

TARTU ÜLIKOOL

Sotsiaalteaduste valdkond

Ühiskonnateaduste instituut

Ühiskonna ja infoprotsesside analüüsi õppekava

Siim Klais

Tehisarurakenduste arendust toetava esmatasandi andmetöö korraldus ja nähtavus Eesti avalikus sektoris

Magistritöö

Juhendaja: Maris Männiste, PhD

Tartu 2026

SISUKORD

SISUKORD.....	2
MÕISTED JA DEFINITSIOONID.....	3
SISSEJUHATUS.....	4
1. TEOREETILINE ÜLEVAADE.....	6
1.1 Tehisaru definitsioon ja olemus.....	6
1.2 Kriitilised andmeuringud ja tehisaru.....	7
1.3 Andmetöö, nähtamatu töö, piiritöö ja tehisaru.....	10
1.4 Tehisaru rakendamine avalikus sektoris.....	13
1.5 Andmetöö tehisarurakenduste arendusel.....	14
2. PROBLEEMIPÜSTITUS JA UURIMISKÜSIMUSED.....	24
3. METOODIKA.....	25
3.1 Andmete kogumine: valim ja intervjuud.....	25
3.2 Andmete analüüs.....	29
3.3 Uuri ja refleksiivsus.....	31
3.4 Tehisaru kasutamisest.....	32
4. TULEMUSED.....	33
4.1 Rollide mitmekesisus tehisaruprojektides.....	33
4.2 Tehisaruandmetöö praktiline korraldus avaliku sektori projektides.....	35
4.3 Tehisaruandmetöö nähtavus ja prioriteet organisatsioonis.....	55
5. JÄRELDUSED JA ARUTELU.....	59
KOKKUVÕTE.....	67
SUMMARY.....	69
KASUTATUD KIRJANDUS.....	71
LISAD.....	78
Lisa 1: Intervjuukava.....	78
Lisa 2: Sisuanalüüsi koodiraamat.....	82
Lisa 3: Intervjueeritava informeeritud nõusoleku vorm.....	84
Lisa 4: Päring asutustele intervjueeritavate leidmiseks.....	85
Lisa 5: Lihtlitsents uuringu reprodutseerimiseks ja avalikustamiseks.....	86

MÕISTED JA DEFINITSIOONID

Tehisaru – katustermin arvutisüsteemidele, mis suudavad peamiselt masinõppe algoritmide toel täita ülesandeid, mis on seni eeldanud inimintelligentsi. Tarkvara, mis genereerib inimeste määratletud eesmärkide alusel väljundeid nagu sisu, prognoosid või otsused.

Tehisarurakendus – arvestataval määral tehisarutehnoloogiat (nt masinõpet) sisaldav konkreetne IT-lahendus, nagu tekstipõhised juturobotid, meediumituvastajad või otsustustoe tööriistad.

Masinõppe – tehnika, mis võimaldab arvutisüsteemil parandada oma tulemuslikkust tänu kogemuslikule õppimisele andmete vormis; masinõppe peamine ülesanne on arendada algoritme, mis aitaksid ehitada sisendandmetest mudeleid.

Algoritm – masinõppe kontekstis on algoritm matemaatiline meetod või reeglistik, mis võimaldab arvutil andmetest õppida, leida neis mustreid ja teha ennustusi või otsuseid, ilma et seda oleks otseselt igaks sammuks programmeeritud.

Mudel – andmete ning algoritmi (ehk teisisõnu õppimisreeglite) koosmõjus kujunev representatsioon, mis ei ole pelgalt tehniline artefakt, vaid kollektiivse töö käigus loodud konstruktsioon.

Andmetöö – tegevus, mis on seotud andmete kogumise, tõlgendamise, töötlemise, märgendamise ja süstematiseerimisega ning mis ei eelda süvatasemel tehnilisi andmeteaduslikke oskusi ning tööriistu.

Tehisaruandmetöö – käsitsi või poolkäsitsi tehtav inimtöö (andmete kogumine, puhastamine, märgendamine ja väljundi hindamine), mis toetab masinõppe algoritmide abil tehisarurakenduste ehitamist ning seda eelkõige treeningandmetike ettevalmistamise kaudu.

Nähtamatu töö – hädavajalik, kuid formaalselt tunnustamata, mõõtmata või alaväärtustatud taustatöö, mis jääb sageli ametlike arvestuste ja juhtimistasandi silme alt välja.

Kriitilised andmeuuringud – uurimistraditsioon, mis rõhutab, et andmed ei ole neutraalsed tõejäädvustused, vaid inimtekkelised artefaktid, mis kujunevad olemasolevate sotsiaalsete struktuuride ja võimuhete raamistikus.

Referentstõde (*ground truth*) – mudeli eesmärgi saavutamiseks vajalik kontekstispetsiifiline (treening)andmestik, mis toimib masinõppesüsteemides reeglite ja "seaduse" loojana.

Märgendamine (annoteerimine) – andmete täiendamine tähendusega (sildid, kategooriad, metaandmed), et aidata masinõppemudelitel infot mõista ja klassifitseerida.

SISSEJUHATUS

Tehisaru on muutunud avaliku sektori efektiivistamise diskursuses üheks keskseks jututeemaks, töötades tõhustada nii haldusprotsesse, parandada teenuste kvaliteeti ning samal ajal ka säästa avalikke ressursse. Katusmõiste *tehisaru* – st arvutisüsteem, mis suudab peamiselt masinõppe algoritmide ja nende ülikiiret arvutamist võimaldavate mikrokiipide toel “täita seni inimintelligentsi eeldanud ülesandeid” (Tangi jt, 2022) – hulka kuuluvate IT-rakenduste edukus sõltub suurel määral (lisaks tehnoloogias endas avalduva induktiivse eelhäälestuse) sisendandmete kvaliteedist (Amershi jt, 2019; Marcus, 2018; Muldoon jt, 2024; Tubaro jt, 2020). Sisendandmete ettevalmistamine ei ole aga üldjuhul automatiseeritud ning kõrgtehnoloogiline protsess, vaid töömahukas inimtöö.

Sisendandmega seotud töömahukaid tegevusi tehisaru kontekstis defineeritakse ka kui *tehisaruandmetööd* (*AI datawork*) – käsitsi või poolkäsitsi tehtav andmete kogumine, puhastamine ja märgendamine ning mudeli väljundi hindamine (Muldoon jt, 2024: 8–11; Girard-Chanudet, 2025: 4-8), mis moodustavad kuni 80% ühe keskpärase tehisaruprojekti töötundidest (Muldoon jt, 2023). Kuigi esile on kerkinud ka andmeanalüütiku ja andmeteadlase rollid, tehakse suur osa eelmainitud töödest endiselt traditsioonilistes ametites, kus neid ei käsitleta sageli keskse tööülesandena (Møller jt, 2020: 52). Samal ajal sunnivad uued, sealhulgas tehisaruga seotud andmevajadused traditsiooniliste ametite esindajaid omandama uusi andmepõhiseid oskusi, mida on praktikas seni alahinnanud nii töötajad ise, nende tööandjad aga ka akadeemilised uurijad (Møller jt, 2020: 52).

Tehisarurakendustes kasutatav masinõppemudel kui “andmete ja õppimisreeglite koosmõjus kujunev representatsioon” ei ole pelgalt tehniline artefakt, vaid kollektiivse töö käigus loodud konstruktsioon, mida kujundavad ühiselt valdkonnaekspertid, andmeteadlased ja arendajad ning mille kaudu on teadmised sinna samm-sammult akumuliseerunud (Girard-Chanudet, 2025: 4). Kriitiliste andmeuuringute (Critical Data Studies, CDS) uurimistraditsioon rõhutab, et andmed ei ole neutraalsed tõejäadvustused, vaid samuti inimtekkelised artefaktid, mis omakorda kujunevad ja toimivad olemasolevate sotsiaalsete struktuuride ja võimuhete raamistikus ning mille kujunemine ja tähendus on tihedalt seotud andmeloomet ümbritsevate protsesside, agentide ja institutsioonidega (Flensburg ja Lomborg, 2023; Hepp jt, 2022; Iliadis ja Russo, 2016; Loukissas, 2019; Richterich, 2018).

Sageli algoritmiliselt neutraalsetena kujutatud tehisarurakendused, olles suuresti andmetöö tulemusena tekkinud, tekitavad aga uusi küsimusi vastutuse ja kontrolli kohta (Girard-Chanudet, 2025: 2). Just seetõttu kaasneb ka tehisaru rakendamist toetava esmatasandi andmetöö nähtamatusega – nähtamatu töö kui hädavajalik, kuid formaalselt tunnustamata ja mõõtmata taustatöö (Star ja Strauss, 1999) – risk, kus eelmainitud rakenduste olulised eeldused, olemus ja väärtused jäävad varju. Avaliku sektori tänased tehisarurakendused, näiteks juturobotid, erinevad meediumituvastajad ja ka otsustustoe tööriistad, asenduvad tehnoloogiatrendide mõjul peagi komplekssemate ning autonoomsemate süsteemidega. Keerukad ning kodanike heaolu ja ka ühiskonda laiemalt, sealhulgas näiteks politikat, otseselt mõjutavad automatiseeritud teenused ja otsustusmehhanismid eeldavad aga tiptasemel andmetöö praktikaid, eriti aga andmete (ja andmetöö) dokumenteerimise, märgendamisotsuste ja kvaliteedikontrolli osas.

Eesti kui digiriik on seadnud sihiks tehisaru laialdase kasutuselevõtu (Justiits- ja Digiministeerium, 2025a), kuid senine fookus pole pööranud piisavalt tähelepanu praktilisele töökorraldusele seoses tehisaruandmetööga. Kes on need inimesed, kes andmeid koguvad ja märgendavad? Milline on nende ettevalmistus? Kuidas seda tööd rahastatakse, hangitakse, juhitakse, kontrollitakse ja väärtustatakse? Just seetõttu ongi uurimistöö eesmärk luua selgust tehisaruandmetööde nähtavuse ja korralduse hetkeseisust Eesti avaliku sektori tehisaruprojektides. Eesmärgi saavutamiseks on püstitatud kaks uurimisküsimust: (1) „Kuidas on Eesti avaliku sektori tehisaruprojektide esmatasandi andmetöö korraldatud?“ ning (2) „Kuidas tajuvad Eesti avaliku sektori tehisaruprojektide juhid esmatasandi andmetöö nähtavust ja prioriteeti oma projektides?“

Kui uurimistöö teoreetiline pool keskendub peamiselt tehisaru ja sealse andmetöö temaatika akadeemilistele käsitlustele, siis empiiriline pool tugineb poolstruktureeritud intervjuudel Eesti avaliku sektori projektijuhtide, tooteomanike ja valdkonnaekspertidega, kel kõigil on oma asutuste näidetel ka vahetud kokkupuuted tehisarurakenduste arendusprojektidega. Uuringu tulemused ning sealsetest leidudest inspireeritud arutelu sektsioon panustavad aga tehisaru ja sealse andmetööga seotud akadeemilistesse uurimissuundatesse (nii andmetööga seotud üldine korraldus, uues võtmes tõe- ja teadmusalade, andmehariduse, aga ka digitaalse sõltumatuse suunad) ning ühtlasi pakuvad praktilist sisendit organisatsioonide vastava töökorralduse parandamiseks. Samuti võib uuring anda sisendi ka strateegiadokumentide, näiteks riikliku andmete tegevuskava (Justiits- ja Digiministeerium, 2025b) edasiarendamiseks, aidates sellega muuta tehisaruga seotud andmetööd selgelt juhitud strateegiliseks varaks.

1. TEOREETILINE ÜLEVAADE

Järgnev teoreetiline ülevaade avab tehisaru mõiste, paigutab selle kriitiliste andmeuuringute raamistikku ning käsitleb andmetööd kui sageli nähtamatuks jäävat, kuid keskset praktikat tehisarurakenduste arendamisel. Lisaks käsitletakse tehisaru rakendamist avalikus sektoris ning andmetöö rolli tehisarutehnoloogial põhinevate rakenduste arendusprotsessis.

1.1 Tehisaru definitsioon ja olemus

Euroopa Komisjoni tehisintellekti määruse (AI Act) kohaselt viitab tehisarurakendus tarkvarale, mis on välja töötatud masinõppel, loogikal- ja teadmusel põhinevate ning statistiliste lähenemisviiside abil, ning mis suudab etteantud, inimeste määratletud eesmärkide alusel genereerida väljundeid, nagu näiteks erinev tekstiline või pildiline sisu, prognoos, soovitus või otsus, mis omakorda mõjutavad keskkondi, millega see kokku puutub (European Parliament, 2025). Tehisaru kui katustermin hõlmab aga mitmeid alamvaldkondi, sealhulgas masinõpe (ML), robotika ja arvutinägemine (Hanassab jt, 2024: 2). Oluline on eristada tehisarurakendust – mida argikeeles üha sagedamini nimetatakse lihtsalt AI-ks ning mis viitab eeskätt praktilisele kasutuslahendusele – ja neid meetodeid ning tehnikaid mille abil selliseid rakendusi luuakse.

Masinõpet võib defineerida kui tehnikat, mis võimaldab parandada arvutisüsteemi tulemuslikkust ja seda tänu kogemuslikule õppimisele arvutuslike meetodite abil – kuna arvutisüsteemides kogemus eksisteerib andmete vormis, on seega masinõppe peamine ülesanne arendada õppivaid algoritme, mis ehitavad omakorda andmetest vastavaid mudeleid (Zhou, 2021: 2). Masinõppe meetodid jagunevad tavaliselt kolme kategooriasse (Hanassab jt, 2024: 2-3; Mirtaheri ja Shahbazian, 2022): (1) juhendatud masinõppes kasutatakse sisend- ja väljundandmetega märgendatud andmeid, et õppida nende vahelist seost ja prognoosida tulemusi uute juhtumite korral (sageli kasutatakse kas klassifitseerimisel või pool-juhendatud õppes ka kõrvalekallete tuvastamisel), (2) juhendamata masinõppes puuduvad väljundmärgendid ning fookus on andmete sees struktuuri leidmisel, näiteks klasterdamisel ning (3) tugevdusmasinõppes õpib algoritm keskkonnaga suheldes eesmärgi poole iseseisvalt liikuma, saades oma tegevuste eest tagasisidet tasu (*reward*) kujul (kasutatakse näiteks järgmise edasiviiva sammu ennustamisel).

Mudelite arendamisel kasutatakse tavaliselt treening-, valideerimis- ja testandmestikke: esimene mudeli õppimiseks, teine sisendandmete ja hüperparameetrite täiendavaks häälestamiseks ja

kolmas sõltumatuks hindamiseks. Sobiva masinõppemeetodi valik sõltub kontekstist. Praktikas järgitakse sageli aga üldisi tavasid ja reegleid.

1.2 Kriitilised andmeuuringud ja tehisaru

Kriitiliste andmeuuringute uurimistraditsioon annab antud uurimistööle olulise teaduslik-koolkondliku lähtekoha, mille kaudu mõista andmete ning sealtkaudu ka esmatasandi andmetöö kriitilist rolli. Uurimistraditsiooni väljatoomine on vajalik ka Eesti avaliku sektori organisatsioonides (aga ka laiemalt) tehisarurakendusi toetavate esmatasandi andmetööde olulisuse rõhutamisel – andmetöodes tehtavad suuremad ja väiksemad valikud määravad olulise komponendina, millist mõju loodavad lahendused konkreetsetele protsessidele ja ka ühiskonnale hiljem avaldama hakkavad.

Kriitilised andmeuuringud on saanud tõuke andmestumise (*datafication*) ja digitaliseerimise protsesside kiirest tõusust. Andmestumisega kaasuvaid (suur)andmeid on sageli kujutatud kui uut ressursi, millele rajatakse täiendav majanduslik ja ühiskondlik väärtus, mida omakorda illustreerib ka kuulus Silicon Valley päritolu loosung “andmed kui uus nafta” (Richterich, 2018: 1). Vastupidiselt aga üldlevinud ühiskondlikule diskursusele ja ka majanduses domineerivale dogmale, ei ole andmed lihtsalt “uus nafta” ega ka muul viisil looduse poolt iseenesest antu, vaid alati kellegi poolt loodud ja vormitud kindlas, juba olemasolevas ühiskonna- ja võimustruktuuris (Hepp jt, 2022: 4).

Just eelmainitud nn „California ideoloogia“ on pikalt edendanud arusaama, et andmed ja digitehnoloogiad viivad meid automaatselt parema elu poole; tegelikkuses aga näeme ka vastupidist mõju – tugevnevat järelevalvet, kapitalistlikku jälgimist, andmerassismi ja muid probleeme (Hepp jt, 2022: 2). Üha enam on hakatud küsima, kelle (ja mille) arvelt see kõik toimub ning millised riskid andmestumisega kaasnevad (Richterich, 2018: 2). Siinkohal kriitilised andmeuuringud vastanduvadki positivistliku koolkonna metodoloogiale andmeanalüüsis (Iliadis jt, 2016: 1; Hepp jt, 2022: 7) ning uurivad, kuidas andmed suhestuvad sotsiaalse transformatsiooniga, et jõuda parema arusaamani digitaalse andmestiku ja taristute rollist tänapäeva ühiskondades (Hepp jt, 2022: 6, 17). Ka kasvav huvi tõenduspõhise poliitika kujundamise vastu seab küsimuse alla, kas andmeteaduse abil toodetud arvulised tõendid peaksid alati olema aluseks poliitika ja uute valitsemisvormide kujundamisele – CDS-i uurijad on hakanud neid küsimusi analüüsima, uurides näiteks ka seda, kuidas andmeid kasutatakse muutuste toetamiseks valitsemispraktikas ja sotsiaalsetes organisatsioonides (Iliadis jt, 2016: 3).

Viimase kümnendi jooksul on ebaõiglus ja kahju, mida andmepõhised süsteemid toodavad, sageli koondatud mõiste alla *bias* ehk kallutatus (Miceli jt, 2022b: 2). Erinevad uuringud on näidanud, et kallutatus võib siseneda ka tehisarurakendustesse ning seda igal tasandil – alates andmetest ja mudeli disainist kuni selle rakendusteni (Miceli jt, 2022b: 2). Kazimzade ja Miceli (2020: 1) rõhutavad, et tehisaru aluseks olevas masinõppes peitub eriline oht just algoritmide näilises neutraalsuses – kuna algoritmid jõuavad järeldusteni iseseisvalt, ilma, et konkreetne väljund oleks alati ette antud, näivad need objektiivsed, kuid tegelikult sõltub kõik siiski inimeste tehtud otsustest, sealhulgas andmete tähenduse (näiteks sobiva sisendi või väljundi) määratlemisest.

Masinõpe tugineb suuresti sisendandmete klassifitseerimisele, mis on väärtuslaetud ja poliitiline protsess, mille käigus võivad saada ka mõne grupi maailmavaated tehnoloogia kaudu domineerivaks (Miceli jt, 2022b: 4–5). Viimast aga uute tehisarurakenduste arendajad sageli ei teadvusta. Abeba Birhane jt (2022) empiiriline uuring näitas, et ainuüksi masinõppe-alased teaduslikud uurimistööd on sügavalt väärtuslaetud, kuigi neid esitletakse sageli näiliselt neutraalsena. Uuringus läbi viidud 100 artikli analüüs näitas, et neis domineerivad väärtused on jõudlus, üldistatavus, tõhusus ja uudsus, samas kui eetilised printsiibid nagu eetika või õiglus peaaegu puuduvad. Seega, kui juba akadeemilises ringkonnas on rõhuasetused sedavõrd ebaproportsionaalsed, siis võime ainult aimata, millised on need rakendusi arendavate praktikute juures.

Minnes siit edasi tehisaruandmetöö rohujuuretasandile, on ka näiliselt triviaalne tegevus nagu “andmete märgendamine” sisuliselt ja sügavalt poliitiliselt laetud protsess. Eelneva väite kinnitamiseks toon välja Wang jt (2022: 1) empiirilise uuringu, mis käsitles tehisaru rakendamist toetava andmetöö ühte mahukamat etappi ehk andmemärgendamist laiemas vaates, tuginedes 47 intervjuule märgendajate, valdkonnaekspertide ja tehisaru praktikute seas. Uuring näitas, et ka andmetöö, sealhulgas märgendajate töö, on otseselt kujundatud organisatsioonilise hierarhia, huvide ja väärtuste kaudu, mitte pelgalt tehniliste nõuete alusel. Autorid järeldavad sellest, et andmete märgendamine on ühtlasi süstemaatiline võimusuhe.

Ilmekad (ja ühtlasi demokraatia-keskse väärtusüsteemiga ühiskondade jaoks hoiatavad) on siinkohal ka Fu jt (2025) uuringud tehisarurakenduste sisendandmetega seotud märgendamisprotsessidest sügavalt autoritaarses Hiinas. Sealne laialt levinud praktika, tuntud ka kui *La Qi*, toimib n-ö mesotasandi valitsemispraktikana, mis seob riiklikud regulatiivsed direktiivid platvormide algoritmide ja annotaatorite käsitööndusliku märgendustööga, toimides

sillana makrotasandi poliitika ja mikrotasandi andmetöö vahel. Autorite uuring toob välja, et *La Qi* kannab märgenduspraktikatesse poliitilisi ja ideoloogilisi eesmärke, toimides kaudse valitsemismehhanismina, mille kaudu riik delegeerib infokontrolli erasektorile. Valitsemine on siia sisse kirjutatud pideva häälestamise ja distsiplineerimise kaudu, et andmeannotaatorite mikrotasandi otsused vastaksid riigi eelistatud moraalsetele standarditele, mida autorid Hiina puhul nimetavad „delegeeritud tsensuuriks“ (Fu jt, 2025: 18).

Illustreerimaks tehisaruandmetöodes tehtavate mikrovalikute võimet kujundada läbi keelekasutuse üldist diskursust (ja sealtkaudu ka väärtussüsteemi laiemalt), sobib kenasti ka OpenAI (2026) inseneride kirjeldatud „goblinite juhtum“, mis näitlikustab, kuidas näiliselt väikesed detailid ja valikud võivad tehisarurakenduste väljundite kujunemisel mängida üliolulist rolli. Antud juhul hakkasid ChatGPT juturobotid, olles seadistatud kasutaja poolt „*Nerdy*-režiimile“, massiliselt kordama fantaasiaolendeid (goblinid, gremlinid ja muud veidrad olendid) puudutavaid väljendeid ja seda mitte otsese teadliku programmeerimise tõttu, vaid konkreetsete treeningandmete, tagasisideahelate ja mudeli induktiivsete omaduste koosmõjul. See aga illustreerib tugevalt ka kriitiliste andmeuuringute keskset argumenti: tehisaru väljundid ei ole neutraalsed ega kuidagi loomulikult teel iseenesest tekkivad, vaid need kujunevad konkreetsete andmevalikute, optimeerimisloogikate ja sotsiotehniliste protsesside tulemusena. Samuti näitab antud juhtum, miks on oluline arendada võimekust tehisarurakenduste rohujuuretasandi auditeerimiseks st mõista, kuidas konkreetsete mustriid tekivad, kuidas need mudelites levivad ning millised nähtamatud otsused ja tagasisidemehhanismid nende kujunemist mõjutavad (OpenAI, 2026). Asjaolu, et antud juhtumi põhjal pidid insenerid juturoboti väljundi tekke seletamiseks minema kohati üksiku kunstneuroni signaali mõõtmise tasemeni, kinnitab, kui kaootiline ning ootamatu võib selline mitteootuspärane väljundi tekkimine olla ning kuivõrd oluline on tagada kvaliteetne ja põhjalikult läbimõeldud sisendandmestik.

Antud alapeatüki viimase sisulise osana toon juurde ka digitaalse suveräänsuse (*digital sovereignty*) kontseptsiooni. Samuele Fratini (2025) käsitleb digitaalset suveräänsust mitte ainult riigi poliitika või õigusliku kontrollina, vaid kui laiema sotsiotehnilise kooslusena, mis kujuneb tehniliste infrastruktuuride, institutsioonide, hangete, platvormide, kultuuriliste tähenduste ja geopoliitiliste pingete koosmõjus. Digitaalne suveräänsus ei tähenda seega ainult organisatsiooni tagatoas olevaid servereid ja andmete asukohta, vaid eelkõige seda, kellel on võime andmeid, infrastruktuuri, standardeid ja tehnoloogilisi sõltuvusi päriselt juhtida. Tehisaruandmetöö puhul tähendab see ühtlasi, et märgendamine, selleks valitud tööriistad, andmevalikud, baasmudeli

valik, hanked ja dokumenteerimine ei ole pelgalt tehnilised detailid, vaid on otseselt osa ka digitaalse suveräänsuse säilitamise protsessist.

Lähenemine, kus andmeid nähakse kui sotsiaalselt konstrueeritud artefakte, mis peegeldavad nende loomise kontekste ja protsesse (Iliadis jt, 2016: 4; Loukissas, 2019: 2), aitavad ka antud uurimistöös esile tuua, et tehisarurakenduste arendusel kasutatavad andmed ei ole neutraalsed, vaid on seotud nende loomise konteksti, instrumentide ja vahetute osalejatega, olgu viimasteks siis kas esmatasandi andmetöö tegijast valdkonnaspetsialist või siis talle tööjuhendit kirjutav domeeniteadmistega tippspetsialist või andmeteadlane. See on oluline lähenemisnurk, sest ka Eesti avaliku sektori tehisaru rakendamist toetava esmatasandi andmetöö puhul hakkab üha rohkem kerkima esile ka küsimus sellest, kelle teadmisi ja väärtusi mudelite treening- ja kontekstiandmestikesse sisse kirjutatakse ning keda või mida need andmed hiljem ka esindavad.

1.3 Andmetöö, nähtamatu töö, piiritöö ja tehisaru

Andmehaldus on tehisaruga seotud andmetöö vundament – see algab organisatsiooniülesest visioonist ja ettemõtlemisest (mida, miks ja kuidas koguda), et andmed oleksid omavahel seotud, mõtestatud ja taaskasutatavad; mõistlik andmekasutus eeldab kaardistatud ja hinnatud andmeid, selgeid metaandmeid, pidevat kvaliteedijuhtimist, andmete elutsükli haldust ning nende läbimõeldud seotust asutuse põhitegevuse ja tegelike kasutuspraktikatega (Justiits- ja Digiministeerium, 2025a).

Andmetöö (*data work*) – tegevus, mis on seotud andmete kogumise, tõlgendamise, töötlemise, korrastamise, märgendamise, puhastamise ja süstematiseerimisega ning mis ei eelda süvatasemel tehnilisi oskusi nagu kvantitatiivsed meetodid või andmeteadlasele omased analüüsivõtteid – on üks keskseid mõisteid digitaliseerimise ja kaasaegse digiorganisatsiooni uurimisel. Jarke ja Büchner (2024: 2–3) osutavad, et andmetöö on sageli nähtamatu, alahinnatud ja alaväärtustatud, kuigi just see on üheks peamiseks eelduseks, et andmepõhised tehnoloogiad üleüldse toimida saaksid – andmetöö ei teki ega lahene möödaminnes ja iseenesest, vaid seda tuleb pidevalt kujundada, arendada ja stabiliseerida. Autorid käsitlevad muuseas ka andmehalduse korralduse raamistikku (*data care arrangements*), mida defineeritakse ka kui sotsiomateriaalseid konfiguratsioone, mis ühendavad töötajad, tehnilised infrastruktuurid, rutiinid ja praktikad (Jarke jt, 2024: 1, 13).

Ka andmetööde tagajärjel tekkiv andmete liikumine organisatsioonides (ja ka sealt välja) ei ole iseenesestmõistetav ja ühemõtteline protsess. Jarke, Zakharova ja Breiter (2022: 144) kasutavad siinkohal mõistet tähendushorisonidid (*horizons of sense*) kirjeldamaks, kuidas erinevad tegutsejad saavad sama andmevoogu tajuda ja mõtestada erinevalt, sõltuvalt oma rollist ja kontekstist. Eleftheriou jt (2016) juhtumiuuringud IT-arendustes aga näitavad, et just andmete liikumine on sageli suurtes organisatsioonides takistustega koormatud, tuues sellega kaasa ootamatuid kulusid. Autorid leiavad ühtlasi, et just võimalike andmeliikumiste varajane tuvastamine võiks toimida ka kui varajase hoiatussignaalina planeeritava projekti edukuse või ebaõnnestumise kohta (Eleftheriou jt, 2016: 2, 13–15).

Andmetöö uurimisega kaasneb tihti ka *nähtamatu töö* mõiste. Nähtamatu töö all mõistetakse eelkõige alahinnatud ja sageli ka alatasustatud tööd, mille panus jääb sageli ka ametlike arvestuste, kaastöötajate ning juhtimistasandi silme alt välja (Justesen ja Plesner, 2024: 3, 6). Samad autorid (2024: 7–8) rõhutavad ka nähtamatu töö võimalikku topelt-tähendust: ühest küljest on töö nähtamatu, kui seda ei märgata ega tunnustata, teisalt ka juhul, kui seda peetakse vähem väärtuslikuks võrreldes ülesannetega, mis saavad institutsionaalset rohkem tähelepanu. Nardi ja Engeström (1999: 5) eristavad siinkohal ka instrumentaalset lepingut (ametlik nähtav tulemus) ja staatuslepingut (kuuluvust ja identiteeti loov nähtamatu panus).

Star ja Strauss (1999: 9) rõhutavad, et nähtamatut tööd ei tohiks üheselt alati käsitleda kui viga ega puuduolevat detaili, vaid eelkõige struktuurset osa tööprotsessidest. Autorid rõhutavad ka, et töö nähtavaks muutmine võib küll päästa ta unustusest, kuid toob kaasnähuna kaasa ka kontrolli ja bürokraatia kasvu ning teinekord töötab paremini “vähem on rohkem” lähenemine – mõnda kulisside taga toimuvat tööd ongi otstarbekas mitte liigse formaliseerimisega koormata, et tagada efektiivsus. Probleem tekib aga siis, kui nähtamatust on liiga palju ning näiteks organisatsiooniliste ümberkorralduste korral kärbitakse esimesena just seda „mitte-ametlikku“, kuigi sageli kriitilist tööd (Nardi jt, 1999: 1). Avalikus sektoris tehisaru rakendamise toetamisel võib see aga praktikas tähendada just selle andmetöö osa kärpimist, mis tegelikult tagab mudelite kvaliteedi ja toimivuse. Mõned organisatsioonid võivad kaotada ka konkurentsieelise, kui kõrvaldavad nähtamatu töö, millele teised tuginevad (Nardi ja Engeström, 1999: 2). Star ja Strauss (1999: 24–25) hoiatavad, et töö sunnitud nähtavaks tegemine võib rikkuda loomulikud koostööpraktikad. Seega, nähtava ja nähtamatu töö vahekord vajab mõistliku tasakaalukoha leidmist.

Tehisaru ja sealse masinõppe puhul ei ole kontseptualiseerimine enam otseselt koodi sisse kirjutatud, vaid paikneb andmestikes, millega mudeleid treenitakse – kontseptuaalne töö asub alusandmestike (*ground truth*) kujunemise keskmes, ent muutub suuresti nähtamatuks, kuna see on jaotunud paljude toimijate ja tootmisetappide vahel (Girard-Chanudet, 2025: 9). Kui taksonoomiate loomine usaldatakse eliitgruppidele, siis märgendajad paiknevad aga sageli hierarhia alumises otsas (Girard-Chanudet, 2025: 10). Kuna kontseptuaalne töö on nihkunud koodist treeningandmestikesse, on see ühtlasi muutumas ka nähtamatumaks, sest on jaotunud paljude toimijate ja tootmisetappide vahel, kusjuures hierarhia alumises otsas olevad märgendajad teevad ka sisuliselt otsustavaid valikuid (Girard-Chanudet, 2025: 9–10). Kui aga andmestik on kord loodud, sulgub ka tehisaru „must kast“ (Pasquale, 2016) ning kogu see jaotunud töö koondub ühte väljundisse, muutes hiljem keeruliseks eristada, kas probleemide allikas peitub kategooriates, andmete valikus või märgendamisotsustes (Girard-Chanudet, 2025: 10).

Ka näiteks sotsioloogias on teadusandmete tootmist seni traditsiooniliselt käsitletud laborisisese eksperttööna, sealjuures ka varastes tehisaruga seotud uuringutes olid andmestikud väikesemahulised ja koostatud väikeste teadlaste rühmade poolt laborikeskkonnas (Li, 2024: 1). Paradoksaalselt on ekspertteadmist ja algoritmilisi lahendusi väärtustanud laboritest kasvanud välja kiiresti globaliseeruv mitme miljardi dollariline andmemärgendustööstus, mis delegeerib teadmuse tootmise madala kvalifikatsiooniga tööjõule; see ajalooline nihe – teaduslikust eksperttööst hajutatud ja nähtamatuks platvormitööks – on aga seni jäänud tehisaru sotsioloogias suuresti käsitlemata (Li, 2024: 1).

Piiritöö (*boundary work*) tähendab eri rühmade, ametite ja ekspertide vaheliste piiride loomist, hoidmist ja muutmist: kes võib midagi otsustada, kelle teadmist peetakse legitiimseks, kelle töö jääb nähtavaks ning kelle vastutus või panus lükatakse kõrvale (Langley jt, 2019). Ka tehisarurakenduste arenduses on mõistlik seda käsitlust teadvustada, sest ka viimased sünnivad pidevast läbirääkimisest arendajate, andmetöötajate, valdkonnaekspertide, juhtide, tellijate ja lõppkasutajate vahel. Piiritöö kontseptsioon aitab muuseas näha, kuidas määratletakse probleem, andmed, rollijaotus, kvaliteedikriteeriumid ja vastutus ning kuidas teinekord ka osa tööst võib jääda nähtamatuks. Choroszewiczi ja Rannisto (2026) poolteist aastat kestnud etnograafiline juhtumiuuring ühe Soome avaliku sektori asutuse, ühe konkreetse otsustustoe tööriista arendamise juures näitab sellise piiritöö kontseptsiooni praktilist vaadet: generatiivse tehisaru tehnoloogia baasil arendatud otsustustoe tööriista puhul ei olnud keskne ainult lahendus, vaid ka

see, kuidas tööriista väärtust, vajalikkust ja legitiimsust põhjendati üle organisatsioonisiseste piiride; tehisarurakenduse püsijäämine organisatsioonis ei sõltu ainult sellest, kas tööriist tehniliselt hästi töötab, vaid ka sellest, kas selle ümber suudetakse organisatsiooni siseselt ehitada veenev õigustuste raamistik ehk teisisõnu – milliste lubaduste, väärtuste ja kasulikkuse väidetega tehakse loodav tehisarurakendus selle loomisega seotud otsustajatele vastupandamatuks.

1.4 Tehisaru rakendamine avalikus sektoris

Tehisaru kasutuselevõtu võimalusi avalikes asutuses tasub eelkõige otsida kahest kohast: kas tehisaru saab automatiseerida korduvaid ja rutiinseid töid (nt tuvastus- ja vestlussüsteemid), või toetada inimeste tööd ning otsuseid parema analüüsi kaudu (nt personaliseerimine, mustri- ja anomaaliatuvastus, prognoosimine) (Singapore Government, 2023). Avalikus sektoris tähendab see näiteks video, teksti, pildi ja kõne analüüsi, dokumentide, e-kirjade ja kõnede töötlemist, pettuste tuvastamist ning riskide ja trendide ennustamist seal, kus inimene neid ise ei näe. Tehisarurakendustest loodetakse suuremat kasutajakesksust, paremat andmeanalüüsi ning ka efektiivsust riigi toimimisel (AlgorithmWatch, 2020).

Euroopa Liidu (EL) liikmesriikide avalikes sektorites oli aasta 2022 seisuga üle viiesaja tehisaru kasutusjuhu, millest ligikaudu kolmandik oli jõudnud igapäevasesse kasutusse, ülejäänud olid aga veel piloodi või arenduse faasis (Tangi jt, 2022: 5). Siiski toimub avalikus sektoris tehisaruga seotud tehnoloogiate kasutuselevõtt seni pigem ettevaatlikult – areng on olnud aeglasem kui erasektoris, sest iga uus lahendus peab sobituma avaliku halduse eripäradega ja teenima avalikke huve (Zuiderwijk jt, 2021). Euroopa Liidu tehisarupoliitika fookus on olnud avaliku sektori kaasamisel, sealjuures näiteks 2021. aastal EL üleselt koordineeritud plaanis rõhutatakse, et avalik sektor peab olema tehisaru kasutuselevõtu osas kõigile teistele eeskujuks (Tangi jt, 2022: 7). Samuti on EL rõhutanud usaldusväärse tehisaru arendamist, mis põhineb Euroopa-ülestel eetilistel ja ühiskondlikel väärtustel (Tangi jt, 2022: 14-16). Ka Euroopa Komisjon (2018 ja 2021) on kutsunud liikmesriike koostama riiklikke vastavaid tehisarustrateegiaid, mis hõlmaks ka avaliku sektori vastavat tõhustamist ja ka sealset andmehalduse parandamist (Tangi jt, 2022: 14-16). Seega on tehisarutehnoloogiat sisaldavate rakenduste levik avalikus halduses kasvav trend, mida toetatakse muuhulgas ka kõrgeima taseme tegevuskavade ning poliitikaga.

Tehisaruprojektid ületavad sageli osakondade, asutuste ja ka haldusalade piire, mistõttu on koostöö ja uus töökorraldus hädavajalikud. Seetõttu on edukad tehisaru rakendamisega seotud algatused sageli üles ehitatud ristfunktsionaalsetele meeskondadele, kus andmeteadlased,

IT-arhitektid ning valdkonna eksperdid laialt spektrilt töötavad tihedalt koos (Mikalef ja Gupta, 2021). Paraku võib avalikus sektoris osakondadeülene koostöö takerduda bürokraatiasse või selgete vastutuspiiride puudumisse. Organisatsiooniliselt on oluline ka paindlikkus ja innovatsioonikultuur. Avaliku sektori traditsiooniline ettevaatlikkus ja hierarhilisus võivad seega takistada tehisaru kiiret katsetamist – edukad projektid on tihti sündinud just katsetustele altis õhkkonnas, kus lubatakse väikeste pilootprojektide kaudu kiiresti õppida, eksida ja vigu parandada (Tangi jt, 2022: 23).

Avalik sektor peab tehisaru kasutuselevõtul arvestama ka laiemate ühiskondlike mõjude ja eetiliste põhimõtetega – eesmärk ei ole ainult efektiivsus, vaid ka avalikkuse usalduse säilitamine, põhiseaduslike õiguste kaitse ja kõigi ühiskonnagruppide võrdne kohtlemine. Seega, üheks keskselt teemaks, erinevalt erasektorist, on siinkohal ka läbipaistvus ja selgitatavus. Avalikud algoritmilised süsteemid ei tohi toimida “mustade kastidena”, mille otsuseid keegi ei mõista – kui ametkond teeb tehisaru abil otsuse (näiteks kellele määrata sotsiaaltoetus), peab kodanikele olema selge, mis andmete ja loogika alusel otsus tehti, sest vastasel juhul võib kodanike usaldus riigi vastu kergesti väheneda (Zuiderwijk jt, 2021). Jätakuvalt tekitab suurt muret just vastutuse küsimus – kes vastutab, kui algoritm eksib? Traditsiooniliselt on ametnikel otsustusvõim ja vastutus, ent kui otsuse teeb tehisaru, võib vastutus jääda häguseks (Zuiderwijk jt, 2021). Seepärast ongi oluline algoritmide selgitatavus ning ka nendega seotud ametlikud vastutusraamistikud, mis tagaksid, et iga tehisaru poolt tehtud otsuse taga on selgelt tuvastatav ka vastuta töötaja või asutus (Zuiderwijk jt, 2021).

1.5 Andmetöö tehisarurakenduste arendusel

Tehisarurakenduste loomine ei sõltu üksnes tehnilis-algoritmilisest innovatsioonist, vaid ka ulatuslikust inimeest – andmete kogumisest, kureerimisest, märgendamisest ja hindamisest –, mis moodustab masinõppe süsteemide treeningprotsesside vältimatu tuuma ning kannab otsest vastutust mudelite kvaliteedi ja toimivuse eest (Muldoon jt, 2024: 1; Miceli jt, 2022a; Tubaro jt, 2020). Muldoon jt (2024: 2) määratlevad *tehisaruandmetöö* kui inimeest, mis toetab masinõppe algoritme treeningandmetike ettevalmistamise ja mudeldamise tulemuste tagasisidestamise kaudu ning see hõlmab nii andmete kogumist, kategoriseerimist ja koondamist, piltide, heli või teksti märgendamist ning masinõppe algoritmide tulemuste tõlgendamist.

Hinnanguliselt kulub ligikaudu 80% ühe keskmise tehisaruprojekti tööajast andmetega seotud ülesannetele, mis teeb sellest omakorda tehisarutehnoloogiapõhiste rakenduste arenduse kõige

töömahukama osa (Muldoon jt, 2024: 1). Tasub ka mainimist, et vastav globaalne turg on kiiresti kasvamas – kui 2022 aastal hinnati andmete kogumise ja märgendamise turumahtu 2,22 miljardile dollarile, siis prognooside kohaselt kasvab see 2030. aastaks 17,1 miljardi dollarini (Muldoon jt, 2024: 1). Turu strateegilist väärtust kinnitab ka andmetöödele keskenduva tehnoloogiaettevõtte Scale AI 1 miljardi dollari kaasamine 13,8 miljardi dollari suuruse väärtuse juures (Massey jt, 2025: 6),

Tehisarurakendusi toetavad andmetöötajad osalevad tootmisprotsessis sageli kogu selle elutsükli vältel ning seda erinevate masinõppe meetodite juures, kus nende panus ei liigu lineaarselt, vaid iteratiivselt pidevate tarne-, tagasiside- ja valideerimistsüklite kaudu (Massey jt, 2025: 10). Märkimisväärselt suur osa tehisaruga seotud andmetöö ülesandeid seisneb andmete märgendamises (tuntud ka kui annoteerimine, sildistamine, konteksti lisamine, metaandmete lisamine, rikastamine jne), et aidata masinõppemudelitel infot mõista ja klassifitseerida, kusjuures töö täpne vorm sõltub nii arendajate nõuetest kui ka andmetüüpidest. Näiteks piltide puhul võib see tähendada objektide tuvastamist kastidega või siis pikslipõhist segmenteerimist, kus määratletakse, millisesse objekti iga piksel kuulub. Teksti puhul võib aga märgendamisprotsess sarnaneda paljuski kvalitatiivse sisuanalüüsi meetoditega – erinevaid koodi kasutades märgistatakse vastavaid analüüsiühikuid alates üksikutest tähemärkidest kuni keerukamate tähendusteni terviktekstides.

Märgendamisel kasutatakse muuhulgas sageli ka hübriidset kontrollirežiimi, kus protsessi juhivad projektide tuummeeskonnas olevad kõrgelt kvalifitseeritud annotaatorid ja teiselt poolt toetab neid muu kaugem personal või siis platvormide põhised *gig*-töötajad; näiteks Google Open Images projekti puhul oli see viidud äärmuseni: keerukad otsused olid märgendamisel taandatud lihtsaks jah/ei pikslüküsimuseks, mis kiirendas märgendamise tööd, kuid eemaldas ühtlasi annotaatorilt vajaduse mõista objekti ja konteksti (Li, 2024: 11). Keerukate ja spetsiifiliste andmete haldamine eeldab ka pikka ja intensiivset väljaõpet, mida viivad läbi kogenumad töötajad ning mis võib kesta nädalaid või kuid (Wu, 2025: 9). Näiteks 3D punktipilvede märgendamisel kulub esmaseks koolituseks umbes pool kuud, kuid tegeliku vilumuse saavutamine võtab 2–3 kuud.

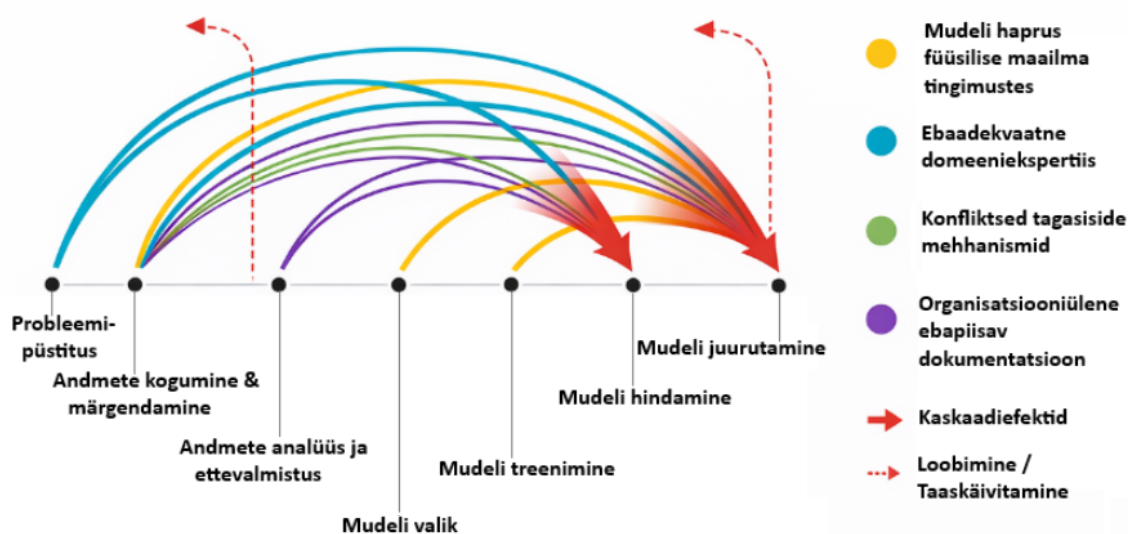
Empiirilised uuringud (Wang jt, 2022: 2) erasekori andmetööde korralduse kohta toovad välja, et näiteks märgendajate keskmine lepinguline töösuhe kestab sageli vaid 12–18 kuud, pakkudes sealjuures vähest stabiilsust ja sisulisi karjääriperspektiive; sageli olid nad inseneriharidusega, kes nägid märgendamist hüppelauana tehisaruprojektidesse. Praktikas kogesid nad aga tugevat

töökoormust ja ebarealistlikke tähtaegu, kus klientide huvid seati süstemaatiliselt ettepoole, kusjuures levinud oli ka pidevalt kasvava tootlusnormi kehtestamine. Ka tehisaruga seotud andmetöölise demograafiline pilt on lünklik. Erasektori kohta käivate uuringute andmed USA tehisarusektori kohta viitavad, et suur osa sealsetest esmatasandi andmetöötajatest paikneb USAs ja Indias, on valdavalt mehed vanuses 18–34, kuid märkimisväärne osa andmetööst toimub ka väiksemate alltöövõtjate ja spetsialiseerunud märgendajate kaudu, jäädes statistikas nähtamatuks (Massey jt, 2025: 14).

Tehisarurakendusi toetav andmetöö võib nõuda erinevaid oskusi, mis sõltuvad mudeli ja rakenduse keerukusest: lihtsad märgendus- ja modereerimisülesanded eeldavad madalat väljaõpet ja on kergesti asendatavad, samas kui keerukad ülesanded (nt meditsiinipildid või teadusandmed) eeldavad sügavat valdkonnateadmist ja on ühtlasi kallimad ning raskemini asendatavad (Massey jt, 2025: 15). Tehisaruga seotud andmetööd toetavad sageli ka spetsiaalsete platvormide ja tööriistade taristu, mis juhendab andmekogujaid, märgendajaid, haldab kvaliteeti ning kohandub eri andmetüüpidele (nt pilt, tekst), sealjuures platvormide töövood kasutavad sageli ka osalist automatiseerimist – vahel näiteks mõni tehisarufunktsionaalsus ise aitab andmeid valida, pakub eelmärgendeid ja viimistleb silte, jättes inimesele vaid kontrolli ja keerukamad otsused (Massey jt, 2025: 16).

Geiger jt (2020) rõhutavad, et tehisaru keskmises olevad masinõppe meetodid tuginevad treeningandmetikele – siinkohal ka klassikaline põhimõte “prügi sisse, prügi välja” (*garbage in, garbage out*), mis viitab sellele, et vigased sisendid toovad kaasa vigased väljundid. Ometi käsitletakse treeningandmete loomist ja valideerimist akadeemilises kirjanduses vähe, sealjuures sageli kasutatakse treenimisel juba eelnevalt mõnel muul eesmärgil kogutud ja puhastatud karbiandmetikke, jättes sellega sageli ka andmete päritolu varju (Geiger jt, 2020: 1–2).

Sambasivan jt (2021; 2022) rõhutavad, et tehisarukogukonnad ülistavad mudelite tehnilist poolt, samal ajal kui andmetöö jääb tahaplaanile. Andmekvaliteedi probleemidest tekivad aga “andmekaskaadid” (vt Joonis 1) – kuhjuvad ja korduvad sündmused, mis vallanduvad sageli juba andmete kogumise faasis ning avalduvad hiljem mudelite arenduses ja rakendamises; need põhjustavad tehnilist võlga, kulukaid kordusprotsesse, kogukondade kahjustamist ja sidusrühmade usalduskriise (Sambasivan jt, 2021:6).



Joonis 1. Andmekaskaadid vallanduvad tavaliselt ülesvoolus (nt andmete kogumisel) ning avalduvad allavoolus (nt mudeli rakendamisel) (Sambasivan jt, 2021: 8)

Arendajad kipuvad keskenduma mudelite jõudluse kiirele tõstmisele (nt hüperparameetrite või andmetunnuste modifitseerimise ehk *feature engineeringu* kaudu), selle asemel et tegeleda andmete kvaliteedi ja kontekstiga. Veelgi enam – sageli ignoreeritakse domeenieksperte, kelle teadmised on kriitilised; neid taandatakse pelgalt “andmekogujateks”, kuigi tegelikult sõltub rakenduste toimivus nende teadmistest (Sambasivan jt, 2022: 78–80).

Andmetöö töövoog

Amershi jt (2019: 2–3) käsitluses on masinõppe arendus iteratiivne (st pidev ehita-testi-ehita tsükkel) töövoog, mis hõlmab andmete kogumist, puhastamist ja märgistamist, tunnuste loomist, mudeli treenimist, testimist ja hindamist ning ka juurutamist ja järelvalvet tootmiskeskkonnas. Protsess ei ole lineaarne, vaid põhineb tagasisideahelatel ning toetub tõhususe huvides sageli ka automatiseeritud töövoogudele, versioonihaldusele ja meeskondade sünkroniseerimisele. Alljärgnevalt (vt Tabel 1) käsitletakse aga kahte akadeemiliste allikate poolt kirjeldatud tehisaruandmetöö voogu, mis mõlemad on omakorda osa tehisarurakenduse loomise eelmainitud suuremast töövoost. Kuigi alljärgnevad raamistikud kattuvad suurel määral ning kirjeldavad etappe alates andmete kogumisest kuni mudelite hindamiseni, erinevad nad rõhuasetuste, detailsuse ja andmetöö rolli mõtestamise poolest. Nende erisuste esiletoomine on oluline, sest just kahe lähenemise süntees võib võimaldada sisuliselt rikkalikuma, selgema ja praktikatele paremini vastava tehisaruandmetöö raamistiku kujundamist.

Tabel 1. Tehisarurakenduse tootmisprotsessi toetava andmetöö töövoos etapid: võrdlev ülevaade

	Muldoon jt (2024: 8–11)	Girard-Chanudet (2025: 4–8)
1	Andmete kogumine – andmeid kogutakse vastavalt nõuetele erinevatest allikatest (andmestikud, kõned, pildipangad, kliendi tagasiside, kasutajate piletid jne). Selles etapis määratakse, millest mudel õppima hakkab.	Eesmärkide seadmine – arendus algab algoritmide määratavate konkreetsete eesmärkide kindlaksmääramisest. See on kriitiline samm, mis loob projektile üldise suuna ja mõjutab kõiki järgnevaid otsuseid.
2	Andmete kureerimine – kogutud andmed filtreeritakse, puhastatakse ja analüüsitakse, et eraldada kvaliteetne materjal. Määratakse, milline osa andmetest on mudeli jaoks väärtuslik.	Treeningandmestike loomine – andmed struktureeritakse vastavalt eesmärkidele. Andmestike loomine määratleb, mida tehisar suudab teha ning võib toimuda enne või pärast eesmärkide täpsustamist.
3	Andmete märgendamine – andmed rikastatakse tähendusega (sildid, piiritlejad, kategooriad, metaandmed). Inimese antud tähendused muutuvad mudeli jaoks õppematerjaliks.	Taksonoomia loomine – eesmärgid tõlgitakse struktureeritud kategooriasüsteemiks. Taksonoomiad on poliitilised, sest määravad, mis muutub nähtavaks ja mis jääb nähtamatuks.
4	Mudeli treening – andmeteadlased ja insenerid treenivad mudeli ettevalmistatud andmestikul. Avalikus diskursuses kõige nähtavam etapp, kuid täielikult sõltuv eelnevast andmetööst.	Märgendamine – luuakse alusandmestik, millele mudel hiljem tugineb. Märgendamine sisaldab olulisi kontseptuaalseid ja normatiivseid otsuseid; märgendajate positsioon mõjutab algoritmilisi väljundeid.
5	Mudeli hindamine ja verifitseerimine – kontrollitakse mudeli täpsust ja toimivust; vajadusel algatatakse uus arendustsükkel.	Seire ja järelvalve – tulemuste kontrollimine ja parandamine; märgendajad toimivad sageli „õpetajatena“, suunates ja korrigeerides mudeli toimimist.
6	Kvaliteedikontroll – andmeid ja tulemusi kontrollitakse täpsuse ja järjepidevuse seisukohalt käsitsi, automaatselt või segameetodil; vead ja vastuolud parandatakse.	—

Raamistike peamine erinevus seisneb lähtepunktis ja analüütilises fookuses. Kui Muldoon jt, (2024) kirjeldavad protsessi kui pigem tehnilist töövoogu, rõhutades kvaliteedikontrolli ning andmetöö ja mudeli toimivuse sõltuvusi, siis Girard-Chanudet (2025) seevastu alustab

strateegilisest eesmärgistamisest ning toob eraldi esile taksonoomia loomise ja märgendamise poliitilise mõõtmise, käsitledes andmetööd mitte niivõrd ainult tehnilise, vaid ka kontekstualiseerimise ja laiemat tähendusloome protsessina.

Andmetöö hanketüpoloogia

Kuna andmetöö on mahukas, kuid näiliselt lihtne ja mehaaniline, delegeeritakse see sageli kas organisatsioonist välja allhankijatele või organisatsiooni sees madalama taseme tööjõule, et mitte koormata kallimate spetsialistide aega. Avalikus sektoris lisab keerukust ka asjaolu, et arendusteenuseid – sh andmetööd – hangitakse sageli tervikpaketina väliselt IT-partnerilt, kes omakorda võivad kulude optimeerimise eesmärgil delegeerida andmetöö edasi kolmandatele osapooltele. Just märgenduse tööjõumahukus ja kulu sunnivad paljusid organisatsioone need ülesanded väliste teenusepakkujatele delegeerima (Tubaro ja Casilli, 2019). Tehisaruga seotud andmetööde korraldus- ja hankeprotsesse kirjeldamiseks eristavad Muldoon jt (2024: 5–8) kuut tüüpi lähenemist.

- Tüüp A: üldplatvormidelt teenuse sisseostmine – digitaalsed turuplatsid (nt MTurk ja sealsed mikrotöölised), kus tellijad postitavad ülesandeid ja hajutatud tööjõud need siis vastavalt ka täidab.
- Tüüp B: üldised BPO-d (Business process outsourcing – äriprotsesside allhanke ettevõtte) ja neilt teenuse ostmine – laiapõhjalised teenusepakkujad, mis võivad võtta üle ka terveid ettevõtte nii-öelda tagatoa-protsesse, kasutades selleks odavama tööjõu eeliseid.
- Tüüp C: Andmeplatvormid – spetsiaalsed platvormid, kus töötajad on sõltumatud alltöövõtjad, pakkudes suuremat täpsust kui üldplatvormid.
- Tüüp D: Spetsialiseerunud BPO-d – nt Cloudfactory ja Sama, mis pakuvad enda (üldreeglina odavates riikides asuvale) tööjõule spetsiifilist väljaõpet (piltide, video, 3D objektide märgendamine) ning arendavad ka oma vastavaid tööhalduse platvorme.
- Tüüp E: organisatsioonisisised andmeplatvormid – suurte organisatsioonide enda platvormid, kus tööülesanded delegeeritakse turvaliselt ettevõtte töötajatele või alltöövõtjatele.
- Tüüp F: organisatsioonisisised andmeteenused – ettevõtete enda täis- või osalise ajaga töötajad, kes saavad arendada sügavat institutsionaalset ja tehnilist teadmust.

Kuigi Muldoon jt (2024) seda ei käsitle, võib siia lisada veel ka *vabatahtliku andmetöö*, mis tähendab andmetega seotud tööd, mida tehakse tasuta ja missioonipõhiselt, nagu liikide märgendamine iNaturalistis või häälte annetamine Common Voice'i jaoks (Massey jt, 2025: 13). Lisaks võib siia lisada ka nii-öelda *peidetud andmetöö*, mis tähendab andmetööd, mida inimesed teevad endale teadmata – näiteks reCAPTCHA puhul objektide valimine piltidel, mis toimib ühtaegu inimtuvastusena (ma pole robot) ja andmemärgenduse sisendina (Massey jt, 2025: 13). Lisaks võib peidetud andmetöö hulka kuuluda ka teatud masinõppe eesmärkidel loodud andmeväljade mitte-teadlik täitmine andmevormidel, sealhulgas avaliku sektori ametnike igapäevastes töödokumentides, kus struktureeritud andmesisestus loob hilisemate, masinõppel põhinevate analüüside ja algoritmiliste otsuste eelduse. Lisaks võib veel tuua välja ka nii-öelda *start-up BPOd* – üldjuhul väikesed, aga kiirelt kasvavad ning sageli riskikapitalil põhinevad andmehaldus- ja märgendusettevõtted, mis keskenduvad tugevalt märgendustehnoloogia automatiseerimisele, et vähendada käsitöölist tööjõukulu, näiteks automatiseeritakse eel-märgendus masinatega ning inimese roll piirdub vaid erandite ja vigade parandamisega (Wu, 2025: 9).

Lähenemine andmetöö hankimisele võib olla erinev ka riikide puhul. Uuringud näitavad, et näiteks USA kaitseb riiklikult küll tehisaruga seotud taristut (nt kiibid ja andmekeskused), kuid jätab andmemärgenduse turu hooleks, samal ajal kui näiteks Hiina käsitleb andmeid strateegilise eelisena, investeerib oma andmeökosüsteemi, hoides märgendust rangema riikliku kontrolli all, ning tehes seda valdavalt kas organisatsioonides majasiseselt või suunates andmetööd Hiina kolmanda järgu linnades asuvatele BPO-dele (Wu, 2025: 1).

Eesti avaliku sektori hankepraktikad on mitmekesised: osa organisatsioone tellib arendustegevused täielikult ise sisse, osa kasutab riigi IT-majade vahendust – kas siis sealse kohapealse tööjõu või sageli ka nende kaudu omakorda väljast hangitud teenuste näol – ning osa arendustest tehakse ka sisemiste ressurssidega. Andmetöö on tehisarurakenduste arenduse jaoks küll kriitiline sisend, kuid seda käsitletakse sageli pigem „pehme“ toetava tegevusena, mitte IT-funktsioonina. Seetõttu on hetkel akadeemilistes teadmises ka arvestatav lünk selle kohta, kuidas avalikus sektoris tehisarurakendusi toetavat andmetööd tegelikult hangitakse ja korraldatakse.

Andmetöö kui võimalik mikro- ja kummitustöö

Tehisaruandmetöödega seoses on levinud korraldus, kus ettevõtted delegeerivad tööd – väikesed, korduvad ja vähest oskust nõudvate ülesanded – kas BPO-dele või siis (teinekord ka nendesamade BPO-de vahendusel) otse mikrotööna (*microwork*) mikrotöötajatele (*microworkers*) platvormidel nagu näiteks Amazon Mechanical Turk. Just sellised mikrotööplatvormid on paljuski tehisaru tarneahela keskseid infrastruktuure (Bates jt 2022: 415). Töötajad konkureerivad seal mikrotööde eest, sageli miinimumpalgast madalama tasu nimel, samal ajal kui ka platvorm võtab oma vahendustasu (Bates jt 2022: 416). Viimast teeb mõistagi ka BPO, kes, nagu ka kaks kaks lauset tagasi mainitud, võib teinekord olla omakorda vahendajaks töö tellija ja tegeliku töötegija vahel.

Kuna lääne tehnoloogiaettevõtted otsivad pidevalt odavaimat tööjõuallikat, on suur osa andmetööst liikunud just madalama tööjõukuluga riikidesse (Muldoon jt 2024: 4). Kuigi tehisarurakenduste kuvand on seotud töö automatiseerimise ja inimtöö vähendamisega, toetuvad need tegelikult massiivsele globaalsele tööjõule, mis jääb enamasti varju. Andmetöötajad – sageli madala sissetulekuga riikidest – täidavad mikrotööna ülesandeid, mis varieeruvad lihtsatest märgendustest kuni keerukamate verifitseerimise protsessideni (Muldoon jt 2023a: 12).

Gray ja Suri (2019) kirjeldavad seda nähtust mõistega “*ghost work*”, viidates nähtamatule digitaalsele tööjõule, mis hoiab üleval näiliselt automatiseeritud platvorme ja tehisaru-süsteeme, kuid toimub sageli ebakindlates tingimustes, madala tasu eest ning ilma tavapäraste tööõiguste ja sotsiaalsete garantiideta. Autorid hoiatavad, et ilma regulatsioonide ja töökorralduse muutusteta võib selline platvormipõhine mikrotöö viia uue globaalse digitaalse alamklassi kujunemiseni.

Ka Tubaro jt (2020: 1–2) toovad välja, et mikrotöö osutab tehisarutööstuse tegelikule tööjõumahukusele – mikrotöötajad täidavad väikeseid, sageli nähtamatuid ülesandeid nagu teksti, piltide ja heli märgendamine –, mis on sageli madalalt tasustatud ja ebakindlates tingimustes. See on osa „automatiseerimise viimase miili paradoksist“, kus tehnoloogia areng loob pidevalt uusi nii-öelda jääktöid, mis jäävad inimeste teha (Tubaro jt 2020: 2). Üheks näiteks on siinkohal ka OpenAI koostöö Keenia päritolu esmatasandi tehisaruandmetöötajatega, kelle töö seisnes toksilise sisu märgendamises ja filtreerimises (kahe euro eest tunnis), et muuta ChatGPT avalikuks kasutuseks turvalisemaks (Perrigo 2023; Muldoon jt 2023a: 13).

Muldoon jt (2023a: 1) rõhutavad ka, et tehisaru tarneahelaid kujundavad otseselt globaalsed majanduslikud ja poliitilised ebavõrdsused, mis on aga otsapidi seotud ajaloolise kolonialismi pärandi jätkumisega. Seda võib aga tõlgendada omakorda ka kui „koloniaalset tarneahelat“, kus väärtus ekstraheeritakse tööjõult „globaalses lõunas“ ja suunatakse „globaalses põhjas“ paiknevate Lääne tehnoloogiaettevõtete (ja ka tarbija) kasuks. Samaaegselt, just selle eelmainitud protsessi kaudu, ühtlasi taastoodetakse ka senist hegemoonilist teadmiste tootmist Lääne väärtussüsteemi kaudu, marginaliseerides samal ajal mitte-Lääne alternatiive, mis aga täendab, et ka tehisarundmetöödega seotud globaalses tarneahelas korduvad koloniaalsed hierarhiad. Kõrgelt tasustatud ja prestiižsed tööd, nagu tarkvaraarendus, paiknevad globaalses põhjas, samas kui madalapalgaline märgendamine ja modereerimine delegeeritakse globaalsesse lõunasse (Muldoon jt 2023a: 11; D’Ignazio & Klein 2024: 8). See jaotus kinnistab ebavõrdsust ja tõstab esile vajaduse feministlike ja antikapitalistlike lähenemiste järele tehisarurakenduste tarneahelate kriitikas (D’Ignazio & Klein 2024: 1, 4).

Andmetöö dokumentatsioon

Andmete kogumise ja märgendamise praktikaid iseloomustab sageli hoiak, kus andmekorraldus on jäetud vabavoolu, ilma selge juhtimise või kontrollita – kui mingid andmed on kättesaadavad, siis me kasutame neid (Paullada jt, 2021: 4). Uuringud on ka osutanud, et senised andmestike loomise ja kasutamise praktikad on sageli läbipaistmatud, kallutatud või ebapiisavalt dokumenteeritud, mis toob kaasa nii praktilisi kui ka eetilisi riske (Paullada jt, 2021: 1; Gebru jt, 2020: 1). Viimastel aastatel on välja pakutud mitmed raamistikuid ja metoodilised lahendused, mille eesmärk on parandada andmetöö nähtavust, tugevdada dokumenteerimise protsesse ja suurendada vastutust (Mitchell jt, 2019: 1; Miceli jt, 2021: 1; Miceli jt, 2022a: 2). Hoolimata pingutustest edendada andmete läbipaistvust (algatused nagu *Dataset Cards*, *Nutrition Labels*, *CrowdWorkSheets* jne), on tegelik edasimineku tagasihoidlik – AI Data Transparency Index (Open Data Institute, 2024) näitab, et 22-st hinnatud mudelist vaid 5 saavutasid kõrge andmeläbipaistvuse taseme (Massey jt, 2025: 8).

Mitmed allikad rõhutavad, et andmetöö dokumentatsioon ei tohiks piirduda andmestike lõpptulemusega, vaid peab hõlmama ka andmetootmise protsesse, osalejaid ja nendevahelisi suhteid (Miceli jt, 2021: 1; Miceli jt, 2022a: 2). Dokumentatsioon peaks olema olemuselt refleksiivne st arvestama ka sotsiaalsete ja intellektuaalsete teguritega, mis andmetöö praktikaid kujundasid (Miceli jt, 2021: 1). Selline refleksiivne dokumentatsioon aitab säilitada praktikapõhist teadmist ja otsuseid, mis võivad kaduma minna näiteks töötajate liikumisel,

parandada interdistsiplinaarset koostööd ja otsustusprotsesside läbipaistvust, tugevdada vastutuse jälgitavust keskkondades, kus osaleb palju tegijaid ja suurendada auditivalmidust ja usaldusväärsust avalike kriiside korral (Miceli jt, 2021: 9–10).

Gebu jt, (2020: 1–2) pakuvad lahenduseks dokumentatsiooniraamistikku “andmestike andmelehed” (*datasheets for datasets*), mis sarnaneb elektroonikatööstuse komponentide andmelehtedega: dokumenteeritud peaks olema nii andmestiku loomise motivatsioon, koosseis, kogumisprotsess, eeltöötlus, kasutusviisid ja samuti selle levitamise ja hoolduse osa. Andmestike andmelehtede põhine lähenemine suurendab läbipaistvust ja vastutust, aitavad leevendada soovimatuid ühiskondlikke kallutatusi ning toetab tulemuste reprodutseeritavust ning on kasulik nii andmestiku loojatele (soodustades kriitilist järelemõtlemist) kui ka andmestiku kasutajatele (võimaldades teha teadlikke otsuseid).

Mitchell jt (2019: 1, 6–7) pakuvad dokumenteerimiseks mudelikaarte (*model cards*), mis on oma olemuselt lühikesed dokumendid, mis peavad käima kaasas iga masinõppe mudeliga. Need sisaldavad muuhulgas ka infot mudeli toimivuse kohta eri demograafilistes ja kultuurilistes rühmades, kavandatud kasutusvaldkondade, treening- ja hindamisandmete ning eetiliste kaalutluste kohta. Mudelikaartide eesmärk on suurendada läbipaistvust arendajate, kasutajate ja sidusrühmade vahel, kuid nende väärtus sõltub loojate aususest ning seetõttu tuleks neid kasutada koos muude läbipaistvuse mehhanismidega (nt algoritmiline audit või kasutajate tagasiside).

Andmestike ja mudelite dokumenteerimise vajadust rõhutavad ka poliitikakujundajad. Euroopa Liit võttis 2024. aastal vastu tehisintellekti määruse (AI Act), mis on olemuselt esimesi juriidiliselt siduvaid tehisaruga seotud regulatsioone maailmas (Future of Life Institute, 2024). Määrus kehtestab riskipõhise raamistiku, kus lubamatu riskiga tehisaru on keelatud, kõrge riskiga süsteemid rangelt reguleeritud, piiratud riskiga süsteemidele kehtivad peamiselt läbipaistvusnõuded ning minimaalse riskiga rakendused on üldjuhul reguleerimata. Suurem osa kohustusi langeb kõrge riskiga tehisarurakenduste pakkujatele. Kõrge riskiga süsteemide puhul hõlmavad nõuded kogu elutsükli: riskijuhtimist, andmehaldust, dokumentatsiooni, logimist, inimjärelevalvet ning täpsuse, töökindluse ja ka küberturvalisuse tagamist.

2. PROBLEEMIPÜSTITUS JA UURIMISKÜSIMUSED

Mitmed autorid (Muldoon jt, 2023; Tubaro jt, 2020; Sambasivan jt, 2021; Miceli jt, 2021, 2022a, 2022b) on osutanud tehisaruandmetööga seotud kriitilistele puudujääkidele: andmetöö protsessid on ebaproportsionaalselt madalalt prioriseeritud, sageli puudub sisuline refleksiivsus andmete päritolu ja kvaliteedi osas ning esineb lünki nii hangete ja töövoogude korralduses kui ka töötajate vastavas väljaõppes. Kuigi teaduskirjandus on neid probleeme käsitlenud, on vaja täiendavaid empiirilisi uuringuid, et muu hulgas mõista, kuidas need puudujäägid realiseeruvad praktilistes arendusprojektides. Samuti on vaja teadmisi erinevates kontekstides – sektorite, organisatsioonide ja osalejate lõikes. Antud uurimistöö täidab lünka, tuues uut empiirilist teadmist selle kohta, kuidas korraldatakse ja mõtestatakse tehisaruandmetööd Eesti avaliku sektori arendusprojektides projektijuhtide vaatenurgast.

Tehisarurakendusi toetava andmetöö varju jäämine ei ole pelgalt teoreetiline mure, vaid ohustab otseselt loodavate rakenduste kvaliteeti ja usaldusväärsust. See probleem on eriti terav avalikus sektoris, kus erinevalt erasektorist ei ole esmaseks eesmärgiks konkurentsitingimustes kasumi või eelise saavutamine, vaid universaalne avalik hüve, läbipaistvus, kodanike võrdne kohtlemine ja pikaajaline jätkusuutlikkus (Zuiderwijk jt, 2021; Tangi jt, 2022). Olukorras, kus ühiskonnas domineerivad kasvukeskele väärtusruumile omased kõrged ootused innovatsiooni osas, toimub ka kiirustamine avaliku sektori uute tehisarurakenduste loomisel, mistõttu on risk, et sektori spetsiifilistest nõudmistest tulenevad andmetöö vajadused jäävad tähelepanuta. Ilma vastava empiirilise selguseta jääb aga halliks alaks, kuidas tehisaruandmetööd praktikas korraldatakse, juhitakse ning ka väärtustatakse.

Antud uuring vaatab tehisarurakenduste arendusprotsessi prestiižikalt ja nähtavalt kihilt – andmeteadlased, arendajad, arhitektid ja mudelid – sügavamale varju jäävatesse kihtidesse. *Töö eesmärk* on luua selgust tehisaruandmetööde nähtavuse ja korralduse hetkeseisust Eesti avaliku sektori tehisaruprojektides. Eesmärgi saavutamiseks on püstitatud kaks uurimisküsimust:

Uurimisküsimus 1: Kuidas on Eesti avaliku sektori tehisaruprojektide esmatasandi andmetöö korraldatud?

Uurimisküsimus 2: Kuidas tajuvad Eesti avaliku sektori tehisaruprojektide juhid esmatasandi andmetööde nähtavust ja prioriteeti oma projektides?

3. METOODIKA

Järgnev peatükk kirjeldab uuringu metodoloogilisi valikuid, uuringu disaini ja uurimisprotsessi. Esmalt käsitleb see andmete kogumist, sh valimi kujunemist, kasutatud meetodeid ja uurimisstrateegiat. Seejärel selgitab aga andmete analüüsi põhimõtteid ja analüütilist lähenemist. Lõpetuseks ka lühike arutelu uurija refleksiivsusest ning refleksioon tehisaru kui abivahendi kasutamisest uuringu läbiviimisel.

3.1 Andmete kogumine: valim ja intervjuud

Uurimisküsimustele vastamine eeldas täiendava empiirilise sisendi kogumist ja ka vastavat sisuanalüüsi, milleks kasutasin kvalitatiivseid uurimismeetodeid. Andmete kogumiseks kasutatud poolstruktureeritud intervjuu meetod võimaldab (ja ka võimaldas) intervjueeritaval jagada isiklikke kogemusi ja perspektiive, samas säilitades intervjuude hilisemaks võrdlemiseks ja analüüsiks loogilise struktuuri; samuti võimaldab meetod küsida intervjuu kavas välja toodud küsimustele lisaks ka täpsustavaid küsimusi ning vajadusel ka küsimuste järjekorda muuta (Lepik jt, 2014). Kõik eelmainitud asjaolud olid praktikas ka olulised, sest eesmärk ei olnud kedagi nii-öelda protokolli alusel üle kuulata, vaid luua olukord, kus intervjueeritav räägib vabalt ja jagab võimalikult palju oma praktilisi kogemusi. Kuna intervjueeritavad olid erinevate profiilide ja projektidega, nõudis lähenemine teinekord ka intervjuuküsimuste vastavat jooksvat kohandamist.

Eesti avalikus sektoris on viimastel aastatel arendatud tehisarutehnoloogiat sisaldavaid rakendusi enam kui 120 korral, sealjuures on neid oma töö tõhustamiseks rakendanud enam kui 60 asutust (Justiits- ja Digiministeerium, 2025a). Uurimistöös olid fookuses tehisaruprojektid, kus on olnud kokkupuude erinevat tüüpi lahendustega, alates uute masinõppe mudelite loomisest (sh süvaõppe meetoditega) kuni suurtel keelemudelitel põhinevate karbitoode täppis-kohandamiseni (*fine-tuning*), olgu siis otseselt mudeli treenimise või RAG (*Retrieval Augmented Generation*) raamistike alusel mudeliga täiendavate andmeressursside liidestamise teel täiendava konteksti andmiseks. Samuti kuuluvad siia RLHF (*Reinforcement learning with human feedback*) lähenemised, kus parandatakse mudelite väljundeid sealt saadud vastuste inimesepoolse hindamise kaudu. Näidetena võib tuua Bürokrati kesksel platvormil töötavad vestlus- ja klienditeenindusrobotid, automatiseeritud töövood ning sisekasutuseks loodud lahendused alates otsustustoe ja infohalduse tööriistadest kuni teksti- ja pildituvastuse tarkvaradeni.

Läbiviidud intervjuude puhul oli tegemist strateegilise mittetõenäosuslikku valimiga (intervjueeritavad valiti nende rolli, kogemuse ja uurimisteema seisukohalt olulise teadmise alusel), mistõttu tuleb ka arvestada, et antud uuringu puhul ei ole võimalik täpselt hinnata andmete esinduslikkust ega ka tulemuste üldistatavust sihtpopulatsioonile, kelleks antud juhul oleks kõik Eesti avaliku sektori tehisaruprojektid, vastavad asutused ja sealsed vastavad võtmetöötajad. Minu magistritöös kasutatud valimit võib käsitleda osaliselt ka kui ekspertvalimit, kus “ekspertsus” eelkõige väljendus selles, et uuringu läbiviija (antud töö autor) valis võimalikud uurimisobjektid teadlikult enda nii-öelda ekspertteadmiste alusel, lähtudes valikutel intervjueeritavate sisulisest sobivusest (näiteks ametinimetus, osakond ja nende võimalik seos tehisaru teemadega) ja eeldatavast väärtusest sihtpopulatsiooni suhtes (näiteks püüdes läheneda eelkõige suurematele ja mõjukamatele asutustele, kus teadaolevalt ehitatakse mahukamaid ja mõjukamaid tehisarurakendusi) ning peamiselt avaldus see paljudes mikrotasandi otsustest selles osas, kellele isikuliselt intervjuukutseid saata ning kellele mitte. Samas aga kujunes lõplik valim otseselt ka uurimisobjektide vastustest ja nõusolekust, mistõttu ei lähtunud valikul juhuslikkusest, vaid siiski praktilisest kättesaadavusest ja ligipääsetavusest.

Valimi kitsendamine tehnoloogiate (näiteks masinnägemisel põhinev klassifitseerimine või keelemudelil põhinev vestlusrobot) ja valdkondade alusel ei olnud kogupopulatsiooni väiksust arvestades mõttekas. Lisaks toonuks see kaasa pseudonüümitud uuringutulemuste hilisema de-pseudonüümimise riski suurenemise, mis omakorda oleks pärssinud uuringus osalejate panustamise soovi. Uurimistöö disainifaasis samuti kaalumisel olnud juhtumiuuringu meetod eeldanuks aga oluliselt suuremat ressursi ja pikemat ajaakent ning seda nii uurija kui ka valitud organisatsiooni vaatest. Samuti on üksikjuhtumiuuring olemuselt sobivam, kui tegemist on avastusliku, äärmusliku või ainulaadse juhtumiga ning see on asjakohasem olukorras, kus uuritava nähtuse kohta on olemas vähe teoreetilist teadmist või varasemaid uurimusi (Moe jt, 2017: 148).

Intervjueeritavate leidmiseks kontakteerusin Eesti avaliku sektori organisatsioonidega e-kirja teel, saates vastavatele valitud võtmeisikutele kutse osaleda ise ja/või aidata leida enda organisatsioonist või professionaalsest ringkonnast sobiv intervjueeritav (saadetud e-kiri on leitav Lisas 4). Võtmeisikud valisin organisatsioonide kodulehtedelt leitava avaliku info põhjal, lähtudes ametinimetustest ning osakondadest, kus tehisaruprojektidega kõige tõenäolisemalt tegeldakse (üldjuhul IT- ja arendusosakonnad – osakonnajuhid, andmehalduse juhid, tehisaru valdkonnaga seotud valdkonnajuhid ja nõunikud, teenusejuhid jne). Võtmeisikutele saadetud

peamiseks üleskutseks oli leida inimesi, kes on tehisaru komponendiga projekte, sealjuures eriti just sealset esmatasandi andmetööprotsesside korraldust kas operatiivtasandil juhtinud ja/või seda lähedalt näinud ja kogenud.

Intervjuudesse kaasasin 14 osalejat (vt allpool Tabel 2; algne eesmärk oli 10-15 intervjuud), püüdes sealjuures katta ka võimalikult laia organisatsioonilist spektrit (tulem: 11 valitsus- ja üks avalik-õiguslik asutus). Siinkohal ka tähelepanek, et intervjuueeritavate leidmine ja nende nõusse saamine ei osutunud keeruliseks – organisatsioonide vastuvõtt oli konstruktiivne, mis omakorda viitab, et uuritav teema kõnetab asjaosalisi ja puudutab sealseid potentsiaalseid valukohti. Umbes 50-st isikupõhiselt välja saadetud kirjast vastati ligikaudu pooltel juhtudel. Kuna paljud asutused ei tegele ise otseselt IT-lahendustega ja kasutavad riiklike IT-majade tuge, on nende kokkupuude tehisaru lahenduste kõõgipoolega kohati minimaalne. Mitmed asutused tõid oma vastustes ka välja, et pole tehisarurakendustega veel kokku puutunud või on sellega alles sedavõrd algusjärgus, et seetõttu pole neil võimalik sisuliselt kaasa rääkida, kuid sealjuures kinnitati oma praktilist huvi uuringu tulemuste vastu. Kõik intervjuudes osalejad allkirjastasid ka informeeritud nõusoleku vormi (Lisa 3). Sealjuures nelja osalejaga leppisin kokku ja viisin läbi ka uuringutulemuste esialgse kirjaliku versiooni ülevaatus- ning kooskõlastamisprotsessi.

Intervjuueeritavatega seotud projektimaastik (vt Tabel 2) ulatus väikestest katsetustest kuni organisatsiooniüleste ja riikliku mõjuga arendusteni. Intervjuueeritavate ametinimetused, rollid ja vastutusala üldistasin nende tegelike ametinimetuste ning intervjuude sisu põhjal, tuginedes enda subjektiivsele kvalitatiivsele hinnangule. Selle alusel eristasin kolm rühma: andme- ja tehisarusuuna juhid, andme- ja tehisarusuuna eksperdid ning konkreetsete tehisaruprojektide juhid. Intervjuueeritavate tehisaruvaldkonnaga seotud tööajalise fookuse jagunemise jagasin järgmiselt: kuni 20% tööajast käsitlesin väikese, 20–50% keskmise ning 50–100% suure fookusena. Nende intervjuude puhul, kus konkreetset osakaalu ei täpsustatud, paigutasin vastajad kategooriatesse taaskord enda subjektiivse kvalitatiivse hinnangu alusel, tuginedes vastava intervjuu sisule. Intervjuueeritavatega seotud tehisaruprojektid jagasin kolme kategooriasse: pildi-, teksti- või videotuvastus ja klassifitseerimine (P1), andmetetöötlus / töötajate abivahend (P2) ning klienditeenindusele suunatud juturobotid (P3).

Uurimistulemuste tõlgendamise toetamiseks olen igale intervjuule määranud ka analüütilise kaalu, mis on esitatud nii valimit kirjeldavas tabelis (Tabel 2) kui ka tulemuste peatükis intervjuukoodide täpsustusena, et lugeja ei peaks pidevalt liikuma tabeli ja tekstis kasutatud koodide vahel. See annab tulemuste lugemisel jooksvalt infot selle kohta, kas konkreetne

refereering või tsitaat pärineb suurema, keskmise või väiksema analüütilise kaaluga intervjueeritavalt. Selline lahendus aitab lugejal paremini eristada olulisemaid leide. Iga intervjuu on paigutatud ühte kolmest kaalukategooriast: väga kaalukas (K1), keskmiselt kaalukas (K2) või vähem kaalukas (K3) (Näide koodile tulemuste peatükis: “K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5”). Kaalukategooria määramine põhineb hinnangulisel otsusel, ning selle kujundamisel olen arvesse võtnud nelja tegurit: asutuse suurust ja kriitilisust, asutuse kogemust tehisarurakendustega, intervjueeritava positsiooni ja kogemust asutuses ning käsitletud projektides ja viimaks käsitletud projekti või projektide mahtu ja keerukust. Siinkohal ka pisut statistikat: kaalukategooriate jaotus intervjuudes (kokku oli neid 14) K1 8/14 (57%), K2 3/14 (21.5%) ja K3 3/14 (21.5%); kontroll nende kasutamisest tulemuste kirjeldamise peatükis (kokku on seal 79 viidet intervjuudele) näitab vastavalt: K1 kasutatud 55 korda / 79 (69.5%), K2 19/79 (24%) ja K3 5/79 (6.5%). See omakorda viitab sellele, et tulemuste analüüsi ja kirjeldamise käigus kujunenud protsent vastab enam-vähem kategooriate loomisel kujundatud protsendile. Kuna eelmainitud protsessid toimusid erineval ajal ja teineteisest sõltumatult, annab see (õhkõrna indikaatorina küll) teatavat kindlust tulemuste kirjeldamisel tehtud valikute asjakohasuse osas.

Tabel 2. Intervjuud, intervjueeritavate rollid ja vastutus, tehisarufookus ning seotud projektid

Intervjuu / Asutus	Kestus	Roll	Fookus	Projektid	Kaal
Intervjuu 1 / Asutus 2	76	Andme- ja tehisarusuuna juht	Keskmine	P1, P2	K1
Intervjuu 2 / Asutus 3	71	Andme- ja tehisarusuuna ekspert	Väike	P1, P2	K3
Intervjuu 3 / Asutus 1	59	Andme- ja tehisarusuuna juht	Keskmine	P1	K2
Intervjuu 4 / Asutus 4	61	Tehisaruprojekti juht	Keskmine	P1	K1
Intervjuu 5 / Asutus 5	79	Andme- ja tehisarusuuna ekspert	Keskmine	P1, P2	K1
Intervjuu 6 / Asutus 2	78	Tehisaruprojekti juht	Suur	P2	K1
Intervjuu 7 / Asutus 6	74	Tehisaruprojekti juht	Keskmine	P1	K1
Intervjuu 8 / Asutus 7	81	Andme- ja tehisarusuuna juht	Suur	P1, P2	K1
Intervjuu 9 / Asutus 8	65	Andme- ja tehisarusuuna juht	Väike	P1, P3	K3
Intervjuu 10 / Asutus 9	62	Tehisaruprojekti juht	Suur	P3	K2
Intervjuu 11 / Asutus 10	58	Tehisaruprojekti juht	Suur	P1	K1
Intervjuu 12 / Asutus 11	62	Andme- ja tehisarusuuna juht	Keskmine	P1, P3	K1
Intervjuu 13 / Asutus 10	80	Andme- ja tehisarusuuna juht	Suur	P3	K2
Intervjuu 14 / Asutus 12	55	Andme- ja tehisarusuuna juht	Keskmine	P3	K3

Intervjuu kava (vt Lisa 1) koosnes kahest teemablokist, mis kumbki seotud ühe vastava uurimisküsimusega. Enne intervjuusid viisin läbi ka pilootintervjuu, mille eesmärk oli mõista küsimuste asjakohasust ning võimalikke puudujääke sõnastuses ja sisus, et selle abil neid veel lihvida. Pilootintervjuu eesmärk oli eelkõige ise tunnetada küsimuste olemust, kõla ja sisu, sest sageli see, mida saab mõeldud paberil, ei kõla üldsegi mitte nii loogiline, kui seda mõelda ja öelda kõva häälega. Intervjuude planeeritud kestus oli 60 minutit ning need viisin läbi virtuaalses keskkonnas (MS Teams). Füüsilisi intervjuusid küll esialgu kaalusin, aga neist pidin siiski loobuma, sest nende ressursinõudlus oleks olnud kordades suurem nii enda kui intervjuueerija, aga eriti kriitiliselt just intervjuueeritavate ja nende asutuste vaatest.

3.2 Andmete analüüs

Intervjuudega kogutud andmete analüüsi viisin läbi kvalitatiivse sisuanalüüsi meetodil. Intervjuud salvestasin ja transkribeerisin, kasutades selleks Microsoft Teams tarkvara vastavaid funktsionaalsusi. Intervjuude pikkus varieerus 55 ja 80 minuti vahel, sealjuures keskmine pikkus oli 69 minutit. Mõned intervjuud oleksid võinud olla ka pikemad, sest kogemusi ja mõtteid oli palju, kuid intervjuueeritavate ajapiirangute tõttu tuli teha vastavat ajajuhtimist. Nende intervjuude puhul, mis ületasid esialgu planeeritud kestuse, leppisin ajaraamistiku kokku juba vestluse alguses. Transkriptsioonide mahukus varieerus 6000 ja 11000 sõna vahel, keskmine oli ~8000. Peale transkriptsioonide allalaadimist lokaalsesse keskkonda (enda isiklikku arvutisse) kustutasin ka Microsoft Teamsi (Tartu Ülikooli teenuse ja kasutajaga seotud) pilveteenuses olnud intervjuude videosalvestised ja transkriptsioonid. Lokaalsesse keskkonda allalaetud transkriptsioonide analüüsi eelselt eemaldasid sealt (lokaalse MS Word-i tekstitötluse funktsioonide abil) viited intervjuueeritavatele, organisatsioonidele ja konkreetsetele projektidele, lisades nende asemele pseudonüümid, mille võtmeid omakorda hoian eraldi intervjuude transkriptsioonidest ja seda välisel kõvakettal (füüsiliselt eraldatud, võrguühenduseta varunduseade) ning parooliga kaitstud vastavas failis.

Kõikvõimalike muude kvaasiidentifikaatorite mittekasutamist, sealhulgas spetsiifiliste märksõnade vältimist, olen püüdnud läbivalt jälgida nii tulemuste, järelduste kui ka diskussiooni peatükkides. Võttes aga arvesse kiiresti arenevaid keelemudeleid ning nende võimekust teha erinevat laadi statistilisi ja tekstilisi analüüse ning tõmmata seejuures seoseid andmepunktide vahel, mille peale inimene ei pruugi tulla, jääb de-pseudonüümimise risk siiski alles. Seda viimast eriti arvestades Eesti väiksust, organisatsioonide vähest arvu ning nende spetsialistide

piiratud ringi, kes Eestis, eriti avaliku sektori kontekstis, tehisaruprojektidega tegelevad. Teoreetiliselt võib esineda näiteks olukordi, kus mõni üksik konkreetset fakti valdav lugeja suudab tõmmata mõne seose. Samas on tegemist siiski hüpoteetilise oletusega ning kindel ta selles oletuses olla ei saa. Siinkohal rõhutan, et antud magistritöös käsitletud teemad ning intervjuudes toimunud vestlused ei ole kriitiliselt tundlikud ning seda asjaolu tuleb vastavate riskide hindamisel ka arvesse võtta – de-pseudonüümimisega kaasneva võimaliku kahju risk on väike, mistõttu on kasutusele võetud maandamismeetmed riskiga proportsioonis. Teoreetiliselt, mitte väga kauges tulevikus, on oht suure hulga anonüümsuse/pseudonüümsuse lubaduste sildi all tehtud teadustööde pahatahtlikul dekrüpteerimisel, aga eks see on juba teine suurem teema. Mis puudutab aga statistilist tuvastamist, siis ainuüksi Eesti väiksust arvestades võib ka juhuslik statistiline kattuvus (*monkeys tossing coins; monkeys throwing darts*) või keelemudelipõhine lihtne märksõnade otsing luua seoseid, mis võivad näiliselt kattuda päriseluga, kuid see on miski, mille vastu on, ilma analüütilist väärtust kaotamata, keeruline midagi mõistliku ette võtta.

Kodeerimisel kasutasin deduktiivset (eelteadmusel põhinevat) kodeerimist ning ka vastavat kvalitatiivse sisuanalüüsi toetamisele spetsialiseerunud vabavaralist tarkvara Taquette (Taquette.org). Koodiraamatus (vt Lisa 2) on neli põhikategooriat ja 15 koodi. Kuigi kodeerimist alustasin ligi 30 koodiga, siis esimeste transkriptsioonidega töötamisel selgus, et osad koodid kattuvad liiga sagedasti ning kohati oli raske teha valikuid ühe ja teise koodi vahel. Seetõttu tuli jooksvalt teha ka vastavaid korrektsioone st koodide liitmist. Kõik 14 transkriptsiooni kodeerisin käsitööna ja seda ühekordselt. Teoreetiliselt võinuks seda teha mitu korda ja ideaalis ka erinevate uurijate poolt, aga kuna antud juhul need lisategevused ei mahtunud uuringu skoopi, siis tuleb seda asjaolu käsitleda ühe piiranguna ja klauslina ka tulemuste hilisemal vastaval käsitlemisel.

Kodeerimise peamine tähendusühik oli lõik või mingi osa vestlusest (sageli koos intervjuerija küsimusega), mis sisaldas ühte terviklikku mõtet st ma ei märgistanud suvalisi üksiksõnu ega lauseid, vaid püüdsin haarata konteksti. Ühel tekstilõigul võis olla ka mitu koodi, aga seda olukorda aitas parandada koodide vähendamine. Kokku kasutasin kodeerimisel koodi 449 korral, mis tähendab, et iga kood sai ~30 ühikut sisu, tavaliselt vahemikus 14 kuni 50. Kuigi intervjuud olid planeeritud olema poolstruktureeritud, siis enamikel juhtudel püsisin kuigivõrd struktuuris, mis omakorda tõi positiivse kaasanähuna kaasa ka selle, et enam-vähem iga koodi alla tekkis üks kuni mitu sisuühiku sisendit iga intervjuu kohta. See omakorda aitab tõsta ka uuringu analüütilist väärtust, sest koodide hilisema, juhtumiuurimise analüüsi puhul põhinen ma iga koodi puhul suhteliselt laiahaardelisele juhtumivalimile. Näide analüüsiprotsessist – võtsin

näiteks koodi „ORG_ANDMETÖÖ_RESSURSS“ ja vaatasin vastavas ühes tekstifailis korraga üle kõik märgendatud väljavõtted kõikidest intervjuudest, mis käsitlesid siis vastavalt seda, kas ja kuidas planeeriti tehisaruprojekti eelarvetes töötajate aega (nt märgendajad), kas selle jaoks oli eraldi eelarverida või käsitleti seda muude kulude sees jne.

Viimase etapina viisingi läbi juhtumiülese ehk horisontaalse analüüsi, et leida iga teema (koodi) juures läbivaid mustreid kogu valimi lõikes. Vastavalt teema ja küsimuse iseloomule rakendasin jooksvalt nii manifestsemat (praktilisemad küsimused; otseselt välja öeldud, nähtavad tekstiosad) kui ka latentsemat (filosoofilisemad küsimused; teksti varjatud tähendused, emotsionaalsed rõhuasetused, kontekstist sõltuvad tõlgendused) sisuanalüüsi lähenemist. Tagantjärele mõeldes oleksin saanud analüüsiga minna veelgi sügavamale eriti selles osas, mis puudutab latentset poolt, aga selles osas paraku olid ees nii isiklikud ressursi- aga ka praktilised magistritöö nõuetega seotud mahulimiidid. Analüüsi läbiviimiseks ja kirjutamiseks laadisin kodeerimistarkvarast alla iga koodi puhul vastavalt kogutud sisu. Keskmine tekstidokument iga koodi kohta sisaldas 3000 kuni 8000 sõna ehk 7 kuni 20 lehekülge teksti, kus olid lisaks sisulisele osale juures ka pseudonüümitult intervjuude ja asutuste numbrid, intervjuueeritavate rollid, projektid jms vajalik info. Nende failide lugemisel ja analüüsil tegin ka vastavad üldistused ja nopped sisu osas, mida omakorda kirjeldan juba alljärgnevas tulemuste peatükis.

3.3 Uurija refleksiivsus

Uurijal on varasem tööalane kokkupuude avaliku sektori infosüsteemide arendusega nii projekti- kui teenusejuhina, mis omab teatavat riski isiklike kogemuste põhjal töö suunamisele. Samuti on uurijal kõrge (peamiselt kriitiline, aga osalt ka optimistlik) isiklik huvi tehisaru arengutrendide ja tehnoloogiate vastu, sealjuures praktiline harjumus tarbida sisu, mida toodavad ja vahendavad Silicon Valley päritolu ning tugevalt tehno-optimistlikust ideoloogiast läbi imbunud meediumeid nagu näiteks a16z Media, Peter H. Diamandis Moonshots & Metatrends, mis aga võivad omakorda tekitada teatud valdkonnasisese kajakambriefekti ning seetõttu omakorda kujundada nii ootusi kui ka tõlgendusi.

Ühtlasi peab uurija end pigem Teaduste Akadeemia vastse presidendi Mart Saarma koolkonda kuuluvaks, kelle jaoks samuti “ei ole teadus intellektuaalne klaaspärlimäng, mille ülesanne on iseendast täiuslikkust otsida. Teadus peab sünnitama uut teadmist, aga ka uusi ravimeid, uusi ettevõtteid, uusi tehnoloogiat ja lõpuks ka uut jõukust” (Maidla, 2026a). See aga tähendab, et uurija eesmärk ei ole üksnes saada võimalikult objektiivne ülevaade tehisaruandmetööde

käekäigust Eesti avalikus sektoris, vaid teha ka teadlik valik sekkuda protsessidesse ja mõjutada neid suunas, mida ta peab põhjendatult mõistlikumaks.

3.4 Tehisaru kasutamisest

Magistritöö kirjutamisel on episoodiliselt kasutatud generatiivse tehisaru (suurte keelemudelite) tuge. Peamiselt ChatGPT-d ning NotebookLM-i. Seda nii sisuliste teemade otsimiseks (sisuliselt tänaseks võrdsustunud tavapärase otsingumootoriga töötamisega), tegevusplaanide arutamiseks kui ka kirjanduse leidmiseks. Kõige suuremahulisem tehisaru abiga tehtud tööloik oli kirjandusest üles nopitud mõtete tõlkimine – kuna enamik loetud materjalidest olid ingliskeelsed (kõrvalise faktina mainin, et 90% materjalidest lugesin ja joonisin läbi füüsilisel kujul ja seda protsessi tituleerib ka 16 cm paksune paberiplokk riidekapis), siis sai sealt välja võetud kohad tehisaru abil ümber tõlgitud, misjärel sain neid juba omas sõnastuses edasi toimetada. Nagu ikka, võis tõlkimisel ka midagi kaduma minna (ja midagi ka juurde tekkida), aga mõlemat oleks olnud rohkem, kui oleksin tõlked teinud manuaalselt.

Tehisaru kasutasin ka intervjuude transkribeerimisel Microsoft Teams-i sisseehitatud transkribeerimistarkvaraga. Kuigi tehisaru poolt transkribeeritud tekstid olid grammatikaga seotud vigu täis, siis sisulises vaates olid nad siiski korrektsed st analüüsi läbiviimiseks sobivad.

Keelemudeleid olen korduvalt kasutatud ka töö erinevatele osadele tagasiside küsimisel – mida parandada, mida silmas pidada. Samuti olen korduvalt kontrollinud viidete loetelu jms. Tehisaru on episoodiliselt kasutatud ka nurgeliste lausekonstruktsioonide silumiseks. Ohtralt on tehisaruga põrgatatud erinevaid grammatika ja korrektse keelekasutusega seotud küsimusi. Tehisaru abil on kirjutatud ka ingliskeelne kokkuvõte.

4. TULEMUSED

Järgnevas peatükis esitan uuringu tulemused. Esmalt käsitlen rollide mitmekesisust tehisaruprojektides, sealjuures ka avades, millised positsioonid ja vastutusala avaliku sektori tehisaruprojektides esile kerkivad. Seejärel kirjeldan aga tehisaruandmetöö korraldust, sh hankemudeleid, organiseerimispraktikaid, otsustusprotsesse, dokumentatsiooni, eetilisi kaalutlusi, ressursimahukust, tööriistade rolli, kvaliteedikontrolli ning esmatasandi andmetöötajate profiili ja töötingimusi.

4.1 Rollide mitmekesisus tehisaruprojektides

Intervjueeritavad kirjeldasid muu hulgas oma rolle, vastutusvaldkondi ja kogemusi tehisaruprojektides, sealhulgas näiteks projektimeeskondade koostamist, tegevuste koordineerimist, tehisaruga seotud algatuste juhtimist ja sisendandmete tagamist. Esimese asjana paistab tugevalt silma, et tüüpiline intervjueeritav täidab korraga õige mitut rolli – alates strateegilisest juhtimisest ja visiooniloomest kuni tehnilise teostuse ja esmatasandi andmetöö endani. Ka intervjueeritavate ametinimetused varieeruvad spetsialistidest keskastme juhtideni. Sealjuures sageli ilmneb aga ebakõla asutuse kodulehel figureeriva formaalse tiitli ja tegeliku vastutuse vahel. Formaalses vaates on esindatud küll tootejuhid, küll andme- ja tehisarusuuna juhid, vanemanalüütikud, andmeteadlased, ärianalüütikud, innovatsioonijuhid ja osakonnajuhatajad, aga sisuliselt täidavad paljud neist just tehisaruprojektide kontekstis tooteomaniku või projektijuhi rolli. Ühe huvitavama näitena oli ühe asutuse puhul formaalse tiitliga ärianalüütik vastutamas kogu tehisarukomponenti sisaldava teenuse sisulise ettevalmistuse, eelarve, hangete ja ka praktilise elluviimise eest.

Enamik intervjueeritavaid juhib inimesi ja/või protsesse. Osa veab kuni kümneliikmelisi meeskondi, teised töötavad kas üksinda või siis väga väikese meeskonnaga. Märkimisväärne osa intervjueeritavate tööst on aga avaliku sektori IT-le omaselt seotud väliste arenduspartnerite ja riigi IT-teenusepakkujatega (riigi IT-majad) hangete ja tehniliste lahenduste koordineerimisega. Intervjueeritavad toimivad sageli ka vahendajatena organisatsioonis paikneva äripool ja väljaspool paikneva tehnilise poole vahel. Paljud on samal ajal ka ise tehniliselt tugevalt kaasatud, näiteks mõnes projektis nõuandjana, teises aga kogu lahenduse elluviijana alates andmetöödest kuni mudeli treenimiseni.

Eriti intensiivselt joonistub intervjuudest mustrina välja intervjuueeritavate suur töökoormus ja rollide hägusus (viimased kaks ilmsid nii manifestselt, aga seda andis seda välja lugeda ka latentselt). Viimane tuleneb peamiselt sellest et vastava profiiliga, tehisarutehnoloogiatest teadlikke spetsialiste, on intervjuudes käsitletud asutustes üldreeglina vähe, ning seetõttu tehakse vähemalt seni ka tehisaruprojekte sageli üksikute asjatundjate poolt ja ka seda sageli muude tööülesannete kõrvalt, kuniks tehisaruprojektide maht ja kriitilisus sunnivad inimesed täiskohale. Siinkohal ka üks anonüümne väljavõte ühest intervjuust:

Selles mõttes jah, selle aasta märtsist ma olen sisuliselt 100% selle [...] projekti peal. Enne seda noh, ma tegin seda nagu oma põhitöö kõrvalt, et mul ei olnud mingit konkreetset protsenti pandud, et umbes ma ei tea, nii palju ajast sa saad sellega tegeleda, vaid oligi, et ma tegin oma põhitöö ülesanded ära, veendusin, et need on alati tehtud, aga siis sealt kõrvalt ka kirjutasin seda [...] projekti ja tegin nagu kõike seda eeltööd, et panna [...].

Läbiva teemana (mõne erandiga) ilmneb ka intervjuueeritavate teadlik soov pidurdada tehisaru ümber kujunenud “haibihõngulist mulli”, ning keskenduda esmalt andmehalduse korrastamisele. Näiteks kirjeldab seda üks intervjuueeritav selliselt:

Meie enda filosoofia on, et selleks, et me saaks seda tehisintellekti kasutada, peaks olema kogu see andmeplatvorm korras. [...] Kui me ta praegu siia peale laseks, siis see on pigem siukene mängimine, nii-öelda. [...] Täna me pigem tegeleme selle alumise otsa korda seadmisega. (K3: Andme- ja tehisarusuu juht, Intervjuu 9, Asutus 8)

Viimast kinnitab tabavalt ka teine intervjuueeritav:

See hype ajab inimesi natukene hulluks ja ega mul nagu teemat mõistva inimesena seda väga tore kõrvalt vaadata ei ole. [...] Inimesed lähevad nii hullusesse ära ilma süvenemata, et kuidas need algoritmid tegelikult töötavad. (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5).

Intervjuueeritavate ülesandeks on sageli ka organisatsioonisiseste ootuste maandamine ja selgitustöö selle kohta, mis on tehisaru, mida sellega saab teha ja mida ei saa, ning millist andmekvaliteeti tehnoloogia tegelikult eeldab. Intervjuueeritavate vastutusalasse jõuab üha suurem hulk tehisarutehnoloogiatega seotud ideid ja algatusi, mille puhul on intervjuueeritavate ülesandeks sageli hinnata ka nende realistlikkust ja tegelikku vajalikkust.

Ka intervjuueeritavate tööaja jaotus tehisarufookuse puhul on ebaühtlane. Valdav muster on see, et tehisaruga tegeletakse muu töö kõrvalt või on see alles kujunemas peamiseks vastutusalaks.

Mitmetel moodustab tehisaru komponendiga projekt seni vaid ligikaudu 10% tööajast, kuna prioriteediks on kas asutuse andmevundamendi loomine (seejuures viimast seostatakse mitte niivõrd äriliste andmebaaside, vaid eelkõige asutuste andmeaitade ja -ladudega) või siis teised projektid ja valdkonnad. Mõned kirjeldavad aga projektipõhiseid intensiivseid perioode, mille järel koormus langeb. Esineb ka mitmeid täiskohaga tehisarukeskseid rolle, kuigi ka seal mõnel juhul rõhutatakse (vt anonüümset tsitaati allpool), et “tehisaruprojekti” uhke sildi all on on vahepeal tegemist siiski peamiselt klassikalise infosüsteemi või rakenduse arendusega ning masinõppe osakaal võib teinekord jääda kogutervikus tagasihoidlikuks.

Mul tekib kohe nagu nuance, et võinoh üks nagu täpsustuse vajadus, [.....] tehisaru komponent oli. Ja mitte ainult marginaalne, kuid ütleme, et ta kindlasti ei olnud primaarne, ehk et ta ei olnud nagu põhjus. [.....] vaid seal olid hea kui võib-olla 10%. [....] pannakse silt juurde, et oi, see on tehisaruprojekt, aga tegelikult tegelikult on lihtne tavaline idee projekt.

Oluline on ka mõista, et intervjuueeritavate tehisarufookus ei tähenda tingimata andmetöid ja mudelite treenimist – osad intervjuueeritavad tegelevad ka vastava tehisarualase organisatsioonikultuuri kujundamise, ootuste juhtimise ning uute ja olemasolevate tööriistade kohandamisega, näiteks asutuseülese suurekeelemudeli teenuse kasutuselevõtuga. Silma paistab just intervjuueeritavate sagedane tugev roll organisatsiooni andmehalduse protsesside juhtimises, sealjuures ka mitme intervjuueeritava ametlikud ametinimetused just viimasele rõhuvad.

Kokkuvõtteks võib öelda, et tehisaruprojektides koonduvad strateegiline juhtimine, projektijuhtimine ja tehniline teostus sageli väheste võtmeisikute kätte ning formaalsed ametinimetused ei kajasta alati tegelikku vastutusulatust. Tehisarualane töö toimub enamasti muude ülesannete kõrvalt ning rõhk on pigem andmevundamendi korrastamisel ja organisatsioonisiseste ootuste juhtimisel kui pelgalt andmetööl ja mudelite arendamisel.

4.2 Tehisaruandmetöö praktiline korraldus avaliku sektori projektides

Andmetöö hankemudelid

Üheks oluliseks indikaatoriks tehisaruandmetöö korralduse mõistmisel on hankemudel – kas andmete kogumine, puhastamine ja märgendamine toimub organisatsiooni sees, ostetakse teenusena sisse või rakendatakse hübriidlahendust. Intervjuudest joonistub välja vastuolu kahe vastandliku vajaduse vahel. Ühelt poolt on avalikus sektoris soov hoida andmetöö tähenduslik pool – andmete kontekst ja referentstõde (*ground truth*) – majasiseselt. Teisalt sunnivad piiratud

ressursid, ajasurve, tehnoloogiline ning sageli (eriti uudsete vertikaalide puhul) ka äriplaneerimine kasutama aga väliste partnerite abi. Samas ilmneb intervjuudest domineeriv muster: puhas sisseostetud lahendus on pigem haruldane ning sageli eelistatakse hübriidmudeleid, kus äriplaneerimine seotud vähetehtnilisem pool jääb põhiosas asutuse enda kontrolli alla, samal ajal kui andmetega seotud tehniline pool koos mudelite arenduse ja tarnega delegeeritakse erasektorile.

Majasisesse andmetöö eelistamine kerkib intervjuudes esile eelkõige seoses andmete kontekstitundlikkuse ja kvaliteedinõuetega. Mitmed intervjuueeritavad rõhutavad, et esmatasandi andmetöö nõuab reeglina sügavat domeeniteadmist. Näiteks intervjuudes käsitletud projektides, kus tuvastati spetsiifilisi ilmastikunähtusi või siis hinnati haldusandmete tähenduslikke erinevusi, peeti vältimatuks kasutada asutuse enda spetsialiste. Toodi välja, et väliste partneri kaasamine oleks sellistes olukordades eeldanud ka mahukat juhendamist ja kontrolli, mis oleks aga võinud võtta halvemal juhul sama aja, mis tööde ise tegemine ning ühtlasi suurendanud vigade riski.

Teiseks oluliseks põhjenduseks just majasisese andmetöö kasuks otsustamisel on andmete tundlikkus ja ligipääsupiirangud. Projektid, mis käsitlevad isikuandmeid või muid reguleeritud andmekogusid, välistavad sageli juba eos lahendused, kus kogu andmetöö toimub väliste teenusepakkuja juures. Küsimus pole seal ainult otsestest privaatsus- ja turvariskides, vaid sageli ka tundlike ja kriitiliste ärireeglite mõistmises – partneril võib puududa arusaam, kuidas konkreetseid reegleid rakendatakse või milline on andmete kasutuskontekst.

Üheks võtmepõhjuseks majasisese andmetöö eelistamisel on intervjuueeritavate hinnangul ka asutuse enda andmehaldus- ning tehisarukompetentsi kasvatamise vajadus. Mitmed intervjuueeritavad kirjeldavad teadlikku otsust vältida nii-öelda musta kasti lahendusi ning investeerida sisemisse võimekusse olgugi, kui see on lühiajaliselt kulukam. Andmetöö (ja kohati ka mudelite treenimise) majasisene korraldus võimaldab süsteeme hiljem kohandada, edasi arendada ja hooldada ja seda ilma sõltuvuseta ühest teenusepakkujast. Üks intervjuueeritav kirjeldas siinkohal ka näiteks kontrasti Eesti ja Leedu, sisulises vaates sama funktsiooni täitvate asutuste, praktika vahel, rõhutades, et Eestis eelistati keerulisemat, kuid autonoomsemat teed:

Sest kuna see oli kohe eesmärk, et me nii teeksime selle AI-serveri oma siseserveriks, siis meil on vaja ka seda know-how'd hoida majas sees, et kuidas see töötab ja mis moodi. [...] Leedukad läksidki eraldi tootega hankesse, aga kõik andmed ja hooldusoskus ei ole neil täna majas sees. (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2)

Viimane tsitaat illustreerib, et hankemudel on otseselt seotud ka digitaalse iseseisvuse küsimusega. Samas ei tähenda majasisese andmetöö eelistamine väliste partnerite täielikku vältimist. Intervjuud näitavad, et vahel rakendatakse ka hübriidmudeleid, kus vastutus andmetööde osas on jaotatud. Sealjuures üks võimalik lahendus on olukord, kus asutus vastutab treeningandmete loomise ja märgendamise loogika eest, samal ajal kui välispartner teostab andmete tehnilisemat laadi eeltöötlust (mis sageli juba andmeteadlase pärusmaa) ja mudeli arenduse. Sellist tööjaotust peetakse kulutõhusaks, kuna andmete sisuline ettevalmistus on asutuse jaoks sageli ka märkimisväärselt odavam.

Väliseid partnereid kasutatakse rohkem ja teadlikult aga innovatsiooniprojektides, kus organisatsioonil puudub varasem kogemus või selge arusaam lõpptulemusest. Näiteks kirjeldab üks intervjuueeritav vestlusroboti projekti kui innovatsioonipartnerlust, kus eesmärk ei olnud ainult toimiva lahenduse loomine, vaid ka organisatsioonisisese teadmise kasvatamine:

Me hankisime selle täiesti teadlikult välise teenusena, sest me ei olnud alguses päris kindlad, kuhu see meid viib. Kui oleksime kõike ise teinud, ei teaks me täna pooli asju nii hästi. (K1: Andme- ja tehisarusuuna juht, Intervjuu 13, Asutus 10)

Intervjuud toovad siiski esile ka mitmeid riske, mis kaasnevad andmetöö täieliku väljast hankimisega. Kõige sagedamini mainitakse kontekstipimedust: väline partner ei pruugi mõista andmete sisulisi eripärasid ega pikaajalisi mõjusid. Üks intervjuueeritav (K3: Andme- ja tehisarusuuna juht, Intervjuu 9, Asutus 8) näiteks kirjeldas olukorda, kus väliste partneritega tehtud projektide tulemusel tekkis asutuse andmelaos struktuurne kaos, mis vajab hiljem veel eraldi korrastamisprojekte.

Probleemina nähakse ka riigihangete jäikust. Uudsed tehisarurakendused eeldavad sageli agiilsemat laadi, iteratiivset lähenemist (ehita → testi → saa tagasisidet → kohanda → korda), samas kui hankereeglid nõuavad täpset lähteülesannet ja lineaarset teostust olukorras, kus lahenduskäik ei ole alati üks-ühele veel selge. Lisaks osutati intervjuudes ka liigsele tarnijasõltuvuse riskile: kui andmetöö on seotud ühe konkreetse partneriga, võib näiteks partneri vahetus muuta andmete võrreldavuse ajas problemaatiliseks.

Kokkuvõttes näitavad läbi viidud intervjuud, et väliseid partnereid kasutatakse peamiselt tehnilise võimekuse täiendamiseks, samas kui andmete sisu, kvaliteet ja tõlgendus hoitakse eelistatult majasiseselt. Nagu üks intervjuueeritav kokku võttis:

Meie poolt on treeningandmed, sest see on meile kõige odavam ja kõige mõistlikum teha oma andmetel – ja meil on selleks spetsialistid olemas. (K1: Tehisaruprojekti juht, Intervjuu 11, Asutus 10)

Intervjuud näitavad, et kuigi väliseid partnereid kasutatakse tehnilise võimekuse ja innovatsiooni toetamiseks, eelistatakse andmete sisuline tõlgendamine, märgendusloogika ja kontekstitundlik osa hoida majasiseselt. Domineerib hübriidmudel, kus strateegiline kontroll ja tähenduslik andmetöö jääb asutuse kätte, samal ajal kui tehnilisem arendus delegeeritakse partneritele.

Andmetöö organiseerimispraktikad ja eripärad

Andmetöö organiseerimisel kohtub kahte tüüpi lähenemist: projektipõhine *ad hoc* lähenemine ja väljakujunenud protsessidega rutiinne töövoog. Enamasti domineerib esimene, kus andmetöö korraldus luuakse projekti käigus käigu pealt ja nullist. Mitme intervjuueeritava kirjeldusest ilmneb, et andmetöö protsesside algatused on sageli üksikute ekspertide või projektijuhtide õlgadel. Näiteks üks intervjuueeritav (K3: Andme- ja tehisarusuuna ekspert, Intervjuu 2, Asutus 3) kirjeldas projekti, kus oli otsustusring kitsas, piirdudes 2–3 inimesega, kes „otsisid viisi, kuidas midagi tehtud saada“. Sarnast mustrit kirjeldab ka teine intervjuueeritav (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5), kus andmetööde loogika tuli partnerile ise nullist selgeks õpetada ja töövoog “tuli kuidagi peast“.

Mõnevõrra selgem struktuur ja rollijaotus on projektides, kus on kaasatud välised arenduspartnerid. Siin rakendatakse sageli tööjaotust, kus asutus tellijana vastutab ärilise sisendi ja andmete kättesaadavuse eest, samas kui partner tegeleb andmete töötlusega. Siinkohal üks intervjuueeritav (K2: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2) toob ühe varasema projekti näitel välja mudeli, kus sisemiste ressurssidega kaardistati organistatsioonis asuvad andmeallikad ja väljastusreeglid, kuid andmete puhastamine ja annoteerimine jäeti partnerile, et vältida konfliktseid olukordi vastutuse jagamisel. Teisalt, ühe projekti puhul (K2: Tehisaruprojekti juht, Intervjuu 10, Asutus 9) on rollid rangelt reguleeritud ja dokumenteeritud protsessijoonistega, kus on määratud kindel isik, kelle igapäevane ülesanne ongi mudeli resultaadi analüüsimine ja mudeli treenimine. Seega, pikaajaliste teenuste puhul asendub *ad hoc* korraldus lihtsamini formaliseeritud rutiiniga.

Andmete kogumise faasi iseloomustab sageli ka mastaapne filtreerimistöö, kus esialgselt suurest andmehulgast jääb sõelale vaid murdosa. See protsess nõuab aga märkimisväärset inimressurssi ja domeeniteadmist, mida ei ole paraku võimalik täielikult automatiseerida. Ilmekas näide

pärineb ühest projektist, kus algsest mitmest miljonist pildist jäi pärast mitmekordset filtreerimist (pimedad pildid, ebasobivad rakursid) järele vaid umbes 10 000 kasutuskõlblikku ühikut.

Tegelikult, kui me esialgu läksime projekti, see teadmine mul juba tegelikult oli eelnevatest kogemustest ja koolitustest, et kui tellija alguses ütleb, et tal on tohutu infopank olemas, siis kasutada lõpuks saad 10%. Ütleme nii, et meie projektiga enam-vähem nii läkski. (K1: Tehisaruprojekti juht 2, Intervjuu 4, Asutus 4)

See illustreerib ühtlasi kenasti ka andmetöö korralduse ühte kesket väljakutset: toorandmete olemasolu ei võrdu treeningandmete olemasoluga.

Tekstipõhiste projektide puhul (K2: Andme- ja tehisarusuuna juht, Intervjuu 13, Asutus 10) on ettevalmistus aga veelgi keerukam, nõudes teinekord ka asutuse olemasoleva dokumentatsioonibaasi sisulist ümberkirjutamist. Näiteks selgus, et inimestele mõeldud kasutusjuhendid ei olnud masinloetavad ega loogilised. Lahendusena kasutati innovaatilist lähenemist, kus teine tehisarurakendus (Google Gemini) kirjutas ekraanipiltide põhjal uued, struktureeritud juhendid, mida seejärel söödeti loodavale mudelile.

Vähemasti aga mõnede erialaspetsiifiliste valdkondade kontekstis näib fookus olevat nihkunud kvantiteedilt kvaliteedile. Kui suured keelemudelid treenitakse tohutute andmemahutude peal, siis spetsiifiliste ülesannete puhul on sealne korralduslik prioriteet luua „väike kogus ülihead“ andmestikku, isegi kui see tähendab „käsitsi trükkimasina teksti sissetoksimist“ (K: Andmete- ja tehisarusuuna juht, Intervjuu 8, Asutus 7).

Märgendamine on esmatasandi andmetöö kõige töömahukam osa ning selle korraldamisel kasutatakse seinast-seina lahendusi – alates täielikust käsitööst kuni keerukate GIS-põhiste kihtide töötluseni. Kõige levinum praktika on endiselt manuaalne märgendamine, mida teostavad sageli asutuse enda töötajad. Näiteks ühes projektis (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4) toimus märgendamine käsitööna, kus töötajad tõstsid lihtsalt pilte Windowsi kaustade vahel, millel olid siis vastavalt erinevad kategooriad. Samuti kirjeldab üks intervjuueeritav (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2) taotluste analüüsi, kus projektijuht töötas käsitsi läbi sadu lehekülgi teksti, värvides olulisi löike roheliseks või kollaseks, et luua sellega võrdlusmaterjal AI tulemuste hindamiseks.

Keerukamate andmete puhul (K1: Tehisaruprojekti juht, Intervjuu 11, Asutus 10) oli aga märgendamine integreeritud juba olemasolevatesse tööriistadesse ja töövoogudesse. Sellisel

juhul kujunesid treeningandmed olemasolevate andmekogude või andmekihtide võrdlemise kaudu. Andmetöö oli korraldatud nii, et see toetus osaliselt töötajate tavapärastele kontrolli- ja hindamisrutiinidele: aluseks võetakse hiljuti üle vaadatud juhtumid või objektid, mis aga vähendas eraldi märgendamistöö vajadust.

Hübriidlahendused ja märgendamise automatiseerimine on samuti tõusuteel, kuid ettevaatlikult. Üks intervjueeritav (K2: Andme- ja tehisarusuuna juht, Intervjuu 3, Asutus 1) kirjeldab protsessi, kus arenduspartner tegi esmase märgendamise, seejärel treeniti seal automaatne mudel ning lõpuks teostas telliv asutus vaid pistelist kontrolli (mõnest kümnest tuhandest pildist kontrolliti käsitsi tuhatkond).

Huvitava erandina toob üks intervjueeritav (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5) välja olukorra, kus andmetöötlus toimub täielikult „musta kastina“ välise teenusepakkuja poolt, ning asutuse roll piirdub vaid vigade parandamise töövoos haldamisega.

Andmetöö kvaliteedijuhtimine on seni korraldatud peamiselt läbi inimese-ahelas (*human-in-the-loop*) meetodi. Kuna tehisaruga seotud mudelite eksimisruum on avalikus sektoris riskantne (nt toetuste jagamine või maksuinfo), on protsessid reeglina üles ehitatud nii, et lõpliku töö kriteeriumiks jääb (vähemalt seni) inimese hinnang. Ühe intervjueeritava (K1: Tehisaruprojekti juht 1, Intervjuu 6, Asutus 2) projektis oli kvaliteedikontrolliks loodud näiteks suur tabel, kus võrreldakse (otsustutsoena funktsioneeriva) tehisarurakenduse hinnanguid inimkomisjoni ajalooliste otsustega. Viimase puhul on huvitav tähelepanek see, et seal polnud eesmärk alati 100% ühtivus, vaid argumentatsiooni kvaliteet – märgati ka juhtumeid, kus tehisarurakenduse argumentatsioon oli komisjoni omast tugevam. See aga viitab omakorda uudsele dünaamikale (mis, ma arvan, et vääriks täitsa eraldi juhtumipõhist akadeemilist uurimist), kus andmetöö korraldus ei ole seotud pelgalt enam mudeli treenimisega ja mingi äriprotsessi tõhustamisega, vaid võib võimaldada paralleelselt ka olemasoleva halduspraktika kui äriprotsessi teatavat auditeerimist st loodud tehisarumudel annab tagasisidet ka algele äriprotsessile endale.

Vaatad neid hindeid, mis need AI-d annavad ja siis sa vaatad selle järgi, et nagu mis ühtib sellega, mis komisjon andis ja mis ei ühti [...] Aga neid lahtreid ma ikkagi vaatasin ka puhtalt sellepärast, et on olnud noh, väga häid näiteid sellest, kus AI ja komisjoni arvamused on täiesti lahku läinud. Aga noh, see argumentatsioon on nagu oluliselt võimsam kui näiteks komisjoni enda oma on. (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2)

Riskide maandamisel on keskne koht andmekaitsel ja juurdepääsuhoodusel. Näiteks tõi üks intervjuueeritav (K1: Andme ja tehisarusuuna juht, Intervjuu 12, Asutus 11) näite praktikast, kus tundlike andmete puhul võib olla korraldus vahel ka selline, et väline partner tuleb füüsiliselt ise kohapeale asutuse sisevõrku andmeid töötlemas, selle asemel et andmed liiguksid asutuse infosüsteemidest välja. See tekitab küll korralduslikke ebamugavusi ja aeglustab protsesse, kuid on turvalisuse seisukohalt teinekord ainuvõimalik.

Peamine andmetöö korraldust pärssiv tegur on ressursipuudus ja ajasurve. Üks intervjuueeritav (K2: Tehisaruprojekti juht, Intervjuu 10, Asutus 9) toob välja karmi reaalsuse: kuigi vestlusi on juturobotis tuhandeid, suudetakse ressursipuuduse tõttu (olukord, kus juht täidab kümnet rolli) süvitsi analüüsida vaid murdosa - puudub eraldi eelarve täiskohaga "treeneri" palkamiseks, mistõttu kvaliteedikontroll on pisteline.

Otsustusprotsessid andmetöös

Kõige selgem muster, mis intervjuudest läbi kumab, on andmetega seotud otsustusõiguse paiknemine sisu- ehk äripoolel. Tehnilised meeskonnad ja arenduspartnerid täidavad pigem nõustavat või täideviivat rolli, samas kui otsused andmete valiku ja kvaliteedikriteeriumide osas teevad need, kes valdavad valdkondlikku teadmist. Näiteks ühe projekti puhul (K2: Tehisaruprojekti juht, Intervjuu 10, Asutus 9) on andmete valik ja vastuste koostamine täielikult delegeeritud asutuse teenusejuhtidele, kes on oma valdkonna tippspetsialistid. Mitmes asutuses määravad andmeotsused ka rahvusvahelised standardid ja metoodikaosakond ning Euroopa Liidu poolt tulevad nõuded.

See praktika aga viitab Eesti avaliku sektori projektide puhul andmetöö korralduse ühele olulisele tunnusele: tehisaruprojektide sisulised juhid ja suunajad asuvad väljaspool IT-osakonda. Siiski, kuigi sisu dikteerib vajaduse, piirab reaalsus võimalusi ning seetõttu kirjeldavad mitmed intervjuueeritavad andmetega seotud otsustusprotsesse pigem kui "edasi-tagasi põrgatamist" analüütikute, tehniliste partnerite ja andmevaldajate vahel.

Sarnast dünaamikat kohtab ka riistvarast sõltuvates projektides. Üks intervjuueeritav (K2: Andme- ja tehisarusuuna juht, Intervjuu 3, Asutus 1) kirjeldab, kuidas otsused pildituvastuse andmete osas dikteeris olemasolev kaameravõrk ja sellega seotud olemasolevad protsessid. Kuigi välised arendajad pakkusid välja võimaluse kasutada videot, otsustati riskide maandamiseks ja võimaliku tehnilise võla vältimiseks jääda siiski piltide juurde. Teisalt võib otsustega seotud

surve olla vahel ka hierarhiline. Näiteks kirjeldab intervjueritav (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5) ülevalt alla innovatsiooni survele, kus juhtkond (ka ministeeriumi tasandil) soovib proovida uusi meetodeid (nt masinlugemine), et kaasas käia rahvusvaheliste trendidega.

Dokumenteerimine ja logimine andmetöös

Kõige levinum muster andmetöö korralduses on kirjalike juhendite ja reeglistike puudumine varajases faasis või väiksemamahulistes projektides. Mitmed intervjueritavad põhjendavad dokumentatsiooni puudumist projekti eksperimentaalse iseloomu, väikese meeskonna ja ajutise loomuga. Domineerib arusaam, et kui andmetööga tegeleb vaid üks või kaks inimest, on formaliseerimine pigem ebaefektiivne bürokraatia. Üks intervjueritav toob välja, et kuna töövoog ei olnud veel paigas ja testiti erinevaid lähenemisi, tundus juhendite loomine ennatlik:

Me lihtsalt sellepärast, et need meeskonnad olid nii väikesed, et see oli nagu tundus, et me ei pea sellele nagu aega raiskama. Täna eks see on nende kahe projekti kontekstis, et üks siis poolteist inimest, üks, üks inimene, et pole polegi vähe nagu juhendit kirja panna otseselt veel. (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2)

Intervjueritavad (K1: Andme- ja tehisarusuuna juht, Intervjuu 8, Asutus 7 ja K1: Tehisaruprojekti juht, Intervjuu 11, Asutus 10) viitavad, et kui märgendajad vahetuvad või projekt venib pikemaks, võivad tekkida ka nihked ja lahknevused, kuna algsed suulised kokkulepped ununevad või tõlgendatakse neid erinevalt. Dokumentatsiooni asendajaks võivad aga teinekord olla ka mitteformaalsed kanalid: näiteks kirjeldab intervjueritav (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4), kuidas ametliku juhendi asemel toimus "Teamsi jutunurk", kus lahendati jooksvalt probleeme, ning Exceli tabel, mis toimus lihtsa registrina tööjaotuse jälgimiseks.

Kui protsessidokumentatsioon (ja tööjuhised) on sageli puudulik, siis tehniline dokumentatsioon ja logid on tänu arenduspartneritele ja ka kasutatavatele tööriistadele paremas seisus. Andmetöö logimine tugineb siin peamiselt tarkvara (nt Label Studio ja muud spetsiifilised tarkvarad) sisseehitatud võimekusele salvestada tegevuste ajalugu. Mitmed intervjueritavad kinnitavad, et süsteemid logivad ise kasutajate tegevusi (tekib auditeerimisjälj), kuid neid andmeid kasutatakse harva. Oluline puudujääk ilmneb aga sisuliste otsuste dokumenteerimisel – süsteem võib küll fikseerida, kes ja millal andmeid muutis, kuid miks ta seda tegi, jääb fikseerimata. Siinkohal toob üks intervjueritav esile praktilise piirangu: sisuliste põhjenduste lisamine (näiteks miks

andmekirje loeti vigaseks) nõuaks töötajalt lisaliigutusi, mida massandmetöötuse puhul välditakse efektiivsuse nimel.

Seda saab näha, et kes mingisugust kirjet kontrollis... Ee aga seda, et miks ta otsustas selle millegiks parandada, seda mis igal alati nagu ei ole. Ja kui see eeldaks seda, et peaks ise midagi sisestama, mis võtaks rohkem kui ühe kliki või ühe vajutuse, siis see läheks juba lihtsalt... Seda ei oleks mõtet rääkida, kui seda automaatselt kätte ei saa. (K1: Andmete ja AI suuna ekspert, Intervjuu 5, Asutus 5)

Samuti toimib väline hange sageli dokumenteerimise osas distsiplineeriva meetmena. Intervjueeritavad (K1: Andme- ja tehisarusuuna juhid, Intervjuu 12, Asutus 11 ja K2: Intervjuu 3, Asutus 1) märgivad, et väliselt partnerilt nõutakse dokumentatsiooni kui osa hankelepingust ja tarnest. Viimane viitab aga teatud paradoksile: avalik sektor suudab nõuda dokumentatsiooni arenduspartneritelt, kuid on iseenda sisemistes protsessides sageli vähem nõudlik.

Intervjuudes (4, 8, 9, 11) puudutati ka Euroopa Liidu tehisintellekti määrusega (*AI Act*) seotud dokumenteerimiskohustusi. Valdav hoiak on hetkel äraootav ja passiivne. Enamik vastajaid hindab oma projekte madala riskiga kategooriasse ega ole astunud konkreetseid samme vastavusdokumentatsiooni loomiseks. Seega näib andmetöö korraldus hetkel olevat suunatud pigem praktilisele tulemusele, mitte regulatiivsele ennetustööle.

Eetilised kaalutlused tehisaruandmetöös

Valdav osa intervjueeritavatest alustab eetika teema käsitlemist andmekaitsest ja privaatsusest. Intervjueeritud praktikute jaoks on eetika tehisaruandmetöös seega eelkõige küsimus sellest, kas andmete töötlemiseks on olemas seaduslik alus ning kas riskid selles osas on piisavalt maandatud. Ühes intervjuus tuuakse näiteks välja, et sealsele valdkonnale omaselt on eetiline kontroll andmetööde osas osaliselt institutsionaliseeritud näiteks eetikakomiteede kaudu. Teistes avaliku sektori valdkondades selline raamistik sageli puudub. Mitmest intervjuust ilmneb ka, et näiteks haldusandmete kontekstis defineeritakse eetikat sageli läbi olemasoleva seadusandluse: kuna andmete kasutamine põhineb ju seadusest tuleneval mandaadil, eeldatakse vaikimisi ka, et küllap see on siis juba iseenesest eetiline. Selline filosoofia vähendab ühtlasi ka projektimeeskonna poolset vajadust arutada andmete võimaliku kallutatuse või sotsiaalse mõju üle.

Intervjuud toovad esile ka olukorrad, kus eetilised riskid ilmnevad alles pärast süsteemi kasutuselevõttu. Vestlusrobotite puhul kirjeldatakse mitmes intervjuus, kuidas kasutajad sisestavad vabatahtlikult ülitundlikku teavet, mida süsteem ei olnud algselt mõeldud töötleva. Sellistes olukordades kohandatakse andmetöö korraldust tagantjärele, täiendades kasutustingimusi ja hoiatusi. Üks intervjuueeritav kirjeldas seda dilemma aga järgmiselt:

Me oleme seda arutanud mitu korda ja see kõik taandub selle taha, et me ei ole mitte kunagi piisavalt leidlikud, et katta ära kõik juhud, kuidas inimesed midagi sinna kirjutavad. Lihtsam on tegelikult kirjeldada see privaatsuseeskirjades ära ja panna robotile hoiatus.
(K2: Andme- ja tehisarusauna juht, Intervjuu 13, Asutus 10)

Kui andmekaitse on projektides selgemalt paigas, siis andmete kallutatuse (*bias*) teemad on andmetöö korralduses märksa vähem pildis. Intervjuueeritavad tunnistavad küll, et näiteks mõned ajaloolised andmestikud võivad sisaldada süsteemseid eelarvamusi, kuid rõhutavad, et nende korrigeerimine ei ole tehniline ülesanne ning et selliseid küsimusi käsitletakse sageli ka poliitiliste või väärtuspõhiste otsustena, milleks projektimeeskonnal puudub mandaat.

Üks intervjuudes välja toodud korralduslik lahendus eetikaga (*bias* kontekstis) seotud riskide maandamisel on “inimene ahelas” (*human-in-the-loop*) põhimõtte rakendamine, kus algoritm ise ei tee lõplikke otsuseid, vaid toetab ametniku tööd. Selline lähenemine võimaldab tehisarurakenduste arendusega edasi liikuda ka siis, kui andmestiku ja sellega seotud tehisarurakenduse neutraalsuses ei olda kindlad. Seega, keerulisemad teemad nagu andmete kallutus või algoritmide sotsiaalne mõju, lahendatakse üldjuhul korralduslikult vastutuse nihutamisega inimesele. Või siis otsuste edasilükkamisega.

Üks intervjuueeritav tõi aga välja, et mõnes olukorras on aga kallutus andmete paratamatu omadus, sest ajaloolised andmed peegeldavad sageli oma ajastu väärtusi ning nende puhastamine oleks sisuliselt ajaloo ümberkirjutamine. Sellistes projektides piirdub andmetöö korraldus pigem teadlikkuse tõstmisega selle kohta, millist materjali mudel kasutab.

Teisalt ilmneb intervjuudest, et eetilised kahtlused ei pruugi alati ka jõuda vajalikule otsustustasandile, sest asutuste hierarhiline struktuur ja teinekord ka kõrgemalt tulev poliitiline tellimus võivad piirata kriitiliste küsimuste eskaleerimist. Üks anonüümne intervjuueeritav kirjeldas näiteks olukorda, kus eetilised vaidlused jäid väikese tiimi tasemele ega jõudnud juhtkonnani:

Ja siis aga siis ongi need eetilised vaidlused olid ongi meie väikse tiimi sees ja minu ja ülemuse vahel. Ja siis ülemus nagu kuidagi ei lase nagu peadirektorile ega siis sellele ministeeriumile kuidagi ausalt öelda, et kuulge, te teete siin midagi ebaeetilist ja rumalat.

Vastutus eetiliste otsuste eest jääb sageli individuaalseks, tuginedes projektijuhtide isiklikele väärtustele, mitte organisatsiooniülesele raamistikule. Intervjuudest selgub ka, et eetikaalane kompetents on projektides ebavõrdne. Juristid ja IT-spetsialistid on tavaliselt kaasatud, kuid spetsiifiline eetikakeskus on harv. Mitmes intervjuus viidatakse ka Justiits- ja digiministeeriumi poolt koostatud tehisarurakenduste arendusega seotud eetikajuhendi läbitöötamisele ning sealt teatava kindlustunde saamisele, et eetilisest vaatest on loodav lahendus nõuetega vastavuses.

Andmetöö osakaal ja ressursivajadus

Analüüsides intervjuueeritavate hinnanguid tööaja jaotusele, on läbivaks mustriks esmatasandi andmetöö (kogumine, puhastamine, märgendamine, kvaliteedikontroll) domineerimine tehnilisema töö ees. Kuigi intervjuueeritavate esmane reaktsioon "80% andmetöö, 20% tehniline töö (Muldoon, 2023)" väitele on mõne intervjuueeritava puhul skeptiline, jõutakse süvenemisel sageli järelduseni, et proportsioon on realistlik või isegi konservatiivne.

Mitmed vastajad kinnitavad, et enne "tõsise arenduse" algust kulub kuude viisi aega andmete korrastamisele ja nendega seotud bürokraatiale. Näiteks ühe projekti puhul (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2) toodi välja, et andmetööks ja testimiseks kulus pooleteise inimese täistööaeg kuue kuu vältel, samas kui teises projektis panustas üks inimene terve aasta.

Andmetöö suur osakaal tuleneb sageli asjaolust, et avaliku sektori andmed, kuigi olemas ja suures koguses, ei ole tehisarutemaatika kontekstis üldjuhul koheselt ja üheselt kasutatavad. Üks intervjuueeritav (K1: Andme- ja tehisarusuuna juht, Intervjuu 11, Asutus 10) märgib, et kuigi tema asutusel on väga head analüütilised andmed, on need tehisarutemaatika kontekstis toored ja nõuavad massiivset puhastamist, mis võib vabalt viia andmetöö osakaalu 80% ligi. Seda ajalist ebavõrdsust illustreerib ilmekalt teise intervjuueeritava kogemus, kus mudeli tehniline arendus oli andmete loomise kõrval sekundaarne:

Ma arvan, jälle äkki siuke 70-30. 70 pigem andmete loomise poole ja 30 siis see keelemudeli pool ja tuvastatud pool... kui kaaluda kogu inimeste aega ja fookust organisatsioonis, siis

laias laastus võib see umbes niimoodi, et üks arendaja... ja kolm-neli inimest, kes tegelevad andmete ettevalmistusega. (K1: Andmete ja tehisaru suuna juht, Intervjuu 8, Asutus 7)

See tsitaat viitab korralduslikule eripärale: tehisaruprojekt vajab väikest, kõrgelt kvalifitseeritud tehnilist tiimi, kuid selle ümber peab olema mitmekordne tagala, kes tegeleb andmete ettevalmistamisega.

Mustrina joonistub välja ka erisus aja ja raha proportsioonides. Kuigi andmetöö võtab enamiku ajast (60–80%), moodustab see eelarveliselt sageli vähemuse (nt 30–40%). Üks intervjuueeritav tõi välja järgmist:

Võib-olla ajalises mõttes on andmete ettevalmistamine kuskil 60% võiks olla, et need andmed oleksid tõesti kvaliteetsed ja neid oleks piisavalt palju. See oli ajaline ja rahalises mõttes võib-olla ma ei tea, 30%, sest need tööd on palju odavamad kui mudelite treenimine ja andmeteadlaste tunnid ja lisaks veel see arvutusressurss on ka väga kallis. (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 11, Asutus 10)

Seda põhjendataksegi üldjuhul tööjõu hinnastamise erinevustega: andmeid märgendavad ja puhastavad sageli asutusesisesed spetsialistid, kelle tunnitasu on madalam kui sisseostetud andmeteadlase või arendaja oma. Seda illustreerib täiendavalt ka tsitaat teiselt intervjuueeritavalt:

Seal on ju ikkagi nagu tulevad kerged käärid sisse, et seal see proportsioon ikkagi mingil määral langeb. Kuna jah, et ja et selles mõttes selliste need [....] palk on ikkagi väiksem kui kui kui andmeteadlase või jah tehisintellekti mingi spetsialisti. Aga ma ei julge sulle öelda mis see on nagu arvudes. (K2: Tehisaruprojekti juht, Intervjuu 7, Asutus 6)

Ka kuues ja seitsmes intervjuueeritav (K1: Asutus 2 ja K2: Asutus 9) toovad otseselt välja, et need tööd on palju odavamad kui mudelite treenimine. See tekitab aga olukorra, kus andmetöö on projektides rahaliselt alahinnatud komponent. Veelgi enam, tihti puudub andmetööl eraldi eelarverida sootuks, kuna seda tehakse olemasolevate töötajate poolt muude ülesannete kõrvalt. Üks intervjuueeritav kirjeldab näiteks olukorda, kus intensiivne andmetöö (hinnanguliselt 40% projekti mahust) tuli mahutada olemasoleva tööaja sisse, ilma lisatasuta.

Rahalises numbris ma mõtlen, et pigem on see arendus puhul suurem, sest seda andmete ettevalmistamist me tegime oma oma jõududega, et minu teada kellegile selle eest lisa lisamaksu ei tulnud. See tuli oma tööde tegemiste vahele ära mahutada... Tuli tuli ka siin õhtul ja nädalavahetusel tunde teistel sisse. (K2: Andme- ja tehisarusuuna juht, Intervjuu 3, Asutus 1)

See aga viitab otseselt varjatud kuludele. Kuna andmetöö on odav ja sisemine, kiputakse selle tegelikku koormust alahindama, mis aga võib viia töötajate ülekoormuse ja varjatud ületundideni.

Intervjuudest selgub ka, et andmetöö osakaal kasvab sageli ootamatult suureks sageli just projekti hilisemates faasides, mida algselt ei osatud ette näha. Eriti puudutab see vestlusrobotite ja LLM-põhiste lahenduste tulemuste järjepidevat ja kohati lõpmatut valideerimise vajadust. Üks intervjuueritav (K2: Tehisaruprojekti juht, Intervjuu 10, Asutus 9) toob välja, et kuigi esmane mudel võib olla lihtne, nõuab selle kvaliteedi tagamine igapäevast käsitööd - keegi peab iga päev tund aega ainuüksi logisid lugema, et aru saada, mis tegelikult toimub ja see on tegevus, mida ei saa automatiseerida.

Tööriistade ja tehniliste lahenduste roll andmetöös

Hoolimata tehisaruprojektide kõrgtehnoloogilisest kuvandist, sõltub esmatasandi andmetöö korraldus paljudel juhtudel endiselt lihtsatest, mitte-spetsiifilistest tööriistadest. Eriti projektide algaasis või seal, kus andmepäringute automatiseerimine on takistatud, muutuvad Excel, Word ja Confluence (viimane on nii avaliku- kui ka erasektori IT-s laialdaselt kasutatav pilvepõhine dokumenteerimise tarkvara) asendamatuteks vahenditeks.

Üks intervjuueritav toob ilmekalt esile olukorra, kus andmekaitse ja süsteemide jäikus sundisid projektijuhti kasutama Confluence'i ja käsitsi andmekorjet. Kuna automaatsete päringute tegemine oleks võtnud juba isegi ülekoormatud andmeanalüütikute piiratud aega, leiti, et lihtsam on neid asju käsitsi kokku koguda.

Noh, et see, et andmed, mis ma sain taotlustelt, et neid ma tihti võtsin käsitsi ja see on puhtalt meie enda tehnilistel põhjustel... Ja siis ma lihtsalt noh, leidsin päeva lõpuks, et mul on lihtsam neid asju nagu käsitsi kokku koguda kui koguda kokku noh, mingite loogilisemate süsteemidega. (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2)

Selline korraldus viitab, et tööriista valikut ei dikteeri alati efektiivsus, vaid ka institutsionaalsed takistused – Excel ja manuaalne kopeerimine on küll ajamahukad, kuid pakuvad mitte-tehnilisele töötajale sageli autonoomiat, mida jäigad süsteemid ei võimalda.

Küpsmates projektides on andmetöö korraldus liikunud spetsialiseeritud tarkvara kasutamisele, kusjuures avalikus sektoris on tugev eelistus vabavaraliste (open source) lahenduste suunas. Üks intervjuueritav (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2) rõhutab Label Studio

kasutamist pildituvastuse projektides, põhjendades seda riigiasutuse praktikaga eelistada vabavara ja lokaalseid (*on-premises*) lahendusi. Selline valik võimaldab eelmainitud intervjueeritava sõnul hoida andmetöö tulem füüsiliselt majas sees ja vältida sõltuvust suurtest kommertspilvedest. Viimased võivad olla kallid ja kätkevad endas ka turva- ja jätkusuutlikkuse riske. Olulise aspektina tuuakse välja, et kaasaegsed andmetöötööriistad ei ole sageli enam lihtsalt passiivsed märgistuskeskkonnad, vaid sisaldavad mõnikord juba ka näiteks mini-mudeleid, mis teevad eelmärgendamist, muutes inimese rolli andmete märgendajast üha rohkem nende automaat-märgendite valideerijaks.

Riikliku vestlusroboti Bürokratt puhul on peamine tööriist aga integreeritud juba teenusehaldusse vaikimisi. Sealne ametniku "tagatuba" võimaldab mitte ainult vestlusi jälgida, vaid neid ka reaajas treeningandmeteks muuta ("lukkan küsimuse vastuse külge", nagu ühes intervjuus tabavalt märgitakse) ja mudelit treenida. See näitab andmetöö korralduse integreeritust igapäevatöösse – tööriist on tehtud mugavaks, et soodustada andmete pidevat rikastamist.

Olukordades, kus standardlahendused jäävad hätta, on asutused asunud ka ise tööriistu looma või kasutama uudseid lähenemisi. Üks intervjueeritav (K1: Andme- ja tehisarusuuna juht, Intervjuu 8, Asutus 7) näiteks kirjeldab unikaalset "vaba märgistustööriista", mis loodi spetsiifilise treeningmaterjali töötlemiseks. Kuna kommertsmudelid eeldavad väiksemahulisi sisendeid, oli vajalik luua tüki oma vahekihti, mis võimaldaks häälestada pilti nii, et see oleks loetav nii inimesele kui masinale.

Andmetöö kvaliteedikontrolli mehhanismid

Üheks peamiseks tehisarurakenduste loomise kvaliteedikontrolli alustalaks peetakse andmeteaduses sageli topelt-märgendamist või ristkontrolle, et välistada inimlikke eksimusi. Intervjuudest ilmneb aga, et Eesti avalikus sektoris on see pigem erand kui reegel. Valdav muster näitab, et ressursside nappuse tõttu usaldatakse sageli ühe spetsialisti otsust, eriti kui tegemist on pikaajalise kogemusega töötajaga. Üks intervjueeritav toob selle dilemma ilmekalt ka esile, tunnistades ausalt nurkade lõikamist - kuigi ollakse teadlikud, et professionaalsetes rakendustes peaks andmeid kontrollima kaks paari silmi, loobutakse sellest, tuginedes töötajate kõrgele kompetentsile.

Päris professionaalsetes rakendustes, kus mitu inimest vaatavad, on lihtsalt... Me oleme selle nurki lõiganud ja tegelikult aru saanud, miks see nii on. [...] Meil on mingis mõttes, et see võib olla nüüd õigustus, aga kuna me tunneme, et meil on need inimesed, kes sellega

tegelevad, on väga pädevad. [...] Siis me oleme nagu sellest möödunud, aga tegelikult oleks muidugi õige jah, dubleerida. (K1: Andme- ja tehisarusuuna juht, Intervjuu 8, Asutus 7)

Siiski leidub ka positiivseid erandeid. Üks intervjuueeritav kirjeldab projekti, kus rakendati süsteemset topeltkontrolli: pärast esimese märgendaja tööd vaatas vähemalt üks inimene faili üle.

Ja siis me tegime nii, et [.....], siis pärast seda ka üks inimene vähemalt vaatab selle üle selle ühe faili, mis [.....] saadab. Vähemalt topeltkontroll oli. Ja peaks olema ka tulevases projektis. (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 11, Asutus 10)

Üks kvaliteedikontrolli keskne element on etalonandmestiku (*gold standard*) loomine, mis toimib allikana nii arendajatele kui ka mudeli testimiseks. Analüüsist selgub, et etalonide loomine on pigem reaktiivne tegevus – vajadus selle järele tekib siis, kui andmemahd kasvab üle pea või tekivad lahknevused hinnangutes. Näiteks tõi üks intervjuueeritav (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4) näitena protsessi, kus piltide suure mahu tõttu ei jõutud kõike kontrollida. Lahendusena loodi etalonid vaid nendele juhtumitele, kus oldi ~100% kindlad (kolme inimese konsensus), ja anti need arendajale ette.

Korralduslikust vaatepunktist on oluline ka küsimus: kes otsustab vaidluse korral? Intervjuudest joonistub välja selge hierarhia, kus tehniline meeskond delegeerib sisulised vaidlused domeeniekspertidele või tellijale. Näiteks tõi üks intervjuueeritav (K3: Andme- ja tehisarusuuna ekspert, Intervjuu 2, Asutus 3) välja, et erimeelsuste korral „pöörduakse alati tellija poole“, kes peab ütlema, mis lahendus on õige. See aga tähendab, et ka andmetöö kvaliteet on otseses sõltuvuses tellija võimekusest oma soove ja definitsioone sõnastada. Huvitava muustrina ilmneb ka ebakindluse haldamine. Üks intervjuueeritav (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4) toob välja, et kui inimene ei suutnud pilti hinnata, oli tal võimalus tõsta see eraldi kausta, et mitte raisata aega mõtisklemisele.

Üks kriitilisemaid teemasid kvaliteedikontrollis on andmete vastavus reaalsele kasutusolukorrale. Siinkohal jagas üks intervjuueeritav valusat õppetundi, kus treeningandmetel saavutati 99% täpsus, kuid töökeskkonnas süsteem ebaõnnestus (enamik valenegatiivseid). Põhjuseks oli see, et inimesed tegid reaalsuses toiminguid teistmoodi kui treeningandmestikus eeldati.

Me saime väga hea õppetunni... treenitud mudelite täpsus oli 98-99 protsenti. Panime asja laivi... Siis me jõudsim sinna, et 90 protsenti pilte said vale negatiivse vastuse. [...] Lihtsalt õpetamise käigus me ei osanud arvestada seda, et inimene võtab selle objekti endale kätte...

teeb kolme kilomeetri kauguselt pilti. (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2)

Intervjuudest ilmneb, et tehisaruandmetöö kvaliteedikontroll avalikus sektoris on tugevalt ressursist (või täpsemalt selle puudumisest) sõltuv ning sageli rajatud usaldusele üksikute pädevate spetsialistide vastu. Otsustusõigus vaidluste korral koondub enamasti domeeniekspertidele või tellijale, mis muudab kvaliteedi sõltuvaks nende kompetentsist.

Esmatasandi andmetöötaja profiil avalikus sektoris

Valdav enamus intervjueeritavatest kinnitab, et andmetöö tegijate valikul on määravaks sisuline ekspertiis, mitte andmetöö alane või vastav tehniline taust. Kuna tehisarutehnoloogiate põhised mudelid vajavad treenimiseks tõestandardit (*ground truth* ehk mudeli eesmärgi saavutamiseks vajalik kontekstispetsiifilisele tõele lähedane andmestik; mitte ajada segi *absoluutse tõe* tähendusega), saavad seda pakkuda vaid inimesed, kes valdavad vastava domeeniga seotud nüansse. Näiteks ilmastikunähtuste tuvastamise projektis ei palgatud väliseid märgendajaid, vaid kasutati asutuse enda vastavaid spetsialiste, sest vaid neil oli silma ja kogemust eristada piltidelt spetsiifilisi nüansse.

Sama loogika kehtib ka näiteks projektides, kus andmete dokumenteerimine on delegeeritud spetsiaalsetele, uuringute dokumenteerimisega seotud personalile ja vastutavatele uurijatele, ning haldusasutustes, kus anomaaliade tuvastamine eeldab sageli kogenud spetsialisti vaistu. Viimast illustreerib ühe intervjueeritava tsitaat:

Need, kes neid pilte hindasid, olid meie enda töötajad tol hetkel. Ja olidki, noh, võib-olla profiili poolest on need nad meil [.....]... miks me neid kasutasime, oli see, et neil oli see kogemus ja neil oli see teadmine, kuidas seda hinnata. See oli see nende profiil tegelikult, et see sobis täpselt nende nende tol hetkel tööülesannetega. (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4)

See tsitaat illustreerib olulist korralduslikku printsiipi: andmetöö on integreeritud vastava domeeni juurde ja pole eraldiseisev - andmetöö viiakse sinna, kus on teadmine, mitte vastupidi. See tagab küll andmete kõrge sisulise kvaliteedi, kuid tekitab samas ka väljakutseid ressursside planeerimisel, kuna spetsialistide aeg on kallis ja piiratud.

Kuigi andmetöötajatel on üldjuhtudel kõrge domeenikompetents, ilmneb intervjuudest, et neil võib siiski olla arvestatav lünk arusaamises sellest, miks ja kuidas tuleb andmeid tehisaru jaoks

teisiti töödelda. Eriti teravalt tuleb see esile asutustes, kus on olemas pikk andmehalduse ajalugu - näiteks töötajad, kes on harjunud klassikalise deterministliku loogikaga (kus igale reale tuleb leida vaste, et statistika ja numbrid ikka klapiks jne), satuvad konflikti masinõppe loogikaga, kus ebakindel andmepunkt on parem jätta märgendamata kui valesti märgendada. Andmetöö korraldus nõuab seega juhtidelt pidevat selgitustööd. Seda illustreerib tabavalt ühe intervjuueeritava tsitaat:

Ma väga palju rõhutasin seda, et et ma tean, et te olete harjunud, et tavaliselt teil on [...] töös, et kõik asjad, mis on seal vaja kodeerida... Aga et nüüd on just väga tähtis see, et kui sulle ei tule pähe, mis mis koodi sinna peaks määrama, et ära siis hakka huupi pakkuma. Et see on masinõppe mudelile ohtlik. (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5)

Ka andmetöötajate motivatsioon varieerub seinast sein, sõltudes muuhulgas sellest, kuidas nad tajuvad tehisaru mõju oma tuleviku tööle. Positiivse näitena toob üks intervjuueeritav (K1: Tehisaruprojekti juht, Intervjuu 11, Asutus 10) välja spetsialistid, kes on ise huvitatud märgendamisest, sest nad näevad otsest kasu – mudel vähendab tulevikus nende rutiinset käsitööd. Siin on andmetöö korraldatud vabatahtlikkuse alusel ja motivatsioon on sisemine, tulenedes professionaalsest huvist.

Teisalt toob teine intervjuueeritav esile aga murettekitava trendi. Spetsialistid, kelle töö on ajalooliselt olnud intellektuaalne ja uuriv, tunnevad, et tehisaruprojektide andmetöö (nt masinlugemise vigade parandamine, punktide nihutamine pildil jms) muudab nende töö rumalamaks ja tehnilisemaks, mistõttu tekib võõrandumine: selle asemel, et töötada teksti sisuga, töötatakse nüüd piksli ja sümboliga.

Minu arust need isegi tänu tehisintellektile tegelikult lähevad nagu rumalamaks see, mida inimene peab tegema... Kui palju auru meil on sinna läinud selle asemel, et mitte nagu tekstiga sisuliselt töötada... vaid selle peale, et nagu mingi pildi peal kuskil mingeid punkte käsitsi ümber nihutada. (K2: Tehisaruprojekti juht, Intervjuu 7, Asutus 6)

See aga viitab riskile, et kui andmetöö korraldus ei arvesta töötajate professionaalse identiteediga, võib see viia motivatsiooni languseni ja ühtlasi tuua kaasa ka vastandumise innovatsiooni suhtes.

Koolitus, tagasiside ja tunnustamine andmetöös

Valdavalt iseloomustab koolitusprotsessi *ad hoc* lähenemine. Selle asemel, et luua mahukaid juhendeid, eelistatakse personaalseid vestlusi ja lühikesi praktilisi sessioone. Näiteks kirjeldas intervjuueritav (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2) protsessi kui lennult leiutamist, kus inimestega istutakse koos maha, arutatakse töö läbi ja lastakse neil töökeskkonnas ise katsetada, et tekiks esmane tunnetus. Selline personaalne ja vahetu juhendamine on võimalik aga eelkõige tänu suhteliselt väikestele projektidele ja meeskondadele. Samuti kirjeldas üks intervjuueritav (K2: Tehisaruprojekti juht, Intervjuu 10, Asutus 9) treeningut kui ühepäevast koosolekut, kus näidatakse ette, kuidas treeningandmeid tekitada ja millised on erandid.

Analüüsist ilmneb ka olukordi, kus koolitus ja konteksti loomine on olnud täielikult puudulikud. Eriti teravalt tuleb see esile intervjuueritavaga meditsiinisektorist (!), kes toob välja, et andmeid kogunud töötajad ei teadnud ega huvitunud asjaolust, et nende töö on sisendiks masinõppele. Nende jaoks oli tegemist järjekordse ankeedi täitmisega.

Tagasiside andmine andmetöötajatele varieerub ka projektiti märkimisväärselt. Parimad praktikad hõlmavad töötajate hoidmist infoväljas ka pärast nende tööloigu lõppu. Näiteks ühe intervjuueritava (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4) sõnul on esmatasandi töötajad siiani kaasatud ja teavad, kuidas mudel päriselus töötab. Ka tunnustamine toimus seal läbi organisatsioonisiseste sündmuste – prototüübi esitlused majasisestel koosolekul või "aasta tegu" nominatsioonidega seoses. Kuigi rahalist boonust üldiselt selliste asjade eest eelmainitud intervjuueritava sõnul ei maksta, on sotsiaalne tunnustus oluline motivaator.

Teisalt on ka projekte, kus tagasiside on minimaalne või olematu. Üks intervjuueritav (K2: Andme- ja tehisarusuuna juht, Intervjuu 3, Asutus 1) märgib, et kuigi seminaril räägiti eesmärkidest, siis töö käigus anti tagasisidet vaid mudeli testimise faasis. Veelgi kriitilisem on olukord seal, kus andmetöö on integreeritud igapäevarutiini nii sujuvalt, et see muutub märkamatuks. Siinkohal on alampeatüki lõpetuseks sobilik tsiteerida ühte teist intervjuueritavat:

"[...] meil ei pane keegi seda tähele. Selles mõttes saa võtta, kui... kui asjad töötavad, siis ei panda tähele lihtsalt ja ongi hästi." (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2)

See tsitaat toob esile andmetöö paradoksi: ideaalne andmetöö on nähtamatu. Kui süsteem töötab, ei saa andmetöötaja kiita ning tähelepanu saadakse vaid siis, kui midagi on valesti.

Peamised õppetunnid ja tagasisivaateline refleksioon

Peamise muustrina toovad intervjuueeritavad esile vajaduse alustada andmetööga – eriti puhastamise ja märgendamisega – oluliselt varem. Tagantjärele tunnistatakse, et projektide fookus oli algselt liigselt tehnilisel teostusel, samas kui andmete kvaliteet jäeti rohkem juhuse hooleks. Mõned intervjuueeritavad (K2: Intervjuu 3, Asutus 1 ja K1: Intervjuu 4, Asutus 4) rõhutavad lausa, et andmevood ja märgendamine peaksid olema läbi mõeldud ja ideaalis isegi teostatud enne projekti ametlikku starti. See peegeldab olulist nihet ka andmetöö korralduse mõtestamises: andmetöö ei ole enam projekti kõrvaltegevus, vaid eeldus. Eriti radikaalse ja selge seisukoha võtab siinkohal intervjuueeritav, kes leiab, et kui ta peaks täna uuesti alustama, suunaks ta põhiressursi mitte mudelite arendamisse (mis on tänaseks muutunud juba kättesaadavaks kaubaks), vaid kvaliteetse treeningandmestiku loomisesse.

Paneks jah, paneks kohe rohkem inimesi, eks ole, seda ground truth'i toksima või seda ülikõrge täpsusega treeningandmestikku looma mingisuguses hästi lahtises formaadis, mis iganes formaati, lihtsalt külma ridu sisse toksima, sest päeva lõpuks ka meile paistab, et see on kõige parema kuluefektiivsusega tegevus. (K1: Andme- ja tehisarusuuna juht, Intervjuu 8, Asutus 7)

Teine läbiv teema on ressursside – nii inimtööjõu, aja kui ka arvutusvõimsuse – krooniline alahindamine. Mitmed intervjuueeritavad kirjeldavad tagantjärele naiivsust töömahu planeerimisel. Eriti teravalt tuleb see esile asutusesisese tööjõu kasutamisel. Näiteks kirjeldab üks intervjuueeritav psühholoogilist barjääri, mis takistas küsimast piisavalt inimesi, kuna kardeti, et see ei paista paberi peal proportsionaalne:

Kuidagi pigem see noh... Kuidas nagu? Selle projekti planeerimise raames meie enda majas nagu nii-öelda sellega seotud inimestega rääkides, et noh, võib-olla endalt nagu tekkis tunne, et siin, et äkki ei ole hea toon nagu siia siia veel peale 2 inimese veel 3. juurde nõuda ja nii edasi... Et kuidagi pigem ongi süüdistab, süüdistab iseennast, et nagu ei julgenud rohkem selle eest seista. (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2)

Paradoksaalsel kombel näevad mitmed intervjuueeritavad ebaõnnestumisi ja segadust vajaliku etapina organisatsiooni küpsemisel. Üks intervjuueeritav sõnastab "pilootprojektide paradoksi": andmehaldust ei saa organisatsioonis korda teha teoreetiliselt; on vaja ebaõnnestunud pilootprojekte, mis tõestaks juhtkonnale, et asjad on valesti:

Aga teisest küljest ma ütlen, et väiksed piloodid annavad sulle organisatsioonis arusaama, mis sul valesti on. Ja sa ei saa mitte kunagi organisatsioonis minna näidata, et meil on need asjad valesti, sest keegi ei usu sind enne, kui sa tegelikult hetk on siiski käes. (K2: Andme- ja tehisarusuuna juht, Intervjuu 13, Asutus 10)

Viimane aga viitab sellele, et andmetöö korraldus on olemuslikult sageli ka eksperimentaalne. Õppetund ei ole siin vältida vigu, vaid eelkõige teha vead kiiresti ja nähtavalt, et saada põhjendused muutuste läbiviimiseks (nt andmesõnastike loomine, taksonoomiate ühtlustamine, andmeomanike määramine, andmehalduse inimeste värbamine jne). Sarnast mõttekäiku arendab ka teine intervjuueeritav (K3: Andme- ja tehisarusuuna juht, Intervjuu 14, Asutus 12), kes rõhutab, et vahel peab juht paratamatult teadlikult alustama projektiga, teades, et tingimused pole ideaalsed, sest alternatiiv – mitte midagi teha ja oodata ideaalset andmekorda – on tagurlik; selline lähenemine on vahel psühholoogiliselt raske, kuid siiski vajalik taktika innovatsiooni käivitamiseks.

Üks intervjuueeritav (K1: Andme -ja tehisaru suuna juht, Intervjuu 1, Asutus 2) on aga mures tehisarukompetentsi organisatsioonisisese ajutise loomu ja peatse kadumise pärast - kuna uued ettevõtmised on sageli rahastatud projektipõhiselt ja neil puudub püsiv meeskond, kaob teadmine sageli koos võtmeisikute lahkumisega (*"minu organisatsioonist lahkumisega kaob isegi see teadmine"*).

Ka väliskeskkonna muutuste kiirus – eelkõige tehisarutehnoloogiate kiire areng – on andnud valusa õppetunni jäikade hangete osas. Üks intervjuueeritav (K1: Tehisaruprojekti juht, Intervjuu 11, Asutus 10) toob välja, et hankeid kirjutades ei arvestatud tehnoloogia arengukiirust, mis viis hilisemate vaidlustusteni. Seega, tulevikuvaates nähakse ka vajadust paindlikumate, eelkõige eesmärgi- mitte vahendikesksete hangete järele.

Samuti tõstetakse esile tööriistade valikut. Intervjuueeritavad (K1: Intervjuu 1, Asutus 2 ja Intervjuu 5, Asutus 5) nendivad, et oleksid pidanud kohe kasutama spetsiaalseid tööriistu (nt Label Studio) ja vältima "iseõppimist" või ebasobivate mudelitega (nt mõned vigased vabavaralised mudelid) katsetamist. See aga näitab, et andmetöö korraldus vajab standardiseeritumat tööriistakasti, et vältida iga projekti puhul jalgratta leiutamist.

4.3 Tehisaruandmetöö nähtavus ja prioriteet organisatsioonis

Tehisaruandmetöö institutsionaalne korraldus

Intervjuudest joonistub läbiva mustrina välja tehisaruprojektidega seotud andmetöö korralduse juhuslikkus ning tugev sõltuvus nii konkreetse intervjuueeritud projekti juhi varasemast kogemustest, aga ka projekti iseärasustele vastavatest vajadustest. Valdav osa intervjuueeritavatest kirjeldab andmetööd kui *ad hoc* tegevust, millel puuduvad organisatsiooniülesed rollid, protsessid või püsivad kokkulepped. See viitab aga olukorrale, kus tehisaruandmetöö nähtavus on pigem madal. Eriti selgelt ilmneb see uuemate tehnoloogiate puhul. Üks intervjuueeritav kirjeldab näiteks kontrasti visuaalsete ja tekstiliste andmete vahel: kui pildipõhistes rakendustes on kujunenud teatav tööpraktika, kuid keelemudelite treenimiseks vajaliku tekstilise andmestiku ettevalmistus on korrastamata ja kokkulepeteta:

Suhteliselt ad hoc, projektipõhine lähenemine, kuna see maailm on nii uus... see, mis puudutab kõike, mis puudutab sellist keelemudelite treenimist ja mis puudutab teksti andmestikku, on ikkagi täielik nagu... Battlefield ma ütles, et seal on ikkagi nagu hullumaja... tekstiga on raskused. Jah, et seal pole ühtegi nagu kokkulepet, kokkulepitud töövoogu. (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2)

Ad hoc lähenemist ei tajuta siiski üheselt negatiivsena. Näiteks toob teine intervjuueeritav esile, et liigne formaliseerimine varajases faasis võib innovatsiooni pärssida ning just paindlikkus võimaldab projektidel kiiremini liikuda:

Mul on endal tunne, et kui see on natukene nagu sellega ujuv, siis ta annab natuke nagu projektis sellist liikumist rohkem, et oleks nagu sujuvam. Kui kirjutame väga rangelt ette asutuse poolt, mida sa tegema pead, siis võib-olla mõni projekt võib sinna taha toppama jääda. (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4)

Tulemused viitavad ka, et organisatsioonilise struktuuri puudumist kompenseerivadki sageli üksikud entusiastid, kelle õlgadele koondub kogu andmetöö korraldus. Intervjuueeritavad kirjeldavad ka korduvalt nn "ühemeheorkestri" fenomeni, kus andmetöö protsessid eksisteerivad ühe indiviidi teadmistes ja tööpraktikates, mitte organisatsiooni kollektiivses mälus. Üks intervjuueeritav toob selle selgelt esile:

Ee noh, selles mõttes on [...] projekt lihtsam sellepärast, et siin projektis on sisuliselt ainult mina, nagu nii-öelda selle põhikohana. [...] ja siis ongi, et nii-öelda selle noh, nagu sa

ütlesid, see esmane andmetöö, et see on nagu kõik, mida mina teen [...]. (K1: Tehisaruprojekti juht, Intervjuu 6, Asutus 2)

Mitmed intervjueritavad osutavad ka ajaloolisele pärandile, kus mahukat käsitööd on aastaid tehtud vaikimisi, ilma seda süstemaatiliselt andmekvaliteedi või masinõppe vajaduste kontekstis teadvustamata. Siinkohal ühe intervjueritava sõnul on see loonud olukorra, kus andmetööd küll tehakse, kuid selle strateegiline tähendus jääb varju.

Et see oligi, et varem neil masinõppe suunda ei olnud, siis oligi nagu default seisund see, et inimesed siis vaatavad peale ja määravad... selliseid tublisid kodeerijaid on aastaid ja aastaid tööl olnud ja seda käsitööd on tohutult palju tehtud... Et siukest süsteemset lähenemist on vähe. (K1: Andme- ja tehisarusuuna juht, Intervjuu 5, Asutus 5)

Üks intervjueritav (K1: Andme- ja tehisarusuuna juht, Intervjuu 12, Asutus 11) kirjeldab muuhulgas olukorda, kus kogu treeningandmestiku loomeprotsessi lähenemine tuli kiiruga projektijuhil ise luua ja kasutades selleks ka organisatsiooniväliseid ja kohati ka personaalses tutvusringkonnas olevaid inimesi, kuna organisatsioonilised protsessid seda ei toetanud.

Tehisaruandmetöö nähtavus ja väärtustamine

Intervjuudes kordub lõhe tellijate ootuste ning tehnilise reaalsuse vahel. Mitmes intervjuus toodi välja, et tehisaru (peamiselt peetakse silmas generatiivset tehisaru) laialdane kasutuselevõtt on loonud paljudele töötajatele (nii tava- kui juhtivtasand) eksliku ettekujutuse, et tehisarurakendused on kiiresti ja minimaalse eeltööga rakendatavad. Selle tulemusel alahinnatakse sageli süstemaatiliselt ka esmatasandi andmetöö mahtu ja keerukust. Üks intervjueritav tõi näiteks välja:

Ja no täna on nagu see ootus on see, et kui kellelgi tuleb nagu tootearenduse poole pealt mingi idee, siis kõik arvavad seda. Et noh, mis seal siis on, paned käima ja töötab... Ma arvan, et see on võib-olla tekkinud nende suurte keelemudelite nii-öelda täna pakutavast pakutavatest teenustest... Et ei saada aru sellest vajadusest, mis tähendab andmete töötluse ennem. (K1: Andme- ja tehisarusuuna juht, Intervjuu 1, Asutus 2)

Andmetööd on kirjeldatud ka kui "pahnaga tegelemist" – nähtamatut ja väheprestiizset tööd, mis ei konkureeri tähelepanu poolest mudeliarenduse või arhitektuuriliste otsustega. Üks intervjueritav kirjeldab olukorda, kus ta pidid oma töö vajalikkust juhtidele aktiivselt tõestama, visualiseerides ajakulu ja tegevusi, mis projekti stseenide taga toimusid.

Aga ütleme selline märgendamise pool tuli ju ikkagi arendaja poole pealt ja kui me räägime kogumise poole pealt, siis need andmed olid olemas. Et aga neid tuli hakata puhastama. [...] tööprotsessid olid paigas selles osas, kus sa hakkad juba märgendama kõiki neid asju nagu tegema, andmeid töötlemas. Aga seal vahepealses osas on tegelikult hästi palju sellist musta tööd, probleeme, mis justkui tekivad. Et noh, et võib-olla seda väärtust jah, inimesed võib-olla ei oska tunnetada. Minu juht küsis ükskord, et aga mida sina teed, et sul on ju meeskond sisse ostetud... Ja siis ma tegin talle sellise päeva pildistamise [...]. Aga reaalselt, jah, et see, mis on stseenide taga toimub. Et noh, kui ma ütlen, et projekt kestis 9 kuud vist. Ma arvan, et sellest 6 kuud me tegelesime igasuguse pahnaga. (K1: Andme- ja tehisarusuuna juht, Intervjuu 12, Asutus 11)

Andmetöö madal prestiiž avaldub ka töötajate motivatsioonis. Märgendamist ja puhastamist kirjeldatakse rutiinse, emotsionaalselt kurnava ja ajasurvel tehtava tööna, mille kvaliteet kannatab just seetõttu, et seda püütakse teha võimalikult kiiresti. Üks intervjuueeritav kirjeldas seda näiteks selliselt:

Pigem on siukene rutiinne töö, mida eriti nagu ei taheta teha, üks ole. Et kui tehakse, tahetakse. Ta kiirustades teha, üks ole, et seal tekivad ka nagu need kvaliteediprobleemid võib-olla tulevikus selles mõttes, et et ise seda märgendust tehes, mul oli küll vahepeal väga kopp ees lihtsalt. (K2: Andme- ja tehisarusuuna juht, Intervjuu 3, Asutus 1)

Samas ilmnes intervjuudes üksikuid vastupidiseid juhtumeid, kus kvaliteetset algandmestikku nähakse tehisaruprojektide keskse väärtusena ning ressursid suunatakse teadlikult märgendamisse.

Minu jaoks on see mudelivalik ja see tehniline pool võrdlemisi triviaalne, aga see alusandmestik, mille pealt sa üldse edasi lähed... kvaliteetne algandmestik on vundament. Ja ega me praegu ka raha kulutame põhiliselt ikkagi sellele, et palgata inimesi, kes märgendavad [...]. (K1: Andme- ja tehisarusuuna juht, Intervjuu 8, Asutus 7)

Viimane näide aga võib ka osutada sellele et andmetöö nähtavus ja väärtustamine sõltub eelkõige tugevalt asutuse põhiülesannetest ja andmete rollist selle missioonis. Intervjuudest ilmneb selge lõhe tehisarü ümber kujunenud ootuste ja andmetöö tegeliku töömahukuse vahel. Tehisarurakendusi tajutakse sageli kui kiiresti rakendatavaid tehnoloogilisi lahendusi, mistõttu alahinnatakse süstemaatiliselt andmete puhastamise, struktureerimise ja märgendamise mahtu.

Ressursside planeerimine tehisaruandmetöös

Tulemustest ilmneb ka, et andmetöö madal nähtavus saab alguse juba projektide planeerimise staadiumist. Enamik intervjuueeritavad tunnistavad tagasiulatuvalt, et nende projektides puudus realistlik arusaam andmetöö tegelikust mahust. Näiteks toob ühe intervjuueeritav selle tabavalt esile:

Selle suhtes, et aimdust väga ei olnud. Esialgu võiks öelda, et olime täiesti nagu pimedas toas... Ja tegelikult, noh, kui see projekti algus oli juba, siis oli aru saada, et see tuleb päris intensiivne ja mahukas hindamisprotsessis piltide poolest. Noh, andmete poolest, et seda puhastada ja teha ettevalmistusi. Tegelikult me alguses ei olnud minu arvates päris valmis selleks, et see töömaht nii suureks läheb. (K1: Tehisaruprojekti juht, Intervjuu 4, Asutus 4)

Eriti problemaatiline ongi just asutuse sisemiste ressursside arvestamine. Kui väliste partnerite töö on üldjuhul eelarvestatud ja nähtav (kas see on ka välise partneri enda majasiseselt eelarvestatud piisavalt ja nähtav? Ei pruugi praktikas nii olla ning sellele viitavad ka mõned kohad intervjuudes, kus viidatakse arenduspartneri sisemistele raskustele), siis asutusesisene andmetöö toimub sageli vaikimisi, muude ülesannete kõrvalt. Intervjuud kirjeldavad sagedasi olukordi, kus üks spetsialist täidab mitut rolli ning andmetöö konkureerib pidevalt teiste tööülesannetega. Kuna sisemisel andmetööl puudub sageli ka eraldi eelarverida, on seda lihtne alahinnata ning juhtimisotsustes eirata. Üks näide puudutab ühe intervjuueeritava kirjeldatud tehisarukomponendiga uuringut, kus viga tehti juba eos:

Viga tuli sisse juba täiesti alguses... kuna mind ei olnud kaasatud sinna planeerimisse ja keegi ise selle peale ei olnud tulnud, et meil on ju vaja kuskilt saada seda treeningandmestikku... Ja noh, kuna see on kärbete aeg ka, siis mina pidin hakkama veenma neid igasugustel erinevatel asutuse tasanditel, et kuulge, teil läheb see uuring pekki, kui te kiiresti ei hakka selle treeningandmestikuga tegelema. (K1: Andme- ja tehisarusuuna ekspert, Intervjuu 5, Asutus 5)

Mitmed intervjuud viitasid muuhulgas ka rahastusmehhanismidele, sh Euroopa Liidu struktuurifondide reeglitele, mis süvendavad seda mustrit, soosides küll teenuste (eriti arendusteenuste) sisseostmist väljast, kuid keelates samade rahade abil katta asutuse enda personalikulusid. Tulemuseks ongi olukord, kus raha on olemas tehniliste lahenduste, kuid mitte inimeste jaoks, kes peavad alusandmestiku ja selle kvaliteedi tagama.

5. JÄRELDUSED JA ARUTELU

Järgnevas peatükis koondan uurimistöö peamised järeldused ning ühtlasi seon olulisemad empiirilised leiud teoreetilises osas käsitletud nähtamatu töö, kriitiliste andmeuuringute ja tehisaruandmetöö käsitlustega. Arutelu keskne väide on, et Eesti avaliku sektori tehisaruprojektides on andmetöö strateegiline ja kriitiline tegevus, kuid organisatsiooniliselt jääb see sageli nähtamatuks, tööjõu ressursiplaanides ja arenduseelarves käsitlemata ja ka nõrgalt dokumenteerituks. Samal ajal aga eristub Eesti juhtum globaalsest mikrotöö ja kummitustöö mudelist, kuna andmetöö tugineb peamiselt kohalikele domeeniekspertidele.

Tehisaruandmetöö korraldus Eesti avaliku sektori tehisaruprojektides

Eesti avaliku sektori tehisaruprojektide esmatasandi andmetöö (andmete kogumine, puhastamine, märgendamine, tõlgendamine jms) korraldust iseloomustab pragmaatiline *ad hoc* lähenemine. Korralduslikult domineerib strateegia, kus tehniline teostus ostetakse sisse, kuid andmetöö sisu ja tähendusloome hoitakse majasiseselt. Viimasel on nii praktilised eelarvega seotud põhjused, aga samas ka sügavamad strateegilised ning eelkõige sisulise kontrolliga seotud tagamaad. Andmetööd ei teosta reeglina spetsiaalsed ja värskest värvatud andmetöötajad, vaid olemasolevad valdkonnaekspertid, kes teevad seda üldjuhul oma põhitöö kõrvalt lisaülesandena. See tagab andmete kõrge sisulise kvaliteedi ja potentsiaalse vastavuse kullastandardile, mida suudabki sageli defineerida vaid domeeniekspert. Samas aga tekitab see senise praktika näitel ressursikonflikte ja varjatud töökoormust. Andmetöö moodustab intervjueeritavate hinnangul 60–80% projekti mahust, kuid on rahaliselt sageli puudulikult eelarvestatud, kuna tugineb sageli just asutuse enda sisemisele ja näiliselt tasuta ressursile.

Ühtlas on tehisaruandmetööga seotud tööprotsessid algfaasis sageli mitteformaalsed ja eksperimentaalsed. Andmetöö juhised ja dokumentatsioon tekivad (kui üldse) töö käigus. Andmetöötajate koolitus on valdavalt praktiline (lennult õppimine, briifingud, praktiline ettenäitamine), keskendudes kohati ka töötajate mõttemalli kohandamisele masinõppe eripäradele vastavaks. Abivahenditena kasutatakse nii lihtsaid kontoritarkvarasid (Excel) kui ka spetsiifilisi tööriistu (nt Label Studio). Kvaliteedikontrollis tehakse ressursside nappuse tõttu järeleandmisi: näiteks topelt-valideerimist rakendatakse harva, usaldades pigem kogenud spetsialisti ainuisikulist otsust.

Uurimistöö tulemused toovad esile ka märkimisväärse ebakõla akadeemilistes käsitlustes pakutavate tehisaruandmetööde struktureeritud korraldusraamistike ja intervjuudes käsitletud projektide tegelikkuse vahel. Kui teooria – nii Muldoon jt (2024) kui ka Girard-Chanudet' (2025) raamistikud – esitleb andmetööde korraldust valdavalt lineaarse jadana (eesmärgid → andmekogumine → puhastamine → märgendamine → ...), siis intervjuueeritavad kirjeldasid reaalselt töövoogu kaootilise edasi-tagasi pörgatamise või töö käigus protsesside lennult leiutamisenä. See aga viitab, et praktikas ei järgi tehisaruandmetöö nii-öelda tehaseleini lineaarset loogikat, vaid toimib pigem iteratiivselt, kus etapid on hägused ja kattuvad. Kuigi eelmainitud autorid (Muldoon jt, 2024; Girard-Chanudet', 2025) on seda ka oma töödes välja toonud, on siiski tähelepanuväärne, et selline töövoo kaootilisus on sedavõrd radikaalne.

Nähtamatu, kuid kriitiline töö: andmetöö prioriteedi ja tegeliku korralduse vastuolu

Eesti avaliku sektori tehisaruprojekte juhtivad osapooled tajuvad esmatasandi andmetööd vastanduvast. Teoreetilisel ja ka strateegilisel tasandil mõistetakse ning tunnistatakse küll andmete keskset rolli tehisaru toimimises, kuid praktikas ja operatiivtasandil puuduvad sageli protsessid, eelarved ja rollijaotused, mis võimaldaksid seda prioriteeti realiseerida. Andmetöö on valdavalt projektipõhine, *ad hoc*, reaktiivne ja institutsionaalselt nähtamatu.

Samas on märgata ka nihkumist. Esimesed projektid ja kogemuslik õppimine on suurendanud teadlikkust, et ilma investeeringuteta andmehalduse vundamenti ei ole võimalik ehitada toimivaid ning jätkusuutlikke tehisarurakendusi. Andmetöö on vaikselt liikumas nähtamatust paratamatusest teadvustatud ja läbimõeldud tegevuseni, kuid see teekond on veel pooleli ning sõltub endiselt tugevalt üksikute võtmeisikute kompetentsist, motivatsioonist ja initsiatiivist.

Uurimistulemused haakuvad siinkohal ka Star ja Straussi (1999) ning ka Jarke ja Büchneri (2024) käsitlustega *nähtamatust tööst*, tehisaruandmetöö sealhulgas. Eesti avaliku sektori tehisaruprojektides avaldub andmetöö nähtamatus eeskätt institutsionaliseeritud andmetööprotsesside puudumise ning ressursiplaneerimise tasandil – kui tehniline arendus on reeglina formaliseeritud ja ka eelarvestatud, siis andmete kogumine, puhastamine ja märgendamine toimub sageli varjatud tööna muude tegevuste kõrvalt. Aga nagu ütlevad nii Star ja Strauss (1999) ja ka mõned intervjuueeritavad, siis mõned protsessid ongi vahepeal mõistlik hoida nähtamatuna, et tagada suurem paindlikkus ning ühtlasem liikumine. Seetõttu tasub siinkohal hoolikalt kaaluda nii võimalikke eeliseid kui ka kaasnevaid riske.

Ühtlasi toetavad uuringu tulemused ka Justeseni ja Plesneri (2024) väidet, et digiorganisatsioonides jääb suur osa andmetööst allapoole ametlikku, nii-öelda *päris töö* lävepakku. Tuleb aga arvestada, et selline korraldus võib muuta andmetöö haavatavaks – kui mõni vastava kogemusega spetsialist lahkub, kaob ka kogemus ja protsess. Mis aga veelgi olulisem, samal ajal jääb ka andmetöö tegelik maht ja keerukus juhtimistasandile sageli hoomamatuks, mis omakorda teeb aga keeruliseks uute projektide planeerimise – kirjutades sinna sageli juba eos sisse ressursiplaneerimise vead – ning sellega halvates ka jätkusuutliku personalipoliitika kujundamise.

Andmetöö poliitilise laetuse vähene teadvustamine

Kriitiliste andmeuuringute (CDS) traditsioon käsitleb andmeid sotsiaalsete artefaktidena, mitte neutraalsete faktidena. Kuigi CDS teooria rõhutab andmetöö poliitilist laetust, siis hinnates intervjuerijana viisi, kuidas intervjueritud praktikud käsitlesid intervjuudes eetika, kallutatuse, andmetööde kvaliteedi ning ka muude teemadega seotud küsimusi, tajuvad nad oma tegevust pigem pragmaatilise töö “ära tegemisena” kui et aktiivses “ühiskonnakujundamise poliitikas osalemisena”. Ühest küljest võib see viidata sellele, et avaliku sektori vastavate praktikute andmekirjaoskus (selle kriitilises tähenduses) on pigem väike. Teisalt aga haakub see kaudselt ka Jarke jt (2022) kirjeldatud tähendushorizontide (*horizons of sense*) lahknevuse temaatikaga: CDS-uurijate vaates andmete ja andmetöödega avalduv mitmekihiline võimusuhe ja andmete kaudu ühiskonna taastootmine on ühe keskmise tehisarupraktiku jaoks selgelt mitte-poliitiline praktiline probleemilahendus. Edasistes uuringutes tuleks aga sügavamalt analüüsida, millised tagajärjed on sellisel tähendushorizontide lahknevusel. Kui tehiarurakendused sõltuvad andmetest, ning andmed on sotsiaalsed artefaktid, mis peegeldavad nende loojate väärtusi ja valikuid, siis tuleb tegeleda täiendavalt sellega, et ka tehisaruprojektidega vahetult seotud võtmeisikud oleksid oma poliitiliselt laetud rollist ka päriselt teadlikud. Laias plaanis viitab see aga vajadusele süstemaatiliselt tõsta avaliku sektori infosüsteemidega seotud töötajate andmekirjaoskuse taset. Sellesama andmekirjaoskuse üks ülioluline komponent on ka kriitiliste andmeuuringute alustõdede tundmine.

Andmetöö kui referentstõe loomine: nähtavad eksperdid aga nähtamatu protsess

Uuringu tulemused asetuvad dialoogi ka Girard-Chanudet’ (2025) käsitlesega, mis laiendab Lowrence Lessigi (2000) kuulsat „kood on seadus“ (*code is law*) maksiimi tehisaruajastule sobivaks printsiibiks „referentstõde on seadus“ (*ground-truth is law*). Girard-Chanudet’ (2025)

väitel on masinõppesüsteemides kontseptuaalne töö ja reeglite loomine nihkunud koodikirjutamisest treeningandmete koostamisse ja märgendamisse. Kuigi ka Eesti avaliku sektori praktika kinnitab eelmainitud teoreetilist nihet, toob see aga lisaks esile ka muid spetsiifilisi organisatsioonilisi dünaamikaid. Kui Girard-Chanudet (2025) osutab, et kontseptuaalne töö muutub andmestikes “sageli nähtamatuks” ja “hajub paljude madalama astme toimijate vahel”, siis antud uuringu empiirika näitab pigem vastupidist, mis väljendub nii konkreetsete nähtavate domeenieksperptide kaasamises (ei saa öelda, et domeenieksperdid on madalama astme töötajad) kui ka vastava tööjõu kõrges kvalifikatsioonis. Mis aga puudutab “nähtamatust” ja “hajuvust”, kus eelmainitud autori mõtted haakuvad ka Tubaro ja Casilli (2019) ja Muldoon jt (2024) käsitlestega, mis keskenduvad tehisaruandmetöö puhul platvormitööle ja kummitustööle (*ghost work*), toob antud uuring esile Eesti avaliku sektori vastupidise eripära: Eesti avalik sektor ei kasuta anonüümset odavtööjõudu, vaid, nagu juba eelnevalt mainitud, eelkõige kõrgelt kvalifitseeritud domeenieksperpte (pigem James Muldooni tehisaruandmetööde hankemudeli tüüp F), kes on projektijuhtidele üldreeglina ka täpselt teada ning on ühtlasi sageli ka osa vahetust füüsilisest projektimeeskonnast.

Eesti avalikus sektoris ei delegeerita „tõe“ defineerimist anonüümsetele mikrotöötajatele kolmandates riikides, vaid seda teevad, nagu eelnevalt korduvalt mainitud, üldjuhul kõrgelt kvalifitseeritud kohalikud domeenieksperdid – vaatelejad, analüütikud, menetlejad, kaardistajad, arstid, arhivaarid aga ka klienditeendusspetsialistid. See küll kinnitab Girard-Chanudet’ (2025) väidet, et tehisaruandmetöö on kollektiivne konstruktsioon, kus teadmine formaliseeritakse samm-sammult, kuid Eesti kontekstis on see protsess, vähemalt seni, tugevalt tsentraliseeritud ja vastavate töötajate tuvastatavuse vaatest igati nähtava ekspertteadmise ümber. Tasub siinkohal aga nüüd kriitiliselt ja teraselt tähele panna – viimane ei tähenda seda, et töö, mida need *nähtavad* isikud andmetega teevad, tingimata alati *nähtav* on.

Uuringutulemused toetavad tugevalt ka Girard-Chanudet’ (2025) hoiatust, et kui andmestik on kord suletud, sulgub ka tehisarurakenduse „must kast“, peites endasse kõik valikud ja poliitilised otsused. Uuringus tuvastatud sagedane puudulik dokumentatsioon – näiteks olukord, kus logitakse küll *kes* tegi, aga näiteks mitte seda *miks* tehti – loob olukorra, kus „uus seadus“ (*ground truth is new law* ehk andmepõhise mudeli käitumisloogika) on küll ekspertide loodud, kuid tagantjärele raskesti auditeeritav. See aga tähendab, et kuigi „seadusandjaks“ (*ground truth* loojaks) on igati pädev ja heatahtlik ametnik, muutub “seadusloome” kui protsess (tehisaruandmetöö) ise läbipaistmatuks *ad hoc* tegevuseks.

Eesti eripära globaalse tehisaruandmetöö tööjaotuse taustal

Eesti avalik sektor hoiab andmetöö strateegiliselt majasisesena mitte ainult andmekaitse, hankereeglite ning kulude kokkuhoiu tõttu, vaid ka kontrolli säilitamiseks oma tuumkompetentsi ehk tähendusloome üle. Kui asetada Eesti avaliku sektori tehisaruandmetöö praktikad laiemasse arutellu mikrotöö, kummitustöö ja ka võimaliku (neo)andmekolonialismi teemadel, on eelmainitud sisemiste ressursside eelistamine oluline ja positiivne leid. Globaalses tehisarutööstuses on laialt levinud töökorraldus, kus andmete märgendamine, puhastamine, modereerimine ja verifitseerimine delegeeritakse madalama tööjõukuluga riikidesse, BPO-dele või mikrotööplatvormidele. Sellist tööd on aga kirjeldatud ka kui tehisaru tarneahela (aga ka laiemalt globaalse platvormimajanduse) varjatud infrastruktuuri: näiliselt automatiseeritud infosüsteemid toetuvad tegelikult suurele hulgale sageli madalalt tasustatud, spetsiifilise kvalifikatsioonita ja ebakindlates tingimustes töötavatele inimestele (Gray & Suri, 2019; Tubaro et al., 2020; Bates et al., 2022; Muldoon et al., 2023a).

Baasmudelid, varjatud tehisaruandmetöö ja digitaalse suveräänsuse risk

Geiger jt (2020: 1-2) sõnul käsitletakse treeningandmete loomist ja valideerimist akadeemilises kirjanduses vähe, sealjuures sageli kasutatakse treenimisel juba eelpuhastatud karbiandmestikke, jättes sellega sageli ka andmete sügavama päritolukihi varju. See haakub otseselt ka antud uuringu leidudega. Eesti avaliku sektori tehisarurakendused ei põhine (vähemalt seni) suurte baasmudelite (olgu selleks pildituvastus või keelemudelid) nullist ise väljatöötamisel, mistõttu ei vajata reeglina ka ülisuuri sisendandmestikke. Kuna üldjuhtudel täppis-töödeldakse (*fine-tuning*) juba olemasolevaid, kas juhtivate platvormide loodud tasulisi või siis muid vabavaralisi baasmudeleid, ning kuna andmevajadused ongi sellisel lähenemisel väiksemad, siis on ka andmetööde puhul omade jõududega piirdumine võimalik. Tõenäoliselt on seega madala kvalifikatsiooniga suuremahulised andmetööd juba eos suurte baasmudelite loomisel kolmandate osapoolte poolt ära tehtud, mis omakorda aga tõstatab vajaduse püstitada juba uued uurimisküsimused – näiteks milliseid teadmisi ja väärtusi ning millises ulatuses on neisse suurtesse baasmudelitesse (tõenäoliselt odava ja vähequalifitseeritud tööjõu poolt) andmetööna varasemalt sisse akumulieeritud –, mis aga antud uuringu skoopi enam ei mahu.

Eelmise lõigu viimase mõttelõnga võimestamiseks veel ka kõige viimane nope uuringu käigus läbiviidud intervjuudest – vähemalt kolm Eesti avaliku sektori asutust tunnistasid Hiina päritolu vabavaraliste baasmudelite kasutamist, põhjendades otsuseid nende kõrge kvaliteediga ning ka

asjaoluga, et RIA (Riigi Infosüsteemide Amet) pole nende kasutamist keelanud. Tasub siinkohal aga meelde tuletada teooria peatükis käsitletud Fu jt (2025) uuringut sügavalt autoritaarse Hiina *La Qi* praktikast, mis kirjeldab keerukat süsteemi, kus valitsemine on tehisaruandmetöösse sisse kirjutatud pideva häälestamise ja distsiplineerimise kaudu selliselt, et andmemärgendajate mikrotasandi otsused vastaksid riigi eelistatud moraalsetele standarditele, mida autorid Hiina puhul nimetavad „delegeeritud tsensuuriks“. Kas ja millises ulatuses võivad sellised vabavaralised mudelid, mille loomist on muuseas toetanud Hiina valitseva režiimi mõjutatud tehisaruandmetöö, kujutada ohtu Eesti digitaalsele suveräänsusele – st võimele sisuliselt kontrollida andmeid, infrastruktuuri, standardeid ja tehnoloogilisi sõltuvusi (Fratini, 2025) – ning sealt edasi ka laiemale julgeolekule, jääb aga edasiste uuringute lahata.

Eetika, dokumentatsioon ja vastutuse edasilükkamine

Kui tehisarurakenduste eetilist külge esile toov dokumenteerimisalane kirjandus (Miceli jt, 2021; Miceli jt, 2022a; Paullada jt, 202; Gebru jt, 2020; Mitchell jt, 2019) kutsub üles laiapidmisele eetilisele refleksioonile andmetike ja mudelite osas, on Eesti avaliku sektori praktikas eetika sageli taandatud eelkõige juriidiliseks vastavuskontrolliks seoses andmekaitse ja privaatsusega (GDPR). Keerulisemad küsimused, nagu andmete sotsioloogiline kallutatus (*bias*), pole üldjuhul päevakorda kerkinud (siinkohal tuleb teha ka mööndus, et mõnede intervjuude käigus käsitletud projektide iseloomust sõltuvalt *per se* kallutatuse küsimus objektiivselt võttes ongi teisejärguline) ning kui ongi, siis lahendatakse neid seni pragmaatiliselt “inimene ahelas” (*human-in-the-loop*) meetodil või siis lükatakse edasi poliitilise otsustusprotsessi ootele. Kes aga selle “poliitilise ja väärtusepõhise” otsustamise eest vastutab ja kes ning kuna need teemad (sh otsustamise) ette võtab, jääb aga esialgu intervjuudest selgusetuks (ühtlasi leian, et ka see teema vajaks tulevikus eraldi akadeemilist uuringut). Arvestades aga tehisaruvaldkonna uudsust, on selline lähenemine tõenäoliselt ka ratsionaalne – seni, kuni projektid on asutuste vaatest piloteerivad ja väiksemahulised, ei ole ressursiplaneerimises (mis on üldjuhul isegi puudulik, nagu ka eelnevas arutelus juba välja toodud) peetud proportsionaalseks eetilist käsitlust toetava dokumentatsiooni süsteemset loomist ning ka juurutamist. Tõenäoliselt jõutakse nende teemadeni siis, kui projektid ja korraldus arenevad, mahud kasvavad ning meeskonnad laienevad.

Tehisaru eetika-, dokumentatsiooni- ja juhtimisküsimustes on viimastel aastatel sagenenud üleskutsed suuremaks koostööks, standardiseerimiseks ja institutsionaliseerimiseks. Ka juhtivad platvormid on vastavaid üleskutseid üha rohkem tegemas – nii OpenAI poliitikadokumentides kui ka Dario Amodei (Anthropic) esseedes rõhutatakse vajadust luua mehhanisme, mis

võimaldaksid tehisaru arendust paremini auditeerida, dokumenteerida ja ühiskondlikult kontrollida (Amodei, 2026; OpenAI, 2026). Ühtlasi kutsuvad eelmainitud platvormid üha sagedamini üles ka laiapõhjalisemale dialoogile ühiskondliku lepingu potentsiaalseks übermõtestamiseks, arvestades eesootavat võimalikku tööturutransformatsiooni. Kuigi eelmainitud platvormidele saab nii mõndagi ette heita ja seda üldjuhtudel õigustatult, siis üha enam näib, et sealsed üleskutsed dialoogiks on seni saanud passiivset tähelepanu nii kodanike ja nende ühenduste, akadeemia, meedia kui ka poliitikakujundajate suunalt. Kaerajaani tantsimiseks on ka partnereid vaja ning vähemalt seni pole võrdväärseid, entusiastlikke ja häälekaid tantsup partnereid veel lavale astunud (kui siis ehk EL ja AI Act mingil määral).

Tehisarurakendused ning juhtimis- ja institutsiooniline risk

On oluline teadvustada, et tehisarurakendustestega (ja ka sealse tehisaruandmetööga) seotud riskid ei tulene ainult suurte rahvusvaheliste tehnoloogiaettevõtete kasumihuvidest või võimuihast, vaid ka institutsionaalsest ja poliitilisest survest võtta tehisarulahendusi kiiresti kasutusele enne, kui nende aluseks olev andmetöö, vastutusjaotus ja auditeeritavus on piisavalt läbi mõeldud. Eriti just tehisaruandmetööde puhul võib selline tempo hakata ületama asutuste suutlikkust mõista andmetega seotud klassifikatsioone, väärtusvalikuid ja referentstõdesid, mida andmetöö kaudu süsteemidesse sisse kirjutatakse.

Seega, kui me räägime tehisarurakendustest ja nendega seotud riskidest ja võimalustest, siis on üha tõenäolisem, et platvormide asemel (nagu näib olevat senine juhtiv narratiiv) muutub peamiseks riskide loojaks ja eirajaks hoopis poliitiline tasand, püüdes haarata valijate hääli sellega, et pakub erinevate tehisarulahenduste abil oma suurtele lubadustele katet ja sisu. Lõpetuseks toon siinkohal esile ka hiljutise karmi hoiatuse TTÜ rakendusliku tehisaru professorilt Tanel Tammetilt (Maidla, 2026b): järgmine suur usalduse kadu ei pruugi tulla mitte süvavõltsingute tulvast, vaid mõnest ebaõnnestunud tehisaru integratsioonist riigiprotsessidesse – kui must kast hakkab valede eelduste alusel inimesi sortima ja karistama.

Uuringu puudused ja piirangud

Üheks uuringu peamiseks piiranguks on strateegiline valim ja sellega seotud piiratud üldistusvõime kogu sektorile. Kuigi 14 intervjuud ja 12 asutust hõlmasid olulisi osapooli, jäid siiski kõrvale mitmed suured ja infotehnoloogia teemadele lähedal asuvad avaliku sektori organisatsioonid, millel on oluline roll ja pädevus Eesti avaliku sektori tehisarurakenduste arendamisel. Teiseks, uurimismeetodina ei pruugi intervjuu vorm anda täielikku pilti – see, mida

räägitakse, ei pruugi alati kattuda sellega, mida tegelikult tehakse. Siinkohal leian, et etnograafiline lähenemine koos vaatluse ning tegelike tööprotsesside ja tulemuste analüüsimisega võiks anda tõepärasema ülevaate kui pelgalt küsimine. Ideaalis võiks meetodeid kombineerida. Kolmandaks, kuna tehisaruvaldkond areneb väga kiiresti ning mitmed intervjuudes käsitletud projektidest ja teemadest olid kas piloodid või katsetused, võivad avaliku sektori arenduspraktikad, sealjuures tehsiaruandmetöölased praktikad, juba lähitulevikus märkimisväärselt muutuda. Seetõttu võib osa töös käsitletust kiiresti aeguda. Neljanda puudusena võib veel lisada ka selle, et kuna analüüsi ei valideeritud teise kodeerija ning uurijatevahelise arutelu kaudu, võib see mõjutada tulemuste rõhuasetusi ja narratiivi kujunemist.

KOKKUVÕTE

Uurimistöö käsitles tehisarurakenduste arendust toetava tehisaruandmetöö (andmete kogumine, puhastamine, märgendamine ja valideerimine) korraldust ja nähtavust Eesti avalikus sektoris. Tehisarurakenduste edukus sõltub otseselt sisendandmetest, kuid nende ettevalmistamine on sageli mahukas käsitsi tehtav tööprotsess, mis moodustab kuni 80% tehisaruprojekti töötundidest. Uurimistöö probleemipüstitus tugines peamiselt seisukohale, et vaatamata andmetöö mahule ja olulisusele on see sageli institutsionaalselt nähtamatu, alahinnatud ning jääb arendusprotsessis tehnoloogiliste lahenduste varju.

Töö teoreetilise raamistiku lähtepunkt asus kriitiliste andmeuuringute traditsioonis ning rõhutas, et esmatasandi andmetöö, sealhulgas tehisaruandmetöö kaudu kujundatakse referentstõde ja selle kaudu ka väärtusi ning valikuid, mis juhivad hilisemaid algoritmilisi ja ka algoritmide poolt mõjutatud inimeste otsuseid. Töö empiiriline osa tugines 14 poolstruktureeritud intervjuule Eesti avaliku sektori tehisaruprojektide juhtide ja ekspertidega.

Uuringu tulemused näitasid, et Eesti avalikus sektoris domineerib tehisaruandmetööde puhul hübriidne hankemudel: kui tehniline arendus koos mudelite loomisega ostetakse sisse, siis andmetööd ja sellega koos ka tähendusloomet hoitakse nii operatiivsetel kui ka strateegilistel kaalutlustel majasiseselt. Erinevalt rahvusvahelistest trendidest ei kasuta Eesti avalik sektor tehisarurakendusi toetavatel andmetöödel anonüümset platvormitööjõudu, vaid üldjuhul kaasatakse kõrgelt kvalifitseeritud majasiseseid domeenieksperthe (nt arhiivispetsialistid, vaatlejad, menetlejad), kelle jaoks aga lisanduv andmetöö on sageli senise põhitöö kõrvalt tehtav uus ja nähtamatu lisakohustus. Uuring kinnitab lõhet andmetööde olulisuse ja nähtavuse vahel: tehisaruandmetöö on ajaliselt mahukaim, kuid nii eelarvelises kui ka laiemas planeerimisprotsessis selgelt alahinnatud etapp, mis tugineb valdavalt *ad hoc* töökorraldusele ning sageli üksikute spetsialistide ning entusiastide initsiatiivile.

Teoreetilise analüüsi ja empiiriliste uuringutulemuste sünteesile tuginedes esitan avaliku sektori otsustajatele neli peamist soovitusi:

1. Tehisaruandmetööd ei tohiks käsitleda nähtamatu taustatööna, vaid see tuleb ametlikult siduda arendusprotsesside, eelarvete, koolitusplaanide ja hankestrateegiatega.

2. Kuna tehisaruandmetöö nõuab kvalifitseeritud majasiseste valdkonnaspetsialistide panust, tuleb nende aeg planeerida ning seda tööd ei tohi käsitleda põhitöö kõrvalt täidetava madalama prioriteediga lisakohustusena.
3. Fookus tuleb laiendada tehnoloogialt andmekvaliteedile – kvaliteetne treeningandmestik ja referentstõde peab olema asutuse strateegiline vara ning andmetöö otsused peavad olema dokumenteeritud, läbipaistvad ja auditeeritavad.
4. Tehisaruandmetöö tegijaid ja nende juhte tuleb koolitada, et nad mõistaksid, kuidas nende töö mõjutab tehisarurakenduste väljundeid ning kuidas andmetöö käigus tehtavate otsuste kaudu osaletakse laiemas ühiskonna kujundamises – nii keele, normide, väärtuste kui ka võimusuhte tasandil.

Kui neile teemadele süsteemselt tähelepanu ei pöörata, jääb uute tehisarurakenduste arendus paratamatult sõltuma üksikute spetsialistide teadmistest, initsiatiivist ja sageli ka kriisipõhisest reageerimisest. See omakorda suurendab riski, et andmekvaliteedi, eetika ja vastutusega seotud probleemid jäävad piisavalt varakult tuvastamata või lahendamata. Tehisaruandmetöö parandamine on peamiselt juhtide ja juhtimistasandi ülesanne, sest probleem ei seisne üksnes spetsialistide töövõttes, vaid selles, kuidas asutuses rolle, ressursse, vastutust ja tööprotsesse tervikuna kujundatakse ning juhitakse.

Edasiste uurimissuundadena (lisaks neile kolmele, mida juba eelnevalt olen põgusalt käsitlenud) pakun välja tehisaruandmetöö mõju analüüsi nii avaliku kui ka erasektori ekspertide professionaalsele identiteedile. Samuti võiks ette võtta uuringu, mis püüaksid siduda tehisaruprojektides andmetöödesse panustatud ressursid projektide edukusega. Lisaks tasuks uurida – näiteks juhtumiuuringute või miks mitte ka simulatsioonide või eksperimentide abil – ning võimalusel ka kvantitatiivselt mõõta tehisaruandmetöö, sealhulgas märgendusotsuste, mõju praktikas juba kasutusel olevate tundlike ja kriitiliste tehisarurakenduste väljunditele.

SUMMARY

The organisation and visibility of foundational data work supporting AI application development in the Estonian public sector

The study examined the organisation and visibility of AI data work supporting the development of artificial intelligence applications in the Estonian public sector. AI data work refers here to data collection, cleaning, labelling and validation. The success of AI applications depends directly on input data, yet preparing such data is often an extensive manual work process that accounts for up to 80% of the working hours in an AI project. The problem statement of the study was based mainly on the position that, despite the scale and importance of data work, it often remains institutionally invisible, undervalued and overshadowed by technological solutions in the development process.

The starting point of the theoretical framework of the thesis was located in the tradition of critical data studies and emphasised that first-level data work, including AI data work, shapes referential truth and, through it, also the values and choices that guide later algorithmic decisions as well as human decisions influenced by algorithms. The empirical part of the thesis was based on 14 semi-structured interviews with managers and experts involved in AI projects in the Estonian public sector.

The findings showed that a hybrid procurement model dominates AI data work in the Estonian public sector: while technical development, including model creation, is outsourced, data work and the accompanying production of meaning are kept in-house for both operational and strategic reasons. Unlike international trends, the Estonian public sector does not use anonymous platform labour for data work supporting AI applications; instead, highly qualified in-house domain experts are generally involved, such as archive specialists, observers and case handlers. For them, however, the additional data work is often a new and invisible extra obligation performed alongside their existing main duties. The study confirms a gap between the importance and visibility of data work: AI data work is the most time-consuming stage, yet it is clearly underestimated in both budgeting and broader planning processes, relying mainly on ad hoc work organisation and often on the initiative of individual specialists and enthusiasts.

Based on the synthesis of the theoretical analysis and empirical research findings, I present four main recommendations for public sector decision-makers:

1. AI data work should not be treated as invisible background work, but must be formally integrated into development processes, budgets, training plans and procurement strategies.
2. Since AI data work requires the contribution of qualified in-house domain specialists, their time must be planned, and this work must not be treated as a lower-priority additional obligation carried out alongside their main duties.
3. The focus must be broadened from technology to data quality: a high-quality training dataset and referential truth must be treated as strategic assets of the organisation, and decisions related to data work must be documented, transparent and auditable.
4. Those carrying out AI data work, as well as their managers, must be trained so that they understand how their work affects the outputs of AI applications and how, through the decisions made during data work, they participate in the broader shaping of society — at the level of language, norms, values and power relations.

If these issues are not addressed systematically, the development of new AI applications will inevitably remain dependent on the knowledge, initiative and often also crisis-based responses of individual specialists. This, in turn, increases the risk that problems related to data quality, ethics and responsibility will not be identified or resolved early enough. Improving AI data work is primarily a task for managers and the management level, because the problem does not lie only in the working practices of specialists, but in how roles, resources, responsibilities and work processes are shaped and managed as a whole within the organisation.

As future research directions, in addition to the three that I have already briefly discussed earlier, I propose analysing the impact of AI data work on the professional identity of experts in both the public and private sectors. It would also be useful to conduct a study that would seek to connect the resources invested in data work in AI projects with project success. In addition, it would be worthwhile to examine — for example through case studies or, why not, through simulations or experiments — and, where possible, also to quantitatively measure the impact of AI data work, including labelling decisions, on the outputs of sensitive and critical AI applications already in practical use.

KASUTATUD KIRJANDUS

1. AlgorithmWatch. (2020). Estonia. In Automating society report 2020.
<https://automatingsociety.algorithmwatch.org/report2020/estonia/> Kasutatud 15.05.2026
2. Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., ja Zimmermann, T. (2019). Software engineering for machine learning: A case study. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE.
<https://doi.org/10.1109/ICSE-SEIP.2019.00042>
3. Amodei, D. (2026). The adolescence of technology: Confronting and overcoming the risks of powerful AI. Anthropic.
<https://www.darioamodei.com/essay/the-adolescence-of-technology> Kasutatud 19.05.2026
4. Bates, J., Checco, A. & Gerakopoulou, E. (2022). Worker Perspectives on Designs for a Crowdwork Co-operative. In A. Hepp, J. Jarke & L. Kramp (toim.), *New Perspectives in Critical Data Studies: The Ambivalences of Data Power*, 415–443. Palgrave Macmillan
5. Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., ja Bao, M. (2022). The values encoded in machine learning research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533083>
6. Choroszewicz, M., & Rannisto, A. (2026). AI innovation at the boundaries: Justifying a generative AI decision support tool. *Big Data & Society*, 13(1).
<https://doi.org/10.1177/20539517261424159>
7. D'Ignazio, Catherine, & Klein, Lauren F. (2024). Data Feminism for AI. FAccT '24: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 100 – 112. <https://dl.acm.org/doi/10.1145/3630106.3658543>
8. Eleftheriou, I., Embury, S. M., ja Brass, A. (2016). Data journey modelling: Predicting risk for IT developments. In J. Horkoff, M. A. Jeusfeld, ja A. Persson (Eds.), *The practice of enterprise modeling: 9th IFIP WG 8.1 Working Conference, PoEM 2016*, Skövde, Sweden, November 8–10, 2016. Springer. https://doi.org/10.1007/978-3-319-48393-1_6

9. European Parliament. (2025). EU AI Act: First regulation on artificial intelligence.
<https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
10. Future of Life Institute. (2024). High-level summary of the AI Act. Artificial Intelligence Act. <https://artificialintelligenceact.eu/high-level-summary>
11. Flensburg, S., ja Lomborg, S. (2023). Datafication research: Mapping the field for a future agenda. *New Media ja Society*, 25(6), 1451–1469.
<https://doi.org/10.1177/14614448211046616>
12. Fratini, S. (2025). The sociotechnical politics of digital sovereignty: Frictional infrastructures and the alignment of privacy and geopolitics. *Big Data & Society*, 12(4).
<https://doi.org/10.1177/20539517251400729>
13. Fu, P., Lin, Z., ja Wang, W. Y. (2025). Operationalizing AI governance: Data annotation, la qi and manual alignment in China. *Information, Communication ja Society*, 1–23.
<https://doi.org/10.1080/1369118X.2025.2526162>
14. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., ja Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
15. Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., ja Huang, J. (2020). Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery. <https://doi.org/10.1145/3351095.3372862>
16. Girard-Chanudet, C. (2025). Ground-truth is law: The invisible conceptual work behind AI. *Big Data ja Society*, 12(2), 1–14. <https://doi.org/10.1177/20539517251352823>
17. Gray, M. L., & Suri, S. (2019). Ghost work: How to stop Silicon Valley from building a new global underclass. Houghton Mifflin Harcourt.
18. Hanassab, S., Abbara, A., Yeung, A. C., Voliotis, M., Tsaneva-Atanasova, K., Kelsey, T. W., Trew, G. H., Nelson, S. M., Heinis, T., ja Dhillon, W. S. (2024). The prospect of

- artificial intelligence to personalize assisted reproductive technology. *Digital Medicine*, 7(1), Article 55. <https://doi.org/10.1038/s41746-024-01006-x>
19. Hepp, A., Jarke, J., ja Kramp, L. (Eds.). (2022). *New perspectives in critical data studies: The ambivalences of data power*. Palgrave Macmillan.
<https://doi.org/10.1007/978-3-030-96180-0>
 20. Iliadis, A., ja Russo, F. (2016). Critical data studies: An introduction. *Big Data ja Society*, 3(2). <https://doi.org/10.1177/2053951716674238>
 21. Jarke, J., ja Büchner, S. (2024). Who cares about data? Data care arrangements in everyday organisational practice. *Information, Communication ja Society*. Advance online publication. <https://doi.org/10.1080/1369118X.2024.2320917>
 22. Jarke, J., Zakharova, I., ja Breiter, A. (2022). Organisational data work and its horizons of sense: On the importance of considering the temporalities and topologies of data movement when researching digital transformation(s). *Historical Social Research/Historische Sozialforschung*, 47(3), 142–170.
<https://doi.org/10.12759/hsr.47.2022.29>
 23. Justesen, L., ja Plesner, U. (2024). Invisible digi-work: Compensating, connecting, and cleaning in digitalized organizations. *Organization Theory*, 5(1).
<https://doi.org/10.1177/26317877241235938>
 24. Justiits- ja Digiministeerium. (2025a). Kratid. <https://kratid.ee/> Kasutatud 15.04.2026
 25. Justiits- ja Digiministeerium. (2025b). Eesti andmete tegevuskava 2025–2026. Eesti Vabariigi Valitsus. Kasutatud 15.04.2026
 26. Kazimzade, G., ja Miceli, M. (2020). Biased priorities, biased outcomes: Three recommendations for ethics-oriented data annotation practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '20)*. Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375809>
 27. Langley, A., Lindberg, K., Mørk, B. E., Nicolini, D., Raviola, E., & Walter, L. (2019). Boundary work among groups, occupations, and organizations: From cartography to process. *Academy of Management Annals*, 13(2), 704–736.
<https://doi.org/10.5465/annals.2017.0089>

28. Lessig, L. (2000). Code is law. On liberty in cyberspace. *Harvard Magazine*.
<https://harvardmagazine.com/2000/01/code-is-law-html> *Kasutatud* 27.05.2026
29. Lepik, K., Harro-Loit, H., Kello, K., Linno, M., Selg, M., ja Strömpl, J. (2014). Intervjuu. Sotsiaalse analüüsi meetodite ja metodoloogia õpibaas, Tartu Ülikool.
<https://samm.ut.ee/intervjuu/> *Kasutatud* 15.04.2026
30. Li, Z. (2024). When being a data annotator was not yet a job: The laboratory origins of dispersible labor in computer vision research. *Socius: Sociological Research for a Dynamic World*, 10, 1–16. <https://doi.org/10.1177/23780231241259617>
31. Loukissas, Y. A. (2019). *All data are local: Thinking critically in a data-driven society*. MIT Press.
32. Maidla, M. (2026a). Teadus, kapital ja ettevõtlus tuleb siduda üheks strateegiaks. *Sirp*.
<https://www.sirp.ee/teadus-kapital-ja-ettevotlus-tuleb-siduda-uheks-strateegiaks/>
Kasutatud 19.05.2026
33. Maidla, M. (2026b). Tehisarul pole käsiraamatut. *Sirp*.
<https://www.sirp.ee/tehisarul-pole-kasiraamatut/> *Kasutatud* 19.05.2026
34. Marcus, G. (2018). Deep learning: A critical appraisal. arXiv.
<https://arxiv.org/abs/1801.00631> *Kasutatud* 16.04.2026
35. Massey, J., Worth, S., ja Simperl, E. (2025). Report: Mapping the role of data work in AI supply chains. Open Data Institute.
<https://theodi.org/insights/reports/mapping-the-role-of-data-work-in-ai-supply-chains/>
Kasutatud 15.04.2026
36. Miceli, M., Posada, J., ja Yang, T. (2022a). Studying up machine learning data: Why talk about bias when we mean power? *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2), Article 34. <https://doi.org/10.1145/3492853>
37. Miceli, M., Yang, T., Alvarado Garcia, A., Posada, J., Wang, S. M., Pohl, M., ja Hanna, A. (2022b). Documenting data production processes: A participatory approach for data work. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2).
<https://doi.org/10.1145/3555623>

38. Miceli, M., Yang, T., Naudts, L., Schuessler, M., Serbanescu, D., ja Hanna, A. (2021). Documenting computer vision datasets: An invitation to reflexive data practices. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2021)*. Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445880>
39. Mirtaheri, S. L. ja Shahbazian, R. (2022). *Machine learning: Theory to applications*. CRC Press. <https://doi.org/10.1016/j.giq.2021.101596>
40. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., ja Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT '19)*. ACM. <https://doi.org/10.1145/3287560.3287596>
41. Mikalef, P., ja Gupta, M. (2021). Artificial Intelligence Capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm performance. *Information ja Management*. <https://doi.org/10.1016/j.im.2021.103434>
42. Moe, C. E., Newman, M., ja Sein, M. K. (2017). The public procurement of information systems: Dialectics in requirements specification. *European Journal of Information Systems*, 26(2), 143–163. <https://doi.org/10.1057/s41303-017-0035-4>
43. Møller, N. H., Bossen, C., Pine, K. H., Nielsen, T. R., ja Neff, G. (2020). Who does the work of data? *Interactions*, 27(3), 52–55. <https://doi.org/10.1145/3386389>
44. Muldoon, J., Cant, C., Wu, B., ja Graham, M. (2024). A typology of artificial intelligence data work. *Big Data ja Society*, 11(1). <https://doi.org/10.1177/20539517241232632>
45. Muldoon, J., Cant, C., Graham, M., ja Spilda, F.U. (2023). The poverty of ethical AI: Impact sourcing and AI supply chains. *AI ja Society*, 2(1). <https://doi.org/10.1007/s00146-023-01824-9>
46. Muldoon, J., ja Wu, B. A. (2023). Artificial intelligence in the colonial matrix of power. *Philosophy ja Technology*, 36(4), 1–24. <https://doi.org/10.1007/s13347-023-00687-8>
47. OpenAI. (2026). Industrial policy for the intelligent age: Ideas to keep people first. OpenAI.

- <https://cdn.openai.com/pdf/561e7512-253e-424b-9734-ef4098440601/Industrial%20Policy%20for%20the%20Intelligence%20Age.pdf> Kasutatud 19.05.2026
48. OpenAI. (2026). *Where the goblins came from*. OpenAI.
<https://openai.com/index/where-the-goblins-came-from/> Kasutatud 19.05.2026
49. Open Data Institute. (2024). *The AI data transparency index*.
<https://theodi.org/insights/reports/the-ai-data-transparency-index/> Kasutatud 15.04.2026
50. Pasquale, F. (2016). *Black box society: The secret algorithms that control money and information*. Harvard University Press.
51. Paullada, A., Raji, I. D., Bender, E. M., Denton, E., ja Hanna, A. (2021). Data and its (dis)contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(12), 100336. <https://doi.org/10.1016/j.patter.2021.100336>
52. Perrigo B. (2023). Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic. *Time Magazine*. Kasutatud 18.06.2026
53. Richterich, A. (2018). *The big data agenda: Data ethics and critical data studies*. University of Westminster Press.
54. Sambasivan, N. (2022). All equation, no human: The myopia of AI models. *Interactions*, 29(2), 78–80. <https://doi.org/10.1145/3516515>
55. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., ja Aroyo, L. M. (2021). “Everyone wants to do the model work, not the data work”: Data cascades in high-stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. Association for Computing Machinery.
<https://doi.org/10.1145/3411764.3445518>
56. Singapore Government. (2023). *Public sector AI playbook*.
<https://www.developer.tech.gov.sg/products/collections/data-science-and-artificial-intelligence/playbooks/public-sector-ai-playbook.pdf> Kasutatud 15.04.2026
57. Star, S. L., ja Strauss, A. (1999). Layers of silence, arenas of voice: The ecology of visible and invisible work. *Computer Supported Cooperative Work (CSCW)*, 8(1-2), 9–30. <https://doi.org/10.1023/A:1008651105359>

58. Tangi, L., van Noordt, C., Combetto, M., Gattwinkel, D., ja Pignatelli, F. (2022). AI watch: European landscape on the use of artificial intelligence by the public sector (EUR 31088 EN). *Publications Office of the European Union*. <https://doi.org/10.2760/39336>
59. Tubaro, P., ja Casilli, A. A. (2019). Micro-work, artificial intelligence and the automotive industry. *Economia E Politica Industriale*, 46(3), 333–345.
<https://doi.org/10.1007/s40812-019-00121-1>
60. Tubaro, P., Casilli, A. A., ja Coville, M. (2020). The trainer, the verifier, the imitator: Three ways in which human platform workers support artificial intelligence. *Big Data ja Society*, 7(1). <https://doi.org/10.1177/2053951720919776>
61. Wang, D., Prabhat, S., ja Sambasivan, N. (2022). Whose AI dream? In search of the aspiration in data annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM. <https://doi.org/10.1145/3491102.3502121>
62. Wu, T., Muldoon, J., ja Xia, B. (2025). Global data empires: Analysing artificial intelligence data annotation in China and the USA. *Big Data ja Society*, 12(2), 1–13.
<https://doi.org/10.1177/20539517251340600>
63. Zhou, Z. H. (2021). *Machine learning*. Springer Nature.
64. Zuiderwijk, A., Chen, Y.-C., ja Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Government Information Quarterly*, 38(3), Article 101577.
<https://doi.org/10.1016/j.giq.2021.101577>

LISAD

Lisa 1: Intervjuukava

FORMAALSUSED (5 min)	
○ Sissejuhatavad tänusõnad ○ Uuringu eesmärk ○ Keskse mõiste tutvustus: tehisaru ja esmatasandi andmetöö ○ Intervjuu ülesehitus: teemablokid, küsimused ja vastused ○ Intervjuu salvestamisest, töötlustest ning pseudonüümimisest ○ Kirjalik nõusolek andmete kasutamiseks uurimistöös ○ Intervjueeritava mured ja küsimused	
SISSEJUHATUS: Intervjueeritava taust (5 min)	
Temaatika	Küsimus
Ametikoht ja vastutusvaldkonnad	1. Mis on Teie ametikoht organisatsioonis ja millised on Teie peamised vastutusalad ja tööülesanded?
Tehisaruprojektide kokkupuude	2. Kirjeldage lühidalt 1-2 tehisaru komponenti sisaldanud projekti, milles olete osalenud. ○ Millist tehisarutehnoloogiat kasutati (nt tekstianalüüs, klassifitseerimine, vestlusrobot)? ○ Hoiame ülejäänud intervjuu fookust peamiselt neil projektidel!
Roll tehisaruprojektides	3. Mis oli Teie roll ja ülesanded nendes projektides? ○ Kuidas need kattusid või erinevad Teie praegusest ametikohast?
Tehisaruprojektide osakaal töös	4. Kui suur osa teie tööajast viimase kahe aasta jooksul on kulunud tehisaruprojektidele?

TEEMAPLOKK 1: RQ1 Tehisaruandmetööde väärtustaju, nähtavus ja prioriteet (20 min)

Andmetöö juhtimine organisatsioonis	1. Millisel viisil on tehisaruandmetöö teie organisatsioonis korraldatud? <i>On teil mingisugused väljakujunenud protsessid või rollid?</i>
Andmetöö väärtus ja eetikadiskussioon	2. Milliseid eetilisi küsimusi olete tehisaruprojektide kontekstis arutanud ning miks on just need teemad teie hinnangul olulised? ○ Kui sellised arutelud tekivad, siis kes on see n-ö tiim kellega seda arutatakse? <i>On teil nt andmekaitse spetsialist, keda siis kaasatakse või on muul moel see korraldatud?</i> ○ Mis teemadel on näiteks tiimis olnud erimeelsusi või erinevaid arusaamu? Ja mida sel puhul tehtud on või kuidas need lahenenud on?
Andmetöö etappide tähelepanu tehisaruandmetöö töövoos	3. Kuidas teie hinnangul jaotuvad tähelepanu ja ressursid erinevate tehisaruandmetöö etappide vahel? <i>Näiteks: kas andmete kogumisele pööratakse rohkem tähelepanu kui märgendamisele, või vastupidi?</i> ○ Meenutage mõnda hetke, kus tähtajad surusid peale. Kas ja milliseid järeleandmisi tuli teha andmetöö arvelt, et mudel õigeaks ajaks valmis saada? Mis oli see koht, kust 'lõigati'? ○ Kas on olnud tegevusi või ressursikohti, mis on tekitanud probleeme – näiteks sobiva inimese puudus, ajasurve või muu takistus? Ja milliseid konkreetsemalt?
Andmetöö etappide tähelepanu tehisarurakenduste arendusprotsesside töövoos	4. Kuidas kirjeldaksite andmetöö nähtavust ja prioriteeti teie projektides võrreldes mudeliarenduse või arhitektuuriliste otsustega? ○ <i>Kas andmetöö roll ilmnes selgelt või pigem vaikselt taustal?</i>
Andmetöö eelarve kujunemine	5. Kuidas on andmetöö teie projektides eelarve ja ajaplaani mõttes planeeritud? ○ <i>Kas andmetöö on projektides eraldi arvestatud tegevus (oma mahu ja ressursiga) või kipub see pigem kujunema jooksvalt koos muude töödega?</i>
Andmetöö nähtavus probleemina	6. Kuidas teie organisatsioonis ülduliselt mõistetakse ja tajutakse tehisaruandmetööd?

	Kui nähtav või nähtamatu see töö teistele projektiosalistele ja töötajatele paistab?
TEEMAPLOKK 2: RQ2 Tehisaruandmetöö vahetu korraldus (25 min)	
Andmetöö hankemudel	1. Mõelge projektidele, mida alguses mainisite. Kuidas seal andmetööd hangiti? Omad sisemised ressursid, väline teenusepakkuja või hübriidselt?
Hankemudeli valiku põhjendused	2. Millistel kaalutlustel te selle valiku tegite? Kui vastus on napp, siis toetada näidetega nagu töömaht, oskused, aeg, turvalisus, raha. Kui otsustasite andmetöö ise teha / sisse osta, siis mis oli selle otsuse suurim üllatus tagantjärele? Kas see osutus arvatust kallimaks, aeglasemaks või keerulisemaks?
Andmetöö mahu tajumine	3. Kui suur osa projekti ressursist läks teie hinnangul andmetööle? Olete ehk kuulnud väidet, et kuni 80% kesmise tehisaruprojekti ajast kulub andmetöödele. Kas see vastab teie kogemusele?
Andmetöö protsess ja juhised	4. Teie osalusega tehisaruprojektides: - Kuidas andmetöö protsess planeeriti? Kas oli selged sammud ja rollid, või kujunes see käigult? - Kuidas otsustati, milliseid andmeid vaja läheb ja kust need tulevad? Kes otsustas? - Kui põhjalikud oli kirjalikud juhised kogujatele, puhastajatele ja märgendajatele? Kes need koostas? - Kuidas märgendamine toimus - käsitsi, automaatselt või kombineeritult? - Kui oli mitu märgendajat, kuidas erimeelsusi lahendati? - Kui palju andmeid läbi töötati ja milliseid tööriistu kasutati? - Kuidas andmete kvaliteeti tagati? Kes selle eest vastutas? - Tehisaru vajab õppimiseks nn kuldstandardit. Kuidas te defineerisite, mis on 'õige' vastus? Kes oli see inimene või ekspert, kelle sõna jäi peale, kui tekkis vaidlus, kuidas mingit andmetükki märgendada?
Andmetöö dokumenteerimine	5. Kuidas on teie organisatsioonis korraldatud tehisaruandmetöö dokumenteerimine? Milliseid põhimõtteid või tavasid selleks kasutatakse? Mida täpselt dokumenteeriti?

Andmetöötajate taust ja motivatsioon	6. Kelle lauale tehisaruandmetöö teie projektides langes? Kas seda tegid konkreetse rolli või ametikohaga töötajad? ○ Milline oli nende eelnev kogemus? Kuidas neid koolitati? ○ Kuidas töökoormust hallati - kas põhitöö kõrvalt või eraldi aeg? ○ Kuidas oli korraldatud tagasiside andmine? ○ Kuidas teile tundub teie organisatsiooni või selle näite varal - kuivõrd osatakse andmetöö tegijate panust hinnata? Kuidas seda teie organisatsioonis tehakse? Kuivõrd erineb see sellest, kuidas hinnatakse nt mudeli ehitajaid või projektijuhte?
Andmetöö peamised õppetunnid	7. Lõpetuseks nüüd pisut tagasi vaadates – Kas teeksite sama projekti uuesti alustades andmetöö korralduses midagi teisiti? ○ Kas on midagi, mida ma siiani küsimata jätsin, kuid mis teie meelest oleks tehisaruandmetööga seotud projektide puhul teistele kindlasti väärt jagada?
Uute intervjuueeritavate kontaktid	Kas oskate soovitada kedagi oma asutusest või mujalt avalikust sektorist, kellega tasuks sel teemal veel rääkida?

Lisa 2: Sisuanalüüsi koodiraamat

Kategooria 1: KONTEKST

Kood	Täisnimetus	Millal kasutada
KONTEKST_ROLL	Roll ja vastutusosalad organisatsioonis ja projektis	Intervjueeritava roll ja seos projektiga (nt projektijuht, tootemanik, arendaja, analüütik). Intervjueeritava vastutusvaldkonnad organisatsioonis.
KONTEKST_tehisarufookus	Tehisaruteemade osakaal tööstajast	Varasem kokkupuude tehisaruprojektidega, sh viimaste aastate praktikad ja õppimiskõver.
KONTEKST_PROJEKT	Projekti kogemus	Milliste projektidega on intervjueeritav kokku puutunud ja mida neis tehti (nt juturobot, tekstivastus, pildianalüüs). Kirjeldused kasutatud tehisarutehnoloogiatest, mudelitest, tööriistadest ja lahendustest.

Kategooria 2: ANDMETÖÖDE NÄHTAVUS JA PRIORITEET

Kood	Täisnimetus	Millal kasutada
ORG_ANDMETÖÖ_KORRALDUS	Andmetöö protsedid organisatsioonis	Kuidas andmetöö on organisatsiooniülevalt läbi mõeldud ja hallatud.
ORG_ANDMETÖÖ_NÄHTAVUS	Andmetöö kuvand, väärtus ja nähtavus	Üldised kirjeldused ja hinnangud andmetöö kohta; töö prestiiž, nähtav väärtus, maine. Viited sellele, et andmetöö toimub taustal, jääb märkamatuks. Olukorrad, kus tähtsajad surusid peale ja tehti kvaliteedis või mahus järeleandmisi. Kas organisatsioon ja juhtkond mõistavad andmetöö vajalikkust või tuleb selle väärtust pidevalt tõestada.
ORG_ANDMETÖÖ_RESSURSS	Ressursside planeerimine: tööaeg ja raha	Kas ja kuidas planeeriti eelarvetes sisemiste töötajate aega (nt märgendajad). kas oli eraldi eelarverida või muude kulude see

Kategooria 3: ANDMETÖÖDE KORRALDUS

Kood	Täisnimetus	Millal kasutada
PROJEKT_HANGE	Hankemudel	Kas tehisaruanandmetöö tehti majas sees, osteti sisse või kasutati hübriidlahendust. Argumendid ja kaalutlused konkreetse hankemudeli valikul.
PROJEKT_ANDMETÖÖ_OSACAAL	Andmetööde osakaal tehisaruprojektides	Hinnangud andmetöö osakaalule projektides; ajalised ja eelarvelised üllatused.

PROJEKT_KORRALDUS	Andmetöö korraldus	Kuidas andmetöö projektides praktiliselt üles ehitati ja juhti.
PROJEKT_OTSUSED	Andmetöö otsustamine	Kes ja kuidas tegi andmetööd puudutavaid otsuseid.
PROJEKT_DOKUMENTATSIOON	Andmetöö logid ja juhendid	Juhiste, reeglite ja töökorralduslike kokkulepete olemasolu ja kasutus. Logid kes mida tegi jms.
PROJEKT_TÖÖRIISTAD	Andmetöö tööriistad	Kasutatud vahendid (nt Excel, märgendustarkvara, platvormid).
PROJEKT_KVALITEET	Andmetöö kvaliteedikontroll	Kuidas kontrolliti andmete ja märgenduse kvaliteeti; topeltkontroll. Kuidas koostati etalonandmestik (standard-, näidiseksemplar); kes oli lõplik otsustaja vaidluste korral.
PROJEKT_EETIKA	Eetika ja vaidlused	Millest selle küsimuse juures esmalt üldse rääkima hakati? Kas arutelud andmekaitse, privaatsuse, kallutatuse üle; kas eetika oli sisuline prioriteet või formaalsus. Millest räägiti. Keda kaasati. Kui oluline see teema on
ANDMETÖÖTAJA_TAUST	Andmetöötajate taust	Andmetööd teinud inimeste taust. Kas andmetöötajad said aru mida ja miks nad pidid tegema ning kas nad tegid seda motiveeritult.
ANDMETÖÖTAJA_KOOLITUS_RETRO	Andmetöötajate koolitus ja retro	Kas ja kuidas andmetöötajaid välja õpetati. Kuidas andmetöötajatele anti tagasisidet töö kvaliteedi ja tulemuste kohta.
PROJEKT_ÕPPETUNNID	Õppetunnid ja järeldused	Tagantjärele refleksioon: mida teeks intervjuueritav andmetöös teisiti.

Lisa 3: Intervjueeritava informeeritud nõusoleku vorm

Informeeritud nõusoleku vorm

Uuring “Tehisaru rakendamist toetava esmatasandi andmetöö korraldus ja nähtavus Eesti avaliku sektori näitel” (2025–2026)

Lugupeetud osaleja,

Teid on palutud osaleda uuringus, mille eesmärk on kirjeldada, kuidas Eesti avaliku sektori organisatsioonides tajutakse ja korraldatakse tehisaru rakendamiseks vajalikku esmatasandi andmetööd - nt andmete kogumine, puhastamine, märgendamine, kvaliteedikontroll.

Mis on informeeritud nõusolek?

Informeeritud nõusolek tähendab uuringus osalemiseks nõusoleku andmist, kuid ainult juhul, kui mõistate täielikult, mis sellega kaasneb. Allkiri antud dokumendil näitab, et olete kogu informatsioonist aru saanud ja soovite vabatahtlikult osaleda.

Informatsioon intervjuu kohta.

Uuringuga seotud intervjuu kestab ligikaudu 60 minutit ning see salvestatakse, et tagada täpne transkriptsioon ja hilisem analüüs. Intervjuu ei käsitle konfidentsiaalselt teavet, nagu andmekogude sisu, isikuandmed või tundlikud süsteemi-detailid. Uuringus kogutud andmeid kasutatakse ainult magistritöö ja teadusartikli koostamiseks. Intervjuus osalejate ning nendega seotud asutuste ja projektide nimed pseudonüümitakse. Intervjuus osaleja identiteet on teada ainult uuringu läbiviijale. Asutuste ja projektide identiteet on teada nii uuringu läbiviijale kui ka juhendajale. Lisaks vahetutele identifikaatorite jälgitakse ka võimalike kvaasi-identifikaatorite kasutamist, et minimeerida osalejaid, asutusi ja projekte kaudselt tuvastada aitava info väljastamist (nt unikaalsed asutusespetsiifilised ametinimetused või spetsiifilised projektikirjeldused). Intervjuudes kogutud andmeid ei jagata kolmandate osapooltega ning neid kasutatakse ainult teadustöö eesmärgil.

Uuritava õigused.

Uuringus osalejale on õigus keelduda vastamast küsimustele ilma põhjendust esitamata, katkestada uuringus osalemine igal ajal ja nõuda pärast intervjuud, et tema andmed eemaldatakse kogutud materjalist.

Küsimuste korral uuringu kohta pöörduge vastutava läbiviija (Siim Klais, Tartu Ülikool – siim.klais@ut.ee) poole.

Mina,, olen informeeritud ülalmainitud uurimusest, olen teadlik läbiviidava uurimistöö eesmärgist ja metoodikast ning kinnitan allkirjaga oma nõusolekut selles osalemiseks.

Uurimuses osaleja allkiri

Kuupäev, kuu, aasta

Lisa 4: Päring asutustele intervjuueeritavate leidmiseks

Saatja: Siim Klais <siim.klais@ut.ee>

Saadetud: teisipäev, 15. november 2025 11:31

Teema: Koostööpakkumine uuringus: tehisaruandmetöö praktikad avalikus sektoris

Tere,

Olen Tartu Ülikooli magistrant ja viin läbi uuringut teemal „**Tehisaru rakendamist toetava esmatasandi andmetöö korraldus ja nähtavus Eesti avaliku sektori näitel**”. Uuring aitab süsteemselt kirjeldada, kuidas tehisaruprojektides tajutakse ja korraldatakse (treening) andmetega seotud esmatasandi andmetööd – protsesse, millest sõltub otseselt tehisarulahenduste usaldusväärsus, toimivus ja mõju.

Otsin intervjuudeks **teenusejuhte, projektijuhte või teiste sarnaste rollide esindajaid**, kes on osalenud tehisarutehnoloogiaid sisaldavate lahenduste arenduses ning on seejuures kokku puutunud **andmetööde korralduse väljakutsetega** – näiteks andmete kogumise, märgendamise, andmetike ettevalmistuse ja mudelite täiustamisele suunatud tagasisideprotsessidega.

Intervjuusid toimub kokku **10–15**, iga osalejaga üks. Need viiakse läbi **detsembrist 2025 kuni jaanuarini 2026**, kestusega **60 minutit**, ning toimuvad **veebikeskkonnas**. Kogutud andmeid analüüsitakse ja tulemused esitatakse **anonüümselt** – nii organisatsioonide, projektide kui ka intervjuueeritavate tasandil. Uuringus ei käsitleta konfidentsiaalset teavet (nt andmete sisu või tundlikku süsteemiarhitektuuri).

Analüüsitulemused esitatakse **suvel 2026** magistritöö ja tõenäoliselt ka teadusartikli vormis ning jagatakse ka osalevatele asutustele.

Kui Teie asutuses võiks leiduda sobiv intervjuueeritav, oleksin tänulik, kui annaksite sellest teada **hiljemalt 27. novembriks**.

Parimate soovidega,

Siim Klais

Magistrant, Tartu Ülikool

Ühiskonna ja infoprotsesside analüüsi õppekava

✉ siim.klais@ut.ee | ☎ +372 53 236 589 | Microsoft Teams [siim86@ut.ee](https://teams.microsoft.com/join/siim86@ut.ee)

Lisa 5: Lihtlitsents uuringu reprodutseerimiseks ja avalikustamiseks

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Siim Klais,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

“Tehisarurakenduste arendust toetava esmatasandi andmetöö korraldus ja nähtavus Eesti avalikus sektoris”,

mille juhendaja on Maris Männiste,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commons'i litsentsiga CC BY NC ND 3.0, 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.

Siim Klais

27.05.2026