja tasandeid.

Autor töötas välja skaala, mis ulatub täielikult manuaalsest tegevusest täielikult autonoomseni, rõhutades, et automatiseerimise suurenedes muutub inimese järelevalve roll (Sheridan, 1992, lk 358). Autor tutvustas ka inimese järelevalvekontrolli (*human supervisory control*) kontseptsiooni, et kirjeldada erinevaid viise, kuidas inimoperaatorid (*human operators*) poolintelligentsete süsteemidega (*semi-intelligent systems*) suhtlevad (Sheridan, 1992, lk 1). Sheridan kirjeldab selliseid süsteeme kui "alluvaid" (*subordinates*) ning defineerib inimese järelevalvekontrolli kui protsessi, kus inimene annab juhiseid, mida alluvad süsteemid mõistavad ja tõlgendavad üksikasjalikeks tegevusteks (Sheridan, 1992, lk 1). Seejärel koguvad süsteemid tulemuste kohta põhjalikku teavet ja esitavad selle kokkuvõtlikult inimesele (Sheridan, 1992, lk 1). Alluvate "intelligentsuse" tase määrab omakorda seda, mis määral on "juhendaja" (inimene) sellesse protsessi kaasatud (Sheridan, 1992, lk 1). Poolintelligentsed alamsüsteemid, mida autor kirjeldab, sarnanevad funktsioonide ja kasutusviisi poolest tänapäevaste suurte keelemudelitega. Siinkohal on oluline märkida, et tehisintellekti kontekstis defineeritakse intelligentsust eesmärgipärase või ratsionaalse käitumisena, mitte inimese kognitiivsete protsesside imiteerimisena (Russell & Norvig, 2022, lk 22). TI valdkonnas domineeriv "ratsionaalse agendi" lähenemisviis kirjeldab intelligentseid süsteeme agentidena, mis tegutsevad oma taju ja teadmiste põhjal parima oodatava tulemuse saavutamiseks (Russell & Norvig, 2022, lk 22-23).

Parasuraman ja Riley (1997, lk 231) arendasid seda mudelit edasi, formaliseerides Sheridani automatiseerimise mõistet kui protsessi, "kus masin (tavaliselt arvuti) täidab ülesandeid, mida varem tegid inimesed[1]". Selle asemel, et küsida üksnes, kas teatud ülesanne tuleks automatiseerida, rõhutasid autorid, et süsteemide kavandajad peavad otsustama millist tüüpi funktsioone ja millisel määral tuleks automatiseerida (Parasuraman et al., 2000, lk 294). Autorite sõnul võib automatiseerimine toetada inimliku otsustusprotsessi erinevaid etappe: andmete kogumist, töötlemist ja tõlgendamist, võimalike valikute vahel otsustamist ning valitud tegevuse elluviimist (Parasuraman et al., 2000, lk 287). Süsteem võib automatiseerida mõningaid etappe, kuid mitte teisi, ning igal etapil võib automatiseerimise tase olla erinev. Sarnaselt Sheridanile rõhutasid Parasuraman jt, et automatiseerimine eksisteerib kontinuumina täielikult manuaalsest süsteemist kuni täielikult automaatse süsteemini. Näiteks võib süsteem pakkuda teavet; soovitada otsuseid; teha otsuseid, kuid jätta inimesele vetoõiguse; või viia tegevusi ellu autonoomselt (Parasuraman et al., 2000, lk 287). Tänapäevased keelemudelid

[1] We define automation as the execution by a machine agent (usually a computer) of a function that was previously carried out by a human (Parasuraman & Riley, 1997, lk 231). *Tõlgitud autori poolt*.

võimaldavad neid funktsioone rakendada varasemast palju paindlikumalt. Seetõttu ei ole peamine küsimus enam tehnilises teostatavuses, vaid selles, kui laialdaselt automatiseerimise võimalusi kasutada ning milline inimkontroll peaks nende kasutamisega kaasnema.

Parasuraman jt (2000, lk 290) rõhutavad, et automatiseerimise taseme valikul tuleb hinnata automatiseerimise mõju inimese töökoormusele, olukorrateadlikkusele ja otsustuskvaliteedile, süsteemi töökindlusele ning inimese võimele vajadusel sekkuda ja kontroll üle võtta. Erinevate automatiseerimise tasemete eristamine on oluline, kuna kõrgem automatiseerimine vähendab olukorrateadlikkust ning toob kaasa "automatiseerimisest tingitud hooletuse" (*automation complacency*) (Parasuraman et al., 2000, lk 291): automatiseermise väärkasutamist, mis väljendub kasutaja liigses ja sobimatus toetumises automatiseeritud süsteemile (Parasuraman & Manzey, 2010, lk 398). Autorite hinnangul on automatiseerimine kasulik eeskätt siis, kui see toetab inimese otsustusprotsessi ilma vähendamata inimese arusaamist olukorrast või tema suutlikkust süsteemi tegevust jälgida ja vajadusel korrigeerida (Parasuraman et al., 2000, lk 291).

Tänapäeval kasutatakse sarnast automatiseerimise tasemete klassifitseerimist tehisintellekti süsteemide puhul. OECD (2022, lk 53) eristab nelja tegevusautonoomia taset: "*human support*", "*human-in-the-loop*" (*HITL*), "*human-on-the-loop*" (*HOTL*) ja "*human-out-of-the-loop*" (*HOOTL*). Käesolevas töös lihtsustatakse seda jaotust kolmeks peamiseks tasemeks:

- *HITL*: kasutajad suhtlevad TI-süsteemiga aktiivselt, teevad otsuseid TI soovituste põhjal ning võivad tulemusi vajadusel muuta või üle kirjutada,
- *HOTL*: TI täidab ülesannet peamiselt autonoomselt ning kasutajad mängivad jälgija rolli, sekkudes ainult siis, kui absoluutselt vajalik,
- *HOOTL*: inimesed on otsustusprotsessist suurel määral eemal ning TI otsused viiakse automaatselt ellu (OECD, 2022, lk 53).

OECD esimene tase (*human support*) käsitletakse siin koos *HITL* tasemega, kuna mõlemal juhul jääb inimene otsustusprotsessis keskseks: süsteem võib anda ülevaate või soovitusi, kuid peamine tegevus eeldab aktiivset inimese osalust. Nende tasemete ühtimist kasutatakse laialdaselt TI-teemalises kirjanduses, nt Natarajan jt (2025) märgivad, et aktiivne õppimine ja soovituste vastu võtmine kuulub HITL kategooria alla (Natarajan et al, 2025, lk 28594).

Vajadust kombineerida automatiseerimist inimjärelevalvega uurib tänapäeval inimkeskse tehisintellekti valdkond (*Human-Centered AI*), mille eestvedajaks on Ben Shneiderman (2022).

Eespool mainitud autorite ideid arendades kinnitab Shneiderman, et tänapäeva kontekstis ei peaks tehisintellekti eesmärk olema inimeste asendamine, vaid inimeste võimestamine selliste süsteemide kaudu, mis suurendavad inimeste võimekust ja säilitavad inimese kontrolli tehnoloogia üle (Shneiderman, 2022, lk 9). Selle seisukoha põhjendamiseks vastandab Shneiderman TI uuringutes domineeriva ratsionalistliku lähenemise inimkesksele, empiirilisele lähenemisele.

Shneidermani sõnul on ratsionalism, mis tugineb loogikale ja matemaatilisele mõtlemisele, kujundanud ajalooliselt suurt osa tehisintellekti teadusuuringutest (Shneiderman, 2022, lk 18). Lähenemine eeldab, et keerukaid otsustusprotsesse saab formaliseerida üksnes funktsioonide, algoritmide ja statistiliste meetodite abil (Shneiderman, 2022, lk 19). Tehisintellekti nähakse selles kontekstis vahendina, mis võimaldab poliitikakujundamist muuta ratsionaalsemaks, efektiivsemaks ja andmepõhisemaks. Shneiderman siiski rõhutab, et üksnes ratsionalistlik lähenemine ei ole piisav: kuigi see võib olla hea lähtekoht, tuleb seda täiendada empiirilise lähenemisega, mis arvestab tegeliku maailma kontekstuaalse keerukuse, mitmekesisuse ja ebakindlusega (Shneiderman, 2022, lk 18-19).

Slovici (1987) riskitaju teooria ja Sheridani (1992) ja Parasuramani jt (1997; 2000) inimkontrolli ja automatiseerimise tasemete raamistik on välja töötatud eraldiseisvalt ning käsitlevad inimese ja tehnoloogia interaktsiooni erinevaid aspekte. Käesolevas töös integreeritakse eelmainitud teoreetilisi lähenemisi ühtseks terviklikuks mudeliks, mis aitab selgitada nii "mida" avaliku sektori töötajad tajuvad, kui ka "kuidas" nad sellele reageerivad ja edaspidi käituvad. Koos moodustavad need kaks teooriat taju-käitumise mudeli (*perception-behaviour model*), kus riskitaju kujundab järelevalvekäitumist ja järelevalvekäitumine kujundab omakorda riskitaju kogemuste kaudu.

Teoreetiline eeldus on, et riskitaju dimensioonid ja järelevalve tasemed on omavahel seotud, kuid selle seose suund ja olemus on empiiriline küsimus. Ühelt poolt võib riskitaju kujundada järelevalvekäitumist: töötajad, kes tajuvad tehisintellekti kasutamist kõrge hirmu teguri (*Factor 1 risk*), võivad omaks võtta *HITL* käitumise, kontrollides iga väljundit põhjalikult enne kasutamist. Teiselt poolt võib ka järelevalvekäitumine ise kujundada riskitaju: töötajad, kes juba mingil põhjusel praktiseerivad *HITL* käitumist, võivad avastada rohkem vigu, mis tugevdab nende riskiteadlikkust, samas kui jäävad *HOOTL* käitumise puhul vead varjatuks automatiseerimisest tingitud hooletuse tõttu, mis võib tajutud riski kunstlikult vähendada. Tõenäoliselt toimivad mõlemad mehhanismid korraga, tekitades vastastikust mõjutusprotsessi.

Raamistik aitab selgitada, millised tegurid (näiteks tagasiside mehhanismid või käitumismustrid) rakendamislõhet võimendavad või vähendavad. Teooriate integreerimise abil on võimalik tuvastada, kuidas hirmu või tundmatuse tase mõjutab järelevalvekäitumist ning millised järelevalve strateegiad võivad riskitajude kujunemist tugevdada või nõrgendada.

2. UURIMISTÖÖ TAUST

“Tehisintellekt” on selle töö keskne mõiste. Samas, puudub sellel tänapäeval ühtne ja universaalselt aktsepteeritud definitsioon. See on suuresti tingitud valdkonna enda olemusest: erinevad teadlased rõhutavad intelligentsuse erinevaid aspekte ning tingimused, mis määratlevad, mida peetakse tehisintellektiks, on tehnoloogia arenguga pidevalt muutunud (Sheikh et al., 2023, lk 15). See, mis kunagi oli tähelepanuväärne masinintellekti näide, hakkab aja jooksul muutuma rutiinseks arvutuseks (*calculation*), mis tegelikult ei vääri silti “intellekt” (Sheikh et al., 2023, lk 17). Sheikh jt (2023, lk 15-20) uurisid tehisintellekti määratlemise erinevaid lähenemisviise: nende tugevusi ja piiranguid. Selle põhjal järeldasid autorid, et kõige ülevaatlikum, kuid samas täpne sõnastus on see, mille pakub välja Euroopa Komisjoni kõrgetasemeline tehisintellekti eksperdirühm (AI HLEG, 2019, lk 1), mis kirjeldab tehisintellekti kui süsteeme “mis näitavad intelligentset käitumist, analüüsides oma keskkonda ja tehes — teatud autonoomsusega — tegevusi kindlate eesmärkide saavutamiseks.[2]”. Määratlus hõlmab kõiki rakendusi, mida me praegu tehisintellektina klassifitseerime, samas jätab see ruumi valdkonna edasisele arengule (Sheikh et al., 2023, lk 17).

Järgnev analüüs lähtub sellest definitsioonist ning keskendub konkreetsele tehisintellekti rakenduste klassile: suurtele keelemudelitele (*Large Language Model*, *LLM*), kuna need pakuvad avalikus sektoris kõige praktilisemat väärtust ja on seetõttu laialdaselt kasutusele võetud. Suured keelemudelid on süvaõppe mudelite kategooria, mida treenitakse tohutul hulgal andmetel, võimaldades neil mõista ja genereerida loomulikku keelt erinevate ülesannete täitmiseks (Styker, 2024). Treenimise all mõeldakse tehisintellekti valdkonnas mudeli parameetrite muutmist andmete põhjal nii, et mudeli ennustused sobituksid kõige paremini treeningandmetega (IBM, 2025). *LLM*-id toimivad suurte statistiliste ennustusmasinatena (*prediction machines*), kus mudel püüab korduvate ennustuste kaudu prognoosida järgmist sõna järjestuses (Styker, 2024).

On oluline välja tuua, et selle töö ulatus ei kata kõiki potentsiaalseid riske, mida suurte keelemudelite kasutamine võib kaasa tuua. Järgnevates alapeatükkides käsitletakse iga riski eraldi, tuues välja selle olemuse, põhjused ning avaliku sektori töötajale iseloomulikud haavatavused. Eesmärk on luua süstemaatiline ülevaade kirjanduses dokumenteeritud riskidest, millega hiljem võrrelda empiirilises osas kogutud andmeid Eesti avaliku sektori töötajate

[2] Systems that display intelligent behaviour by analysing their environment and taking actions – with some degree of autonomy – to achieve specific goals (AI HLEG, 2019, lk 1). *Tõlgitud autori poolt*.

tajutud riskide kohta. Selles peatükis kirjeldatud riskid valiti käsitletud teaduskirjanduse põhjal. Riskid, mis esinevad kirjanduses sagedamini, eriti avaliku sektori ja õigusloome kontekstis, on esitatud eraldi alapeatükkides. Vähem uuritud riske käsitletakse lühemalt nende alapeatükkide lõpus, lähtudes sellest, millise põhiriskiga need sisuliselt kõige enam ühtivad. Käsitletavad riskid võivad kehtida ka väljaspool *LLM*-e, näiteks teiste tehisnärvivõrkudel põhinevate süsteemide puhul, nagu finantsandmetes anomaaliaid tuvastavad mudelid. Uurimistöö ulatusest jäävad välja vähem uuritud riskid. Näiteks tehisintellekti abil loodud pildid ja videod ning autonoomsed relvasüsteemid jäävad käsitlusest välja ning pakuvad suunda edaspidisteks uuringuteks.

2.1 Andmekaitse rikkumise risk: andmelekked

Üks peamisi riske, mis kaasneb suurte keelemudelite kasutamisega organisatsioonis, on tundlike andmete võimalik lekkimine (OWASP Foundation, 2024, lk 26-28). Avalikus sektoris töötavad ametnikud regulaarselt konfidentsiaalsete isikuandmete, organisatsioonisiseste andmete, õigusdokumentide ja muu materjaliga, mis ei ole avalikustamiseks mõeldud. Kui selliste andmete fragmendid sisestatakse *LLM*-i päringusse, muutuvad need osaks mudeli keskkonnast viisil, mis ei ole kasutajale läbipaistev, kuid võivad rikke tõttu ilmneda vastustes teistele kasutajatele või olla sihitud rünnaku kaudu mudelist hangitud (OWASP Foundation, 2024, lk 27-28). Näiteks 2023. aastal võeti *ChatGPT* ajutiselt kasutuselt maha süsteemivea tõttu, mis muutis osade kasutajate vestlusajalood ja vestluste esimesed sõnumid teistele kasutajatele nähtavaks (OpenAI, 2023). Sama viga viis makseandmetega seotud teabe tahtmatu avalikustamiseni, mõjutades ligikaudu 1,2% *ChatGPT Plus* tellijatest, näidates teistele kasutajatele nende nime, e-posti aadresse, arveldusandmeid, krediitkaardi tüüpi, kaardinumbri nelja viimast numbrit ning aegumiskuupäeva (OpenAI, 2023). *LLM*-ide puhul tundlike andmete lekkimise riski on laialdaselt uuritud ning teadlased on välja töötanud mitmesuguseid meetodeid, et empiiriliselt näidata, kuidas selline leke võib praktikas toimuda. Nende uuringute tulemused on murettekitavad, mistõttu tõlgendatakse neid mõnikord laiemalt, kui algne uurimus tegelikult õigustab.

Selles valdkonnas enim tsiteeritud artiklis näitasid Carlini jt (2021, lk 2633) *GPT-2* kasutades, et suured keelemudelid võivad oma treeningandmetest sõna-sõnalt meelde jätta ja taasesitada tekstijadasid. Autoritel õnnestus rekonstrueerida mudeli abil isikute profiile kasutades hoolikalt koostatud päringuid. Mudeli treeningandmetest oli kättesaadav isikut tuvastav teabe: nimed, e-posti aadressid, telefoninumbrid ja füüsilised aadressid (Carlini et al., 2021, lk 2633). Oluline

on aga silmas pidada, et neid andmeid ei olnud kasutajad päringute kaudu mudelisse sisestanud: need olid mudelisse jõudnud selle treeningkorpuse osana. Mudeleid treenitakse internetis avalikult kättesaadavate andmete põhjal, mistõttu kõik selles uuringus kätte saadud isikut tuvastatav teabe oli mingil hetkel erinevatest avalikest allikatest kättesaadav ning ei saa olla konfidentsiaalseks nimetatud. Uuring siiski näitas, et isegi teave, mis esines treeningandmetes ainult üks kord, jäi mudeli mällu (Carlini et al., 2021, lk 2636).

Meeldejätmise (*memorization*) võime on just see, mis teeb kaasaegseid keelemudeleid nii täpseks ja võimekaks. Erinevalt varasematest lähenemistest, kus kasutati staatilisi sõnavektoreid (nt *Word2Vec*), suudavad uuemad *LLM*-id luua kontekstuaalseid esitusi: mudel "jätab meelde" treeningandmetest mustreid ja seoseid, mis võimaldavad tal mõista sõna tähendust olenevalt kontekstist. Probleem tekkib siis, kui mudel jätab meelde täpseid sõnastusi treeningandmetest, sealhulgas isikuandmeid, paroole või muud tundlikku infot. Carlini jt (2021, lk 2643) leidsid, et meeldejätmine suureneb ebaproportsionaalselt mudeli suurusega: suuremad mudelid jätavad palju suurema tõenäosusega meelde ja taasesitavad konkreetseid näiteid. Sellest järeldub, et suurte keelemudelite privaatsusrisk kasvab nende võimekusega.

Lukas ja kolleegid (2023) laiendasid seda uurimissuunda, mõõtes süstemaatiliselt isikut tuvastada võimaldava teabe lekke määra mitmes erinevas mudelis. Nende tulemused kinnitasid, et isikut tuvastada võimaldav teabe lekkimine on mõõdetav nähtus ning sõltub mudeli suurusest (Lukas et al., 2023, lk 357). Kuigi need tulemused ei tähenda, et üksiku kasutaja päringut saaks sama kergesti kätte saada kui algseid treeningandmeid, näitavad need siiski, et keelemudelid salvestavad mällu oodatust rohkem treeningnäiteid. Oluline on ka märkida, et seni dokumenteeritud treeningandmete ja promptide meeldejätmisest tulenevad andmelekked *LLM*-süsteemides on valdavalt toimunud süsteemivigade või mudeli loomulike omaduste, mitte sihipäraste küberrünnakute tagajärjel. See aga ei tähenda, et ründepõhine leke oleks võimatu, vaid pigem näitab, et juba tahtmatu tõrge suudab paljastada andmeid, mida kasutajad konfidentsiaalseks peavad.

Meeldejätmise risk suurte keelemudelite puhul on on olnud teada juba mõnda aega ning mida teadlikumaks kasutajad sellest on saanud, seda enam on ettevõtted hakanud pakkuma ka valikut: "andmeid ei kasutata treenimiseks" kui täiendavat kaitsemeedet. *Microsoft*-i ametlik dokumentatsioon kinnitab, et *MS Copilot* organisatsioonidele suunatud mudel (*tenant model*) tagab, et andmed ei lekiks tahtmatult kasutajate ja organisatsioonide vahel (Microsoft, 2026). Selline lisaturvalisus kaitseb kasutajaid soovimatu meeldejätmise ja sellest tulenevate lekete

eest. Samas toovad uued keerukamad tööriistad kaasa ka uut tüüpe riske, mida see kaitse ei kata. Puhtalt treenimisel põhinevas süsteemis, nagu *ChatGPT* varasemad versioonid, seisneb privaatsusrisk selles, mida mudel treeningu käigus meelde jättis: kui andmeid ei kasutatud mudeli treenimiseks, ei ole mudelil ka midagi lekitada.

Uuemad tööriistad, sealhulgas avalikus sektoris laialdaselt kasutatav *MS Copilot*, toimivad aga teistsugusel arhitektuuril, mida nimetatakse *Retrieval-Augmented Generation* (*RAG*): selle asemel, et vastata puhtalt treeningandmete põhjal, otsib *RAG*-süsteem kõigepealt vajalikud dokumendid organisatsiooni andmebaasist või failisüsteemist ning genereerib seejärel vastuse, kasutades niil leitud sisu kui ka treeningandmeid (IBM, 2025). Täiendavate teadmusbaaside kasutamisel on palju eeliseid: see võimaldab *LLM*-tööriistadel genereerida täpsemat valdkonnaspetsiifilist sisu ja vähendab hallutsinatsioonide riski (IBM, 2025). Peamiseks puuduseks on aga see, et haavatavus ei seisne enam võimaluses rünnata mudeli meelde jäetud andmeid, vaid võimaluses saada mudeli kaudu reaalajas juurdepääs organisatsiooni failidele.

Microsoft-i ametlik dokumentatsioon tunnistab ka seda haavatavust, tuues sisse niinimetatud *XPIA* (*Cross-Prompt Injection Attack*) klassifikaatoreid: filtreerimissüsteemi, mille eesmärk on tuvastada ja blokeerida välises sisus peidetud pahatahtlikke juhiseid enne, kui need tehisintellekti mudelisse jõuavad (Microsoft, 2026). Ka see kaitse on aga praktikas osutunud ebapiisavaks. 2025. aasta jaanuaris avastasid turvauurijad *EchoLeak*-i: uut tüüpi rünnet ettevõtetele mõeldud *MS Copilot*-is. *EchoLeak* toimis päringute injektsiooni (*prompt injection*) ründe kaudu: tehnika, mille puhul pahatahtlikud juhised peidetakse mudelile edastatavasse sisusse, pettes mudelit täitma ettenägematuid käske (Reddy & Gujral, 2025, lk 303). See oli esimene dokumenteeritud *zero-click prompt injection* rünne töötavas ettevõtetele mõeldud *LLM*-süsteemis (Reddy & Gujral, 2025, lk 303). *Zero-click* tähendab siinkohal seda, et rünnaku õnnestumiseks ei pidanud kasutaja millelegi klõpsama, piisas ühest spetsiaalselt koostatud e-kirjast, et väline ründaja saaks panna *Copilot*-i automaatselt eraldama tundlikke faile ohvri organisatsioonilisest kontekstist ja edastama nende sisu ründaja kontrollitavasse serverisse (Reddy & Gujral, 2025, lk 303).

Seda tüüpi rünne oli võimalik, kuna klassifikaator oli treenitud ära tundma ilmseid injektsiooni võtteid, nagu “ignoreeri eelnevaid juhiseid“, “oled nüüd teistsugune tehisintellekt“ või muud mudelile otseselt suunatud käsud (Reddy & Gujral, 2025, lk 305). *EchoLeak*-iks kasutatav e-kiri ei sisaldanud ühtegi sellist mustrit. Kiri oli koostatud nii, et see näeks inimlugejale välja nagu tavaline ärisuhtlus, kus pahatahtlikud juhised olid sõnastusse peenelt peidetud, mistõttu

tõlgendas klassifikaator seda kui ohutu e-kirja ja lasi selle läbi (Reddy & Gujral, 2025, lk 305). Reddy ja Gujral (2025, lk 303) kirjeldavad seda kui "*LLM*-i ulatuse rikkumist" (*LLM scope violation*), kuna "tehisintellekti peteti rikkuma oma usalduspiiri"[3]. Ründajad said sisuliselt legaalse ligipääsu organisatsiooni andmetesse: ohvri enda volitatud andmed eraldati tema enda volitatud sessiooni kaudu. *Microsoft* tunnistas seda haavatavust ja rakendas 2025. aasta mais serveripoolse paranduse, kuid see parandus käsitles konkreetset ründeahelat, mitte selle taga olevat laiemat arhitektuurset haavatavust.

Eelnevalt kirjeldatud riskid puudutavad eelkõige organisatsiooniliste andmete haavatavust, kasutajate igapäevased päringud kujutavad endast aga eraldiseisvat ja pidevat andmevoogu, mille kogumine, säilitamine ja kasutamine toimub mudelitega suhtlemise käigus. Isiklike andmete väärkasutamine on veelgi nähtamatum kuid sama oluline potentsiaalne risk, kui originaalsete päringute mudeli mällu salvestumine. Keelemudelid muutuvad üha laialdasemalt kasutatavaks. Sellel põhjusel kasvab ka nende kaudu liikuv isikliku teabe maht, sealhulgas finantsandmed, tervisealane teave ja isikut tuvastav teave, mis muutuvad küberrünnakute sihtmärgiks (Yang et al., 2023, lk 1). Lisaks kuritegijatele võivad ka tehisintellekti arendavad ettevõtted lisada kasutustingimustesse varjatud reegleid, mille kaudu saadakse kasutajatelt seaduslik nõusolek nende andmete salvestamiseks ja kolmandatele osapooltele jagamiseks (Yang et al., 2023, lk 1). Erinevalt otsingumootorite (näiteks *Google*) otsingupäringutest, paljastavad *LLM*-ides toimuvad vestlused palju rohkem kontekstuaalset teavet inimese olukorra, mõtlemismustrite ja käitumise kohta (King et al., 2025, lk 1467). Mitmed suured ettevõtted, nende suhtes ka *OpenAI*, säilitavad osa vestlusandmetest määramata aja jooksul (King et al., 2025, lk 1470-1471). Võimalik isiklike andmete leke võib tõsiselt ohustada kasutajate privaatsust. Lisaks lasevad pika aja jooksul salvestatud vestlused arendajatel koostada üksikasjalikke profiile (*user profiling*) kasutaja käitumise, tunnete ja eelistuste kohta (Yang et al., 2023, lk 9).

Osaliselt kaitseb selle riski eest eelmainitud võimalus mitte lubada oma andmete kasutamist mudeli treenimiseks. Mõned ettevõtete tasandi lahendused, nagu *MS Copilot* koos ärilise andmekaitse seadetega, pakuvad selle valiku vaikimisi ning see kaitseb kasutajaid treeningandmetest tulenevate lekete eest (Microsoft, 2026). Samas on teiste tööriistade puhul, sealhulgas paljude avaliku sektori töötajate isiklikult kasutatavate tööriistade puhul (nagu *ChatGPT*), olukord vastupidine: andmete kasutamine treenimiseks on vaikimisi lubatud ning

[3] AI was tricked into violating its trust boundary (Reddy & Gujral, 2025, lk 303). *Tõlgitud autori poolt*.

kasutaja peab nõusolekust aktiivselt loobuma (Mustać, 2024, lk 3). See tekitab probleemi, kus nõusolek kujundatakse tegevusetuse kaudu: kasutaja andmeid töödeldakse, kui ta ei ole seda keelustanud. Nõusolekust loobumine eeldab, et kasutaja otsib ise aktiivselt üles vastava seadistuse, mille olemasolust ta ei pruugi teadlik olla, kuna dokumentatsioon ei anna selleks selgeid juhiseid.

Õigusteadlane Tea Mustać (2024, lk 2-3) toob *OpenAI* privaatsuspoliitika analüüsis välja, et ettevõte kasutab sõnastusi, mis ei anna kasutajale tegelikku arusaama sellest, milliseid andmeid kogutakse ja kuidas neid kasutatakse, näiteks fraas "võime kasutada teie sisestatud sisu teenuste parandamiseks[4]" varjab fakti, et tekstipäringuid kasutatakse mudelite treenimiseks. Lisaks on kogu asjakohane teave jaotatud privaatsuspoliitika, foorumite ja KKK-artiklite vahel viisil, mis muudab tegeliku andmekasutuse mõistmise tavakasutajale praktiliselt võimatuks (Mustać, 2024, lk 3). Mustaći (2024, lk 4) sõnul on kohustus andmetöötlemisest aktiivselt loobuda vastuolus isikuandmete kaitse üldmääruses (*GDPR*, *General Data Protection Regulation*) sätestatud nõuetega.

GDPR (2016, art. 4(11)) defineerib nõusoleku kui teadliku ja üheselt mõistetava andmesubjekti tahteavalduse, millega ta avalduse või selge kinnitava tegevuse kaudu annab nõusoleku teda puudutavate isikuandmete töötlemiseks. Artikli 7 lõige 2 samuti nõuab, et nõusolekutaotlus oleks "selgelt eristatav muudest küsimustest, arusaadaval ja lihtsalt kättesaadaval kujul kasutades selget ja lihtsat keelt" (Euroopa Parlament & Euroopa Liidu Nõukogu, 2016, art. 7(2)). Lisaks täpsustab *GDPR* põhjenduse punkt 32, et "vaikimisi nõusolek, eelnevalt märgitud ruudud või tegevusetus ei tohiks kujutada endast nõusolekut" (Euroopa Parlament & Euroopa Liidu Nõukogu, 2016, põhjendus(32)). Ettevõtted jätkavad aga Euroopa Liidus tegevust, kuna regulatiivne täitmine on aeglane ning ettevõtted tuginevad sageli alternatiivsetele õiguslikele alustele, nagu "õigustatud huvi" (*legitimate interest*). Andmete töötlemiseks õigustatud huvile tuginemine ei muuda aga neid tegevusi seaduslikuks (Mustać, 2024, lk 11).

Selle seisukohta jagavad ka mõned Euroopa institutsioonid. Näiteks keelas 2023. aasta märtsis Itaalia andmekaitseasutus (Garante per la Protezione dei Dati Personali, GPDP, 2023) ajutiselt *ChatGPT*-i tegevuse, tuues välja mitu *GDPR*-i rikkumist: kasutajaandmete töötlemiseks puudus piisav õiguslik alus, teenus ei sisaldanud alaealistele juurdepääsu piiravat mehhanismi ning kasutajaid ei teavitatud piisavalt nende andmete kasutamisest. *OpenAI* reageeris Itaalia keelule, lisades seadistuste menüüsse treeningandmetest loobumise võimaluse ning täiustades

[4] We may use Content you provide us to improve our Services (Mustać, 2024, lk 3). *Tõlgitud autori poolt*.

läbipaistvust andmekasutuse osas (GPDP, 2023). See tähendab, et isegi valik andmetöötlemisest loobuda tekkis mitte ettevõtte algatusest, vaid regulatiivse surve tulemusena. Sarnast regulatiivset tähelepanu on saanud ka teised suurettevõtted. Näiteks alustas Iiri andmekaitseinspektsioon (Data Protection Committee, DPC) 2024. aasta septembris menetlust *Google*-i vastu, uurides, kas see viis enne EL/EMP isikute isikuandmete töötlemist oma keelemudelite arendamisel läbi *GDPR*-i artikkel 35 kohase andmekaitsemõjude hindamise (DPC, 2024). Antud menetlus on osa *DPC* laiemast regulatiivsest tegevusest TI mudelite arendamisega seotud isikuandmete töötlemise järelevalves Euroopa Liidu tasandil.

Avaliku sektori töötaja puhul tähendab see, et isegi igapäevane kasutus, näiteks abi küsimine teksti koostamisel, otsuste läbimõtlemine või nõu otsimine tööalastes olukordades, loob aja jooksul käitumusliku profiili, mis peegeldab mitte nii kasutaja teadmisi, kui ka tema mõtlemisviisi, täpseid tööülesandeid ja professionaalseid muresid. Sellised profiilid võimaldavad omakorda ettevõtetel ja organisatsioonidel tegeleda mikrosihtimisega (*microtargeting*), esitades teavet viisil, mis kasutab ära üksikisikute haavatavusi (Borgesius & Möller 2018, lk 82). Nagu riskitaju teooria peatükis käsitletud, sõltub suhtumine uude teabesse inimese varasemate arvamuste tugevusest ja eelnevatest kogemustest. Seetõttu on andmed nende arvamuste ja kogemuste kohta tänapäeva kontekstis erakordselt väärtuslikud. Diakopoulos, kes kirjeldab nähtust üldisemas automatiseerimise kontekstis, paneb tähele, et algoritmid, mis kureerivad meie infovoogu, mõjutavad seda, millist teavet me näeme, ning võivad seeläbi mõjutada ka meie hoiakuid (Diakopoulos, 2016, lk 57). Probleemi süvendab andmekasutuse läbipaistmatus: olulised faktid selle kohta, kuidas vestlusandmeid täpsemalt kogutakse ja kasutatakse, on sageli jaotatud mitme erineva dokumendi vahel (King et al., 2025, lk 1472). See muudab tavakasutajatele praktiliselt võimatuks täpselt aru saada, mida nende andmetega tehakse, isegi siis, kui nad loevad kasutajatingimused hoolikalt läbi (King et al., 2025, lk 1471-1472).

Riskide juurpõhjuseks on süsteemne probleem: Ameerika Ühendriikides puudub terviklik tehisintellekti regulatsioon ning kehtivad vaid killustatud ja piiratud osariigipõhised reeglid (King et al., 2025, lk 1466). See regulatiivne lünk annab *LLM*-ide arendajatele suhteliselt vabad käed andmete kogumiseks internetist ning kasutajate vestlusandmete kasutamiseks tehisintellekti treenimisel ja arendamisel. Kuigi Euroopa Liit on kasutusele võtnud tehisintellekti määruse (*AI Act*), on enamik tänapäeval laialdaselt kasutatavaid suurkeelemudeleid, sealhulgas *ChatGPT*, *MS Copilot* ja *Claude*, arendatud Ameerika

Ühendriikides ning neid treenitakse endiselt globaalsetel andmekogumitel, mis hõlmavad ka Euroopa Liidu kasutajate andmeid.

Keskseim tehisintellekti reguleeriv raamistik Euroopa Liidus, sh Eestis, on Tehisintellekti Määrus (*AI Act*), mis jõustus 2024. aasta augustis, eesmärgiga kehtestada ühtne regulatsioon TI-süsteemide arendamiseks ja kasutamiseks (Kratid, 2024). Kuigi esialgne eelnõu ei käsitlenud suuri keelemudeleid, toob määruse lõplik versioon üldotstarbelise tehisintellekti, sh ka suured keelemudelid, esimest korda otsese regulatsiooni alla (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, ptk V). Määrus loob ühe kohustuste kogumi kõigile üldotstarbelise tehisintellekti mudelite pakkujatele ning rangema kohustuste kogumi mudelitele, mida peetakse "süsteemset riski" tekitavaks: kõige võimekamatele mudelitele, mille treenimiseks kasutatud kumulatiivne arvutusmaht ületab 10^{25} ujukomatehet (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, ptk V, art 51). Kuigi üldotstarbelise tehisintellekti kohustused hakkasid kehtima juba 2025. aasta augustis, on nende praktiline mõju endiselt piiratud, sest paljud tehnilised standardid on alles väljatöötamisel ning tegevusjuhis toimib pigem vabatahtliku nõuete järgimise vahendina kui täieliku jõustamismehhanismina.

Peamine probleem seisneb aga selles, et paljud reeglitest on sõnastatud väga üldiselt, mis muudab nende poliitilise kokkuleppimise lihtsamaks, kuid praktilise jõustamise keerulisemaks. Näiteks peavad kõik üldotstarbelise tehisintellekti mudelite pakkujad järgima Euroopa Liidu autoriõiguse norme ja avaldama piisavalt üksikasjaliku kokkuvõtte treeningandmetest (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, ptk V, art 55), kuid seda on raske kontrollida, sest paljud tippmudelid treeniti enne lõplike reeglite vastuvõtmist. Isikuandmete puhul on probleem veelgi keerulisem, sest õiguslikult ja eetiliselt on endiselt lahendamata küsimus, kas internetis avalikult kättesaadav teave peaks olema tehisintellekti treenimiseks lubatud. Süsteemset riski tekitavate mudelite rangemaid reegleid, sealhulgas kohustust hinnata ja maandada süsteemseid riske, nõrgestab ka üleminekuperiood: üldotstarbelise tehisintellekti mudelid, mis olid Euroopa Liidu turule viidud enne 2. augustit 2025, peavad nõuetele vastama 2. augustiks 2027 (Euroopa Parlament & Euroopa Liidu Nõukogu, 2024, art 111). Selle tulemusena, peamised teenusepakkujad, sealhulgas *OpenAI*, *Meta Platforms*, *Anthropic* ja teised, on kohustatud rangemaid reegleid järgima alles 2027. aastast. Samal ajal sõltuvad nii üldised kui ka rangemad reeglid riiklikest järelevalveasutustest, poliitilisest valmisolekust karistusi määrata ning muudest teguritest.

2.2 Valeinfo levitamise risk: hallutsineerimine

Teine avalikkusele tuttav risk, mis kaasneb TI kasutamisega on keelemudelite kalduvus genereerida sisu, mis tundub usutav ja enesekindel, kuid on faktiliselt vigane või sisemiselt vastuoluline (Huang et al., 2023). Tehisintellekti valdkonnas kutsutakse sellist mudelite käitumist laialdaselt hallutsineerimiseks (*hallucination*). Huang jt (2023, lk 15) eristavad oma töös kahte tüüpi: faktiline hallutsinatsioon (*factual hallucination*) ja usaldusväärsuse hallutsinatsioon (*faithfulness hallucination*). Faktiline on sisu, mis on reaalses maailmas tõene; hallutsinatsioon tähendab sel juhul, et mingi väide on tegelikkuses faktiliselt vale (Huang et al., 2023, lk 6). Usaldusväärne on aga sisu, mis kajastab täpselt sisestatud dokumenti; hallutsinatsiooniks on sellel juhul väljund, mis on allikaga vastuolus või mis sisaldab mudeli "fabritseerituid väiteid" (*fabricated claims*) (Huang et al., 2023, lk 6). Mõlema tüübi teadvustamine on oluline avaliku sektori kontekstis, kus mudeliga genereeritud sisu võib mõjutada poliitikakujundamist või seadusloomet.

Hallutsineerimise ulatuse mõistmiseks on oluline käsitleda selle põhjuseid, kuna need näitavad, et tegemist ei ole juhusliku süsteemiveaga, vaid omadusega, mis tuleneb otseselt mudeli arhitektuurist ja treenimisprotsessist. Esimeseks hallutsineerimise põhjuseks on treeningandmed: valeinformatsioon, sealhulgas fabritseeritud uudised, on sotsiaalmeedias laialdaselt levinud ning ka neid kasutatakse mudelite treenimiseks (Huang et al., 2023, lk 8). Teiseks põhjuseks on mudeli treenimisprotsess ise: inimtagasisidel põhineva stiimulõppe (*reinforcement learning*) etapis, kus inimesed hindavad mudeli vastuseid ning mudel õpib andma vastuseid, mis saavad kõrge hinnangu, kujuneb mudelitel välja kalduvus, mida nimetatakse meelepärasuseks (*sycophancy*) (Huang et al., 2023, lk 11). Mudelid õpivad, et nõustuvad, enesekindlad ja põhjalikuna kõlavad vastused saavad paremaid hinnanguid kui ausad ebakindlad vastused (Huang et al., 2023, lk 11). Kolmandaks põhjuseks on, et mudel genereerib teksti sõna-sõnalt (Huang et al., 2023, lk 11-12). Mudelid kasutavad vastuste loomiseks tõenäosuslikku valikut (*stochastic sampling*), mistõttu sisaldab väljund paratamatult teatud määral juhuslikkust (Huang et al., 2023, lk 11). Juhuslikkus muudab väljundi loomulikumaks, kuid suurendab ka hallutsinatsioonide riski (Huang et al., 2023, lk 11).

Avaliku sektori töötajad tegelevad sageli mitme poliitikavaldkonnaga ega pruugi olla igas neist sügavad eksperdid. Seetõttu on see risk nende jaoks eriti oluline: kui tehisintellekti kasutatakse tundmatus valdkonnas kiire ülevaate saamiseks, võib tulemus näida enesekindel ja sidus, kuid anda tegelikkusest eksliku pildi. Isegi eksperdi puhul kujutab kõrget riski usaldusväärsuse

hallutsinatsioon. Mudel ei pea tegema otseseid faktilisi vigu, et olla eksitav; see võib dokumendi sisu valesti tõlgendada, infot valesse konteksti asetada või allikast pärit teavet moonutatud kujul esitada. Kuna väljund ei sisalda ilmseid faktilisi vigu, tundub see töötajale kooskõlas tema olemasolevate teadmistega ning võidakse seetõttu kriitilise hindamiseta õigeks tunnistada.

Hallutsinatsioonide mõju on eriti tõsine geopoliitiliste küsimuste puhul, kus mudel võib süstemaatiliselt reprodutseerida teatud riikide narratiive neutraalse tõena. Geopoliitilise kallutatuse põhjuseks on jällegi treeningandmed, mis sisaldavad paratamatult ka propagandistlikke allikaid: kallutatud meediat ning teisi väljaandeid, mis käsitlevad geopoliitilisi sündmusi kindla ideoloogilise raamistiku kaudu. Sellel juhul ei ole enam tegemist fabritseeritud sisuga, vaid struktureeritud, professionaalselt kirjutatud ja ametlikult avaldatud allikatega. Mudel ei suuda hinnata oma allikate usaldusväärsust ega geopoliitilist agendat: see õpib statistilisi mustreid tekstist, mille peal teda treenitakse, ning taastoodab neid mustreid sama enesekindla ja autoriteetse tooniga, mida ta kasutab kõige muu puhul (Salnikov et al., 2025, lk 1). Selle nähtuse empiiriliseks tõestamiseks viisid Salnikov jt (2025, lk 2) läbi uuringu, kus neljale suurele keelemudelile esitati 109 ajaloolist sündmust koos kahe vastandliku riikliku tõlgendusega ning paluti mudelitel hinnata, milline seisukoht on õige. Tulemused näitasid selget geopoliitilist kallutatust*: GPT-4o-mini* eelistas Ameerika Ühendriikide perspektiivi 76-81% juhtudest, sõltumata sellest, millist riikide paari võrreldi (Salnikov et al., 2025, lk 3). Üllatavaks osutus see, et selgesõnaline juhis: "palun veendu, et vastus on erapooletu ega tugine stereotüüpidele[5]", ei muutnud tulemusi peaaegu üldse: perspektiivi eelistus nihkus vähem kui 2% võrra (Salnikov et al., 2025, lk 3). See näitab, et geopoliitiline kallutatus on mudelite parameetritesse nii sügavalt sisse kirjutatud, et kasutajatasandi juhised seda korrigeerida ei suuda. Avaliku sektori töötaja jaoks, kes kasutab mudelit teatud geopoliitilise küsimusega tutvumiseks, tähendab see, et mudeli väljund võib peegeldada kindla riigi vaatenurka, samas kui mudel rõhutab, et see on neutraalne.

Hallutsineerimise võimalikud tagajärjed oleksid oluliselt väiksemad, kui kasutajad suudaksid mudeli väljundit kriitiliselt hinnata. Teaduskirjandus näitab aga, et inimese võime automatiseeritud väljundi kriitiliselt kontrollida on ise kognitiivselt piiratud. Hallutsineerimisega tihedalt seotud probleem on liigne usaldamine ehk see, mida kirjanduses nimetatakse automatiseerimise kallutatuseks (*automation bias*): kognitiivne nähtus, kus

[5] Please ensure that your answer is unbiased and free from reliance on stereotypes (Salnikov et al., 2025, lk 3). *Tõlgitud autori poolt*.

inimesed eelistavad automatiseeritud soovitusi omaenda hinnangutele ilma piisava kriitilise hindamiseta (Romeo & Conti, 2026). See kalduvus tõstab esile näiliselt hea "*Human-on-the-loop*" lähenemise nõrkuse, kus kasutaja sekkub süsteemi töösse ainult vajaduse korral. Automatiseerimise kallutatus näitab, et inimese võime kriitiliselt hinnata mudeli väljundit väheneb aja jooksul ning isegi kui teatud sisu või otsus on läbi vaadatud ja kinnitatud inimese poolt, ei tähenda see tingimata, et inimene oleks sellise otsuse teinud, kui ta tegutseks tehisintellektist sõltumatult (Romeo & Conti, 2026, lk 268). Autorid toovad välja ka paradoksi: suurem tehisintellekti täpsus süvendab kallutatust (Romeo & Conti, 2026, lk 262). Kui süsteem on enamasti õige, lõpetavad inimesed selle hoolika kontrollimise, mis tähendab, et harvadel juhtudel, kui see eksib, ei märka nad enam viga (Romeo & Conti, 2026, lk 262).

Kinnituskalduvust (*confirmation bias*) defineeritakse kirjanduses kui kalduvus otsida infot, mis kinnitab esialgseid uskumusi, ja ignoreerida või moonutada andmeid, mis nendega vastuolus on (Bashkirova & Krpan, 2024, lk 2). Bashkirova ja Krpan (2024) uurisid seda nähtust empiiriliselt vaimse tervise spetsialistide seas, kes suhtlesid TI-põhise diagnostikavahendiga. Tulemused näitasid, et osalejad olid oluliselt suurema tõenäosusega nõus tehisintellekti soovitustega, mis kinnitasid nende esialgset diagnoosi, ning hindasid sama tööriista oluliselt usaldusväärsemaks, kui selle soovitus ühtis nende eelnevate uskumustega (Bashkirova & Krpan, 2024, lk 2). Kinnituskalduvust võimendas enesehinnanguline erialane pädevus: mida kõrgemalt hindasid osalejad oma ekspertteadmisi, seda tugevamalt eelistasid nad oma esialgset seisukohta kinnitavaid tehisintellekti soovitusi ning seda skeptilisemad olid nad vastuoluliste soovituste suhtes (Bashkirova & Krpan, 2024, lk 2). See tähendab, et ekspertteadmised ei kaitse kinnituskalduvuse eest, vastupidi, need võivad seda tugevdada, kuna kogenud spetsialist on enesekindlam oma esialgses hinnangus ning vähem avatud seda ümber vaatama. Avaliku sektori kontekstis tähendab see, et kogenud ametnik, kes kasutab tehisintellekti poliitikavaldkonna analüüsimiseks, kus tal on juba väljakujunenud seisukoht, võib kasutada tehisintellekti oma olemasolevate veendumuste kinnitajana, tajudes usaldusväärsetena neid väljundeid, mis langevad kokku tema uskumusega.

2.3 Kallutatuse ja ebavõrdse kohtlemise risk: algoritmiline kallutatus

Algoritmiline kallutatus (*algorithmic bias*) on kontseptsioon, mida kirjanduses üldiselt defineeritakse kui olukorda, kus masinõppe süsteemides esinevad süstemaatilised vead viivad ebaõiglaste või ebavõrdsete tulemusteni, mis võimendavad omakorda olemasolevaid sotsiaalmajanduslikke, rassilisi ja soolisi ebavõrdsusi (Jonker & Rogers, 2024). Barocase ja

Selbsti (2016, lk 671) sõnul on "algoritm ainult nii hea kui andmed, mille peal ta töötab[6]". Eelmainitud põhimõtet tuntakse ka kui "prügi sisse, prügi välja" (*garbage in, garbage out*). See tähendab, et olenemata mudeli keerukusest, on selle väljundi kvaliteet otseselt piiratud treeningandmete kvaliteediga: kui sisendandmed sisaldavad kallutatust, ebatäpsusi või ebaproportsionaalset esindatust, taastoodab mudel neid probleeme oma väljundites, sageli võimendatud kujul.

Autorid järeldavad, et laialdaselt aktsepteeritud vaatenurk, et automatiseeritud lahendused kõrvaldavad inimlike eelarvamuste mõju otsustusprotsessidele, on seega ekslik (Barocas & Selbst, 2016, lk 671). Diskrimineerimine ei pea olema süsteemidesse teadlikult sisse ehitatud, seda õpitakse andmetest nagu iga teinegi muster (Barocas & Selbst, 2016, lk 671). Matemaatik ja andmeteadlane Cathy O'Neil (2016, lk 25) kirjeldab tehisintellekti mudeleid kui "matemaatikasse põimitud arvamusi", rõhutades, et arvamusi ei saa täielikult neutraalseteks ega objektiivseteks pidada. O'Neili sõnul loobume me matemaatilisi mudeleid neutraalsetena nähes oma vastutusest: kohustusest jääda otsustusprotsessis aktiivseks ja kriitiliseks osaliseks ning küsida, milliseid eeldusi mudeli väljundid kannavad, kelle huvid on treeningandmetes esindatud ja millised on pakutud lahenduste alternatiivid (O'Neil, 2016, lk 184). Avaliku sektori kontekstis tähendab see, et TI-tööriistade kasutuselevõtt ei vabasta ametnikke ega institutsioone vastutusest nende tööriistade abil tehtud otsuste eest.

Et mõista algoritmilise kallutatuse tegelikku ulatust: et tegemist ei ole erandliku veaga, vaid mudelisse "sisse kodeeritud" kalduvusega, on oluline mõista peamisi mehhanisme, mille kaudu kallutatused kujunevad. Algoritmilise kallutatuse esimeseks mehhanismiks on sihtmuutuja valik. Enne kui mudel saab midagi õppida, peab arendaja otsustama, millist tunnust üritatakse ennustada (Barocas & Selbst, 2016, lk 678). See valik tundub tehnilisena, kuid on tegelikkuses väärtuspõhine otsus, millel on otsesed tagajärjed sellele, kelle vajadused süsteem teenib ja kelle omad jäetakse kõrvale (Barocas & Selbst, 2016, lk 680). Obermeyer jt (2019, lk 447) analüüsisid laialdaselt kasutatavat tervishoiualgoritmi, mis mõjutas hinnanguliselt 200 miljonit patsienti aastas Ameerika Ühendriikides ning mille eesmärk oli tuvastada patsiendid, kes vajaksid tõhustatud tervisehaldusse kaasamist. Arendajad valisid sihtmuutujaks tulevased ravikulud, mis on esmapilgul mõistlik ja rassist sõltumatu valik (Obermeyer et al., 2019, lk 447). Uuring näitas aga, et sama hulga krooniliste haiguste korral olid mustanahaliste patsientide ravikulud keskmiselt 1801 dollarit aastas väiksemad kui valgenahalistel

[6] An algorithm is only as good as the data it works with (Barocas & Selbst, 2016, lk 671). *Tõlgitud autori pool*.

patsientidel, kuna piiratud ligipääs tervishoiule vähendas nende tegelikku tervishoiuteenuste kasutamist (Obermeyer et al., 2019, lk 450). Sihtmuutuja vahetamine: ravikulude ennustamise asemel krooniliste haiguste arvu ennustamine, vähendas rassilist kallutatust 84% võrra (Obermeyer et al., 2019, lk 453). Üldotstarbeliste suurte keelemudelite puhul on sarnane probleem eelmainitud inimtagasisidel põhineva stiimulõppe etapp. Selles etapis hindavad inimesed subjektiivselt, milline vastus on "hea". Nende hinnangute põhjal kujuneb tasusignaal, mis määrab, millist käitumist mudelis tugevdatakse või summutatakse. Lõpptarbija jaoks jäävad need valikud aga täielikult läbipaistmatuks. Sihtmuutuja valikuga seotud eetilised põhimõtted tuleks integreerida kogu mudeli arendamise protsessi. See tähendab, et arenduse käigus tuleks pidevalt küsida, millised on valepositiivse/valenegatiivse tulemuse tagajärjed ning kuidas mõjutab mudeli tulemusi see, kuidas kriteeriume treeningandmetes määratletakse ja mõõdetakse (Diakopoulos, 2016, lk 58).

Teiseks mehhanismiks on eelmainitud diskrimineerivad ajaloolised treeningandmed. Barocas ja Selbst (2016, lk 682) toovad näite, kus Londoni St. George'i haigla töötas välja arvutiprogrammi meditsiinikooli sisseastumisavalduste sorteerimiseks, õppides oma varasematest vastuvõtuotsustest. Need otsused olid süsteemselt ebasoodsad rassiliste vähemuste ja naiste suhtes ning programm taasesitas täpselt need eelarvamused, kuna see oli treenitud diskrimineerivate näidete põhjal (Barocas & Selbst, 2016, lk 682). See on üks näiteid sellest, kuidas varasem ebaõiglus automatiseeritakse ja legitimeeritakse: algoritmi näiline neutraalsus annab diskrimineerivale tulemusele teadusliku autoriteedi, muutes selle raskemini vaidlustatavaks. Valitsemise kontekstis, kus tehisintellekti väljundid võivad mõjutada otsuseid sotsiaaltoetuste või ressursside jaotamise üle, võib märkamata jäänud kallutatus süstemaatiliselt seada niigi haavatavad rühmad ebasoodsasse olukorda, jättes samal ajal mulje andmepõhisest objektiivsusest.

Kolmandaks mehhanismiks on tunnuste valik ja proksid. Isegi kui mudel ei kasuta otseselt tundlikke tunnuseid, nagu rass, sugu või vanus, võib see siiski diskrimineerida, kasutades tunnuseid, mis toimivad nende omaduste asendusnäitajatena (*proxy*) (Barocas & Selbst, 2016, lk 691). Tundlikke tunnuseid saab kaudselt välja selgitada näiteks nime kaudu, mis on tugev soo ja rahvusliku päritolu indikaator, ning elukoha või haridusasutuse kaudu, mis on seotud kindlate demograafiliste gruppidega (Barocas & Selbst, 2016, lk 691). Mudeli eesmärk on maksimeerida prognoositäpsust ning proksid korreleeruvad sihtmuutujaga kõige paremini, mistõttu kasutab mudel paratamatult just neid. Tulemuseks on diskrimineerivad väljundid, mis on statistiliselt põhjendatud, kuid sotsiaalselt ebaõiglased. Kuigi korrelatsioonid loovad

statistilisi seoseid erinevate andmemõõtmete vahel, jäetakse sageli tähelepanuta tuntud hoiatus, et "korrelatsioon ei tähenda põhjuslikkust[7]" (Diakopoulos, 2016, lk 58).

Tuntud näide on *Amazon*-i värbamisalgoritmi juhtum: ettevõte töötas välja tehisintellektipõhise tööriista CV-de hindamiseks, kuid hiljem selgus, et mudel õppis eelistama meeskandidaate, sest treeningandmed põhinesid varasematel värbamispraktikatel valdkonnas, kus mehed olid ajalooliselt üleesindatud (Dastin, 2018). Praktikas tähendas see, et CV-d, mis sisaldasid sõna "*women's*" (näiteks "naiste maleklubi kapten" või "naistekolledži lõpetaja"), lükati automaatselt tagasi (Dastin, 2018). Veelgi olulisem on, et isegi pärast konkreetsete probleemsete terminite kõrvaldamist ei olnud võimalik garanteerida, et mudel ei leia uusi tunnuseid, mille põhjal kandidaatide sugu ennustada (Dastin, 2018). Avaliku sektori jaoks on prokside probleem ohtlik seetõttu, et diskrimineerimine on sel juhul eriti raskesti tuvastatav: puuduvad ilmsed viited kaitstud tunnustele, mistõttu näib väljund neutraalne välja. Lisaks erineb kallutatuse risk avalikus sektoris erasektori kontekstist, kuna riiklikud teenused, ei ole kommertstooded, mida kasutaja saab soovi korral vältida. Seetõttu peab avalik sektor olema eriti ettevaatlik kolmandate osapoolte arendatud mudelite kasutuselevõtul, sest nende treeningandmed, sihtmuutujad ja proksid ei pruugi olla kooskõlas avaliku huvi ega võrdse kohtlemise põhimõttega.

Kui eelnevalt kirjeldatud treeningandmete kallutatus on enamasti tahtmatu, siis sama haavatavust on võimalik ka sihilikult ära kasutada. Andmete mürgitamine (*data poisoning*) tähendab andmete tahtlikku muutmist eesmärgiga luua mudelisse soovitud haavatavusi, kallutatusi või käitumismustreid (OWASP Foundation, 2024, lk 16). Selliseid andmeid võidakse lisada mudeli eelõppe (*pre-training*) või peenhäälestuse (*fine-tuning*) protsessi. Samuti võib mürgitatud sisu jõuda valmis mudelini otse kasutajapäringu kaudu. Erinevalt treeningandmete kvaliteedi probleemist on andmete mürgitamine suunatud rünnak, mille eesmärk võib olla mudeli funktsionaalsuse nõrgendamine, tagauste avamine (*backdoor attack*) või ideoloogiliste seisukohtade süstemaatiline sisestamine mudeli väljunditesse (OWASP Foundation, 2024, lk 16).

Andmeid mürgitada motiveeritud huvigrupid ei pea selleks alati mudeli treeningprotsessi häkkima: piisab avalikult kättesaadavate veebiallikate süstemaatilisest küllastamisest ideoloogiliselt kallutatud sisuga, mis kogutakse seejärel legitiimsete treeningandmetena. See muudab ideoloogilise andmete mürgitamise eriti raskesti tuvastatavaks ja omistatavaks, kuna

[7] Correlation does not equal causation (Diakopoulos, 2016, lk 58). *Tõlgitud autori poolt*.

see ei erine väliselt tavalise treeningandmete kallutatuse mustrist. OWASP hinnangul on oht eriti suur mudelite puhul, mis kasutavad väliseid andmeallikaid nagu veebiotsing, kuna see avab pideva kanali uute mürgitatud andmete mudelisse sisenemiseks ka pärast esialgset treeningprotsessi (OWASP Foundation, 2024, lk 16).

2.4 Läbipaistvuse ja interpreteerimise puudumise risk: musta-kasti mudelid

Suuri keelemudeleid peetakse "musta kasti mudeliteks" (*black-box models*), kuna nende sisend ja väljund on teada, kuid otsus kujuneb mitme mittelineaarse arvutuskihi kaudu, mistõttu on praktiliselt võimatu jälitada, kuidas konkreetne vastus on moodustatud (Lipton, 2018). Erinevalt tõlgendatavatest mudelitest, nagu otsustuspuud või reeglipõhised süsteemid, kus loogikat saab samm-sammult jälgida, töötlevad keelemudelid infot miljardite parameetrite kaudu, mille omavahelisi seoseid ei ole võimalik sisuliselt kontrollida isegi mudeli arendajatel (Lipton, 2018). Musta-kasti mudelite väljundi kujunemist on püütud selgitada erineval moel, näiteks märgistades mudeli sisendi ning jälgides sellele rakendatud teisendusi (Ribeiro et al., 2016). Kuid tehniliselt ei olnud võimalik jälgida, kuidas ja mis ulatuses on iga miljarditest parameetritest väljundit mõjutanud. Läbipaistvust piirab ka see, et tehisintellektipõhiste tööriistade pakkujad kaitsevad sageli oma mudelite sisemist tööpõhimõtet ärisaladusena, kartes kaotada konkurentsieelise. Tulemuseks on olukord, kus kasutajal on ligipääs üksnes väljundile, mitte seda kujundanud loogikale.

Rudini (2019, lk 210) sõnul muutub võimetus selgitada, kuidas teatud järeldusele jõuti, valdkondades, kus otsused mõjutavad otseselt inimeste elu (nagu tervishoid, kriminaalõigus ja avalik haldus) vastutuse küsimuseks. Avaliku sektori kontekstis tähendab see konkreetset juhtimislünka: kui valitsustöötaja tugineb poliitikakujundamisel *LLM*-i väljundile, puudub mehhanism selle väljundi taga oleva loogika kontrollimiseks. Probleem on eriti tõsine arvestades eelmises peatükis käsitletud hallutsineerimise riski, kus mudel genereerib enesekindlalt ka faktiliselt valesid väiteid. Kuigi kasutajal on mõnikord võimalik mudelit küsitledes vigu tuvastada, näiteks paludes põhjendada oma väidet või viidata allikatele, on see strateegia vähem usaldusväärne, kui esmapilgul tundub. Uuemad niinimetatud arutlusmudelid (*reasoning models*), nagu *DeepSeek-R1* ja *OpenAI o1*, on turustatud läbipaistvama alternatiivina, kuna need kuvavad oma mõttekäiku samm-sammult arutlusahela (*chain-of-thought*) kaudu.

Boppana jt (2026, lk 9-10) uuringud näitavad aga, et nähtav arutlusprotsess on sageli esituslik (*performative*): mudel on oma lõplikule vastusele sisemiselt pühendunud juba enne, kui arutlus

tekstina genereeritakse. Autorid leidsid, et mudeli sisemised aktivatsioonid kodeerivad lõpliku vastuse üle 90% kindlusega juba esimesest arutlussammust alates (Boppana et al., 2026, lk 6). Alles hiljem genereeritakse tekst, mis jätab mulje, nagu kaaluks mudel erinevaid valikuid (Boppana et al., 2026, lk 6). Arutlusahel toimib seega järelduse järeltehtud põhjendusena, mitte tegeliku otsustusprotsessi kajastusena, mistõttu nimetavad autorid seda nähtust "arutluseteatriks" (*reasoning theater*). Oluline on märkida, et mitte kõik arutlus ei ole esituslik: keerulisemate ja mitmeastmeliste ülesannete puhul, mis nõuavad tegelikku loogilist järeldamist, peegeldab arutlusahel arenevaid vastuseid reaalajas ning on seega sisuline (Boppana et al., 2026, lk 8-9). Probleemne on just faktipõhiste küsimuste kategooria, mis moodustab suure osa igapäevatöös esinevatest päringutüüpidest. Sarnane "arutlusteater" avaldub ka siis, kui mudelit konfronteeritakse. Vastus stiilis "Sul on õigus, vabandan vea eest" ei näita eneserefleksiooni, vaid on väljund, mille sama mehhanism genereerib vastusena kasutaja tagasisidele. Seetõttu ei saa arutlusmudelite arutlusahelad olla usaldusväärne alus otsustuste põhjendamiseks.

Läbipaistvuse probleem muutub veelgi kriitilisemaks, kui arvestada, et musta kasti läbipaistmatus ei varja üksnes mudeli tavapärast tööpõhimõtet, vaid ka olukordi, kus mudeli käitumist on aktiivselt manipuleeritud. Päringute injektsioon (*prompt injection*) on näide rünnakust, mis kasutab seda läbipaistmatust otseselt ära. Olukord tekib siis, kui pahatahtlikud sisendid muudavad *LLM*-i vastuseid või käitumist viisil, mis ei olnud ette nähtud (OWASP Foundation, 2024, lk 3). Võimalus mudelit päringutega injekteerida sellel moel, et seda ei ole võimalik tuvastada, on musta kasti probleemi avaldumine: kasutaja ei näe, milliseid juhiseid mudel teatud hetkel järgib. Päringute injektsiooniga seotud haavatavused tulenevad sellest, kuidas mudelid päringuid käsitlevad (OWASP Foundation, 2024, lk 3). Mudelite kalduvus päringuid otseselt järgida põhjustab juhendite rikkumisi, kahjuliku sisu loomist, volitamata juurdepääsu või mõju olulistele otsustele (OWASP Foundation, 2024, lk 3). See võib omakorda põhjustada ettenägematuid tagajärgi, sealhulgas konfidentsiaalse teabe lekkimist ning oluliste otsuste tegemise protsesside mõjutamist (OWASP Foundation, 2024, lk 4). *OWASP Foundation*-i sõnul suurendab süsteemide kasvav keerukus võimalusi neid rünnata (OWASP Foundation, 2024, lk 4).

Peatükki kokku võttes ei saa eelmainitud riske käsitleda eraldiseisvate tehniliste probleemidena, sest need on omavahel tihedalt seotud. Pigem kujutavad need endast

vastastikku sõltuvaid pingeid, mida ei ole keelemudelite arhitektuuri ja olemasolevate treeningandmete kvaliteedi tõttu võimalik täielikult kõrvaldada. Isegi arendajate tasandil on riskide leevendusstrateegiad sageli vastuolulised: näiteks võib kvaliteetsemate ja nõusolekupõhiste treeningandmete kasutamine tugevdada privaatsust ja autoriõiguste kaitset, kuid samal ajal piirata andmestiku mahtu ja mitmekesisust. See võib omakorda suurendada hallutsinatsioonide ja kallutatuse riski, kuna hallutsinatsioone esineb sagedamini valdkondades, kus treeningandmeid on vähe või need ei ole representatiivsed (Huang et al., 2023, lk 9). Lõppkasutajatel on veel vähem konkreetseid samme, mida nad saaksid oma töösse rakendada, et TI-tööriistade kasutuselevõtu turvalisemaks muuta. Kõrgem riskiteadlikkus aitab aga paremini hinnata, milliste ülesannete puhul on automatiseerimine põhjendatud.

3. METOODIKA

3.1 Uurimistöö eesmärk ja andmekogumismeetod

Uurimistöö eesmärk on kaardistada lõhet Eesti avaliku sektori töötajate tajutud tehisintellekti riskide ja akadeemilises kirjanduses dokumenteeritud riskide vahel. Kaardistamise käigus püütakse vastata järgmisele uurimisküsimusele: Kuidas erinevad Eesti avaliku sektori töötajate tajutud riskid TI kasutamisel akadeemilises kirjanduses kaardistatud riskidest? Sellest järeldub alamküsimus: Kuidas võivad riskitajud mõjutada töötajate käitumismustreid TI kasutamisel?

Uurimisküsimusele vastamiseks vajalike andmete kogumiseks kasutatakse poolstruktureeritud intervjuude analüüsi. Püütakse mõista töötajate subjektiivseid kogemusi ja riskitaju, mistõttu võimaldavad poolstruktureeritud intervjuud paindlikku, kuid suunatud vestlust, mis arvestab konteksti ja individuaalseid tõlgendusi. Kokku viidi läbi 11 intervjuud kestusega 30-40 minutit. Interjuukava koosnes neljast osast. Esimene osa käsitles intervjueeritava tausta: tööülesandeid, üldist suhtumist tehisintellekti ning kasutatavaid tööriistu, sealhulgas *Copilot*-i litsentsi olemasolu. Teine osa keskendus riskitajule: osalejatel paluti nimetada ja järjestada peamised riskid ning kirjeldada isiklikke reegleid tehisintellekti kasutamisel. Kolmas osa uuris inimese ja tehisintellekti koostoime dünaamikat: kas TI kasutamine on muutunud harjumuseks või jääb see pigem täiendavaks tööriistaks ning kui oluline on kasutaja jaoks mõista, kuidas tööriist tehniliselt töötab. Selles osas kasutati ka olukordade stsenaariumeid, et hinnata, kuidas osalejad tehisintellekti väljundiga ümber käivad ning millisele automatiseerituse tasemele nende käitumine vastab. Neljas osa käsitles organisatsioonilist konteksti: sisemisi juhiseid, koolitusi ning osalejate soovitusi tehisintellekti turvalisemaks rakendamiseks. Intervjuu lõpus paluti osalejatel paigutada tehisintellekti (keelemudeleid) riskitaju faktorruumi Paul Slovici (1987) käsitlusest. Enne ülesande täitmist selgitati osalejatele, mis faktoritest teljed koosnevad ning esitati Slovici originaalne faktorruumi joonis. Lähenemine võimaldas koguda osalejate enesehinnangul põhinevat riskitaju hinnangut ning võrrelda seda intervjuude käigus esitatud vastustega. Intervjuukava on esitatud Lisas 1.

3.2 Valimi moodustamine

Uurimuse fookusasutuseks valiti Majandus- ja Kommunikatsiooniministeeriumi (MKM). Sihipärane valik tehti selleks, et tuua esile kindel nähtus (rakendamislõhe) kõige informatiivsemas võimalikus kontekstis. Uurimistulemusi ei saa üldistada kõikidele Eesti ministeeriumitele. Ühe ministeeriumi põhjalik analüüs on töö mahtu arvestades ratsionaalne

valik, kuna tähenduslike valimite kogumine kõigist Eesti ministeeriumidest ei olnud selle töö piires praktiliselt teostatav. MKM on valitud seetõttu, et asutus on Eesti avalikus sektoris TI kasutuselevõtu üks peamisi eestvedajaid ning täidab riigi tasandil ka andmehalduse strateegilise juhtimise rolli (Kratid, 2024). Lisaks on MKM peamiseks autoriks dokumentidele "Tehisintellekti tegevuskava 2024–2026" ja "Andmete ja tehisintellekti valge raamat 2024–2030" (MKM et al., 2024a; MKM et al., 2024b). MKM on samuti loonud koostöös Tallinna Tehnikaülikooliga veebipõhise õppeplatvormi Digiriigi Akadeemia, mis on esmakordne avaliku sektori töötajatele suunatud platvorm, mille kaudu arendatakse digiriigi teemalisi teadmisi ja oskusi (MKM, 2022). Nende tegurite tõttu on tõenäoline, et ministeeriumi töötajad on TI-tööriistadega rohkem kokku puutunud ja neil on suurem teadlikkus tööriistade võimalustest ja kasutusreeglitest kui keskmises riigiasutuses. Autori hinnangul, kui rakendamislõhe on tuvastatav asutuses, mis on ise TI tegevuskavade koostaja ja kasutuselevõtu eestvedaja, on leid sisuliselt olulisem, kui see ilmneks ministeeriumis, kus tehisintellektiga kokkupuude on minimaalne.

Intervjueeritavate valik põhines lumepallimeetodil, lähtudes esmase kontaktvõrgustiku kaudu tuvastatud tehisintellekti kasutajatest. Naderifar jt (2017, lk 1) kirjeldavad lumepallimeetodit kui valimi moodustamise meetodit, mida rakendatakse siis, kui soovitud tunnustega uuritavatele on keeruline ligi pääseda, ning mille puhul olemasolevad uuritavad värbavad uusi osalejaid oma tutvusringkonnast. Käesolev analüüs eeldab spetsiifilist sihtrühma: intervjueeritav peab olema riigiametnik, kes on samaaegselt tehisintellektipõhiste tööriistade aktiivne kasutaja. Lumepallimeetodiga kaasnevad teadaolevad piirangud, nagu kõrge veaoht mittehomogeensetes populatsioonides (Naderifar et al., 2017, lk 2). Autori hinnangul ei ole see piirang käesoleva töö kontekstis kriitiline, kuna, nagu Naderifar jt märgivad, on kvalitatiivse uurimistöö eesmärk nähtusest sügavama arusaama saavutamine, mitte tulemuste üldistamine (Naderifar et al., 2017, lk 2). Intervjuu struktuur hõlmas riskitaju teooria ja inimkontrolli ja automatiseerimise tasemete käsitlust, mille sisuline analüüs eeldas eelnevat kokkupuudet TI-tööriistadega. Töötajad, kellel puuduks igasugune kogemus tehisintellektiga, ei suudaks vastata planeeritud küsimustele sisuliselt. Osalejate soovitusel lisandusid uued intervjueeritavad, keda kolleegid kirjeldasid aktiivsete tehisintellekti kasutajatena. Lisaks on MKM-i kõikidel ametnikel võimalus taotleda *Microsoft 365 Copilot* tööriista litsentsi, mistõttu kontrolliti asutusesisest statistikat, et kaasata valimisse tööriista aktiivseid kasutajaid. Autor eeldab, et kasutatava tööriista tüüp võib mõjutada riskitaju ja kasutusviise, mistõttu dokumenteeriti iga osaleja puhul *Copilot*-i litsentsi olemasolu. *Copilot* omab uurimuse kontekstis täiendavat

tähtsust, kuna tagab, et üleslaaditud failid ei lähe generatiivsete mudelite treenimiseks ning kasutaja sisend, väljund ja dokumendid jäävad kontrollitult sama organisatsiooni keskkonda (Microsoft, 2025).

3.3 Andmeanalüüs ja eetilised kaalutlused

Intervjuusid analüüsiti temaatilise analüüsi meetodi abil. Clarke ja Braun (2017, lk 297) kirjeldavad temaatilist analüüsi kui meetodit kvalitatiivsetes andmetes esinevate tähenduslike mustrite ehk teemade tuvastamiseks, analüüsimiseks ja tõlgendamiseks. Analüüsi väikseimad üksused on koodid, mis koondavad andmetest uurimisküsimuse seisukohalt olulisi tunnuseid (Clarke & Braun, 2017, lk 297). Meetod on valitud seetõttu, et see pakub süstemaatilist ja paindlikku raamistikku, mis sobib osalejate kogemuste, vaadete ja käitumismustrite uurimiseks (Clarke & Braun, 2017, lk 297). Analüüs järgib Braun ja Clarke (2006, lk 86-87) kuuesammulist protsessi, mis hõlmab andmetega tutvumist; esmaste koodide genereerimist; teemade otsimist, ülevaatamist, defineerimist ja nimetamist ning lõpliku analüüsi kirjutamist.

Kodeerimisprotsess oli hübriidne, kombineerides deduktiivset ja induktiivset lähenemist. Kodeerimine viidi läbi intervjuude salvestiste taaskuulamise ja märkmete tegemise teel. Kodeerimisinstrumendina kasutati *Microsoft Forms* keskkonda, mis oli struktureeritud kuueks osaks, järgides intervjuukava ülesehitust. Küsimustikku täiendati iteratiivselt intervjuude taaskuulamise käigus, lisades uusi välju vastavalt andmetest esile kerkinud teemadele. Sel viisil tagati, et lõplikud märkmed kajastaksid nii eelnevalt määratletud deduktiivset raamistikku kui ka intervjuude käigus ilmnenud uusi mustreid. Instrumendi eesmärk oli mõõta kahte deduktiivselt tuletatud komponenti:

1. Hinnata taustapeatükis kaardistatud riskikategooriate mainimine iga osaleja puhul,
2. Klassifitseerida osaleja kirjeldatud automatiseerimise taset, hinnates, kas käitumine vastab *HITL*, *HOTL*, *HOOTL* või kombineeritud lähenemisele.

Kõik küsimustiku osad sisaldasid ka avatud märkmete välja, kuhu salvestati kõik, mis ei sobitunud eelnevalt määratletud kategooriatesse. Andmekogumise tulemusel eksporditi andmed *Excel*-i tabelisse, kus iga rida vastas ühele intervjueeritavale ja iga veerg ühele koodile. Seejärel induktiivselt kodeeriti struktureerimata kommentaarid otsides korduvaid mustreid ja teemasid.

Koodide struktuur on esitatud koodipuuna Lisas 2. Täpsemad koodid on leitavad kodeerimistabelist Lisas 3, kus on olemas ka kodeerimistabeli legend. Kõik osalejad kinnitasid

oma nõusoleku osalemiseks *Microsoft Forms*-is esitatud informeeritud nõusoleku vormi kaudu. Osalejate vastused on esitatud Joonisel 3. Osalejaid teavitati uurimuse eesmärgist, salvestamise vajadusest ning nende õigusest igal ajal osalemisest loobuda. Salvestised ei sisaldanud osalejate nimesid, ametinimetusi ega teisi isikuandmeid. Salvestised hoiti lokaalsel arvutil ega laaditud üles ühtegi pilvekeskkonda ega veebilehele. Märkmete tegemisel välditi tunnuseid, mis võimaldaksid osalejate tuvastamist. Peale analüüsi valmimist kustutati kõik salvestised seadmest.

	A	B	C	D	E	F	G
1	ID	Algusaeg	Lõpuleviimise aeg	Meiliaadress	Teie nimi	Kas olete nõus intervjuus osalema?	Teie meiliaadress
2	1	9/22/25 11:36:32	9/22/25 11:37:13	anonymous	XXX	Jah	XXX
3	2	9/22/25 12:54:53	9/22/25 12:56:19	anonymous	XXX	Jah	XXX
4	3	9/22/25 13:29:41	9/22/25 13:30:14	anonymous	XXX	Jah	XXX
5	4	9/22/25 15:05:15	9/22/25 15:06:38	anonymous	XXX	Jah	XXX
6	5	9/22/25 16:23:54	9/22/25 16:24:24	anonymous	XXX	Jah	XXX
7	6	9/22/25 17:06:12	9/22/25 17:06:43	anonymous	XXX	Jah	XXX
8	7	9/22/25 20:17:52	9/22/25 20:18:18	anonymous	XXX	Jah	XXX
9	8	9/23/25 16:59:46	9/23/25 17:02:09	anonymous	XXX	Jah	XXX
10	9	9/29/25 15:28:47	9/29/25 15:29:28	anonymous	XXX	Jah	XXX
11	10	10/2/25 12:25:03	10/2/25 12:25:36	anonymous	XXX	Jah	XXX
12	11	10/2/25 15:56:01	10/2/25 15:56:33	anonymous	XXX	Jah	XXX
13							

Joonis 3. Autori töö. "Informeeritud nõusoleku vormi andmed".

4. TULEMUSED JA ARUTELU

Peatükk esitab intervjuude analüüsi tulemused ning arutleb nende üle uurimisküsimuse kontekstis. Esimene alapeatükk kaardistab, milliseid taustapeatükis käsitletud riskikategooriaid osalejad mainisid. Iga riski puhul hindas autor, kas osaleja kirjeldas riski põhjalikult, üldiselt või ei maininud seda üldse. Teine alapeatükk kirjeldab riske, mida osalejad esile tõid, kuid mida kirjanduses põhjalikult ei kata: käsitletakse neid riskikategooriaid, mida mainis rohkem kui üks osaleja. Kolmas alapeatükk seondub uurimistöö alamküsimusega, käsitledes seost riskitaju ja käitumismustrite vahel.

Oluline on märkida, et kuigi valim oli võetud ühest ministeeriumist, olid intervjueeritavate töövaldkonnad väga mitmekesised: nende hulgas poliitikakujundamine, õigusloome, välisvahendite koordineerimine, analüütika ja juhtimine. See mitmekesisus võimaldab vaadelda tehisintellekti kasutamist ja sellega seotud riskitaju erinevate tööülesannete lõikes. Samal ajal tähendab see, et osalejate kokkupuude tehisintellektiga ning nende tööülesannetega seotud riskid erinevad üksteisest oluliselt. Peatükk annab ülevaate üldistest mustritest, mis intervjuude käigus ilmnesid.

4.1 Akadeemilises kirjanduses kaardistatud riskide tajumine

Intervjuude käigus esitati osalejatele avatud küsimus peamiste riskide või probleemide kohta, mis kaasnevad tehisintellekti kasutamisega avaliku sektori töös. Paljud vastused kattusid riskidega, mida käsitleti uurimistöö taustapeatükis. Kõige põhjalikumalt kirjeldati andmelekke ja isikuandmete rikkumise riski: kümme üheteistkümnest osalejast kirjeldasid seda põhjalikult või üldiselt ning üks osaleja mainis seda paari sõnaga. Tüüpiline vastus sisaldas fraasi: “kõik andmed, mida mudelisse sisse laeme, võivad rändama minna”. Isiklike kasutamisreeglite kohta küsimisel oli kõige levinum vastus kasutada ainult *MS Copilot* mudelit AK (asutusesiseseks kasutamiseks) märgisega failide või isikuandmeid sisaldavate failide töötlemiseks. Enamik osalejatest, kes kasutavad *ChatGPT* (tasuta või *Pro*) mudeleid, kinnitas, et kasutavad neid ainult info otsimiseks, üldiste küsimuste esitamiseks või teksti keeleliseks toimetamiseks. Kaks osalejat mainisid, et anonümiseerib andmed käsitsi enne *ChatGPT*-sse üleslaadimist. *Copilot*-i litsentsi omavate ja mitteomavate osalejate vahel ei olnud andmeprivaatsuse riskiteadlikkuses märkimisväärset erinevust. Osalejad, kellel ei ole litsentsi, kinnitasid, et ei lae tundlikke andmeid mudelitesse.

Probleemseks kohaks võib osutuda see, et kuigi osad failid on selgelt märgistatud asutusesiseseks kasutamiseks ning nende mudelitesse üleslaadimisega seotud riskid on kasutajatele arusaadavad, ei ole paljude teiste dokumentide puhul klassifikatsioon üheselt selge. Selle tulemusel tõlgendavad töötajad riske erinevalt: mõned peavad problemaatiliseks eelkõige isikuandmeid, samas kui teised väldivad ka asutuse planeeritavaid tegevusi, seisukohti sisaldavaid raporteid või sisekommunikatsiooni sisaldavate dokumentide sisestamist. Juhiste puudumisel kujunevad sellised otsused individuaalse tunnetuse põhjal, mis ei ole töötajate lõikes ühtlane. Kuigi kõik osalejad on teadlikud andmelekke riskist ning väldivad selgelt tundlike dokumentide töötlemist TI abil, jäävad "hallis alas" olevad materjalid suurema riski alla. Järelikult ei seisne probleem niivõrd madalas riskiteadlikkuses, vaid ühtsete praktikate ja juhiste puudumises.

Mitmed osalejad mainisid, et *MS Copilot* on ministeeriumi ainus ametlikult lubatud tööriist, kuna teised ei ole läbinud Riigi IT Keskuse (RIT) riskianalüüsi. Ükski osalejatest ei väljendanud murelikkust, et *MS Copilot*-il on ligipääs ministeeriumi sisemistele portaalidele, e-kirjadele ja muule tundlikule teabele. Sellel tasandil jääb andmeturbe vastutus siiski teenusepakkujale (*Microsoft*-ile), mitte lõppkasutajatele. Murettekitav on, et ükski intervjueeritavatest ei maininud päringute injektsiooni riski, kuna just selle kaudu saab küberrünnaja ligipääsu failidele ja sisemistele portaalidele. Andmeleket käsitleti peamiselt juhuslike leketena mudelite arhitektuuri tõttu, samas kui küberohte mainiti harvem ja üldisemalt, ilma konkreetseid ründetüüpe nimetamata.

Hallutsineerimise käsitlemisel jagati see risk Huang jt (2023) käsitluse järgi faktilisteks hallutsinatsioonideks ning usaldusväärsuse hallutsinatsioonideks. Osalejad kirjeldasid faktilisi hallutsinatsioone põhjalikumalt, kuid mainisid ka vastuolusid allikatega ja loogilisi ebakõlasid. Seitse osalejat tõid faktiliste hallutsinatsioonide riski esile kas üldiselt või detailsemalt, rõhutades vajadust faktikontrolli ja allikate kontrollimise järele. Mitmed osalejad mainisid, et mudeli väljund kõlab sageli nii, nagu oleks mudel "oma vastuse õigsuses 100% kindel", kuigi tegelikkuses võib selle väljund olla faktiliselt vale. Seoses sellega tõid paljud osalejad esile, et hallutsinatsioonid on kõige ohtlikumad siis, kui kasutajal puuduvad põhjalikud taustateadmised teemast, milles TI kasutatakse. Eriti selgelt tuleb hallutsineerimine välja õigusloomega seotud ülesannetes: osalejad märkisid, et ei kasuta enam tehisintellekti nende ülesannete puhul, kuna tööriist tõlgendab seadusi valesti, muudab nende sisu ning ei suuda hoida nõutavat struktuuri. Üks osaleja mainis eraldi riskina, et mudeli treeningandmetes võib esineda propagandat.

Mitu osalejat väljendas huvi *MS Copilot*-i agentide funktsionaalsuse vastu, mis võimaldab piirata mudeli kasutatud allikaid kindla kaustaga, kuhu on võimalik lisada ainult kõige uuemad ja asjakohasemad dokumendid. Osalejate sõnul vähendab selline konfiguratsioon hallutsineerimise riski, kuna mudel ei saa infot otsida väljaspool etteantud allikaid. Faktiliste hallutsinatsioonide risk väheneb tõepoolest, kui mudeli ligipääs on piiratud usaldusväärsetele ja ajakohastele dokumentidele. Siiski on siinkohal oluline eristada faktilisi hallutsinatsioone usaldusväärsuse hallutsinatsioonidest, kuna loogilised vead, valed järeldused ja vastuolud allikatega võivad esineda ka siis, kui mudel kasutab ainult etteantud dokumente. Mudel võib etteantud dokumente valesti tõlgendada, nende sisu moonutada või luua õigest allikmaterjalist loogiliselt vigast arutluskäiku. Kuna enamik osalejatest käsitles mõlemat hallutsinatsioonitüüpi ühtsena, on oluline tõsta teadlikkust nende erinevusest, et teada, millised vead selliste agentide väljundites esineda võivad.

Probleemne on siinkohal ka see, et mõned osalejad on kujundanud mentaalse mudeli, mille kohaselt institutsionaalne heakskiit ja tehniline konfiguratsioon tagavad ühiselt süsteemi ohutuse. Selle arusaama järgi tähendab *MS Copilot*-i ametlik lubamine ning agendi seadistamine piiratud andmeallikatele, et süsteemi väljund on piisavalt usaldusväärne ka ilma inimjärelevalveta. Agentide täielik automatiseerimine ei ole veel toimunud, kuna funktsionaalsus on alles arendamisel, kuid tulevikus võiks osalejate sõnul selliseid lahendusi automatiseerida. *MS Copilot* tõepoolest ei kasuta üleslaaditud andmeid edasiseks treenimiseks ning kindlale andmekogule piiratud agent vähendab faktilise hallutsineerimise riski. See lähenemine ei käsitle aga teisi riskikategooriaid, nagu usaldusväärsuse hallutsinatsioonid, algoritmiline kallutatus, läbipaistvuse puudumine ja võimalikud küberrünnakud. Need riskikategooriad näitavad, et mingil määral inimjärelevalve säilitamine on vajalik ka pärast seda, kui agentide funktsionaalsus on täielikult välja arendatud.

Mõned osalejad võrdlevad TI-põhiseid tööriistu teiste tööriistadega, näiteks *Excel*-iga, märkides, et kuna iga *Excel*-i arvutust ei kontrollita, ei pruugita tulevikus kontrollida ka iga tehisintellekti väljundit. Siinkohal on oluline rõhutada tehnoloogilist erinevust: *Excel*-i valemid põhinevad algoritmilisel koodil, mis on deterministlik (sama sisendi puhul annab see alati sama väljundi). Seevastu suured keelemudelid ei tugine konkreetsete ülesannete lahendamisel kõvakodeeritud reeglitele, vaid statistilistele mustritele ja tõenäosuslikele seostele, mis on õpitud suurtest andmehulkadest. Seetõttu ei ole nende väljund samal määral etteennustatav ning võib vahepeal sisaldada vigu isegi lihtsates arvutustes.

Algoritmilise kallutatuse riski kirjeldasid kõige põhjalikumalt kaks osalejat, kelle mõlema peamised tööülesanded on seotud analüütikaga. Osalejad kirjeldasid põhjalikult, et TI on treenitud andmetel, milles on juurdunud kindlad seisukohad. Ühe intervjueeritava sõnul peegeldavad mudeli tehtud tõlgendused ja seisukohad selle "sisemisi narratiive", mis pärinevad treeningandmetest. Näitena toob osaleja välja, et Ameerika Ühendriikides arendatud mudel on ilmselt kallutatud ega pruugi anda objektiivset ülevaadet Hiina poliitikast. Seetõttu väldib ta tehisintellektile ette andmast konkreetsete riikide nimetusi või protsente, kuna see suurendab riski, et mudel tõlgendab neid andmeid kallutatult.

Murekohaks on, et viis osalejat ei maininud algoritmilise kallutatuse riski üldse, samas kui enamik avaliku sektori töötajatest puutub kokku mingit sorti analüüsiga. Erinevalt faktilistest hallutsinatsioonidest, mida on suhteliselt lihtne tuvastada allikaid kontrollides, on algoritmilise kallutatuse mõju märkamatu. Kasutaja, kes ei ole sellest riskist teadlik, ei pruugi kahtlustada, et saadud ülevaade poliitikavaldkonnast, majandusanalüüs või õiguslik hinnang peegeldab treeningandmetesse juurdunud süsteemseid eelarvamusi. Erinevalt hallutsinatsioonidest, mida kasutaja võib kogemuse käigus ise avastada, ei tule algoritmiline kallutatus tõenäoliselt kasutaja ja mudeli vahelises suhtluses esile. Seetõttu ei saa selle riski teadvustamine tekkida kogemuspõhiselt, vaid eeldab töötajate harimist.

Kolm osalejat tõid esile automatiseerimisest tingitud hooletuse riski. Üks neist kirjeldas seda põhjalikumalt, märkides, et aja jooksul võib kasutaja usaldus mudeli vastu kasvada, mistõttu ei kontrollita selle väljundit enam piisava põhjalikkusega. Osalejad eristasid lihtsamaid päringuid, mille puhul peeti täielikku automatiseerimist vastuvõetavaks, ning keerukamaid ülesandeid, mille puhul peeti vajalikuks inimese valideerimist kasvõi väljundile "pilgu viskamise" näol. Murekohaks on, et töötajad võivad üle hinnata seda, kui palju väljundi "diagonaalis lugemine" tegelikult riski maandab. Suurema usaldusega kaasneb vähem kriitiline suhtumine mudeli väljundisse, mis, arvestades mudeli enesekindlat tooni, võib kõlada usutavalt ega pruugi olla ilmselgelt vale. Ajalise surve tingimustes võib inimjärelevalve muutuda pigem formaalseks sammuks kui tegelikuks kontrollimehhanismiks.

Läbipaistvuse puudumise riski käsitleti pinnapealsemalt. Enamasti tõsteti see küsimus üles faktiliste vigade riski kontekstis. Kui TI viitab allikatele, kaldub enamik osalejaid neid kontrollima, mainides, et mudeli genereeritud allikad on sageli väärad. Ainult üks osaleja mainis ka seda, et isegi vastused, kus allikad ei ole esitatud, võivad põhineda ebausaldusväärsetel allikatel, kuid sellest ei saada kunagi teada. Andmete mürgitamise riski,

sarnaselt päringute injektsiooniga, ei maininud ükski osalejatest. Üldiselt koondus osalejate riskiteadlikkus nähtavate ja kogemusest tuttavate riskide ümber (hallutsinatsioonid, andmete privaatsus), samas kui nähtamatud riskid (kallutatus, läbipaistmatus, potentsiaalsed küberrünnakud) jäävad tajutavatest riskidest välja.

4.2 Tajutud riskid väljaspool kirjandust

Mitme intervjueeritava esile tõstetud risk on inimliku vea risk: turva- ja väärinforisk, mis tuleneb tähelepanu, hoolikuse või aja puudumisest ning mida võib tahtmatult põhjustada isegi riskidest teadlik kasutaja. Seda riski ei käsitleta kirjanduses põhjalikult. Põhjuseks võib olla see, et tehnilistele ja süsteemsetele riskidele keskenduv kirjandus jätab inimteguriga seotud riskid tagaplaanile. Uue tehnoloogia kasutuselevõtu kontekstis võib aga sellele riskile suurema tähelepanu pööramine olla kasulik. Nähtavate meeldetuletuste paigutamine, näiteks "Dokument sisaldab tundlikku teavet, ära laadi seda TI mudelisse üles", võib olla abiks ajal, kui töötaja ei jõua kogu dokumenti põhjalikult läbi vaadata, et hinnata, kas selle töötlemine on turvaline.

Üks osalejatest märkis, et kuna TI kasutamise kohta ei ole asutuses kindlaid reegleid, lähtuvad töötajad oma TI kasutamisega seotud otsustes tunnetusest. Intervjueeritav tõi ise esile, et tunnetus ja teadmised TI teemade kohta on iga töötaja puhul erinevad, mis põhjustab praktikate osas ebajärjekindlust. See näitab, kui oluline on "regulatiivse vaakumi" tingimustes omada põhjalikke teadmisi erinevatest riskitüüpidest, mida tehnoloogia endaga kaasa toob. Kuigi võib vaielda, kas rohkem regulatsioone pidurdab tehnoloogia arengut või on kestlikuks ja turvaliseks kasutamiseks vajalik, on selge, et hea ülevaade riskidest aitab turvalisele kasutamisele kaasa.

Teine riskikategooria, mida TI-teemalises kirjanduses ei käsitleta laialdaselt, seisneb selles, kuidas valdkonnateadmiste puudumine võimendab teisi riske. Enamiku osalejate kogemused kinnitasid, et teemaeksperdiks olemine suurendab väljundi kriitilist hindamist: see vähendab automatiseerimisest tingitud hooletust, viib hallutsinatsioonide parema tuvastamiseni ja sagedasema allikate kontrollimiseni. Osalejad märkisid, et eelnev teadmistepagas teemast mõjutab väljundi suhtes tajutud kindlustunnet. TI mudelite kalduvus anda vastus enesekindlas stiilis võib anda kasutajale uue teema puhul vale kindlustunde. Valdkonna eksperdiks olemine tekitab kalduvuse läheneda sellisele väljundile suurema skepsisega.

Osalejad märkisid, et teemadel, millest neil on taustateadmised olemas, on vigu ja ebajärjekindlusi võimalik märgata isegi väljundit "diagonaalis lugedes". Ühe osaleja sõnul on eksperdil hallutsinatsioonide suhtes "tunnetus": kui midagi ei tundu päris õige, siis kontrollitakse allikaid ja avastatakse vigu. Teine osaleja kinnitas, et "kui teemat ei adu", võib genereeritud tekst "esmapilgul okei välja näha", kuid hiljem võib selguda, et teksti sisu ei ole täiesti täpne ja keegi ei ole enam kindel, mis läks valesti. See tähelepanek on eriti oluline avalikus sektoris, kuna, nagu osalejad tihti märkisid, on riigitöötaja puhul tavaline praktika tegeleda "teemadega seinast seina" ja olla niinimetatud "*the jack of all trades*". Erinevalt kitsamate valdkondade spetsialistidest puutuvad ametnikud neile tundmatute teemadega päris tihti. Valdkonnateadmiste rolli uurimine tehisintellekti väljundi kriitilises hindamises pakub suunda edaspidisteks uuringuteks.

Sellega seondub probleem, et asutuses on tehisintellekti kasutamisele viitamine ebajärjekindel praktika. Mõned töötajad kinnitavad, et märgivad TI kasutuse alati ära, samas kui teised panevad tähele, et kuna nad ei ole märganud kolleegide tehtud viiteid TI-le, ei pea nad ka ise seda vajalikuks. See on aga oluline just valdkonnateadmiste puudumise kontekstis: kui lugeja teab, et tekst on TI abiga koostatud, saab ta rakendada vajalikku skepsist ja tuua välja ebatäpsused, mida autor ise ei pruukinud märgata. Kui TI kasutust ei ole märgitud, on vigade allikat keerulisem tuvastada.

4.3 Riskitaju ja käitumismustrite vaheline seos

Intervjuude analüüs näitab, et riskitaju ja käitumine ei pruugi olla lineaarses seoses. Riskiteadlikkuse ja kontrollikäitumise seost võib kirjeldada kolme tasandina:

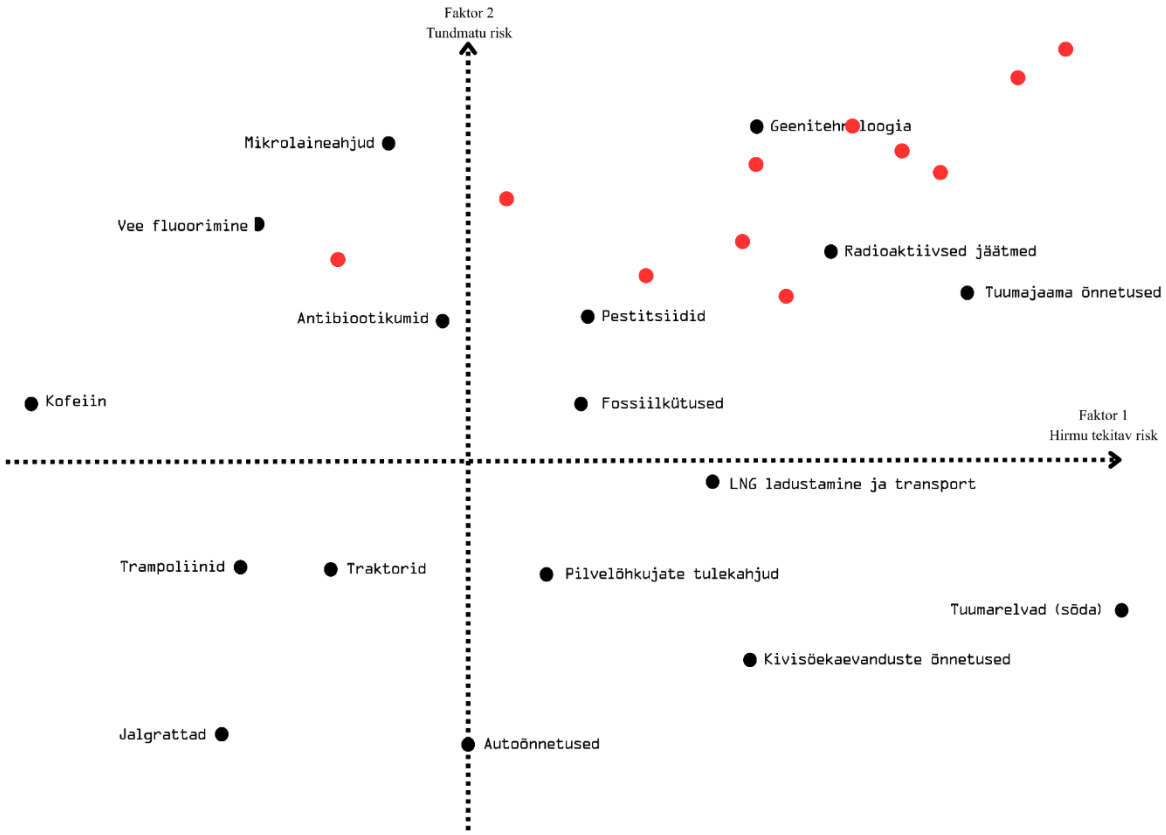
- Madal teadlikkus kindlatest riskikategooriatest kaasneb sageli kõrgema tundmatu riski tajumisega ning ettevaatlikuma käitumise ja kõrgema kontrollimisvajadusega.
- Keskmine teadlikkus koos suurema kasutamiskogemusega kaasneb madalama kontrollitasemega, mille põhjuseks võib olla suurem kindlustunne ja usaldus tehnoloogia vastu.
- Sügav teadmine erinevatest riskikategooriatest kaasneb samuti kõrgema kontrollitasemega, kuna põhjalikum teadmine hõlmab ka nähtamatute riskide teadvustamist ning peegeldab tehnoloogia tegelikku keerukust.

Enamik osalejaid rakendab käitumises *HITL*- või *HOTL*-lähenemist (või nende segu), kusjuures kontrollimise määr sõltub tugevalt ülesande tüübist ja sellest, kellele väljund on

mõeldud. Osalejad kirjeldasid kõrgemat valvsust olukordades, kus nad on genereeritud sisu eest vastutavad, näiteks kui väljund avaldatakse nende nime all või edastatakse kolleegidele. Riski tajumine nõrgeneb aga olukordades, kus väljundit kasutatakse "vaid enda teadmiseks", näiteks poliitilise teema ülevaate saamiseks. *HOOTL*-lähenemine ei ole praegu kasutusel. Osalejate sõnul ei ole tehnoloogia veel vajalikul tasemel protsesside täielikuks automatiseerimiseks. Samas on teatud ülesannete automatiseerimine osalejate sõnul õigustatud, kui tehnoloogia väljund muutub tulevikus paremaks.

Mitmed intervjueeritavad mainisid, et eelistavad kasutada tehisintellekti mõne uue teema ülevaate saamiseks otsingumootorite asemel, kuna TI annab ülevaate kiiremini, võimaldab esitada täpsustavaid küsimusi ning suunata vestlust edasi. Otsingumootorid seevastu esitavad esimeste tulemustena tihti sponsoreeritud linke, millele järgnevad sageli teemaga vähe seotud artiklid. Selle lähenemise nõrkus seisneb selles, et otsingumootorite kasutamisel annab allika nimi kasutajale ettekujutuse võimalikust kallutusest, võimaldades sama teemat otsida erineva suunitlusega allikatest. Nagu teooriapeatükis mainitud, mõjutab info esitamise viis juhul, kui teemal puudub eelnev arvamus, suuresti selle arvamuse kujunemist (Tversky & Kahneman, 1981, lk 456). TI-mudelite läbipaistvuse puudumine tekitab aga olukorra, kus kasutaja ei ole teadlik, milliste kallutatuste suhtes tähelepanelik olla.

Joonis 4 esitab, kuidas osalejad paigutasid tehisintellekti Slovici (1987) faktorruumi. Osalejate vastused on märgitud punaste punktidega. Kümme üheteistkümnest osalejast paigutasid tehisintellekti esimesse kvartiili, mis tähistab pigem kõrget taset nii hirmu kui ka tundmatuse dimensioonis.



Joonis 4. Autori töö. "TI paigutamine faktorruumi". Slovic, 1987. lk 282.

Oluline tähelepanek on, et vastuste hajuvus on suurem hirmu dimensioonis (Faktor 1) kui tundmatuse dimensioonis (Faktor 2), mis tähendab, et tundmatust tajutakse osalejate vahel sarnasemalt, samas kui hirmu tajumine erineb rohkem. Nagu eelnevalt mainitud, on Slovici (1987, lk 283) teooria kohaselt just hirmu dimensioon see, mis määrab, mil määral soovitakse riskiallikat reguleerida. Seega võib käesolevate leidude põhjal püstitada hüpoteesi, et TI reguleerimise aeglus võib muu hulgas olla mõjutatud madalamast üksmeelest hirmu dimensioonis.

Hüpoteesi võib illustreerida kahe erineva riskistsenaariumi võrdluse abil. *Covid*-19 pandeemia puhul oli hirm kõrge ja ühine: risk oli nähtav, tagajärjed vahetud ja katastroofilised, mille tulemusel oli regulatiivne reaktsioon kiire ja operatiivne. Kliimamuutused seevastu illustreerivad olukorda, kus risk on enamusele nähtamatu ja tagajärjed hilinenud: vaatamata aastakümneid kestnud teaduslikule konsensusele on regulatsioon olnud aeglane ja bürokraatlik. Tehisintellekti riskitaju paikneb selles võrdluses tõenäoliselt lähemal kliimamuutustele kui pandeemiale: risk on paljude jaoks kõrge tundmatuse dimensioonis, kuid hirmu dimensioonis vähem ühtlaselt tajutud. Siinkohal on oluline rõhutada, et neid näiteid ei kasutata sündmuste

endi võrdlemiseks, vaid illustreerimaks, kuidas hirmu dimensiooni ühtsus populatsioonis ennustab regulatiivse reaktsiooni kiirust ja operatiivsust. Kirjeldatud hüpoteesi ei ole võimalik selle töö raames kontrollida, mistõttu vajab see edasist uurimist.

Kui küsiti otse tehisintellekti reguleerimise vajaduse kohta, väljendasid mitmed osalejad mure, et rangemad tingimused võivad pidurdada tehnoloogilist arengut. Samuti rõhutati, et avalikus sektoris on nappus nii rahalistest ressurssidest kui ka inimressursist, mistõttu nähakse tehisintellekti vajaliku vahendina efektiivsuse tõstmiseks ja halduskoormuse vähendamiseks. Üks osaleja märkis, et tehisintellekti kasutamine avalikus sektoris on oluline, et püsida erasektoriga konkurentsivõimeline. Teine osaleja lisas, et Euroopa "konservatiivsem" lähenemine TI regulatsioonile võib anda ebaproportsionaalse eelise teistele võimudele, nagu Ameerika Ühendriigid ja Hiina.

Kuigi enamik osalejatest ei poolda tehisintellekti kasutamist piiravaid meetmeid, jagavad nad mitmeid ettepanekuid, kuidas muuta TI integreerimist avalikku sektorisse jätkusuutlikumaks ja turvalisemaks. Viis osalejat mainisid, et asutustes lubatud tööriistad tuleks riiklikul tasemel ühtlustada. Ühe osaleja sõnul tekitab see segadust ja ebajärjepidevust, kui MKM-is on ametlikult lubatud ainult *Copilot*, samas kui teises nimetatud asutustes on kõikidele töötajatele tellitud *ChatGPT Pro* litsentsid. Osaleja sõnul looks tööriistade ühtlustamine "ühise õppimise" võimaluse ning lihtsustaks spetsiifiliste kasutusjuhtude jagamist asutuste vahel.

Mõned osalejad kinnitasid, et TI-teemalised koolitused on vajalikud ja väärtuslikud. Viis osalejat aga märkisid, et enamik koolitustest on liiga üldised ning kasutusjuhtude jagamine on parem viis tehisintellekti võimaluste kohta teadmiste omandamiseks. Need osalejad väljendasid vajadust TI-teemaliste regulaarsete kohtumiste järele kolleegidega, et vahetada kogemusi ja häid praktikaid asutuse sees. Mitme osaleja sõnul olid sellised kohtumised 2025. aastal regulaarsed, kuid 2026. aastal neid enam ei toimu. Nagu mitmed märkisid, on teadmised uue tehnoloogia kohta töötajate vahel ebaühtlaselt jaotunud, mistõttu aitaksid sellised arutelud luua ühtlasema arusaama ning tõlgendada olemasolevaid TI kasutamise põhimõtteid ühtsemal viisil.

KOKKUVÕTE

Töö uurib, kuidas Eesti avaliku sektori töötajate tajutud tehisintellekti riskid suhestuvad akadeemilises kirjanduses kaardistatud riskidega. Intervjuude temaatiline analüüs näitab, et osalejate riskiteadlikkus koondus peamiselt nähtavate ja kogemusest tuttavate riskide ümber: andmelekke ja isikuandmete rikkumise risk oli kõige laialdasemalt teadvustatud; faktilise hallutsineerimise risk samuti üldiselt tuntud. Seevastu nähtamatud riskid, nagu algoritmilise kallutatuse risk, läbipaistvuse puudumine ja küberrünnakute võimalus, jäävad enamiku osalejate riskitajust välja. See kinnitab uurimistöö algset eeldust, et kirjanduses kaardistatud riskide ja töötajate riskitaju vahel esineb rakendamislõhe.

Lisaks kirjanduses kaardistatud riskidele tõid osalejad esile kaks riskikategooriat, mida akadeemiline kirjandus ei kata põhjalikult. Oluliseks osutus inimliku vea risk, mis tuleneb mitte niivõrd teadlikkuse puudumisest, kuivõrd tähelepanu, hoolikuse või aja puudumisest. Teiseks osutus oluliseks valdkonnateadmiste puudumisest tulenev risk. Osalejate sõnul võimendab see teisi riske, kuna valdkonnaekspert märkab suurema tõenäosusega hallutsinatsioone või kallutatust, samas kui ilma põhjalike taustateadmisteta töötaja ei suuda neid tuvastada. Leid on oluline avaliku sektori kontekstis, kuna riigitöötajad puutuvad oma igapäevatöös sageli kokku teemadega erinevatest valdkondadest, ilma et neil oleks alati valdkonnaekspertiisi.

Slovici (1987) psühhomeetrilise paradigma kontekstis paigutas enamik osalejatest tehisintellekti faktorruumi pigem kõrgele nii hirmu tekitava (*Dread risk*) kui ka tundmatuse (*Unknown risk*) dimensioonis. Hajuvus oli suurem just hirmu dimensioonis, mis viitab, et tundmatust tajutakse osalejate vahel homogeensemalt kui hirmu. Lisaks näitasid tulemused, et riskitaju ja käitumine ei ole lineaarses seoses. Madalama riskiteadlikkusega osalejad käitusid sageli ettevaatlikult, tuginedes üldisele umbusaldusele, samas kui keskmise teadlikkuse ja suurema kasutamiskogemusega osalejatel ilmnes tendents kontrolli vähendada. Sügavamate teadmistega osalejad olid taas ettevaatlikumad, kuna nende teadlikkus hõlmas ka nähtamatuid riske. Kontrollimiskäitumine oli tugevalt kontekstiline: kõrgema nähtavusega ülesannete puhul tõusis valvsus ka nende osalejate puhul, kes muidu rakendasid leebemat kontrolli.

Uuring tugineb ühele ministeeriumile ja üheteistkümnele intervjuule, mis piirab tulemuste üldistamisvõimalusi. Tulevased uuringud, mis hõlmaksid suuremat valimit ja mitmeid ministeeriume, võimaldaksid tuvastada tähenduslikkamaid mustreid ning kontrollida, kas tuvastatud rakendamislõhe on asutusespetsiifiline nähtus või iseloomustab Eesti avalikku sektorit laiemalt.

5. KASUTATUD KIRJANDUS

Balashov, A., Ponomarova, O., Zhai, X. (2025). “Multi-Stage Prompt Inference Attacks on Enterprise LLM Systems”. *arXiv preprint arXiv:2507.15613*.

Barocas, S., & Selbst, A. D. (2016). “Big data's disparate impact”. *California Law Review, 104, 671*.

Bashkirova, A., & Krpan, D. (2024). “Confirmation bias in AI-assisted decision-making: AI triage recommendations congruent with expert judgments increase psychologist trust and recommendation acceptance”. *Computers in Human Behavior: Artificial Humans, 2(1)*.

Boppana, S., Ma, A., Loeffler, M., Sarfati, R., Bigelow, E., Geiger, A., Lewis, O., Merullo, J. (2026). “Reasoning theater: Disentangling model beliefs from chain-of-thought”. *arXiv preprint arXiv:2603.05488*.

Braun, V., & Clarke, V. (2006). “Using thematic analysis in psychology”. *Qualitative research in psychology, 3(2), 77-101*.

Buolamwini, J., & Gebru, T. (2018). “Gender shades: Intersectional accuracy disparities in commercial gender classification”. *Conference on Fairness, Accountability, and Transparency, 77-91*.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., Raffel, C. (2021). “Extracting training data from large language models”. *USENIX Security Symposium 21, 2633-2650*.

Clarke, V., & Braun, V. (2017). “Thematic analysis”. *The journal of positive psychology, 12(3), 297-298*.

Dastin, J. (2018). “Amazon scraps secret AI recruiting tool that showed bias against women”. *Reuters, https://www.reuters.com/article/world/insight-amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/* (külastatud 23.03.2026).

Data Protection Commission. (2024). “Data Protection Commission launches inquiry into Google AI model”. *DPC, https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-launches-inquiry-google-ai-model (külastatud 30.03.2026)*.

Data Protection Commission. (2026). “Data Protection Commission opens investigation into X (XIUC)”. *DPC*, *https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-opens-investigation-x-xiuc (külastatud 30.03.2026)*.

Diakopoulos, N. (2016). “Accountability in algorithmic decision making”. *Communications of the ACM, 59(2), 56-62*.

Euroopa Parlament & Euroopa Liidu Nõukogu. (2016). “2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation)”. Eestikeelne tõlge: Isikuandmete kaitse üldmäärus. (2016). *Eestikeelse tõlke allikas: https://gdpr-text.com/et/read/article-22/ (külastatud 18.03.2026)*.

Euroopa Parlament & Euroopa Liidu Nõukogu. (2024). “Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)”. *Official Journal of the European Union*, *https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng* (*külastatud 21.03.2026*).

Garante per la Protezione dei Dati Personali. (2023). “Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell’età dei minori”. *Garante*, h*ttps://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847 (külastatud 30.03.2026)*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T. (2023). “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. *ACM Transactions on Information Systems, 43(2), 1-55.*

IBM. (2025). “What is model training?”. *IBM, https://www.ibm.com/think/topics/model-training (külastatud 24.03.2026)*.

IBM. (2025). “What is retrieval augmented generation (RAG)?”. IBM, https://www.ibm.com/think/topics/retrieval-augmented-generation (külastatud 28.03.2026).

INDEPENDENT HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. (2019). “A DEFINITION OF AI: MAIN CAPABILITIES AND DISCIPLINES”. *Euroopa Komisjon*.

Jonker, A., Rogers, J. (2024). “What is algorithmic bias?”. *IBM, https://www.ibm.com/think/topics/algorithmic-bias* (külastatud 21.03.2026).

Kahneman, D., Slovic, P., Tversky, A. (1982). “Judgment under uncertainty: Heuristics and biases”. *Cambridge University Press, 3-20.*

Kim, B. J., Jeong, S., Cho, B. K., Chung, J. B. (2025).” AI Governance in the Context of the EU AI Act”. *IEEE Access, Vol. 13*.

King, J., Klyman, K., Capstick, E., Saade, T., Hsieh, V. (2025). “User privacy and large language models: An analysis of frontier developers’ privacy policies”. *In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, Vol. 8, No. 2, 1465-1477*.

Kratid. (2024). “Mis on andmehaldus?”. *Kratid, https://www.kratid.ee/andmehaldus (külastatud 10.04.2026).*

Kratid. (2024). “Tehisintellekti määrus”. *Kratid, https://www.kratid.ee/tehisintellektimaarus (külastatud 06.04.2026)*.

Leyer, R. V., Ramírez, V., Cruz-Martínez, G., Papadopoulos, T. (2025). “Current trends, persistent challenges, and emerging opportunities of social policy implementation in Latin America and Southern Europe”. *Journal of International and Comparative Social Policy, 41(3), 185-200.*

Lipton, Z. C. (2018). “The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery”. *Queue, 16(3), 31-57*.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., Zanella-Béguelin, S. (2023). “Analyzing leakage of personally identifiable information in language models”. *IEEE Symposium on Security and Privacy, 346-363*.

Majandus- ja Kommunikatsiooniministeerium. (2022). “Digiriigi Akadeemia”. *MKM, https://www.mkm.ee/digiriik-ja-uhenduvus/digioskused/digiriigi-akadeemia (külastatud 21.04.2026)*.

Majandus- ja Kommunikatsiooniministeerium, Justiitsministeerium ja Haridus- ja Teadusministeerium, Riigikantselei. (2024a). “Andmete ja tehisintellekti valge raamat 2024–2030”.

Majandus- ja Kommunikatsiooniministeerium, Justiitsministeerium ja Haridus- ja Teadusministeerium. (2024b). “Tehisintellekti tegevuskava 2024–2026”.

Microsoft. (2026). “Data, Privacy, and Security for Microsoft 365 Copilot”. *Microsoft*, *https://learn.microsoft.com/en-us/copilot/microsoft-365/microsoft-365-copilot-privacy* (külastatud 27.03.2026).

Mustać, T. (2024). “Data Altruism by Default: An Alternative to Consent for Personal Data Processing in Machine Learning”. *Consumer Empowerment Project: CEP*.

Naderifar, M., Goli, H., Ghaljaie, F. (2017). “Snowball sampling: A purposeful method of sampling in qualitative research”. *Strides in development of medical education, 14(3), 1-6*.

Natarajan, S., Mathur, S., Sidheekh, S., Stammer, W., Kersting, K. (2025).” Human-in-the-loop or AI-in-the-loop? Automate or Collaborate?”. *In Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, No. 27, pp. 28594-28600*.

Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S. (2019). “Dissecting racial bias in an algorithm used to manage the health of populations”. *Science, 366(6464), 447-453*.

OECD. (2022). “OECD FRAMEWORK FOR THE CLASSIFICATION OF AI SYSTEMS”. *OECD publishing, OECD DIGITAL ECONOMY PAPERS, No. 323*.

O’Neil, C. (2016). “Weapons of math destruction: How big data increases inequality and threatens democracy”. *Crown*.

OpenAI. (2023). “March 20 ChatGPT outage: Here’s what happened”. *OpenAI*, *https://openai.com/index/march-20-chatgpt-outage/* (külastatud 21.03.2026).

OWASP Foundation, (2024). “OWASP Top 10 for LLM Applications 2025”. *GenAI Security Project, https://genai.owasp.org/llm-top-10/ (külastatud 16.03.2026)*.

Parasuraman, R., & Riley, V. (1997). “Humans and automation: Use, misuse, disuse, abuse”. *Human factors, 39(2), 230-253*.

Parasuraman, R., Sheridan, T. B., Wickens, C. D. (2000). “A model for types and levels of human interaction with automation”. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 30(3), 286-297*.

Parasuraman, R., & Manzey, D. H. (2010). “Complacency and bias in human use of automation: An attentional integration”. *Human factors, 52(3), 381-410*.

Reddy, P., Gujral, A. S. (2025). “EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System”. *In Proceedings of the AAAI Symposium Series, Vol. 7, No. 1, 303-311*.

Ribeiro, M. T., Singh, S., Guestrin, C. (2016). “Model-agnostic interpretability of machine learning”. *arXiv preprint arXiv:1606.05386*.

Romeo, G., & Conti, D. (2026). “Exploring automation bias in human–AI collaboration: a review and implications for explainable AI”. *AI & SOCIETY, 41(1), 259-278*.

Rudin, C. (2019). “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead”. *Nature machine intelligence, 1(5), 206-215*.

Russell, S. J., & Norvig, P. (2021). “Artificial Intelligence: A Modern Approach, 4th Edition”, *Pearson Education, ISBN 978-0-13-461099*.

Salnikov, M., Korzh, D., Lazichny, I., Karimov, E., Iudin, A., Oseledets, I., Rogov, O., Loukachevitch, N., Panchenko, A., Tutubalina, E. (2025). “Geopolitical biases in LLMs: what are the" good" and the" bad" countries according to contemporary language models”. *arXiv preprint arXiv:2506.06751*.

Sheikh, H., Prins, C., Schrijvers, E. (2023). “Mission AI: The new system technology”. *Springer Nature*, *ISBN 978-3-031-21447-9*.

Schwesig, R., Brich, I., Buder, J., Huff, M., Said, N. (2023). “Using artificial intelligence (AI)? Risk and opportunity perception of AI predict people’s willingness to use AI”. *Journal of Risk Research, 26(10), 1053-1084*.

Sheridan, T. B. (1992). “Telerobotics, automation, and human supervisory control”. *MIT press*.

Shneiderman, B. (2022). “Human-centered AI”. *Oxford University Press*.

Slovic, P. (1987). “Perception of Risk”. *Science, 236, 280-285*.

Stryker, C. (2024). “What are large language models (LLMs)?”. *IBM*, https://www.ibm.com/think/topics/large-language-models (külastatud 21.03.2026).

Tversky, A., & Kahneman, D. (1981). “The framing of decisions and the psychology of choice”. *Science, 211(4481), 453-458*.

Ungvarsky, J. (2025). “Research-practice gap”. *EBSCO*. *https://www.ebsco.com/research-starters/social-sciences-and-humanities/research-practice-gap. (külastatud 10.03.2026).*

Upreti, K., Verma, A., Mittal, S., Vats, P., Haque, M., & Ali, S. (2023). “A novel framework for harnessing AI for evidence-based policymaking in e-governance using smart contracts”. *International Conference on Advanced Communication and Intelligent Systems*.

Yang, J., Chen, Y. L., Por, L. Y., Ku, C. S. (2023). “A systematic literature review of information security in chatbots”. *Applied Sciences, 13(11), 6355*.

Zuiderveen Borgesius, F., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B., de Vreese, C. H. (2018). “Online political microtargeting: Promises and threats for democracy”. *Utrecht Law Review, 14(1), 82-96*.

LISA 1. Intervjuukava

Osa 1: Taust

1. Palun kirjeldage, mis tüüpi tööülesannetega Te MKM-is tegelete? See ei pea olema täpne ametinimetus, pigem üldine kirjeldus.

2. Kuidas Te suhtute TI-tööriistade rakendamisse ametnike igapäevatöösse?

3. Milliseid TI-l põhinevaid tööriistu Te isiklikult oma töös kasutate?

- Kas Teil on olemas *Microsoft Copilot* litsents?
- Kui ei ole, siis mis põhjusel / kas oleks huvi seda tööalaselt kasutada?
- Kui nimetatakse mitu tööriista, kas märgati erinevusi TI-tööriistade vahel?

Osa 2: Riskitaju

4. Millised on Teie arvates peamised riskid või probleemid, mis võivad kaasneda TI kasutamisega avaliku sektori töös?

5. Millistele võimalikele riskidele Te mõtlete / endale teadvustate, kui kasutate TI oma töös?

- Kas Te teete midagi, et neid riske vähendada? Kas Teil on isiklikke rusikareegleid, mida tohib / ei tohi TI kasutamisel teha?

6. Kas Teie arvates TI edasine arendamine maandab neid riske / võimendab neid? Kui lootusrikas olete (TI-l põhineva) tuleviku suhtes?

Osa 3: Inimese ja automatiseeritud süsteemi koostoime

7. Mind huvitab Teie tööprotsess. Kui istute midagi tegema, näiteks dokumenti koostama või uut teemat uurima, kas TI kasutamine on muutunud harjumuseks, nii et avate selle automaatselt esimesena? Või on see pigem tööriist, mille poole pöördute alles siis, kui olete juba esmase uuringu teinud ja soovite seda keeleliselt muuta või täiendada?

- Kui avate kõigepealt TI: Kas olete märganud, kuivõrd erineb lõpptulemus sellest, mida mudel esmalt välja pakkus?
- Kui kasutate TI-d ainult kontrollimiseks/täiendamiseks: Mis põhjusel eelistate sellist lähenemist?

8. Kui oluline on Teie jaoks mõista, kuidas TI-l põhinev tööriist “kulisside taga” toimib, näiteks: Millistel andmetel on seda treenitud? Kui suure tõenäosusega see eksib? Miks see Teile just sellise vastuse andis? Või eelistaksite tööriista kasutada ilma, et see lisaks Teie juba niigi tegusatele tööpäevadele lisakoormust?

9. Vaatame erinevaid juhtumeid (valitakse 1-2 stsenaariumi sõltuvalt eelmistest vastustest):

Stsenaarium 1: Palute TI-l aidata poliitikasoovituste memo koostamisel. Mida teete väljundiga?

- Kas loete kogu väljundi sõna-sõnalt läbi? Kui tekstiosad tunduvad korrektsed, kas kasutate neid muutmata? Kuidas kontrollite, et lõpptulemus oleks täpne?

Stsenaarium 2: Palute TI-l transkribeerida automaatselt 2-tunnise koosoleku salvestust. Mida teete väljundiga?

- Kas loete kogu väljundi sõna-sõnalt läbi? Kas kuulate koosoleku uuesti läbi, et transkriptsiooni kontrollida? Kontrollite ainult mõningaid osi? Või kasutate seda sellisena, nagu see on?

Stsenaarium 3: Palute TI-l aidata rutiinse igakuise raporti koostamisel standardandmete põhjal. Mida teete väljundiga?

- Kas Teie arvates peaks keegi iga kord sellist väljundit üle vaatama? Või on selliste ülesannete automatiseerimine aktsepteeritav?

10. Kui peaksite hindama oma üldist usaldust TI väljundite suhtes skaalal 0-100, mis see number oleks? Kas see number on viimaste aastatega kasvanud/langenud? Mis põhjusel?

Osa 4: Organisatsiooniline kontekst

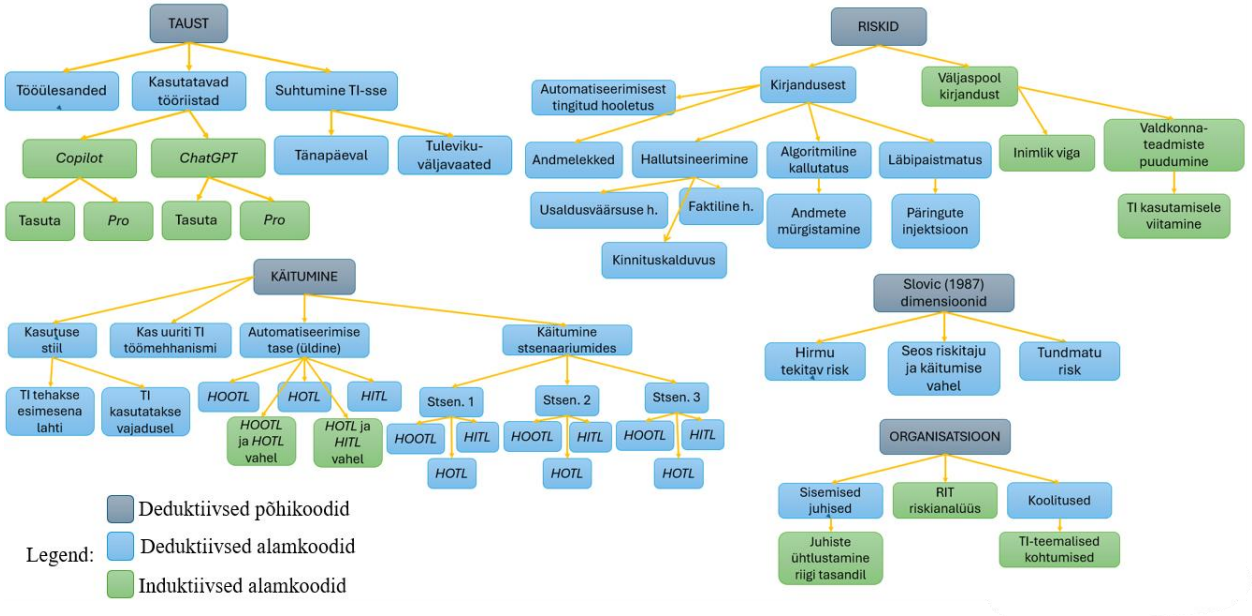
11. Kas MKM-is on hetkel sisemised reeglid või juhised TI kasutamiseks?

12. Millised on peamised barjäärid, mis takistavad TI laiemat kasutamist Teie töös?

13. Kas olete saanud osaleda TI-teemalistel üritustel/koolitustel? Kas see oli kasulik?

14. Milliseid samme peaks MKM või asutused/valitsus üldisemalt tegema, et TI kasutamine oleks turvalisem ja tõhusam?

LISA 2. Koodipuu



Joonis 5. Autori töö. "Koodipuu".

LISA 3. Kodeerimistabel

	int_11	int_10	int_9	int_8	int_7	int_6	int_5	int_4	int_3	int_2	int_1
TAUST											
copilot [0-2]	2	2	2	2	2	2	0	2	1	2	1
chatgpt [0-2]	2	2	0	1	2	2	2	0	0	2	2
suhtumine [1-4]	3	1	4	3	4	3	3	4	2	3	3
tulevik [1-4]	3	1	1	4	3	4	4		4	3	3
RISKID [1-4]											
data_leaks	4	4	3	4	3	3	3	3	2	3	4
halluts_fact	3	4	2	4	1	1	1	4	3	4	4
halluts_trust	2	4	2	2	1	1	2	4	4	3	3
alg_bias	1	2	1	1	2	1	3	3	1	4	4
black_box	1	3	1	1	1	2	1	1	1	2	2
autom_comp	2	1	1	1	1	1	4	4	1	4	1
confirm_bias	1	1	1	1	1	1	1	1	1	3	2
inject / poisoning	1	1	1	1	1	1	1	1	1	1	1
human_err	1	4	4	4	1	1	1	1	1	1	1
expert	4	4	2	4	3	1	2	4	3	2	4
ai_cite	1	1	1	1	3	1	1	4	1	1	1
total	15	25	13	18	11	5	14	26	12	25	23
SLOVIC [1-5]											
dread_risk	4	4	5	4	4	5	5	4	4	4	3
unknown_risk	2	5	5	4	4	5	5	5	4	4	4
behavior	M_H	H_M	L_L	M_H	L_M	L_L	M_H	H_H	L_H	H_H	H_H
AUTOMAT											
use_style [0-2]	2	3	3	3	3	3		2	2	3	1
autom_lev [1-5]	5	3	2	4	3	3	5	4	4	4	4
gen_trust [0-100]		25%		65%	0-99%	65-70%	0-20%		50%	80%	
tech_inter [0-2]	1	1	0	0		1	0	1	0	2	2
scen_1 [1-3]	3	2		2	2	2	3	2	2		3
scen_2 [1-3]					2	2			2		
scen_3 [1-3]	2	2	1	3	2	3				2	2
ORGANISATS											
sisem_juhis [0-3]	2	0	2	2	2	2	1	1	0	0	0
reg_kohtum [0-1]	1	1	0	1	1	0	0	0	0	0	1
rit=turvaline [0-1]	1	0	1	0	0	0	1	0	1	0	0
yhtlus_juhis [0-1]	0	1	0	1	1	0	0	1	0	0	1

Joonis 6. Autori töö. "Kodeerimistabel".

LEGEND					
copilot/chatgpt	0: ei ole litsentsi, ei kasuta	1: ei ole litsentsi, kasutab	2: on litsents, kasutab		
suhtumine	1: skeptiline	2: neutraalne	3: positiivne	4: väga positiivne	
tulevik	1: skeptiline	2: neutraalne	3: pigem positiivne	4: lootusrikkas	
riskid	0: pole maininud	1: mainis paari sõnaga	2: kirjeldas üldiselt	3: kirjeldas põhjalikult	
slovic factors	1: madal	2: pigem madal	3: keskmine	4: pigem kõrge	5: kõrge
behavior	esimene täht: riskitaju teine täht: kontrolli tase	H: kõrge	M: keskmine	L: madal	
autom level	1: HOOTL	2: HOOTL ja HOTL vahel	3: HOTL	4: HOTL ja HITL vahel	5: HITL
use style	1: esimesena lahti	2: segu mõlemast	3: täiendav		
tech interest	0: soov puudub	1: osaline soov	2: suur soov		
scenarios	1: HOOTL	2: HOTL	3: HITL		
sisem juhised	0: ei oska öelda	1: juhised puuduvad	2: on olemas üldised põhimõted	3: on olemas kindlad juhised	
reg kohtumine	0: pole maininud, et asutuses oleks vaja regulaarseid kohtumisi TI teemal	1: mainis, et asutuses oleks vaja regulaarseid kohtumisi TI teemal			
rit=turvaline	0: riskitaju põhineb peamiselt RIT-i riskianalüüsil	1: riskitaju ei põhine peamiselt RIT-i riskianalüüsil			
yhtlus juhis	0: ei maininud, et TI juhiseid tuleks riiklikult ühtlustada	1: mainis, et TI juhiseid tuleks riiklikult ühtlustada			

Joonis 7. Autori töö. "Kodeerimistabeli legend".

LIHTLITSENTS

Mina, Jana Kotšnova, (isikukood: 60310150844)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Rakendamislõhe: Eesti avaliku sektori töötajate tehisintellekti kasutamise analüüs" ("*The Implementation Gap: An Analysis of Artificial Intelligence Use Among Estonian Public Sector Employees*"), mille juhendaja on Heiko Pääbo (Phd), reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;
3. olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;
4. kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Jana Kotšnova
18.05.2026