CYBERNETICA

Kaasrahastanud Euroopa Liit

ECCC

Tehisintellekti ja masinõppe tehnoloogia riskide ja nende leevendamise võimaluste uuring

Aruanne

Version 1.0

27.02.2024

D-16-377

AVALIK

Date	Version	Description
27.02.2024	1.0	Esimene avalikustatud versioon

Projektijuhid: Liina Kamm (Cybernetica AS)
Hendrik Pillmann (RIA)

Autorid: Dan Bogdanov
Paula Etti
Liina Kamm
Andre Ostrak
Fedor Stomakhin
Maria Toomsalu
Sandhra-Mirella Valdma
Anto Veldre

Cybernetica AS, Mäealuse 2/1, 12618 Tallinn, Estonia.

E-post: info@cyber.ee, Veebileht: https://www.cyber.ee, Telefon: +372 639 7991.

Kaasrahastatud Euroopa Liidu poolt. Avaldatud seisukohad ja arvamused on ainult autori(te) omad ega pruugi kajastada Euroopa Liidu või Euroopa küberpädevuskeskuse seisukohti või arvamusi. Euroopa Liit ega Euroopa küberpädevuskeskus nende eest ei vastuta.

Sisukord

1 Sissejuhatus

1.1 Eesmärk

Eesti ühiskond on võtnud kasutusele palju arvutustehnilisi lahendusi töö tõhususe tõstmiseks. Meie e-riik on tuntud oma efektiivse asjaajamise poolest. Riigiasutuste vahel liiguvad tehingud üle X-tee. Nii avalik kui erasektor on võtnud kasutusele digitaalse identiteedi lahendused. Digiühiskond on Eesti jaoks miski, mida me arendame.

Tehisintellekti tehnoloogia areng on tänu arvutustehnika jõudluse kiirele kasvule jõudnud uuele kvaliteeditasemele. Tehisintellektisüsteemid, mis suudavad luua loomuliku keele kirjelduste põhjal teksti, pilti, häält, muusikat ja filme, on muutnud tehnoloogia inimestele mõistetavaks ning tekkimas on veendumus, et infotehnoloogia abil saab luua uue põlvkonna süsteeme, mis täidavad ülesandeid inimestest paremini.

Nii Eesti avalik kui erasektor on arendamas tehisintellektisüsteeme. Selle aruande eesmärk on toetada tehnoloogia rakendamist, andes ette juhised, kuidas tagada küberturvalisus, seaduse nõuete täitmine ning ohutus ühiskonnale.

Aruanne on mõeldud laiale sihtrühmale. Sellest saavad kõige rohkem kasu väikesed ning keskmised organisatsioonid ja eraisikud, kellel ei tarvitse olla koosseisulisi juriste, infoturbe või tehisintellekti eksperte. Nemad saavad kasutada aruande lõpus toodud kiirjuhist tehisintellektisüsteemi riskikontrolliks ja meetmete valikuks. Meie eesmärk on, et kõik Eestis rakendaksid tehisintellekti õiguspäraselt, turvaliselt ja ilma ühiskonda ja keskkonda kahjustamata.

Küpsemad organisatsioonid, kes rakendavad kvaliteedijuhtimise süsteeme ja töömahukamaid riskihalduse protsesse, saavad aruandest nõuandeid tehisintellekti rakendamisel. Neile anname soovitusi, milliseid standardeid ja aruandeid kasutada, et tagada nõuetele vastav küpsustase.

1.2 Mõisted ja lühendid

AGI, *artificial general intelligence*
: Tehislik üldintellekt.

AI, *artificial intelligence*
: Tehisintellekt.

AI-süsteem
: Tehisintellektisüsteem.

AI HLEG
: Euroopa Liidu kõrgetasmeline tehisintellekti eksperdirühm (*High-Level Expert Group on AI*)[1]

API, *application programming interface*
: Rakendusliides.

ASI, *artificial superintelligence*
: Tehislik üliintellekt.

BERT, *Bidirectional Encoder Representation from Transformers*
: Transformerpõhine keelemudel.

[1] Euroopa Komisjon, High-level expert group on artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai (viimati külastatud 27.02.2024).

CaaS, *compute as a service*
Arvutus kui teenus.

CNN, *convolutional neural network*
Konvolutsiooniline neurovõrk. Pildituvastuses kasutatav mudelarhitektuur.

CPU, *central processing unit*
Arvutiprotsessor.

CUDA, *Compute Unified Device Architecture*
NVIDIA väljatöötatud arendusplatvorm arvutusülesannete kiirendamiseks GPU abil.

DPO, *direct preference optimization*
Otsene eelistuse optimeerimine. Peenhäälestustehnika.

FLOP, *floating-point operation*
Ujukomatehe. Mudelite treenimiseks kuluvat arvutuslikku ressurssi mõõdetakse tehetes ujukomaarvudega.

GAN, *generative adversarial network*
Generatiivne vastandvõrk. Pildisünteesis kasutatav mudelarhitektuur.

GPT, *generative pretrained transformer*
Generatiivne eeltreenitud transformer. Tehisintellekti mudelarhitektuur.

GPU, *graphics processing unit*
Graafikaprotsessor.

IaaS, *infrastructure as a service*
Taristu kui teenus.

Intellektitehnika
Tehisintellekti uuriv ja arendav valdkond.

IPO, *identity preference optimization*
Identiteedi eelistuse optimeerimine. Peenhäälestustehnika.

LLM, *large language model*
Suur keelemudel. Tehisintellekti mudel, mida kasutatakse loomuliku keele töötlemiseks ja mis paistab silma oma suure parameetrite arvu poolest.

LSTM, *long short-term memory*
Pikk lühiajaline mälu. Keelemudelites kasutuses olnud mudelarhitektuur, mis oli levinud enne transformerite kasutuselevõttu.

ML, *machine learning*
Masinõpe.

MoE, *Mixture of Experts*
Ekspertide segu. Mudelarhitektuur.

NPU, *neural processing unit*
Neuroprotsessor. Eeskätt telefonides kasutatav tehisintellekti kiirendi.

OWASP, Open Worldwide Application Security Project
Veebirakenduste ja tarkvara turvalisuse ressursse koondav ja tootev veebikogukond.

PaaS, *platform as a service*
Platvorm kui teenus.

RAG, ***retrieval-augmented generation***
: Päringgenereerimine. Tehisintellekti rakenduse levitamises kasutatav meetod, kus keelemudel pärib andmebaasist või muust välisest allikast kasutaja viiba põhjal täiendavat konteksti, et parandada vastuse kvaliteeti.

RLHF, ***reinforcement learning with human feedback***
: Inimtagasisidega stiimulõpe. Stiimulõpet rakendav peenhäälestustehnika.

RNN, ***recurrent neural network***
: Rekurrentne neurovõrk. Keelemudelites kasutuses olnud mudelarhitektuur, mis oli levinud enne transformerite ja LSTM kasutuselevõttu.

SaaS, ***software as a service***
: Tarkvara kui teenus.

SFT, ***supervised finetuning***
: Juhendatud peenhäälestus. Tehisintellekti mudeli treenimismeetod, mis erinevalt eelõppest on juhendatud ning mida kasutatakse mudeli töö täiendavaks suunamiseks.

TPU, ***tensor processing unit***
: Tensorprotsessor. Google'i väljaarendatud tehisintellekti kiirendi.

VAE, ***variational autoencoder***
: Variatsiooniline autokooder. Pildisünteesis kasutatav mudelarhitektuur.

XAI, ***explainable AI***
: Seletatav tehisintellekt. Kogum meetodeid tehisintellekti mudelite töö ning selle tulemite seletamiseks, tõlgendamiseks ja valideerimiseks.

Teiste infotehnoloogiliste terminite tõlke allikaks on andmekaitse ja infoturbe portaal AKIT[2].

1.3 Aruande struktuur

Aruanne algab tehisintellekti ajaloo ja põhiliste tehnoloogiate tutvustamisega (peatükk 2). Seejärel jõuame rakendusteni ja toome näiteid, millistes eluvaldkondades tehisintellektist lisaväärtust loodetakse saada. Uuringu ajal on valdkond ka kiirelt arenemas, seega toome ülevaate praegustest trendidest.

Riigid üle maailma on hakanud tehisintellekti seadustega reguleerima. Peatükis 3 teeme ülevaate õigusloome hetkeseisust. Peatükis 4 uurime tehisintellektisüsteemide ehitust ning pakume välja kolm üldist mudelit tehisintellektirakenduste levitamiseks. Nendele kolmele levitusmudelile tuginedes on organisatsioonil lihtsam riskihaldusmetoodikat rakendada.

Tehisintellekti tehnoloogia rakendamisel tuleb lisaks seadusele järgida ka küberturbe ja ühiskondliku ohutuse nõudeid. Vastava riskikontrolli kohta anname juhised peatükis 5. Ning kui on olemas riskid, peame rakendama leevendusmeetmeid. Neist annab ülevaate peatükk 6.

Peatükis 7 võtame kokku uuringu käigus tekkinud soovitused tehisintellektisüsteemide kasutuselevõtu edendamiseks Eestis.

Aruande viimane osa on kõige praktilisem ning mõeldud neile, kes pikka juttu lugeda ei soovi. Esitame konkreetsed ja lihtsalt järgitavad juhised, kuidas tehisintellektisüsteemi luues või arendades leida olulisemad ohud ja kuidas nendega toime tulla. Vastava juhise ja toetavad joonised leiate peatükis 8.

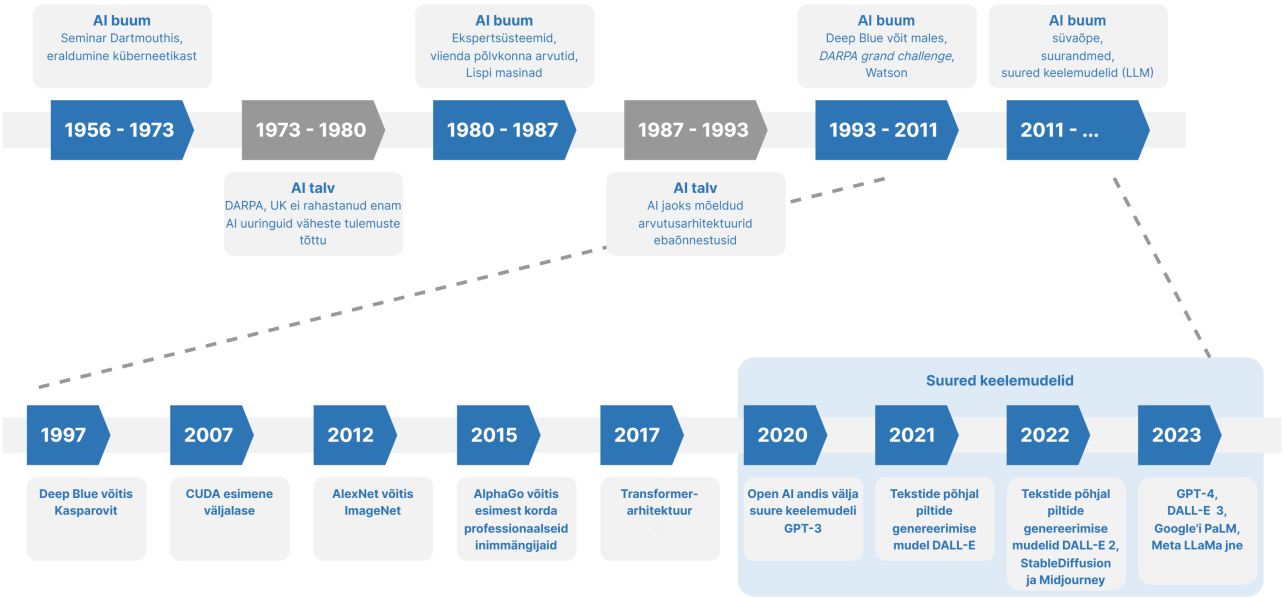
[2] Andmekaitse ja infoturbe portaal. https://akit.cyber.ee Külastatud 27.02.2024

2 Tehisintellekti rakenduste ülevaade ja kasutusjuhud

2.1 Intellektitehnika ajalugu

Tehisintellekti (*artificial intelligence*, AI) all mõistame mistahes süsteemi, mis suudab sooritada ülesandeid pealtnäha inimintelligentsi kasutades. Joonis 1 kujutab olulisi järke intellektitehnika ajaloos. Uurimisvaldkonnana kasvas intellektitehnika välja küberneetikast, mille eesmärgiks oli uurida tagasisidestatud süsteeme, kaasa arvatud bioloogilisi, tehnilisi ja sotsiaalseid süsteeme. Kuigi tehisneuronite idee ja ülesehitus pakuti välja juba 1940ndatel, loetakse intellektitehnika alguseks 1956. aastal Dartmouthis korraldatud suveseminari, mille jaoks tehisintellekti termin ka esimest korda välja pakuti.

Seminaril jõuti järeldusele, et masinaid on võimalik panna sooritama kõiki tegevusi, mida inim-intelligentsiga seostakse. Sealhulgas arvati, et arvuti on suuteline ise õppima, keelt kasutama ning loovust demonstreerima. Kaks kuud kestnud suveseminari jooksul ei saavutatud suuri läbi-murdeid, kuid osalejad olid järgmised 20 aastat intellektitehnika suurimad edasiviijad. Selle aja jooksul loodud tehisintellektid suutsid lahendada matemaatilisi probleeme, mängida kabet või tõlkida teksti ühest keelest teise.



Joonis 1. AI arengu ajalugu

Aastal 1958 pandi alus kõrgkeelele Lisp, millest sai peamine AI programmide kirjutamiseks mõeldud keel kolmeks aastakümneks. Pealtnäha suured edasiminekud ja lahendused olid siiski piiratud. Tõlkeprogrammid kasutasid otsetõlget ega suutnud seega edasi anda väljendite mõtet. Programmid, mis tõestasid matemaatilisi teoreeme või mängisid kabet, suutsid läbi vaadata väga piiratud arvu võimalusi ning põrusid keerulisemate ülesannete korral.

Probleemide lahendusi demonstreeriti väikestes mängukeskkondades, mida nimetati mikromaa-ilmadeks. Tõenäoliselt kuulsaim mikromaailm oli virtuaalne klotside maailm, mida kasutaja sai mõjutada inglisekeelsete käskude abil, näiteks läbi SHRDLU-nimelise keeleparseri. Kuigi genee-

tilised algoritmid ja tehisneurovõrkude algelised põhimõtted pakuti välja juba 1960ndate lõpus, ei olnud need algoritmid võimelised saavutama erilisi edusamme vähese optimeerituse ja arvutustehnika madala jõudluse tõttu.

Lootuste põhjal, mis tekkisid algsete AI-süsteemide loomisega, andsid mitmed teadlased lubadusi, mida polnud võimalik ellu viia. AI rahastajad pettusid ning tehnoloogiate arendamine ja uurimine aeglustus 1970ndatel. Ühendkuningriik ja Ameerika Ühendriigid vähendasid märkimisväärselt AI rahastust ülikoolidele ning AI projektide rahastamise lõpetas ka Ameerika Ühendriikide Kaitseministeeriumi teadusagentuur DARPA. Seda ajastut, aastaid 1974-1980, nimetatakse esimeseks AI talveks.

Vaatamata rahastuse vähenemisele jätkati samal ajal tehisintellekti arendamist, kuid suurte ja keeruliste probleemide lahendamise asemel keskenduti rohkem süsteemidele, mis koondasid teavet valdkondade ekspertidelt ja suutsid neid rakendada kitsaste probleemide lahendamisel. Niinimetatud ekspertsüsteeme kasutati näiteks meditsiinis ja analüütilises keemias. Ekspertsüsteeme uurisid edukalt ka Eesti teadlased (teiste seas Enn Tõugu ja Leo Võhandu).

Ekspertsüsteemide edu tõi 1980ndate algul intellektitehnikale taas suure avaliku huvi. Üks esimestest reeglipõhistest kommertssüsteemidest oli R1, mis aitas klientide soovidele vastavaid arvuteid konfigureerida. 1981. aastal kuulutati Jaapanis välja nn viienda põlvkonna arvutite (*Fifth Generation Computer Systems*) projekt. Tegemist oli kümnendipikkuse plaaniga arendada välja intelligentseid arvuteid. See taastas huvi tehisintellekti vastu ka Ameerika Ühendriikides ja Ühendkuningriigis.

Uus AI buum jõudis tippu 1980ndate teisel poolel. USA suurettevõtetes loodi AI-süsteemidele keskendunud rühmad. Taas hakati uurima tehisneurovõrke ning nende treenimist tagasilevialgoritmidega (*back propagation algorithms*). AI algoritmide arendamiseks hakati üha rohkem kasutama matemaatilisi ja statistilisi optimeerimismeetodeid ning eraldi keeli ja vastavat riistvara. Tuntuimad AI-spetsiifilised keeled kuulusid programmikeelte perre Lisp. Nende keeltega kirjutatud programmide efektiivseks jooksutamiseks loodi eraldi arvutid – Lispi masinad.

Suurtele edusammudele vaatamata algas 1987. aastal teine AI talv. Eriotstarbeliste tehisintellektide ülalpidamine ja uuendamine oli keeruline, samas ei saanud nad ise hakkama uute seninägemata sisenditega, mistõttu jäid nad kiiresti ajale jalgu. IBM ja Apple tootsid aina võimsamaid üldotstarbelisi lauaarvuteid. Sihtotstarbelised masinad (sh Lispi masinad) muutusid üleliigseteks. Viienda põlvkonna arvutite projekt ei saavutanud loodetud tulemusi. Näiteks pidi aastaks 1991 valmima tehisintellekt, mis suudab kasutajaga pidada igapäevaseid vestlusi – eesmärk, mida ei saavutatud veel mitukümmend aastat. Ekspertsüsteemide piiratud võimetes pettunud DARPA vähendas jälle drastiliselt AI-süsteemide uurimiseks mõeldud rahastust.

Intellektitehnika edasine areng tugines üha enam varasematele rangetele matemaatilistele meetoditele. Taas vaadati range loogika poole ja otsiti lahendusi küberneetikast välja arenenud juhtimisteooriast. Teiselt poolt hakati kasutama tõenäosuslikke mudeleid ja hägusloogikat, mis võimaldavad kirjeldada seoseid ja tinglikke tõenäosusi tunnuste vahel ning, erinevalt puhtast loogikast, väljendada ennustamisel ka teadmatust ja ebakindlust.

1990ndatel muutusid populaarsemaks andmekaeve ja masinõppe algoritmid. Süsteeme ei kirjeldanud ainult programmeerijad ja ekspertid, arvuti suutis suuri andmekogusid analüüsides ise õppida. Intellektitehnikat ja tõenäosuslikke meetodeid aitasid kokku siduda Bayes'i võrgud, mis võimaldavad esitada muutujate vahelisi tinglikke tõenäosusi suunatud graafide abil. Intellektitehnikas tekkis paradigma, mis käsitles tehisintellekte kui agente, mis saavad signaale keskkonnast ja üritavad optimaalselt käituda, et saavutada teatud eesmärke. Tehisintellekti suurimaks saavutuseks 1990ndatel võib pidada malet mängiva süsteemi Deep Blue võitu 11. mail 1997. aas-

tal tollase valitseva maailmameistri Garri Kasparovi üle. Selleks ajaks oli AI-süsteeme hakatud juurutama ka igapäevastes teenustes, eriti veebipõhistes lahendustes. Loomuliku keele töötlust rakendas näiteks otsingumootori Google PageRank algoritm, mis loodi samuti aastal 1997. Algoritm järjestas kasutaja otsingul kuvatud lehti ning seda peetakse üheks olulisemaks funktsionaalsuseks, millega Google eristus teistest olemasolevatest otsingumootoritest.

Loomuliku keele töötlust kasutasid ka kõnesünteesi mudelid nagu DECtalk, mida kasutas kõnerobotina näiteks Stephen Hawking, ja Bell Labs'i natuke keerulisem TTS (*Text-to-Speech system*), mis suutis kõnet sünteesida mitmes eri keeles. Alates 1990ndate algusest, pea kakskümmend aastat, domineerisid masintõlke valdkonnas ettevõttes IBM välja arendatud statistilised mudelid. Kõnetuvastusel muutusid enimkasutatuks Markovi peitmudelid. Näotuvastusel oli 1990ndatel peamiseks lähenemiseks omavektornäo (*eigenfaces*) algoritmid, mis kasutasid lineaaralgebra meetodeid näokuju analüüsimiseks.

Hoolimata tehisintellektisüsteemide edusammudest oli 1990ndate lõpus AI mõiste siiski halva maiguga. Teadlased vältisid mõistet, eelistades rääkida pigem statistilistest meetoditest, masinõppest, juhtimisteooriast. Teise AI talve lõpp pole selgelt defineeritud, kuid üldiselt nõustutakse, et see oli läbi saanud aastaks 2005, mil Stanfordi ülikoolis ehitatud isejuhtiv auto Stanley suutis DARPA Grand Challenge'i raames läbida Nevada kõrbes 212 km pikkuse raja vähem kui seitsme tunniga. Tegemist oli tohutu edasiminekuga, aasta varem ei suutnud kümnetunnise võistluse jooksul ükski isejuhtiv auto läbida 12 km. Kaks aastat hiljem korraldas DARPA sama võistluse linnaliikluse tingimustes, võitjaks osutus Carnegie Melloni Ülikooli robot Boss, mis suutis nendes tingimustes vähem kui kuue tunni jooksul läbida 96 km.

Aastal 2011 demonstreeris IBM küsimustele vastavat robotit Watson USA telesaates Jeopardy! (Eestis tuntud kui Kuldvillak). Kahes järjestikuses saates võistles Watson kahe inimmängija vastu, kellest üks oli ajaloo parimaks mängijaks peetav Ken Jeggins, ning võitis mõlemad mängud kindla eduga. Watsoni edu taga oli paljudest erinevatest keelemudelitest kokku pandud ideed ja suur arvutusvõimsus, mis võimaldas suurte andmemahtude peal treenimist. Treenimisel sooritati pidevalt vigade analüüsi ja programmi parendati jooksvalt. Sellegipoolest ei esinenud Watson vigadeta. Näiteks esimese saate viimases küsimuses otsitavaks USA linnaks pakkus ta vastuseks Toronto.

Tehisintellektide ajastu üks suurimaid läbimurdeid saabus aastal 2012, kui konvolutsioonilistel tehisneurovõrkudel põhinev AlexNet suutis ülekaalukalt võita pildituvastuse võistluse ImageNet Large Scale Visual Recognition Challenge (ILSVRC). AlexNet polnud esimene konvolutsiooniline neurovõrk, arhitektuuri pakkus tegelikult juba aastal 1989 välja Yann LeCun. Läbimurdeni viisid graafiliste piltide renderdamiseks mõeldud eriotstarbelistel protsessoritel optimeeritud treenimisalgoritmid, mis võimaldasid suuremate ja sügavamate neurovõrkude treenimist kui varem. ImageNet andmebaasis oli 15 miljonit pilti enam kui 22 tuhandest kategooriast. Järgmiste aastate ImageNeti võistlustel põhinesid kõik võidukad ideed samuti konvolutsioonilistel tehisneurovõrkudel ja AlexNeti tulemust suudeti ületada mitu korda. Praegu peetakse ImageNeti ülesannet lahendatuks.

Pärast AlexNeti läbimurret on tehisneurovõrkude areng olnud tormiline. Lisaks konvolutsioonilistele neurovõrkudele hakati suurt tähelepanu pöörama suurtele keelemudelitele, rekurrentsetele tehisneurovõrkudele, pika lühimäluga mudelitele. Need viisid kõnetuvastuse, -sünteesi ja tõlkemudelite kiire arenguni. Tehisintellekte hakati laiemalt kasutama meditsiini-, tööstus-, finantsvaldkonnas. Rekurrentsed võrgud võeti kasutusele aegridade uurimisel, robootikas, mängude mängimisel. Eriti suurt tähelepanu sai AlphaGo, mis suutis 2015. aastal võita professionaalseid inimmängijaid lauamängus Go.

Selle aruande koostamise ajal pakuvad tehisintellektidest enim kõneainet generatiivsed mudelid, mis suudavad inimkeeles suhelda, vastata küsimustele, pealtnäha loogiliselt arutleda, luua pilte, muusikat, aidata programmeerijatel koodi kirjutada. Generatiivsete masinõppe mudelite kontseptsioon pole uus, kuid sügavate generatiivsete neurovõrkude olulisemad saavutused jäävad viimasesse aastakümnesse. Aastal 2014 tutvustati vastandgeneratiivseid mudeleid ja variatsioonilisi autokoodereid, mis on olulised tööriistad pildisünteesil. Vastandgeneratiivsed mudelid võimaldasid esmakordselt sünteesida kõrge resolutsiooniga pilte inimeste nägudest.

Aastal 2015 näidati, et statistilise füüsika võtteid saab kasutada generatiivsete difusioonmudelite treenimiseks. Arvatavasti suurima edasiminekuni viisid aga tähelepanu mehhanisme kasutavad transformerid, mille baasarhitektuuri pakkus välja Google 2017. aastal. Transformeritel põhinevad mitmed tuntud generatiivsed keelemudelid, näiteks GPT ja BERT, koodi kirjutamiseks GitHub Copilot.

Transformerid võimaldavad ehitada laia kontekstiaknaga paralleliseeritavaid mudeleid, mida on võimalik juhendamata treenida suurtel andmehulkadel. Juhendamata mudelit saab siirdeõpet kasutades ümber treenida ka kindla ülesande jaoks. See on oluline, sest võimaldab aega- ja ressurssenõudvat universaalset mudelit treenida vaid korra. Seda mudelit saab hiljem vähese vaevaga mugandada spetsiifilise probleemi jaoks väiksemal andmehulgal ja väiksemate ressurssidega.

Ka pildisünteesi, täpsemalt tekstide põhjal piltide genereerimise (*text-to-image*), mudelid kasutavad transformereid, kuid nende arhitektuur on üldiselt keerulisem. DALL-E 3 ja Stable Diffusion kasutavad piltide kodeerimiseks ja dekodeerimiseks autokooderit, kodeeritud andmete peal treenitakse difusioonmudeleid, mis omakorda koosnevad konvolutsioonilistest neurovõrkudest.

2.2 Tehisintellekti algoritmid ja taksonoomiad

Intellektitehnika mõiste on lai ja hõlmab erineva keerukuse, seletavuse ja sügavuse ning erinevate rakendusvaldkondade ja treenimisalgoritmidega meetodeid. Kõrgel tasemel jagatakse tehisintellekti algoritmid reeglipõhisteks süsteemideks, traditsioonilise masinõppe algoritmideks ja tehisneurovõrkudeks.

2.2.1 Reeglipõhised süsteemid

Reeglipõhised süsteemid on lihtsaimad tehisintellekti süsteemid. Sellised süsteemid koosnevad üldiselt inimekspertide poolt koostatud reeglitest, mida järgides arvuti suudab lahendada pealtnäha inimintelligentsi vajavaid probleeme. Näiteks suudavad reeglipõhised süsteemid hästi lahendada teatud sorti loogilise mõtlemise ülesandeid ja mõistatusi (näiteks niinimetatud Einsteini mõistatusi ja sebramõistatusi).

2.2.2 Masinõpe

Masinõppe korral õpib arvuti ülesannet lahendama andmete põhjal (mis võivad olla sensorite, varasemate sündmuste vms masinloetavad esitused). Masinõppes rakendatakse matemaatilisi optimeerimistmeetodeid, mille abil programm otsib lähteprobleemile võimalikult täpset lahendit. See võimaldab lahendada ülesandeid, mille lahendusalgoritmi on inimesel keeruline täpsete juhiste abil kirjeldada.

Masinõppe meetodite liigitamiseks on erinevaid võimalusi. Rakenduste ja treeningandmete põhjal saab eristada näiteks juhendatud ja juhendamata masinõpet ning stiimulõpet.

2.2.2.1 Juhendatud ja juhendamata masinõpe, stiimulõpe

Juhendatud masinõppes on treeningalgoritmi ülesandeks koostada mudel, mis suudab sisendi põhjal ennustada väärtust või vektorit, mida nimetatakse märgendiks. Juhendatud õppel antakse mudelile treenimisel ette treeningandmed, mille hulka kuuluvad nii mudeli sisendid kui ka neile vastavad märgendid. Mudel saab jooksvalt ennustusi võrrelda korrektsete märgenditega ning sellele vastavalt oma ennustusvõimet parandada. Juhendatud masinõpet kasutatakse peaaegu igas masinõppe rakendusvaldkonnas, näiteks meditsiiniuuringutel, piltide, teksti ja hääle tuvastamisel või töötlemisel ning otsingumootorite ja spämmifiltrite treenimisel.

Juhendatud masinõppe ülesanded jaotatakse klassifikatsiooni ja regressiooni ülesanneteks. Klassifikatsiooni korral on mudeli ülesanne kirje põhjal ennustada, millisesse kahest või enamast klassist kirje kuulub. Regressioonimudelid püüavad võimalikult täpselt ennustada kirjele vastavat arvulist väärtust.

Juhendamata masinõppe korral ei ole olemas kirjetele vastavaid märgendeid või mudel ei näe neid. Sellistel juhtudel on algoritmi eesmärk leida andmetest seoseid või struktuuri ilma treeningmärgenditeta. Juhendamata algoritmid võimaldavad vähendada lähteandmete mõõdet (peakomponentanalüüs) või rühmitada sarnaseid kirjeid (klasterdamine). Juhendamata masinõppe meetodeid kasutatakse näiteks geneetikas alampopulatsioonide tuvastamiseks ning generatiivsete mudelite nagu autokooderite treenimiseks. Juhendamata meetodeid kasutatakse sageli ka enne juhendatud masinõppe meetodite rakendamist.

Juhendatud ja juhendamata masinõppe algoritmide kõrval eristatakse ka stiimulõpet. Stiimulõppe puhul ei leidu igale sisendile vastavat väljundit. Selle asemel õpib algoritm valima vastavalt oma keskkonnale tegevusi nii, et nende eest saadav preemia oleks lõpuks võimalikult suur. Stiimulõpet saab kasutada näiteks kõne töötlemisel või selleks, et õpetada arvutit mänge mängima. Muuhulgas kasutati stiimulõpet AlphaGo treenimiseks.

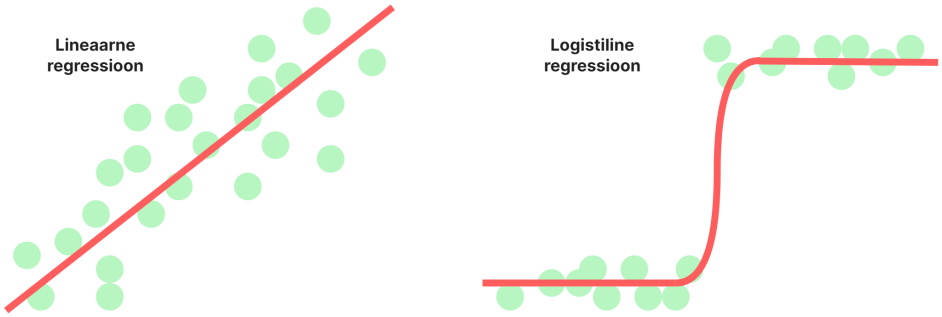
Siirdeõpe ehk ülekandeõpe on masinõppe tehnika, mille puhul ühe ülesande täitmiseks omandatud infot kasutatakse teiste ülesannete täitmiseks. Näiteks võib eelõpetatud üldotstarbelisi keelemudeleid kasutada erinevate keeleliste ülesannete täitmiseks mudeli täiendava peenhäälestamiseta (vt alamjaotis 2.2.4.1).

2.2.2.2 Masinõppe algoritmid

Lineaarne regressioon (joonis 2) on üks lihtsamatest juhendatud masinõppe mudelitest, mis tegelikult on statistilise meetodina kasutusel olnud juba aastasadu. Mudelit kasutatakse reaalarvulise väljundväärtuse ennustamiseks sisendandmete pealt. Nagu nimigi viitab, modelleeritakse lineaarregressiooni abil sisendi ja väljundi vahelist lineaarset seost. Selle tõttu on treenitud mudel kergesti seletatav, sest mudeli põhjal on võimalik täpselt öelda, kuidas sisendväärtuse muutus mõjutab ennustust.

Logistiline regressioon (joonis 2) on lineaarse regressiooniga väga sarnane algoritm, mis erinevalt nimetusest on mõeldud klassifikatsioonianalüüsiks. Binaarse logistilise regressiooni korral rakendatakse ennustamisel esmalt lineaarfunktsiooni, mille väljundit võib tõlgendada kui märgendi tõenäosuse logaritmi. Seejärel rakendatakse väljundile sigmoidfunktsiooni, mis teisendab väärtuse tõenäosuseks vahemikus $[0, 1]$. Logistilist regressiooni võib kergelt mugandada ka juhuks, kui väljundklasse on enam kui kaks.

Tugivektormasin on juhendatud masinõppe meetod, mis töötati algul välja klassifitseerimiseks. Lihtsaim tugivektorklassifitseerija on lineaarne klassifitseerija, mille ülesanne on leida hüpertasandid, mis eraldavad erinevate klasside kirjeid. Lineaarne klassifitseerija eeldab, et andmete

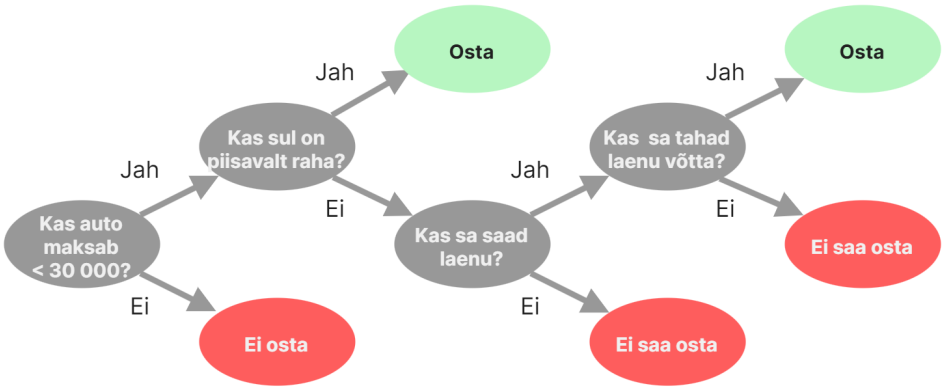


Joonis 2. Lineaarne ja logistiline regressioon

klassid on lineaarselt eraldatud, see aga enamasti ei kehti. Seetõttu on aja jooksul välja töötatud mugandusi, mis võimaldavad tugivektormasinaid treenida mittelineaarseks klassifitseerimiseks, regressioonianalüüsiks, erindite leidmiseks ning dimensionaalsuse vähendamiseks.

Tugivektormasinaid kasutatakse pildi- ja tekstiklassifitseerimisel, aga ka näiteks bioloogias. Tugivektormasinate peamiseks nõrkuseks on nende raske seletavus ja suurem arvutuslik keerukus treenimisel.

Otsustuspuu (joonis 3) on juhendatud hierarhilise andmestruktuuriga mudel, mida rakendatakse rekursiivselt sooritatud otsuste jadana regressiooni- ja klassifikatsioonianalüüsiks. Puu koosneb lahknemissõlmedest ja otssõlmedest ehk lehtedest. Lahknemissõlmedes rakendatakse sisendile teste ja valitakse nende põhjal järgmised oksad. Lehed väljastavad tehtud testide põhjal sisendile vastava väljundi.



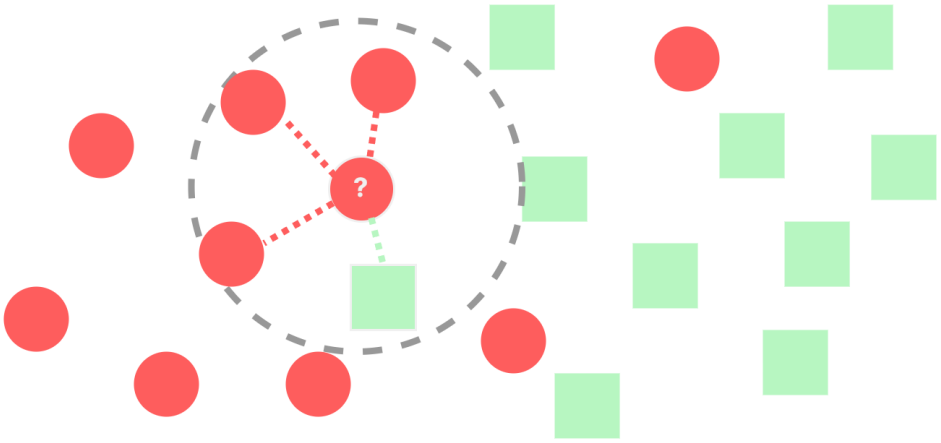
Joonis 3. Otsustuspuu näide auto ostuks

Otsustamisest võib mõelda kui jah/ei küsimuste jadast, kus iga järgmine küsimus sõltub eelmisest ning lõplik ennustatud väärtus sõltub igast antud vastusest. Otsustuspuud on lihtsasti seletatavad ja intuitiivselt arusaadavad mudelid, mistõttu on nad ajalooliselt väga populaarsed.

Naiivne Bayes'i meetod on klassifikatsioonialgoritm, mis kasutades Bayes'i teoreemi, võimaldab sisendi põhjal ennustada kõige tõenäolisemaid märgendeid. Meetod eeldab mudeli treenimisel, et sisendtunnused on üksteisest sõltumatud. Sellegipoolest on naiivne Bayes'i meetod ajalooliselt populaarne, sest see on küllalt võimas, aga ka lihtsasti seletatav ja treenitav. Erinevalt paljudest teistest masinõppe algoritmidest ei pea naiivse Bayes'i meetodi lahendit leidma iteratiivsete sammudena, sest selle suurima tõepära hindamise valemi saab esitada ilmutatud kujul.

k-lähima naabri algoritm (joonis 4) on juhendatud algoritm, mida saab kasutada nii regressiooni- kui klassifikatsiooniülesannete lahendamiseks. Nagu meetodi nimigi ütleb, toimub ennustamine ühe kirje k lähima naabri järgi, kus k on positiivne täisarv. Klassifikatsioonis vaadatakse, milline klass on uue sisendi k lähima naabri seas enim esindatud. Regressiooni korral võetakse ennustatud väärtuseks k lähima naabri väärtuste keskmine. Modifikatsioonina saab naabritele määrata kaalud vastavalt nende kaugustele kirjest, võttes rohkem arvesse lähemal asuvaid punkte. Kaugust erinevate punktide vahel saab vastavalt lähteülesandele mõõta erinevate meetrikate abil.

Lähima naabri meetod on populaarne, sest mudeli treenimist pole vaja, ennustamine käib treeningandmete põhjal. Lisaks on mudel lihtsasti seletatav. Probleemina tuuakse peamiselt välja, et meetod on lokaalne, st ennustamisel võetakse arvesse ainult paari üksikut kirjet, kogu ülejäänud treeningandmestikku eiratakse.

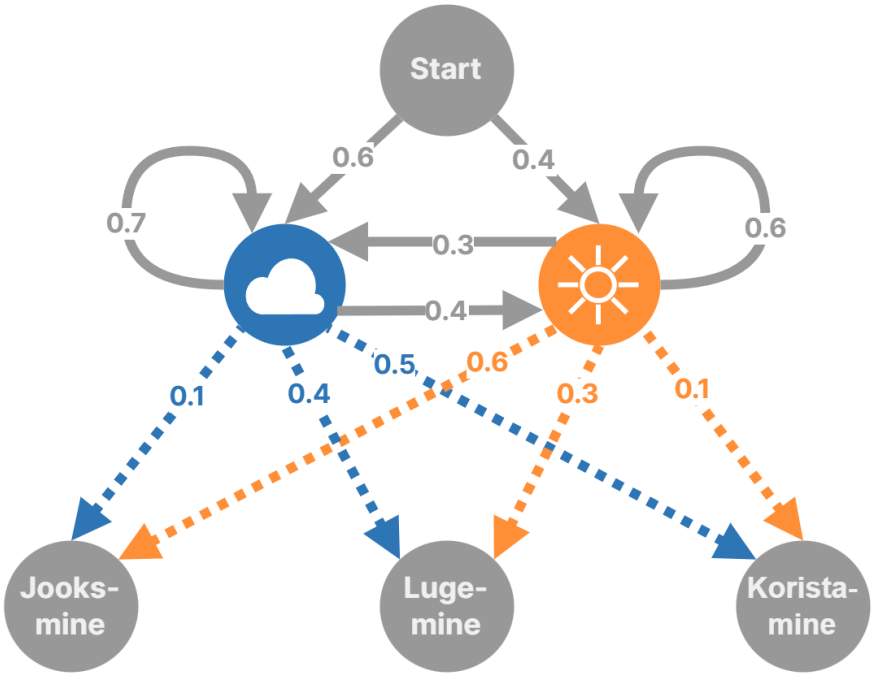


Joonis 4. k-lähima naabri algoritmis vaadeldakse tundmatule lähimaid naabreid

Peakomponentanalüüs on juhendamata algoritm, mis võimaldab lineaarsete teisendustega viia andmed koordinaatsüsteemi, mis on paremini seletatav. Peakomponentanalüüsi kasutatakse sageli andmestiku mõõdete vähendamiseks. See on eriti kasulik kui andmestiku paljud tunnused on omavahel tugevalt sõltuvad. Esimesed peakomponendid on vektorid, millele andmestiku projitseerimisel säilib võimalikult suur osa andmete varieeruvusest. Kui projitseerime andmed esimestele peakomponentidele, saab andmete klasterdust ka visuaalselt uurida.

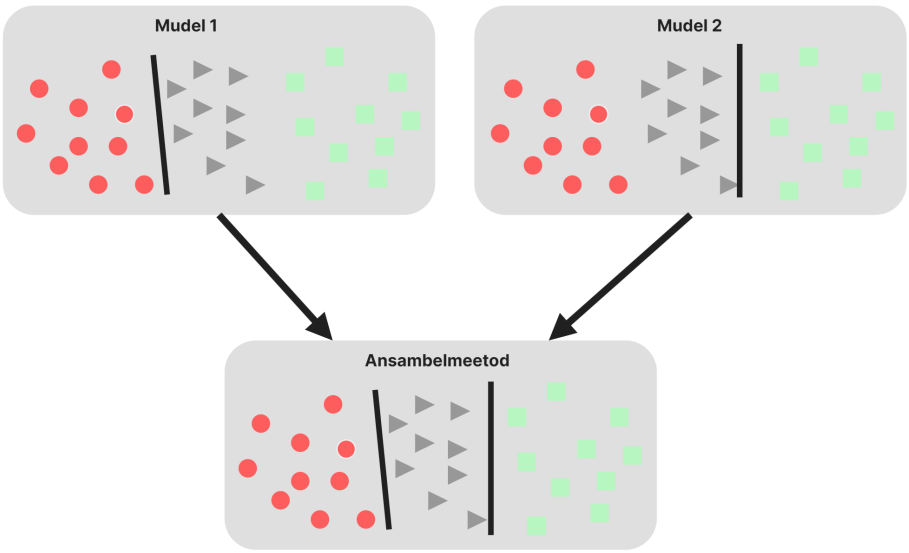
k-keskmise meetod või k-keskmise klasterdamise meetod on juhendamata masinõppe algoritm, mis jaotab andmekirjed k erinevasse klastrisse, kus k on positiivne täisarv. k-keskmise meetodit ei tohi segi ajada k-lähima naabri meetodiga, mis on juhendatud meetod. Kui k-lähima naabri meetodi juures piisab ennustamiseks ainult kirjele lähimate punktide vaatamisest, siis k-keskmise meetod otsib optimaalset klasterdust kõigile punktidele, seega on treenimine palju keerulisem ning väljundi tõlgendamiseks tuleb vaadata, millised kirjed kokku klasterdati. Klastrite põhjal saab otsida andmestikust seoseid. Klasterdamisel on igal klastril keskpunkt, mida saab näiteks signaalitöötlusel kasutada klastri punkti esindajana. Meetodit saab kasutada ka tunnuste automatiseeritud õppimiseks, mis võimaldab sisendandmed viia muu masinõppe meetodi jaoks sobivale kujule.

Markovi peitmudel (joonis 5) on statistiline algoritm, mis modelleerib Markovi protsessi, st võimalike sündmuste jada, milles iga järgmise sündmuse tõenäosus sõltub ainult olekust, milleni protsess jõudis eelmise sündmuse järel. Markovi peitmudelis ei ole Markovi protsessi olekud jälgitavad. Jälgitavad on sündmused, mida peidetud olekud/sündmused otseselt mõjutavad. Ülesandeks on jälgitavate sündmuste kaudu uurida peidetud olekuid ja sündmusi.



Joonis 5. Markovi peitmudeli näide tegevusteks erinevate ilmastikuolude korral

Ansambelmeetodid (joonis 6) on tehnikad, mis kombineerivad masinõppemudeleid. Kombineeritud mudelid on tihti paremad ja stabiilsemad kui üksikud mudelid eraldi. Mudelite kombineerimiseks on erinevad võimalused: valimist korduvate juhuslike andmete võtmine (*bootstrap aggregating* ehk bagging), virnastamine (*stacking*), võimendamine (*boosting*). Tuntumad ansambelmeetodid nagu otsustusmets ja gradientvõimendatud puud (*gradient boosted trees*) kombineerivad otsustuspuid. Samuti on kasutatud difusioonimudeleid omakorda neurovõrgu parameetrite genereerimiseks [1]. Ansambelõpet nimetatakse ka metaõppeks (*meta-learning*).



Joonis 6. Ansambelmeetodiga kombineeritakse erinevaid masinõppemudeleid

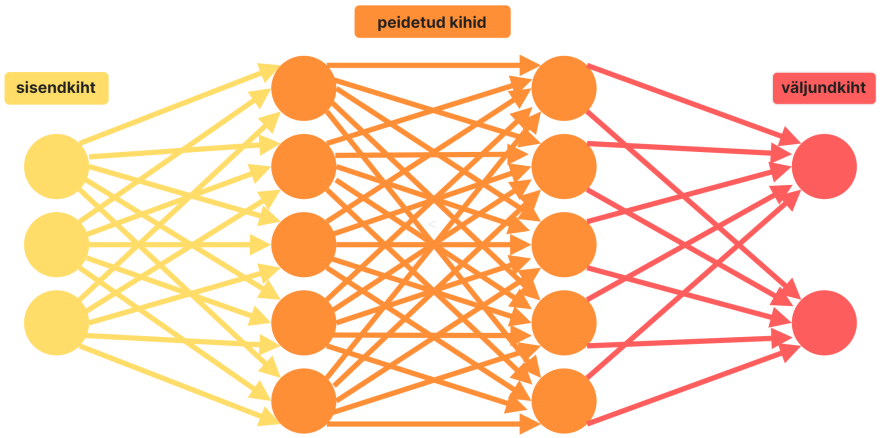
2.2.3 Tehisneurovõrgud

Tehisneurovõrgud on masinõppemudelid, millega üritatakse jäljendada inimaju mõtlemisvõimet. Neurovõrgud koosnevad kihtidesse jaotatud sõlmededest, mille käitumine peaks sarnanema ajus olevate neuronitega. Kuigi esimesed tehisneurovõrgud loodi juba 1950ndatel, algas nende võidukäik alles umbes kümme aastat tagasi, kui suudeti ehitada esimesed konvolutsioonilised neurovõrgud, mis suutsid pilditöötluses ja näotuvastuses saavutada paremaid tulemusi kui mistahes muu olemasolev algoritm.

Arvutusjõudluse kasvamisega ja odavnemisega on treenitavaks muutunud suured ja keerulised tehisneurovõrgud, mis on nii nende uurimises kui rakenduses kaasa toonud võidujooksu. Praegu suudavad tehisneurovõrke kasutavad mudelid lahendada ülesandeid, mida peeti veel paari aasta eest võimatuks. Neurovõrgud on üldiselt raskesti seletatavad ning treenitud mudeleid peetakse mustadeks kastideks. Seetõttu eelistatakse sarnase ennustusvõimega mudelite korral neurovõrkudele tihti lihtsamini seletatavaid masinõppemudeleid. Neurovõrkude seletavuse uurimine on aktiivne teadusharu.

2.2.3.1 Tehisneurovõrkude arhitektuurid

Täissidus neurovõrk (*fully connected network*, joonis 7) on üks esimestest välja töötatud neurovõrguarhitektuuridest. Täissidus võrk koosneb järjestikustest täissidusatest kihtidest, mis omakorda koosnevad lineaarsetest sõlmedest, mille väljunditele rakendatakse mittelineaarseid aktivatsioonifunktsioone.



Joonis 7. Tehisneurovõrgud koosnevad erinevatest kihtidest ja sõlmedest.

Konvolutsiooniline neurovõrk (*convolutional neural network*, CNN) on neurovõrk, mille üheks või mitmeks peidetud kihiks on konvolutsiooniline kiht. Võrreldes täissidusate kihtidega, kus igale sisendväärtusele seatakse vastavusse lineaarne sõlm ehk kaal, koosneb konvolutsiooniline kiht pisikestest tuumadest/filtritest, mistõttu on kihid väiksemad ja nende abil saab ehitada sügavamaid (rohkemate kihtidega) neurovõrke.

Konvolutsiooniliste neurovõrkude tuntuim rakendus on tehisnägemine. Näotuvastusel suudab konvolutsiooniline võrk leida kihi haaval erinevaid tunnuseid, alustades joontest ja nurkadest, seejärel silmadest ja suust ning lõpetades inimnäoga. Konvolutsiooniliste võrkude võidukäik sai alguse 2012. aastal AlexNetiga, mis suutis pildituvastuse võistlusel ImageNet Large Scale Visual Recognition Challenge teisi võistlejaid suure edumaaga võita. Sellest saati on konvolutsioonilised võrgud olnud peamised tehisnägemise tööriistad. Võrke kasutatakse edukalt ka tekstitöötlusel

ja vähemal määral muudel eriülesannetel.

Nii täissidusad kui konvolutsioonilised neurovõrgud on näited pärilevivõrkudest, milles peitekihi väljund on järgmise kihi sisendiks ehk informatsioon liigub mööda võrku läbi kihtide ainult ühes suunas. Kui neurovõrgus võib informatsioon liikuda ka tsükliliselt, st kihi väljund suunatakse võrku tagasi ning mõjutab sama kihi hilisemat sisendit, siis nimetatakse vastavat neurovõrku rekurrentseks neurovõrguks. Rekurrentseid neurovõrke kasutatakse peamiselt jadaliste andmete analüüsiks, sest nad suudavad treenimisel arvestada sisendi korral ka sellele jadas eelnevaid sisendeid. Rekurrentseid neurovõrke kasutatakse laialdaselt näiteks keelemudelites, teksti genereerimisel, kõnetuvastusel, tehisnägemisel, videote sildistamisel.

Rekurrentsete neurovõrkude treenimise muudab ebastabiilseks gradientide plahvatuslik kasv või kahanemine tagasilevi jooksul. Probleemi leevendamiseks on kasutusele võetud pika lühiajalise mälu (*long short-term memory*, LSTM) neurovõrgud, mis moodustavad rekurrentsete võrkude jaotise. LSTM-ide keskmes on mälurakud (*cell*), mida läbivat infovoogu mõjutavad sisend-, väljund- ja unustusventiilid (*input*, *output*, *forget gate*), mis aitavad tagasilevi jooksul vältida gradientide plahvatuslikku kasvamist või kahanemist.

Transformerid on süvaõppe mudelid, mis kasutavad tähelepanu mehhanisme järjestikuliste andmete analüüsimiseks. Transformerite võidukäik algas 2017. aastal, mil näidati, et nad suudavad leida loomuliku keele töötlusel ühele sõnele vastava konteksti eelneva jada põhjal ilma seda jada iteratiivselt analüüsimata. Teatud pikkusega sisendit vaadatakse korraga ning tähelepanu mehhanismi kasutades leitakse eelnevast sõnede järjendist igale sõnele olulisemad signaalid. See võimaldab mudelit treenida paralleelselt, mis vähendab arvutusnõudlust võrreldes näiteks LSTM-idega.

Transformer ei suuda erinevalt LSTM-idest pikkade sisendite korral võtta arvesse kogu eelnevat jada vaid ainult selle kindla pikkusega osa, mis võib pikkade tekstide analüüsimisel osutuda probleemiks. Transformerid koosnevad üldiselt kooderist ja dekooderist, millest esimene analüüsib sisendit ja teine genereerib samm-sammult väljundi. Kooder ja dekooder on kasutatavad nii koos kui eraldi. Näiteks on GPT vaid dekooderi-põhine ja BERT vaid kooderi-põhine mudel, kuid leidub ka mudeleid nagu T5, mis kasutab nii kooderit kui dekooderit.

Transformereid kasutatakse nii juhendatud, juhendamata, kui ka hübriidmudelite treenimisel. Suuri keelemudeleid nagu BERT ja GPT treenitakse esmalt juhendamata suure tekstihulga peal. Seejärel treenitakse mudelit kindla ülesande jaoks väiksema märgendatud andmehulga peal. Transformeritel põhinevad mudelid on saavutanud viimaste aastate jooksul keelemudelite seas peaaegu täieliku ülemvõimu. Teistes rakendusvaldkondades pole samasugust edu veel saavutatud. Näiteks tehisnägemises eelistatakse transformeritele veel konvolutsioonilisi neurovõrke, kuigi ka nendes kasutatakse juba tähelepanu mehhanisme.

Autokooder (*autoencoder*) on juhendamata tehisneurovõrk, mis koosneb nii kooderist kui dekooderist. Kooder saab sisendi ja transformeerib selle teisele kujule ja dekooder üritab transformeeritud sisendist esialgse sisendi rekonstrueerida. Treenitud kooderit saab kasutada sisendandmete mõõdete vähendamiseks, dekooderit aga andmete genereerimiseks. Üldiselt on autokooderi generatiivsed võimed piiratud, sest dekooderi sisendite lähedus ei garanteeri väljundi sarnasust.

Andmete genereerimiseks on välja pakutud **variatsiooniline autokooder** (*variational autoencoder*, VAE), mida saab kasutada näiteks pildisünteesiks. VAE erineb tavalisest autokooderist selle poolest, et kooder ei projekteeri sisendit punktiks vaid hoopis jaotuseks, näiteks väljastab kooder normaaljaotuse keskväärtuse ja kovariatsioonimaatriksi, dekooder saab treenimise ajal sisendina ette juhusliku vektori sellest jaotusest ning üritab kooderi algse sisendi rekonstruee-

rida. Erinevalt autokooderitest annavad treenitud VAE dekooderid tavaliselt lähedaste sisendite korral sarnased väljundid.

Vastandgeneratiivne võrk (*Generative adversarial network*, GAN) on generatiivne mudel, mille treenimisel võistlevad omavahel kaks neurovõrku – generatiivne ja diskriminatiivne. Mõlemaid neurovõrke treenitakse samaaegselt, generatiivne mudel saab sisendi mõnest lihtsast jaotusest ja üritab selle põhjal genereerida väljundit keeruliselt kirjeldatavast jaotusest, diskriminatiivne mudel üritab eristada generatiivse mudeli väljundeid tõelistest andmetest, mille jaotust generatiivne mudel imiteerida üritab. GAN-e saab kasutada näiteks pildisünteesiks, kus generatiivne mudel genereerib inimeste pilte ja diskriminatiivne mudel üritab eristada päris pilte genereeritutest. Vastandgeneratiivseid võrke kasutatakse ka kõne- ja tekstisünteesis.

Difusioonmudel on generatiivne mudel, mis põhineb Markovi protsessidel. Difusioonmudelid sarnanevad mõneti autokooderitega, nad koosnevad pärisuunalisest protsessist, kus pärisandmetele lisatakse sammhaaval müra, ja vastassuunalisest protsessist, kus algset sisendit üritatakse taasluua müra järk-järgult eemaldades. Üldiselt genereeritakse difusioonmudelite treenimisel müra normaaljaotusest ning piisavalt sellist müra lisades kaob algne sisend täielikult, väljundina jääbki järele juhuslik müra.

Kui sisendile lisada palju müra korraga, siis on algse sisendi ennustamine ülimalt keeruline, osutub aga, et kui müra lisatakse piisavalt väikeste sammudena, siis saab viimati lisatud müra ennustada ja seejärel eemaldada, ennustamiseks saab kasutada näiteks tehisneurovõrku. Treenitud mudelit saab järjestikuselt rakendada täiesti juhuslikule sisendile ning seeläbi genereerida pärisandmete sarnane väljund.

Treenimise teeb efektiivseks teadmine, et kui genereerida normaaljaotusest müra ja seda järjestikuselt lisada, siis on kogu lisatud müra samuti normaaljaotusest. Seega saab treenimise ajal lisada algsele sisendile mitme väikse sammu summeeritud müra korraga ja lasta neurovõrgul ennustada ainult vähene viimasel sammul lisatud müra.

Difusioonmudelid said alguse statistilisest füüsikast. 2015. aastal näidati, et neid saab kasutada ka pildisünteesiks. Neid edasi uurides on jõutud arusaamale, et difusioonmudelid on võimsad ja stabiilsemad, kuid samas vähem ressursse nõudvamad, kui näiteks vastandgeneratiivsed mudelid, mis olid seni parimad pilte genereerivad mudelid. Praegu on difusioonmudelid ja transformerid peamised komponendid tekstist piltide (*text-to-image*) sünteesi mudelites nagu DALL-E 3 ja Stable Diffusion.

2.2.4 Suured keelemudelid

Suured keelemudelid (*large language model*, LLM) on enamasti transformeripõhised tekstisünteesimudelid, mis eristuvad oma parameetrite suure arvu ja treeningandmete mahu poolest. On ka keelemudeleid, mis ei ole transformeripõhised. On välja töötatud mitmeid arhitektuure, näiteks RetNet[2], RWKV[3] ja Mamba[4], mida võib samuti rakenduda keelemudelite loomisel ning mis pakuvad lahendusi transformerarhitektuuri nõrkustele. Suur osa viimaste aastate innovatsioonist masinõppes ja tehisintellektis on tulnud just LLMide arengust ja LLMide peale ehitatud toodete (sealhulgas ChatGPT) levikust.

On esitatud hüpotees, et juhul kui tehislik üldintellekt (*Artificial General Intelligence - AGI*) on võimalik, võib see tekkida multimodaalsete suurte keelemudelite baasil[5]. Demis Hassabis tehisintellekti arendavast ettevõttest DeepMind on öelnud, et ”multimodaalsed alusmudelid saa-

vad olema AGI võtmekomponendiks”[1]. AGI käsitlused ja definitsioonid siiski varieeruvad ning mõne järgi on selleks vajalik tase juba saavutatud[6].

2.2.4.1 Treenimine

Nagu teistegi tehisintellekti mudelite puhul, tuleb mudeliarhitektuuri paikapanemise järel mudelit treenida. LLMide treenimine on harilikult mitmesammuline, kuid ükski samm ei ole rangelt kohustuslik. LLMide treenimisprotsess ja selles tehtud valikud on tihedalt seotud nende peale ehitatud AI rakenduste levitusmudelitega.

Eelõpe ehk *pre-training* on treenimise esimene, juhendamata etapp, kus mudelile söödetakse maskeeritud elementidega tekstijadad ning lastakse neid elemente ennustada. Maskeeritud elemente valitakse automaatselt. Eelõpe on kõige arvutusnõudlikum protsess, kus kasutatakse väga suurt hulka ($\sim$triljon sõnet) märgendamata, madala kvaliteediga andmeid, mis on harilikult saadud veebisorimise (*crawling*) läbi. Eelõppe tulemusel valmib eeltreenitud mudel, mis suudab genereerida sisendile treeningandmetes nähtu põhjal jätku. See jätk ei pruugi olla kasulik: esitades eeltreenitud mudelile küsimuse, võib ta genereerida sellele vastuse, kuid ka täienduse või jätkuküsimused.

Juhendatud **peenhäälestus** (*supervised finetuning*, SFT) on treenimise teine etapp, mis häälestab mudelit spetsiifiliseks otstarbeks. Näiteks vestlusrobotite puhul soositakse küsimustele just vastuste, mitte muude väljundite genereerimist. Peenhäälestuses kasutatavad treeningandmed on sageli, kuid mitte alati, inimeste poolt koostatud ja märgendatud. Need on eelõppes kasutatud andmetest kõrgema kvaliteediga ja palju väiksema mahuga ($\sim$kümned tuhanded näidispaarid).

Inimtagasisidega stiimulõpe (*reinforcement learning with human feedback*, RLHF) on treenimise kolmas, stiimulõppel põhinev etapp, kus mudel häälestatakse inimeelistuste järgi. Selleks luuakse preemiamudel (*reward model*), mida rakendatakse seejärel peenhäälestatud mudeli väljundite hindamiseks. Preemiamudelit treenitakse kasutades inimese kaasabil koostatud andmestikku, kus igale päringule seatakse vastavusse 1 või enam $(hea\_vastus, halb\_vastus)$ paari, seejuures üritatakse iga paari puhul maksimeerida preemiamudeli poolt antava hea ja halva vastuse hinnangu vahet. Kui preemiamudel on õppinud eristama soovitavaid vastuseid mittesoovitavatest, kasutatakse seda stiimulõppe käigus SFT läbinud mudeli täiendavaks peenhäälestamiseks.

Otsene eelistuste optimeerimine ning identiteedi eelistuse optimeerimine (*direct preference optimization - DPO, identity preference optimization - IPO*) on alternatiivsed lähenemised peenhäälestamisele, kus sarnaselt RLHFile rakendatakse eelistusõpet inimeelistuste andmestiku põhjal. Lähenemiste eripära seisneb selles, et erinevalt RLHFist ei ole DPO/IPO korral vaja luua preemiamudelit, sest preemiamudeli funktsiooni täidab LLM ise[7, 8], võttes kaofunktsiooniks heade ja halbade vastuste hinnangute vahe. Kui vaid eeltreenimise läbinud mudel võib anda asjassepuutumatuid või ohtliku sisuga vastuseid, siis SFT ja RLHF/DPO/IPO on treenimise sammud, kus mudelit saab treenida inimjuhendamisega turvalisemaks ja ärivajadustega kooskõlalisemaks.

2.2.4.2 Inferents ja kontekstisõpe

Viip (*prompt*) on kasutajapoolne sisendsõne, mille põhjal koostab generatiivne pilt- või keelemudel väljundi. Seda protsessi nimetatakse **inferentsiks**. Viip koosneb tavaliselt loomulikus keeles koostatud tekstist. LLMidel põhinevate vestlusrobotite viibale liidetakse ka eelviip (*pre-prompt*),

[1] The Guardian: 'Google says new AI model Gemini outperforms ChatGPT in most tests”. https://www.theguardian.com/technology/2023/dec/06/google-new-ai-model-gemini-bard-upgrade Külastatud: 11.12.2023

mis sisaldab täiendavat infot vestluse konteksti, kasutaja ning ka keelemudeli kohta. See on oluline muuseas selleks, et juturobot lähtuks väljundis oma rollist juturobotina, kes vastab küsimustele, selle asemel et genereerida jätku kasutaja sisendile. Eelviibaga saab kaasa anda teavet välismaailma kohta, näiteks kuupäeva, kellaaja, kasutajanime, dokumendi või tekstifaili sisu või muid kasutaja või keskkonna tunnuseid.

Mudelid ei suuda eristada viipa eelviibast, ning see on asjaolu, mida kasutavad ära paljud viibasüstimise tehnikad. Et eelviipa on oskusliku viipamisega võimalik kasutajal kergesti pärida, ei tohi see kanda infot, millele kasutajal ei tohiks ligipääsu olla. Viip peab transformerarhitektuuri korral koos eelviibaga mahtuma mudeli **kontekstaknasse**, mille laiust mõõdetakse *tokenite* ehk sõnede arvus, ning mis sisaldab väljundi genereerimiseks vajalikku (eel-)infot. Selle lahenduse keerukamaks vormiks on päringgenereerimine ehk RAG (*Retrieval-Augmented Generation*), mille puhul keelemudel koostab kasutaja päringu ja eelviibas oleva rakendusliidese info põhjal andmebaasipäringu ning tugineb vastuse koostamisel selle tulemitele. See võimaldab lahendada probleemi, mis tekib siis, kui kasutaja andmed on liiga mahukad, et manustada neid viiba kaudu kontekstaknasse. On ka mudeliarhitektuure, kus viiba suurus ei ole piiratud, nagu näiteks Mamba[4] ja RWKV[3].

Kui lihtsamaid keelemudeleid on vaja iga uue ülesande jaoks kas ümber treenida või peenhäälestada, siis LLMide keeleoskus ja üldistusvõime võimaldab piirduda paljude ülesannete korral ülesande sõnastamise ja paari näite lisamisega viibale [9]. Et ülesannet puudutav info söödetakse mudeli kontestaknasse, nimetatakse sellist lähenemist **kontekstisõppeks** (*in-context learning*).

Kontekstisõpe jaguneb mitmeks alamlähenemiseks: multinäideõppeks (*few-shot learning*), kus viipa lisatakse lisaks juhistele mitmeid näiteid, ainunäideõppeks (*one-shot learning*), kus lisatakse vaid üks näide ning näitetuks õppeks (*zero-shot learning*), kus päringut teostatakse näiteid andmata. Mida rohkem on keelemudelis parameetreid, seda vähem näiteid on ülesande edukaks täitmiseks vaja tavaliselt viipa lisada.

2.3 Tehisintellekti rakendused

Pildisüntees. Pildisüntees on etteantud omadustega, nt tekstilisel kirjeldusel (või teisel pildil ja tekstilisel kirjeldusel) põhineva pildi automaatne tekitamine. Pildisünteesi alamülesanded on pildirikastus (*inpainting*), pildilaiendus (*outpainting*), stiiliülekanne, sügavõppe-põhine müraeemaldus, videosüntees ja peenendus (teralisuse suurendamine). Tänapäeval kasutatakse pildisünteesis vastandgeneratiivseid võrke [10] ning üha enam difusioonipõhiseid mudeleid[11, 12].

Tehisnägemine. Tehisnägemise ülesanne on piltidelt info automaatne eraldamine. See hõlmab pildi klassimaskimist (*class segmentation*) ja isendimaskimist (*instance segmentation*), märgendamist ning objektituvastust. Tehisnägemises kasutatakse enamasti konvolutsioonilistel neurovõrkudel (CNNidel) ja transformeritel põhinevaid sügavõppe mudeleid [13, 14]. Levinud kasutused on kariloomade või põllumasinate jälgimine, teeolude ja ümbruse jälgimine isesõitva auto või pakiroboti poolt, näotuvastus ning liitreaalsus.

Kõnesüntees. Kõnesünteesi ülesanne on inimesele mõistetava kõne tekitamine etteantud tekstilõigu põhjal. Algelisemad kõnesünteesi mudelid toimisid jadamisi ühendades eelnevalt salvestatud foneeme või sõnu, kuid nüüd kasutatakse enamasti transformeritele tuginevaid neurovõrke[15, 16]. Kõnesünteesi kasutatakse vestlusrobotites, automatiseeritud sõnumite ettelugemisel, ekraanilugerites, arvutimängude lokaliseerimises ning dubleerimises. Kõnesünteesi alamülesandeks on kõnestiili ülekanne, mis tähendab näidiskõne kõla ja kõnemaneeri jäljendamist.

Kõnetuvastus. Vastandina kõnesünteesile on kõnetuvastuse eesmärgiks info eraldamine inim-

kõnest. Kõnetuvastuse alla käib kõnetranskriptsioon, mille puhul on kõnest eraldatav info tekstiline. Kui varasemad kõnetuvastuse mudelid kasutasid statistilisi meetodeid, siis tänapäeval kasutatakse peamiselt CNNe ja transformereid rakendavaid tehisneurovõrke[17]. Kõnetuvastust kasutatakse nutikodudes ja käed-vabad seadmetes käskude andmiseks ning dikteerimisel.

Loomuliku keele töötlus. Loomuliku keele töötlus on lai valdkond, mis puudutab nii teksti genereerimist, liigitamist kui ka tõlgendamist. Tekstigenereerimine tähendab üldiselt järgmise sõne ennustamist, kus eelnevad sõned on ennustuse kontekstiks. Teksti liigitamine ja tõlgendamine on kasutusel semantilises otsingus, kus dokumendi või tekstilõigu otsimisel ei võrrelda kandidaate sõnalise klappivuse, vaid tähendusliku läheduse alusel. Varasemalt kasutati tekstisünteesiks rekurrentseid neurovõrke (RNN) ja pikka lühimälu (LSTM) sisaldavaid sügavõppe neurovõrke. Valdkonnas toimus läbimurre suurte keelemudelite (LLM) tekkega, mis kasutavad oma arhitektuuris peamiselt transformereid[18, 19, 20]. LLMid on kasutusel nt. turundustekstide kirjutamises, vestlusrobotites, neurotõlkes, tundeanalüüsis ja koodigenereerimises.

Üldine andmetöötlus ja -analüüs. Masinõppe meetodeid kasutatakse ka eelmainitud rakendustega mitteseotud andmeanalüüsis. See võib hõlmata mitmekesiseid liigituse, klasteranalüüsi, diskreetse või pideva tunnuse ennustamise ülesanded, näiteks aktsia turuhinna kõikumise ennustamine, aju-arvuti liidese kogutud ajutegevuse signaalide töötlemine või klientide klasteranalüüs tarbijakäitumise põhjal. Olenevalt ülesande iseloomust on neile võimalik läheneda nii sügavõppe neurovõrkude kui statistilise masinõppe meetoditega.

2.4 Tehisintellekti kasutusvaldkonnad

Eelmainitud tehnikad on leidnud rakendust paljudes eluvaldkondades: e-riigis, erasektoris, hariduses ja teaduses, tervishoius ning selgelt määratlemata isiklikus kasutuses. Järgnevalt on toodud mõned selliste tehnikate rakendused valdkonniti.

E-riik ja e-valitsemine. Majandus- ja Kommunikatsiooniministeeriumi kratikavad on näinud ette tehisintellekti ulatuslikku rakendamist avalikus sektoris. Loomuliku keele töötlusel põhinev virtuaalabiline Bürokratt võimaldab suhelda avaliku sektori teenustega vestlusakna kaudu. Kõnetuvastuse abil transkribeerib riigikogu digistenografist Hans istungisaalis kõneldu. Tekstianalüütika töövahendi Texta Toolkit abiga on mitmed ministeeriumid viinud läbi oma dokumentide auditeid. Rahvusarhiivi teenus Ilme võimaldab tehisnägemise abil leida ajalooliste fotode seast kasutaja üleslaetud piltidele sarnaseid isikuid.

Haridus. Tehisintellektil on palju rakendusi hariduse[2] valdkonnas. Haridusliku suunitlusega mittetulundusasutus Khan Academy kasutab GPT-4 keelemudelile tuginevat vestlusrobotit õppe isikupärastamiseks. Keeleõpperakendus Duolingo sisaldab sarnast GPT-4-põhist interaktiivset vestlusrobotilahendust, loomuliku keele töötluse meetodeid kasutatakse keeleõpperakenduses Lingvist.

Teadus. Tehisintellekti ja masinõpet on teaduses rakendatud nii uute teadmiste avastamiseks kui olemasoleva info otsimiseks ja süstematiseerimiseks[3] : otsinguportaal SemanticSearch kasutab loomuliku keele töötlust ja tehisnägemist teadustööde kokkuvõtmiseks, indekseerimiseks ja otsimiseks, Alphabeti tehisintellektiprogrammi AlphaFold abiga on suudetud ennustada senitundmata struktuuriga valkude kuju. Masinõppel ja tehisintellektil tuginevad mudelid on kasutusele võetud osakestefüüsikas andmeanalüüsis ja simulatsioonide loomisel ning biomeditsiinis

[2] Haridus- ja Noorteamet, Hariduse tehnoloogiakompass. https://kompass.harno.ee/tehisintellekt Külastatud: 10.08.2023

[3] OECD, Artificial Intelligence in Science. https://www.oecd.org/publications/artificial-intelligence-in-science-a8d820bd-en.htm Külastatud: 10.08.2023

uute ravimite väljatöötamisel.

Tervishoid. Tehisintellekti on edukalt rakendatud nii personaalmeditsiinis, kliinilistes uuringutes kui ka ravimite väljatöötamisel[4]. Masinõppel põhinevad suurandmete analüüsimeetodid võimaldavad kasutada patsiendi geeniandmeid parema ravi tagamiseks. Tehisnägemine on abiks meditsiiniliste ülesvõtete tõlgendamisel ja patsiendi diagnoosimisel. Loomuliku keele töötluse ja tekstianalüüsi meetodid võimaldavad otsida ja korrastada patsiendiandmeid. Masinõppemeetodeid kasutatakse ravimite väljatöötamisel, seda nii molekulisimulatsioonides, ravimiomaduste ennustamisel ja molekulistruktuuri ning sünteesiradade genereerimisel.

Erasektor. Telekommunikatsioonis kasutatakse masinõpelahendustel põhinevaid helitöötluse, müraeemalduse ning heli- ja piltvoo pakkimistehnikaid (Skype). Tehisnägemist kasutatakse näiteks robootikas (Milrem, Cleveron), põllumajanduses, isikutuvastuses (Veriff). Klienditeeninduses on levinud loomuliku keele töötlusel põhinevad vestlusrobotid.

Isiklik kasutus. Juba enne LLMide ja difusioonipõhiste pildisünteesimudelite teket olid levinud tehisintellektipõhised personaalabilised, nagu Google Assistant, Amazon Alexa ja Siri. LLMide ning difusioonipõhiste pildisünteesimudelite levik ja tavatarbijale kättesaadavamaks muutumine on põhjustanud selles valdkonnas arenguhüppe, sealhulgas AIaaS (*artificial intelligence as a service*, tehisintellekt teenusena) ärimudeli leviku. Isiklikuks tarbeks mõeldud mudelid ja nende ümber ehitatud pluginad ja rakendused suudavad analüüsida koodi (GitHub Copilot), lugeda dokumente või veebilehti ja eraldada nendest vajalikku infot (Bing Chat), genereerida tekste sünnipäevakutsetest turundusmaterjalideni (ChatGPT).

Pildisünteesimudelite abil saab kasutaja luua meelepärases stiilis illustratsioone, genereerida ideid sisustuse jaoks, suurendada piltide või fotode lahutusvõimet (StableDiffusion, Midjourney) ning tunda metsas ära mõningaid seeneliike.

2.5 Seletavus masinõppes

Sügavõppe meetodite tekkimisega ja masinõppemudelite keerulisemaks muutumisega on kerkinud küsimus mudelite seletavusest. Mudeli seletavus tähendab võimet inimesele arusaadaval moel seletada, kuidas sõltub mudeli väljund selle sisendist. Kui automatiseeritud otsuseid tehakse masinõppemudeli abil, siis on Euroopa Liidu andmekaitseõiguse seisukohast oluline kasutatava tehisintellekti tehnoloogia läbipaistvus [21]. Just mudeli seletavus on omaduseks, mis aitab seda saavutada.

Seletuslikku intellektitehnikat (*explainable AI*, XAI) on pakutud lahendusena, mis võib aidata liikuda läbipaistvama tehisintellekti poole ja vältida seeläbi tehisintellekti kasutuselevõtu piiramist kriitilistes valdkondades [22]. Globaalsel tasandil puudub aga uuringu koostamise ajal konsensus, milline peaks olema soovitud algoritmilise selgitatavuse lävend [23].

Seletavus on tihedalt seotud tehisintellektisüsteemide läbipaistvuse ja usaldusväärsuse aspektidega. Seega on seletavusnõuete süstemaatiline määratlemine läbipaistvate ja usaldusväärsete tehisintellektisüsteemide väljatöötamise oluline samm [24]. OECD on leidnud [25], et läbipaistvuse ja seletavuse tagamiseks peaksid tehisintellektisüsteemi osalised andma sisulist, kontekstile ja tehnikatasemele vastavat teavet järgmistel eesmärkidel:

- edendada üldist arusaamist tehisintellektisüsteemist;
- teavitada sidusrühmasid interaktsioonidest tehisintellektiga;

[4] Tervise Arengu Instituut, Tehisintellekt kui personaalmeditsiini alus onkoloogias. https://www.tai.ee/et/personaalmeditsiini-uudiskirjad/tehisintellekt-kui-personaalmeditsiini-alus-onkoloogias Külastatud: 11.08.2023

- võimaldada tehisintellektisüsteemist mõjutatud osapooltel mõista süsteemi tulemusi;
- võimaldada tehisintellektisüsteemist negatiivselt mõjutatud isikutel süsteemi tulemusi vaidlustada, andes lihtsal ja arusaadaval viisil teavet süsteemi tegurite ja tulemuse aluseks olnud loogika kohta.

Uuritud on ka seda, kuidas on seletavusnõudeid praktikas defineeritud [24]. Leitud on, et tehisaru seletavusel on muuhulgas abiks süstemaatiliste definitsioonide kehtestamine ning selgituste ja tulemusnäitajate formaliseerimine ja kvantifitseerimine [22]. Esitatud [24] on neli seletavuse komponenti:

- adressaat – kellele seletada?
- aspekt – mida seletada?
- kontekst – millises olukorras seletada?
- selgitaja – kes seletab?

Seletatav mudel on usaldusväärsem, seda on kergem arendada, testida ja auditeerida, samuti on kergem tuvastada selle erapoolikust ja seletada anomaalset käitumist. Seletavus on elutähtis meditsiinis, kus nt vähkkasvajat tuvastav pildimudel on usaldusväärsem, kui ennustusega kaasneb ka seletus, milliste pildiomaduste (kontrast, kuju) tõttu tuvastati või ei tuvastatud kasvajat. Samuti võib pakkuda pangast negatiivse otsuse saanud laenutaotlejale huvi, mida ta peaks tegema, et pank talle laenu ikkagi annaks (nö kontrafaktuaalne seletus). Kui roppuste filter tõstab esile need sisendsõnad, mis panustavad enim sõnumi ebatsensuurseks klassifitseerimiseks, on seda sellevõrra kergem arendada ja testida.

Seletavust ei ole alati vaja. Kui riskid on madalad ja probleem on põhjalikult juba uuritud, võib see osutuda ülearuseks. Samuti esineb mudeli võimekuse ja seletavuse vahel reeglina lõivsuhe[26]. Kui lineaarse regressioonmudeli puhul piisab väljundi ja sisendi seose seletamiseks regressioonikoefitsentidele pealevaatamisest, siis keerulisemad ja võimekamad mudelid, nt sügav neurovõrk, on inimesele nö ”mustaks kastiks” [27], kus mudeli ennustamise või otsustamise põhimõtteid ei ole enam võimalik mudeli ülesehituse ega parameetrite põhjal hoomata.

Seletavust saab jagada siseseletavuseks (*intrinsic*) ja järelseletavuseks (*post-hoc*). Siseseletavuse (ka läbipaistvuse) korral piiratakse mudeli keerukust, et ta ei muutuks mustaks kastiks ja et selle parameetrid jääksid algusest peale ning kogu mudeli ulatuses seletatavaks. Iseseletavaiks peetakse lihtsa struktuuriga mudeleid, nagu otsustuspuid või lihtsaid regressioonimudeleid. Kui ülesanne nõuab keerukama mudeli kasutuselevõttu, kasutatakse selle läbipaistvamaks muutmiseks järelseletavuse meetodeid.

Järelseletavuse meetodid on üldiselt mudeli suhtes agnostilised - need ei sõltu mudeli arhitektuurist ega eelda enamasti ülevaadet selle sisemistest komponentidest. Need seletused peavad kõiki, ka oma lihtsuse tõttu olemuslikult tõlgenduvaid mudeleid mustadeks kastideks. Niinimetatud kohaliku järelseletavuse meetodid näitavad, kui palju ja mis suunas nihutavad väikesed, individuaalsed muutused sisendtunnustes mudeli väljundit, või millised on kõige väiksemad vajalikud muutused sisendtunnustes, et mudel ennustaks mõnda muud klassi. Globaalse järelseletavuse meetodid aitavad mõista juba treenitud mudeli vahekihte - näiteks OpenAI on loonud visualisatsioonide kogu Microscope[5], mille abil saab ülevaate erinevate pildimudelite vahekihtidest, neis olevatest neuronitest ja nende omadustest. Lisaks saab uurida, millised sisendandmetes esinevad pildid aktiveerivad vaadeldavat neuronit enim.

[5]OpenAI Microscope https://microscope.openai.com Külastatud: 10.12.2023

2.6 Rahvusvahelised trendid

2.6.1 Kiiremaks ja suuremaks

Mudelite suuruste kasv. Nii nagu arvutite jõudlus, on ka tehisneurovõrkude suurused kasvanud eksponentsiaalselt. Aastal 1989 kasutas Yann LeCuni meeskond piltidelt numbrite tuvastamiseks konvolutsioonilist neurovõrku. Võrk koosnes kahest konvolutsioonilisest ja ühest täissidusast kihist, kokku vähem kui kümnest tuhandest treenitavast parameetrist. Aastal 2012 esitletud AlexNeti mudel koosnes viiest konvolutsioonilisest ja kolmest täissidusast kihist ning sellel oli treenitavaid parameetreid juba 61 miljonit.

Joonis 8. Mudelite parameetrite kasv on olnud eksponentsiaalne

Treenitavate parameetrite arvu kasv jätkus transformerarhitektuuri levikuga (Joonis 8): keelemudelid BERT-base ja GPT-1 (2018) sisaldasid juba ∼110 mln, GPT-2 (2019) 1,5 mld ning GPT-3 (2020) 175 mld treenitavat parameetrit. GPT-4 parameetrite arvu pole avalikustatud, kuid spekuleeritakse, et tegemist on nn. ”ekspertide segu” (*mixture of experts* - MoE) mudeliga, kus on ∼1,76 triljonit parameetrit. Parameetrite arvuga kasvavad ka nõudmised arvutusvõimsusele ja mälumahule, mis on vajalikud nii mudeli treenimiseks kui ka juba treenitud mudeli rakendamiseks (inferentsiks). Samuti on mudeli tõhusamaks treenimiseks vajalikud suuremad treeningandmete mahud.

Parameetrite arvu kasvuga on keelemudelites täheldatud emergentseid (*emergent*) oskusi, mille all peetakse üldiselt silmas võimekusi, mis on olemas parameetrite arvult suuremates mudelites, kuid puuduvad väiksemates [28]. Näiteks on suuremad keelemudelid võimelised võtma kokku ning tõlkima tekste, genereerima koodi, leidma tekstist mustreid ja mõistma huumorit, sellal kui väiksemad piirduvad lihtsatele küsimustele vastamisega või grammatiliselt korrektsena näiva teksti genereerimisega. Emergentseks peetavaid oskusi on püütud seletada ka suuremate mudelite parema meeldejätmisvõime ning parema viiba kaudu juhendatavusega [29]. Mudelikaalude kvantimise ja mudeli hõrendamise (*pruning*) kasutuselevõtuni arvati, et sellised omadused tekivad keelemudelil ∼7 mld parameetrist alates ehkki teatuid emergentseid omadusi oli märgatud ka 1,5 mld parameetriga GPT-2 juures. Tänaseks on selge, et ka väiksematel või kokkupakitud mudelitel võivad need omadused teatud määral siiski olla.

Rohkemate parameetrite arvuga keelemudel vajab nende parameetrite tõhusaks ärakasutamiseks suuremat treeningandmestikku. Suuremad ingliskeelsed treeningandmestikud koosnevad mitmest triljonist sõnest, seevastu jääb eestikeelsete andmestike maht miljarditesse. See tähendab, et eesti keele põhjal treenitud keelemudel on tavaliselt väiksem ja vähem võimekas. Eesti keele

osakaal mitmekeelsetest andmestikest on väga väike, seetõttu ei pruugi sellel treenitud mudel keelt alati omandada. Üks strateegia selle erinevuse leevendamiseks on ingliskeelsel andmestikul treenitud mudeli peenhäälestamine eestikeelsete andmetega.

Riistvaranõuete kasv. Graafikaprotsessorites on kasutusele võetud SIMD (*single instruction multiple data*) arhitektuur, mis võimaldab üht tehet erinevatel andmetel läbi viia samaaegselt. See võimaldab oluliselt kiirendada renderdamise töövoogu ja muid arvutigraafikaga seotud ülesandeid, kus mingit tehet on vaja korrata iga puhvri elemendi kohta. See omadus ei jäänud märkamata sügavõppe neurovõrkude uurijatel, kes pakkusid 2009. aastal välja, et neurovõrkudes tihti esinevaid maatriksitega tehteid on graafikaprotsessorite abil võimalik kiirendada [30].

Transformeripõhised suured keelemudelid vajavad iga järgmise sõne genereerimisel juurdepääsu kõikidele mudeli kaaludele ja tähelepanuvektoritele (q, k, v), et liigutada neid muutmälust graafikaprotsessori registritesse. Piisavalt arvukate ja suurte kaalumaatriksite korral kasvab selle laadimisprotsessi ajakulu. See tähendab, et FLOPSide[6] kõrval on oluline ka mälu maht ja läbilaskevõime.

Erinevalt peenhäälestamisest ei nõua kontekstõpe inferentsile (ennustamisele) lisaks mudeli kaalude arvutuslikult kallist uuendamist. Mõningate LLMide kontekstõppe funktsionaalsust on võimalik kasutada ka võimsama personaalarvuti peal[7]. Et mudeli kaalud mahuksid personaalarvuti graafikaprotsessori mällu, kasutatakse kvantimist [31] – vähendatakse mudeli parameetrite täpsust ja mälunõudlikkust. Näiteks 32-bitiste ujukomaarvude asemel kasutatakse 16-bitiseid; kõige võimsamad kvantimistehnikad kodeerivad parameetrid ümber nii, et üks parameeter võtab pisut üle 2 biti mälu [32]. Kvantimise miinuseks on see, et mudeli võimekus võib selle arvelt kannatada.

Paralleelarvutust kasutavate valdkondade (masinõpe, simulatsioonid, teaduslik modelleerimine, krüptoraha kaevandamine) levik on tõstnud nõudlust nii ülesandeks sobiva riistvara kui püsivara järele. Näiteks on Nvidia arendanud välja CUDA platvormi, mis hõlmab nii riistvaralisi komponente kui ka programmeerimismudelit ja tarkvararaamistikku graafikaprotsessori kasutamiseks parallelarvutusega seotud ülesannetes. Apple oli loonud OpenCL paralleelarvutuse standardi, mis erinevalt CUDAst ei tuginenud konkreetsele riistvarale, kuid nüüdseks on nemadki läinud üle enda arendatud riistvaraspetsiifilise Metal raamistiku peale.

Tehisintellekti pilvteenusena pakkumiseks ei piisa enam tavapärastest serveriarhitektuuridest. Väga suured andmemahud tähendavad ka seda, et andmete hoiustamiseks ja töötlemiseks rajatakse eraldiseisvaid andmekeskuseid või kasutatakse pilvteenuseid. Nii inferentsiks kui treenimiseks on teenuse skaleerimisel mõistlik kasutada pilvetaristut ning spetsiaalset riistvara. Seejuures ei tähenda spetsiaalne riistvara enam pelgalt graafikaprotsessorit, vaid neurovõrkudele veelgi spetsiifilisemaid lahendusi nagu Google väljaarendatud tensorprotsessor (TPU) või nutitelefonides ja asjade interneti seadmetes kasutatav neuroprotsessor (NPU).

2.6.2 Üldotstarbelisest eriotstarbeliseks

Alusmudelitest rakenduste poole. LLMide puhul räägitakse sageli alusmudelitest (*foundation model*), mis on üldotstarbelised mudelid, mida saab kasutada paljude eri ülesannete täitmiseks. Juturobot on üks lihtsamaid alusmudeli rakendusviise, sest selleks piisab loomuliku keele mõist-

[6] FLOPS (*floating point operations per second*) on mõõtühik, mis loeb ujukomatehete arvu sekundis

[7] llama.cpp on vabavaraline rakendus, mis võimaldab kvantimise kaasabil LLaMA, LLaMA 2 ja teiste keelemudelite peal inferentsi käivitamist personaalarvuti peal

misest ning üldteadmistest, mis on tuletatavad mudeli kaaludest ega vaja eraldiseisvat andmebaasiliidest. Samuti on juturoboti puhul vastuvõetav mudeli mittedeterministlik väljund. Domeenispetsiifilistel rakendustel ei pruugi alusmudeli üldistusvõime ja teadmised olla ülesandega toimetulemiseks alati piisavad. Seetõttu on suurte alusmudelite peale ja kõrvale tekkinud eraldiseisvad lahendused ja mudelid. Need tulevad eriti hästi toime meditsiiniliste või juriidiliste tekstidega, võttes kokku pikki dokumente[8], programmeerimiskeelte ja programmeerimismustritega[9], piltide sisu äratundmisega[10] ja oskavad hinnata, kui tõenäoliselt on pilt või tekst pärit mõnest generatiivsest mudelist[11].

On tekkinud ka lihtsamaid rakendusi, mis liidestuvad rakendusliidese kaudu mõne olemasoleva AI mudeliga, näiteks PDF-failidega suhtlemiseks ja nende sisu kokku võtmiseks. Selliste õhukeste rakenduste äriline risk seisneb selles, et rakendusliideste ja mudelite pakkujad saavad selle funktsionaalsuse enda pakutavatesse toodetesse ise kergelt lisada, nagu on OpenAI teinud ChatGPTs PDF-failide analüüsiga[12].

Ühte tüüpi sisu sünteesist mitmekülgse sisu loomise suunas. Kui mudel puutub kokku erinevate sisendi või väljundi modaalsustega, siis seda saab nimetada multimodaalseks. See tähendab, et ka lihtsat piltide klassifitseerijat saab pidada multimodaalseks, sest ta võtab sisendiks pildi ja väljastab tekstimärgendi. Seda terminit kasutakse siiski eelkõige mudelite kohta, kus erineva modaalsusega sisendid projitseeritakse samasse sängitusruumi (embedding), näiteks OpenAI CLIP[13] ja GPT-4V[14]. On ka multimodaalseid tekstist-videoks mudeleid, kust viibale genereeritakse vastav pildijada, seda kas lähtepiltidele tuginedes[33] või ilma nendeta[34, 35].

Kui multimodaalne sisend on seni olnud lihtne, siis erineva modaalsusega väljundi tootmine on keerukam. Seni levinum (ja lihtsam) lahendus on eri mudelite väljundite ja sisendite ühendamine. Näiteks ChatGPT sisaldab pildigenereerimise funktsionaalsust, milles keelemudeli GPT-4 abil kasutaja päringu põhjal genereeritud tekstilised juhised söödetakse DALL-E 3 pildisünteesimudelile ning sealt saadud pildid tagasi kasutajale.

Invideo AI teenus[15] (ja mitmed teised samasugused) koostab tekstist videod: genereerib kasutajasisendi põhjal stsenaariumi ning otsib selle põhjal oma andmebaasist klippe ja pilte ja paneb nad videoks kokku, seejärel genereerib sinna peale heli.

Üks variant AI teenuste kokkuühendamiseks on AI agent (mõnikord ka generatiivne agent), mis on võimeline liidestuma erinevate teenustega, nt tegema internetipäringuid talle etteantud ülesande täitmiseks. AI agenti iseloomustab pidev tagasisidetsükkel päringute tegemise (väliskeskkonnaga liidestumise) ja oma oleku uuendamise vahel. Seetõttu on AI agendi jaoks väga oluline võime planeerida järgmisi samme, pidades samal ajal meeles eelmiste sammude tulemusi, oma olekut ning laiemat ülesande sisu ja eesmärki [36]. Isejuhtivat autot saab pidada AI agendiks.

Tänapäeval peetakse nende all silmas eelkõige suurtele keelemudelitele tuginevaid lahendusi, mis võimaldavad automatiseerida alamülesanneteks jaotamist, täiendavat planeerimist ning pidevat tagasisidet nõudvaid mitmesammulisi toiminguid loomulikus keeles antud juhiste põhjal.

[8] Claude 2: https://www.anthropic.com/index/claude-2
[9] GitHub Copilot X: https://github.com/features/preview/copilot-x
[10] Gpt-4Vision: https://openai.com/research/gpt-4v-system-card
[11] Stable Signature: https://ai.meta.com/blog/stable-signature-watermarking-generative-ai/
[12] ChatGPT Plus members can upload and analyze files in the latest beta. https://www.theverge.com/2023/10/29/23937497/chatgpt-plus-new-beta-all-tools-update-pdf-data-analysis Külastatud: 25.02.2024
[13] CLIP: Connecting text and images. https://openai.com/research/clip
[14] GPT-4V(ision) system card. https://openai.com/research/gpt-4v-system-card
[15] Invideo AI. https://invideo.io/

Mõned levinumad AI agentide loomise ja haldamise raamistikud on näiteks AutoGPT, BabyAGI ning AiAgent.App.

2.6.3 Suletumast avatumaks

Suletud mudelite pakkumise mudelid. Mida suuremaks tehisintellekti mudelid arenesid, seda kulukamaks muutus nende treenimine, haldamine ja levitamine. Mida võimsamaks nad muutusid, seda suuremaks kasvasid nende generatiivsete võimekuste kuritarvitamise riskid. OpenAI asutati 2015. aastal mittetulundusühinguna, mis seadis eesmärgiks tehisintellekti uurimise, rõhuasetusega sügavõppe neurovõrkudele[16]. Organisatsiooni alguspäevadel rõhutati avatust ja väärtuse loomist kogu ühiskonnale.

2019. aastal, mõni kuu pärast keelemudeli GPT-2 väljakuulutamist ja avalikustamist, võeti vastu otsus jagada organisatsioon "kasumi ülempiiriga" ettevõtte (OpenAI LP) ning senise mittetulundusühingu (OpenAI Nonprofit) vahel, mille juhatus jäi kahe uue partnerorganisatsiooni kontrollorganiks[17]. Sammu põhjendati kaasaegsete tehisintellektisüsteemide haldamise kulukusega: nende treenimine on arvutusnõudlik, samuti on kulukas mudelite treenimiseks kasutatavate suurandmete infotaristu ülalpidamine ning mittetulundusühingu võimalused kapitali kaasamiseks jäävad ettevõtete omadele alla. Sellele järgnes partnerlus Microsoftiga, mis investeeris ettevõttesse 1 miljard Ameerika dollarit ning 2023. aastal veel täiendavad 10 miljardit dollarit.

GPT-2 jäi OpenAI viimaseks täielikult avatud keelemudeliks. 2020. aastal kuulutas OpenAI välja GPT-3, kuid ei teinud treenitud mudeli parameetreid avalikult kättesaadavaks, vaid piiras ligipääsu rakendusliidese OpenAI API[18] kaudu ning litsentsis GPT-3 Microsoftile[19] eelnevalt sõlmitud koostööleppe raames. Otsus luua rakendusliides oli põhjendatud turvalisuse vajadusega ning majanduslikke kaalutlusi arvestades. Rakendusliidese haldajatena jääb OpenAIle otsustusvõime mudeli ligipääsu piiramiseks selle kuritarvitajaile, samuti oli rakendusliides OpenAI LP esimeseks kommertstooteks, mis aitas rahastada edasist uurimistööd ning pidada üleval kulukat serveritaristut.

Avalike mudelite teke. 2023. aastal kuulutas Meta välja oma keelemudelite seeria LLaMA[20], teh es mudelid kõigi üllatuseks täiel määral, sh kommertskasutuseks avalikult kättesaadavaks. Mõni kuu hiljem väljakuulutatud LLaMA 2 mudelisarja litsents piiras kommertskasutust vaid enam kui 700 miljonise aastase kasutajabaasiga ettevõtetele, kaitstes end seeläbi suurimate konkurentide eest. Samal ajal avaldas stability.ai generatiivse piltmudeli StableDiffusion[21] lähtekoodi ja parameetrid. GPT-2st oluliselt paremate avatud mudelite, nagu LLaMA 2 teke on vallandanud väiksemate, aga mingtes aspektides võimsamate ning valdkonnaspetsiifiliselt peenhäälestatud AI mudelite ulatusliku leviku, mis jäävad oma jõudluselt vaid napilt alla palju suurema parameetrite arvuga alusmudelitele. Headeks näideteks on Mistral-7B[22] või SSD-1B[23].

[16] OpenAI. https://openai.com/blog/introducing-openai Külastatud: 20.10.2023
[17] OpenAI LP. https://openai.com/blog/openai-lp Külastatud: 23.10.2023
[18] OpenAI API. https://openai.com/blog/openai-api Külastatud: 23.10.2023
[19] OpenAI licenses GPT-3 technology to Microsoft. https://openai.com/blog/openai-licenses-gpt-3-technology-to-microsoft Külastatud: 23.10.2023
[20] Introducing LLaMA: A foundational, 65-billion-parameter large language model. https://ai.meta.com/blog/large-language-model-llama-meta-ai/ Külastatud: 24.10.2023
[21] Stable Diffusion Public Release . https://stability.ai/blog/stable-diffusion-public-release Külastatud: 24.10.2023
[22] Mistral AI. https://mistral.ai/ Külastatud: 24.10.2023
[23] Announcing SSD-1B: A Leap in Efficient T2I Generation. https://blog.segmind.com/introducing-segmind-ssd-1b/ Külastatud: 24.10.2023

Asjaarmastajad, väikeettevõtted ja uurimisasutused ei saa endale lubada OpenAI, Google või Meta infotaristu- ja treenimiseelarveid, mistõttu on rõhuasetus nihkumas parameetrite arvult nende tõhusale kasutamisele, treeningandmete kvaliteedile ning alternatiivsetele mudeliarhitektuuridele. Nagu selgub Google'i lekkinud "We Have No Moat"[24] memost, on nende edu suurettevõtetele muret tekitanud. Tõhusamate ja odavamate peenhäälestusmeetodite teke, nagu LoRA [37] on võimaldanud asjaarmastajatel tehnoloogiahiidudega ebavõrdsest investeerimisvõimekusest sõltumata sammu pidada.

Ajendatuna ühelt poolt tehnoloogiahiidude soovist kasutada tehisintellekti kaasaskantavate seadmetega ning teiselt poolt väikeste ettevõtete ning vabavaralise kogukonna ressursinappusest, on nüüdseks tekkinud hulk piiratud parameetrite arvuga "väikeseid keelemudeleid", nagu Microsoft Phi-1.5 [38] ja Phi-2, Google Gemini Nano[25] ja Gemma [39], samuti Mistral 7B [40] ning väikesed Qwen1.5 perekonna mudelid [41], mis jäävad oma jõudluselt palju suurematele mudelitele vaid napilt alla.

2.6.3.1 Arengud levitusmudelites

Selleks, et AI mudel lahendaks mõnda ärilist ülesannet, ei piisa pelgalt mudelist. See mudel peab olema ühendatud sisendandmetega ja tagastama õigel kujul ja kvaliteetseid väljundandmeid. Levitusmudeli (*deployment model*) all peame silmas seda, kuidas on üles ehitatud AI rakendus, kuidas tehisintellekti mudel liidestub rakenduse teiste komponentidega ning kuidas voolavad nende vahel andmed (sh kasutajate isikuandmed).

Esimesed, lihtsamad tehisintellekti mudelid (nt. lineaarregressioon, pertseptron või reeglipõhised ekspertsüsteemid) ei olnud arvutusnõudlikud, mistõttu oli mudelit toetav infotaristu andmete omast vähemoluline. AI rakenduste levitusmudelitest on mõtet rääkida alates tehisintellekti laialdasest kasutuselevõtust 2010ndatel, kui kasvasid andmemahud, levisid tehisnärvivõrgud ning tekkis vajadus kiirendada nende treenimist ja inferentsi graafikakiirendite abil, mis ei olnud AI mudeli treenijale või kasutajale alati füüsiliselt kergesti kättesaadavad. Pilvetaristu pakkujad hakkasid andmete ja võrgunduse haldamise kõrvale pakkuma ka tehisintellekti mudeliteks vajalikku riistvara ning pilvetöötluskeskkondi (näiteks Google Colab, Amazon SageMaker), kuid kasutaja pidi veel vastutama mudelite ehitamise, treenimise ja kasutamise eest.

Mõni aeg hiljem tekkinud suurte tekstisünteesi- ja pildisünteesimudelite üldotstarbelisus tähendas, et teatud tööülesannete täitmiseks ei olnud oma mudeli nullist treenimine enam vajalik. Selle tulemusena sündis AIaaS ehk "tehisintellekt teenusena", mis võimaldab ettevõtetel ja eraisikutel kasutada suuri AI mudeleid ilma, et tuleks investeerida riistvarasse, AI mudelite treenimisse ja muusse infotaristusse.

ChatGPT ja AI rakendusliideste teke on vallandanud õhukeste "API ümbrisrakenduste" laviini, mis kasutavad ChatGPT või muu AI tekstisünteesilahenduse üldistusvõimet domeenispetsiifiliste ülesannete lahendamiseks. Mõned sellised rakendused ei paku midagi peale mugava kasutajakogemuse ja hoolikalt koostatud eelviiba, kuid selliste lahenduste reprodutseeritavus kujutab endast ümbrisrakenduse loojale suurt ärilist riski. See risk realiseerus OpenAI Dev Dayl, kus OpenAI tuli välja nö "kohandatud GPT" teenusega, mille abil on võimalik ehitada eriotstarbeline juturobot ühegi koodireata[26].

[24] Google: "We Have No Moat, And Neither Does OpenAI". https://www.semianalysis.com/p/google-we-have-no-moat-and-neither Külastatud: 26.10.2023

[25] Google Blogi: Gemini mudeli tutvustus https://blog.google/technology/ai/google-gemini-ai/ Külastatud: 14.12.2023

[26] Introducing GPTs. https://openai.com/blog/introducing-gpts Külastatud 20.11.2023

AI teenuspakkujate äriline nišš ei tugine reeglina innovatiivses mudeliarhitektuuris, sest need on enamasti avalikud, vaid mudeli ümber ehitatud infotaristus, lahenduse kasutajakogemuses ning valdkonnaga seotud treeningandmete kvantiteedis ja kvaliteedis. X (endise Twitteri) AI-teenus Grok omab reaalajas ligipääsu kasutajate postituste andmebaasile; Microsofti AI-koodiabiline Copilot X poleks GitHubi koodivaramu pidevalt täienevate andmeteta nii efektiivne; ChatGPTs on võimalik anda tagasisidet igale juturoboti vastusele, OpenAI on kogunud seeläbi palju väärtuslikke andmeid kasutaja ja juturoboti vahelisest suhtlusest, mis aitab veelgi parandada keelemudelite kvaliteeti.

Treeningandmete kvaliteedi haldamine on oluline, kuna andmete kvaliteedihaldamine võimaldab oluliselt vähendada võrdväärse mudeli treenimiseks vajalike andmete mahtu [42], samuti kuna sünteetilise sisu osakaal internetis on viimasel ajal hüppeliselt kasvanud ning ekspertide hinnangul kasvab 2026. aastaks 90 protsendini [43].

2.6.4 Reguleerimatusest reguleerituks

2.6.4.1 Tehisintellekti eetika

Arvutiteaduse eetika on mitmetahuline, käsitledes nii moraalseid kui ka eetilisi kaalutlusi, mis on seotud arvutustehnoloogiate, nagu näiteks tehisintellekti arendamise, juurutamise ja kasutamisega. Oluline on tagada, et selliseid tehnoloogiaid töötatakse välja ja kasutatakse viisil, mis on kooskõlas inimlike väärtustega, edendades ühiskonna heaolu [44]. Eetikapõhimõtted on dünaamilised, mis tähendab, et need võivad ajas muutuda, kohanedes ühiskonna ja teaduse arenguga [45].

Tehisintellekti tehnoloogiate rakendamine on tõusutrendis – selle turu suurus peaks 2027. aastaks ulatuma 407 miljardi dollarini [46]. Ka Eesti ettevõtted kasutavad üha enam selliseid tehnoloogiaid – 2023. aasta I kvartali seisuga on olnud kasv 2% võrreldes 2021. aastaga. Statistikaameti andmetel kasutavad tehisaru tehnoloogiaid Eestis enim finants- ja kindlustus, info- ja side ning energeetika valdkonna ettevõtted [47].

Kuigi tehisaru tehnoloogiatel on tohutu potentsiaal, tõusetuvad selle rakendamisega ka mitmed küsimused ja kartused. Näiteks 2023. aastal Inglismaal läbiviidud uuringu kohaselt on inimesed kõige rohkem mures isejuhtivate autode ja autonoomsete relvade pärast. Samuti kardetakse, et kui tehisintellekti kasutatakse professionaalsete otsuste puhul, ei pruugi tehismõistus suuta tegelikkuses arvestada individuaalsete asjaoludega ning otsuse tegemisel võib puududa läbipaistvus ja vastutus [48].

Aastatel 2018 – 2021 lahvatas Hollandis skandaal, kuna sealne maksuamet kasutas otsuste tegemisel vigast riskianalüüsi algoritmi, mille tõttu süüdistati alusetult tuhandeid lapsehooldushüvitiste saajaid pettuses [49]. Selle tõttu langesid vaesusesse kümned tuhanded pered, kes olid sageli madalama sissetulekuga või kuulusid etnilistesse vähemustesse. Mõned ohvrid sooritasid enesetapu ja üle tuhande lapse paigutati hooldusperedesse [50].

Selliseks professionaalseks otsuseks võib olla ka kohtuotsus. Kerkivad küsimused, kas tehisintellekti poolt tehtav otsus on samaväärse kvaliteediga kui kohtuniku poolt tehtud otsus, ning kas asjaomast süsteemi on treenitud kvaliteetse andmete pealt ja välistatud on diskrimineerimine mistahes alustel, näiteks soo, rassi või sissetulekute põhjal. Kirjanduses on arutletud selle üle, et tehisintellekti mudelid, mis põhinevad varasematest sisendandmetest saadud teabel, järgivad suure tõenäosusega konservatiivsemaid lähenemisviise ja need ei pruugi kohanduda aja jooksul oluliste poliitiliste muudatustega [51]. Leitud on ka, et tehisintellekti kasutamine kohtuotsuste tegemisel võib kujutada endast ohtu andmetele, mille olemusest tulenevalt nõuaksid need aga

kõrgeimat kaitsetaset [52].

Avastatud on, et suured keelemudelid (LLMid) võivad kalduda kinnistama ebaõigeid õiguslikke eeldusi ja tõekspidamisi, mis omakorda tekitavad tõsist muret tulemuste usaldusväärsuse pärast õiguslikus kontekstis [53], [54]. Samuti on kohtupidamise puhul olulised tehisintellekti mudeli läbipaistvuse ja täpsuse küsimused [55].

Tehisintellekti arendamisel, juurutamisel ja kasutamisel kerkivate eetiliste probleemidega tegeleb tehisintellekti eetika, mida loetakse rakenduseetika üheks alavaldkonnaks. Tehisintellekti eetika eesmärk on välja selgitada, kuidas on võimalik tehisintellekti süsteemil edendada või halvendada inimese heaolu läbi elukvaliteedi või sõltumatuse ja vabaduse muutuste. Erinevate tehisintellekti eetiliste raamistike loomisel on üldjuhul aluseks võetud põhiõigused [45].

2019. aasta 8 aprillil esitles Euroopa Liidu kõrgetasemeline tehisintellekti eksperdirühm (ingl k *High-Level Expert Group on AI*, edaspidi nimetatud AI HLEG või EL AI ekspertrühm) [56] usaldusväärse tehisintellekti eetikasuuniseid [45, 57] eesmärgiga pakkuda juhiseid eetilise ja töökindla tehisintellekti edendamiseks ja kindlustamiseks. Vähem keskendutatakse selles süsteemi seaduslikele aspektidele. Dokument annab esmase usaldusväärse tehisintellekti raamistiku, käsitledes ka tehisintellekti teostamise kui ka hindamisega seonduvaid aspekte [45].

2.6.4.2 Tehisintellekti regulatsioon Euroopa Liidus

Euroopa Komisjon avalikustas 2021. aasta aprillis esimese tehisintellekti reguleeriva raamistiku [58]. Viidatud ettepanek on kantud riskipõhisest lähenemisviisist ehk tehisintellekti süsteeme tuleb analüüsida ja klassifitseerida vastavalt sellele, millist ohtu need kasutajatele kujutavad [59]. Õigusakti läbirääkimised said punkti 8. detsembril 2023. 2024. aasta alguses on oodata tehisintellekti määruse avaldamist Euroopa Liidu Teatajas.

Samas ei tohi ära unustada ka kehtivat õigusraamistikku. Nimelt juba 2016. aasta aprillis vastu võetud isikuandmete kaitse üldmäärus (IKÜM või üldmäärus) [60] peab oluliseks kaitsta füüsilisi isikuid isikuandmete automatiseeritud töötlemise puhul[27]. Lisaks eelnevale peab tehisaru arendamisel, rakendamisel ja kasutamisel arvestama ka muude nõuetega, näiteks intellektuaalomandi õigusega. Tehisaru õiguslike aspektide kohta vaata täpsemalt käesoleva aruande peatükki 3.

[27] Üldmäärus reguleerib isikuandmete automatiseeritud töötlusi, sh profiilianalüüsi tegemist, ning annab andmesubjektile õiguse esitada sellistel töötlustel põhinevate üksikotsuste osas vastuväiteid (vt IKÜM artiklid 2, 21 ja 22 ning pp-d 15 ja 71).

3 Õiguslikud aspektid

3.1 Rahvusvahelised õiguslikud algatused

3.1.1 Õigusaktid

Viimaste aastate praktika näitab, et tehisintellekti teema on globaalselt kiires arengus. Allolevalt on toodud vaid mõned näited tehisintellekti süsteeme reguleerivatest riikidest.

Ameerika Ühendriikide president Joe Biden andis 30. oktoobril 2023. aastal välja täidesaatva korralduse tagamaks, et Ameerika oleks juhtpositsioonil maailmas tehisintellekti süsteemide puhul. Kõnealune korraldus kehtestab uued tehisintellekti ohutuse ja turvalisuse standardid, kaitseb ameeriklaste privaatsust, edendab võrdsust ja kodanikuõigusi, seisab tarbijate ja töötajate eest, edendab innovatsiooni, konkurentsi ja palju muud [61].

Suurbritannias avalikustati tehisintellekti tehnoloogiate töökohal kasutamise reguleerimise eelnõu, millega sätestatakse töötajate ja ametiühingute õigused seoses tehisintellekti tehnoloogiate kasutamisega. Eelnõu esimene lugemine toimus 17. mail 2023 [62], [63]. 2023. aasta kevadel avaldas Ühendkuningriigi valitsus poliitikadokumendi innovatsioonilise lähenemise kohta tehisintellekti reguleerimisel (A pro-innovation approach to AI regulation). Selle aluseks on viis aspekti: (1) ohutus, turvalisus ja töökindlus; (2) läbipaistvus ja seletatavus; (3) õiglus; (4) aruandlus ja juhtimine; (5) vaidlustatavus ja kahju hüvitamine [64].

Tehisintellekti reguleerimise arutelud on alanud ka Austraalias [65]. 2022. aastal avalikustati konsultatsioon tehisintellekti ja automatiseeritud otsuste tegemise eeskirjade osas. Konsultatsioon on ajendatud Austraalia valitsuse digimajanduse strateegiast, mis seab ambitsioonika visiooni, et Austraalia oleks 2030. aastaks 10 parima digimajanduse ja -ühiskonna seas [66, 67]. 8. septembril 2023. aastal avalikustati, et Austraalia uue otsingumootoreid hõlmava õigusakti eelnõu kohaselt peavad interneti otsingumootorite teenuseandjad vaatama üle ja korrapäraselt täiustama oma tehisintellekti tööriistasid, tagamaks seda, et 1A klassi materjale (nt laste seksuaalset ärakasutamist, terrorismi toetavat ja äärmuslikku vägivalda puudutavat materjali) ei tagastataks otsingutulemustes. Viidatud eelnõu nõuab ka seda, et kasutajatel peab olema võimalik kindlaks teha, kas otsingumootori kaudu juurdepääsetav pilt on süvavõltsing [68, 69, 70].

2023. aasta septembris avalikustati Kanadas generatiivsete tehisintellektisüsteemide vastutustundliku arendamise ja juhtimise vabatahtlik tegevusjuhend [71]. Samuti on kavandamisel tehisintellekti ja -andmete seadus (Artificial Intelligence and Data Act, AIDA), mis paneks aluse kanadalaste elu mõjutavate tehisintellektisüsteemide vastutustundlikule kavandamisele, arendamisele ja kasutuselevõtule [71]. Seadus tagaks, et Kanadas kasutusele võetud tehisintellektisüsteemid oleksid turvalised ja mittediskrimineerivad ning paneks ettevõtted vastutama selle eest, kuidas nad neid tehnoloogiaid arendavad ja kasutavad. Lisaks eelnevale kuulutas Kanada valitsus 12.10.2023 välja avaliku konsultatsiooni generatiivse tehisintellekti mõjude kohta autoriõigusele [72].

Tehisintellekti süsteemidega seotud õiguslikke algatusi on lisaks eelpool mainitud riikides ka näiteks Iisraelis, Jaapanis, Hiinas, Tšiilis, Mehikos, Peruus, Singapuris ja mitmel pool mujal [73]. Euroopa Liidu tehisintellektisüsteeme käsitlevate õigusaktide kohta loe käesoleva aruande peatükist 3.3.

3.1.2 Standardid

Kui me räägime rahvusvahelise pehme õiguse lähenemisviisidest, siis on avaldatud erinevaid mittesiduvaid suuniseid ja juhiseid, mis võiksid edendada eetilise, vastutustundliku ja usaldusväärse tehisintellekti arendamist ja kasutuselevõttu. Need on peamiselt kantud põhimõtetest nagu privaatsus, seletatavus, erapooletus, turvalisus ja inimkesksus.

Üks selline standard on ISO/IEC 22989, milles esitatakse tehisintellekti valdkonnaga seotud terminoloogia ja selgitatakse seotud kontseptsioone [74]. Ühtne terminoloogia tagab tehisintellektisüsteemist parema arusaamise ja omab olulist rolli koostöö tegemisel, reguleerimisel, vastutustundliku tehisintellektisüsteemi kasutuselevõtul ja teabe jagamisel [75]. ISO/IEC 23053 standardis on käsitletud tehisintellektisüsteeme, mis kasutavad masinõpet [76]. Viidatud standardis selgitatakse masinõppe süsteemi komponente ja nende funktsioone tehisintellekti ökosüsteemis [75].

Lisaks on standardis ISO/IEC 5259 esitatud tingimused analüüsil ja masinõppes andmete kvaliteedi tagamiseks [77, 78]. ISO/IEC 4213 kirjeldab nõudeid masinõppe klassifikatsiooni toimivuse hindamiseks [79]. Erinevaid standardeid ja raamistikke on veelgi, näiteks BSI valideerimisraamistik BS 30440:2023 tehisintellekti (AI) kasutamiseks tervishoius [80], IEEE eetilise disaini standard [81], Google tehisintellekti printsiibid [82] ja vastutustundliku tehisintellekti praktikad [83] ning Microsofti vastutustundliku tehisintellekti standard [84].

Standardite järgimine panustab toodete või teenuste ohutuse, kvaliteedi ja töökindluse tagamisse, samuti võivad need parandada ja tõhustada ettevõtte süsteeme või protsesse. Tehisintellektisüsteemide erinevate elutsüklite puhul rakendatavate standardite kohta on võimalik lugeda ENISA publikatsioonist heade küberturvalisuse praktikate kohta tehisintellekti süsteemide puhul [85].

3.2 Euroopa Liidu usaldusväärse tehisintellekti algatus

EL AI ekspertrühm esitles 2019. aasta 8. aprillil usaldusväärse tehisintellekti eetikasuuniseid [86], mis käsitleb usaldusväärse tehisintellekti raamistikku, teostamist ja hindamist [87]. Viidatud eetikasuuniste kohaselt peaks usaldusväärse AI-süsteemi elutsükkel olema:

1. seaduslik – vastav kohaldavatele õigusnormidele,
2. eetiline – eetilisi põhimõtteid ja väärtuseid järgiv ning
3. töökindel – nii tehnilisest kui sotsiaalsest vaatenurgast, et vähendada soovimatute tagajärgede realiseerumist [87].

Eetikasuuniste I peatükis antakse kolm peamist põhiõigustest lähtuvat põhimõtet. Esiteks, tehisintellekti süsteemide puhul tuleb austada inimeste sõltumatust, tagada süsteemi õiglus ja seletatavus ning hoiduda kahju tegemisest. Teise põhimõttena tuuakse välja tähelepanu pööramine esiteks haavatavamate sihtrühmadele (nt lapsed ja erivajadustega inimesed) ja teiseks olukordades, kus esineb võimu tasakaalutus või teabe asümmeetria. Kolmandaks juhitakse tähelepanu tehisintellekti süsteemidega kaasnevatele riskidele ja riske leevendavate meetmete kasutuselevõtmisele [87].

Eetikasuuniste II peatükis antakse ülevaade sellest, kuidas saab luua usaldusväärse tehisintellekti süsteemi, ja pakutakse selleks välja seitse kriteeriumi.

1. Esiteks soovitatakse kindlaks teha, et kõnealuste süsteemide arendamine, kasutuselevõtmine ja kasutamine arvestab järgmiste aspektidega: *”(1) inimese toimevõime ja järelevalve,*

(2) tehniline töökindlus ja ohutus, (3) privaatsus ja andmehaldus, (4) läbipaistvus, (5) mitmekesisus, mittediskrimineerimine ja õiglus, (6) keskkonnaalane ja ühiskondlik heaolu ning (7) vastutuse võtmine" [87].

2. Eelnimetatud aspektidega arvestamiseks on soovituslik kasutada nii tehnilisi kui ka organisatoorseid meetodeid.
3. Soodustada tuleks teadusuuringuid ja innovatsiooni, et tehisintellekti süsteemide kohta oleks rohkem teadmust, mis võimaldaks mh koolitada ka uusi tehisintellekti eetika eksperte.
4. Tuleks anda selget teavet tehisintellekti süsteemi suutlikkuse ja piirangute kohta, mis võimaldab seada realistlikke ootusi.
5. Arendama peaks seletatavaid süsteeme, mis hõlbustaks nende auditeeritavust, mis võib osutuda vajalikuks eeskätt kriitilistes olukordades.
6. Kaasata tuleks seotud osapooli kogu tehisintellekti süsteemi elutsükli jooksul, koolitada inimesi ja tõsta nende teadlikkust usaldusväärsest tehisintellektist.
7. Arvestada tuleb, et usaldusväärse tehisintellekti põhimõtete ja nõuete vahel võivad tekkida probleemkohad. Soovituslik on kaalutlused, kompromissid ja vastuvõetud otsused dokumenteerida [87].

Eetikasuuniste III peatükis esitatakse kontrollnimekiri usaldusväärse tehisintellekti kasutuslikule kujule viimiseks, mida tuleb kohandada vastavalt tehisintellekti süsteemi otstarbele. Kogu tehisintellekti süsteemi elukaare ajal tuleb tegeleda nõuetele vastavuse hindamise, osapoolte kaasamise ja tulemuste pideva parendamisega [87]. Tehisintellektisüsteemi usaldusväärsus sõltub selle kõikidest omadustest, mille vaheliste kompromisside täielik mõistmine on aga paraku endiselt oluline uurimisprobleem [88].

Eetikasuuniste viimane osa täpsustab olulisemaid dokumendis adresseeritud küsimusi, tuues näiteid kasulikest võimalustest, mille poole tuleks püüelda, ja toob välja tehisintellekti süsteemide kriitilisi probleeme, mis vajavad suuremat tähelepanu [87].

Lisaks on EL AI ekspertrühm avaldanud ka poliitika- ja investeerimissoovitused usaldusväärse tehisintellekti kohta, milles selgitatakse, kuidas usaldusväärset tehisintellekti Euroopas arendada, juurutada, edendada ja laiendada, maksimeerides kasu ning minimeerides ja ennetades riske [56, 89]. 17. juulil 2020 avaldas EL AI ekspertrühm täiendavalt usaldusväärse tehisintellekti (ALTAI) hindamisnimekirja [90]. See on tööriist, millega on võimalik hinnata tehisintellektisüsteemi vastavust usaldusväärse tehisintellekti nõuetele. Juhis on kättesaadav ka veebipõhise tööriista versioonina [91].

Lisaks avaldati dokument valdkondlike kaalutluste kohta poliitika- ja investeerimissoovituste osas, kus analüüsitakse EL AI ekspertrühma poolt varem avaldatud soovituste võimalikku rakendamist kolmes konkreetses rakendusvaldkonnas: (1) avalik sektor, (2) tervishoid, (3) tootmine ja asjade internet [92].

19.02.2020 avalikustas Euroopa Komisjon aruande selle kohta, milline on tehisintellekti, asjade interneti ja robootika mõju ohutusele ja vastutusele [93]. Kõik tooted ja teenused peavad töötama ohutult, usaldusväärselt ja stabiilselt ning tekkinud kahju tuleb heastada – see on ohutuse ja vastutuse õigusraamistike eesmärk. Komisjoni hinnangul on uute tehnoloogiate kujunemise ajal selge ohutus- ja vastutusraamistik vajalik tagamaks tarbijakaitset aga ka ettevõtjate õiguskindlust [93].

Samal päeval avalikustas Euroopa Komisjon ka tehisintellekti valge raamatu [94], mis kajastab andmepõhise majanduse kõige olulisema väljundi – tehisintellektiga seotud aspekte, milles põimuvad andmed, algoritmid ja andmetöötlusvõimsus. Valges raamatus nenditakse, et digitehno-

loogiate kasutus põhineb usaldusel, ja selgitatakse, kuidas tuleks hoogustada meetmeid erinevatel tasanditel, et tõhustada tehisintellekti kasutuselevõttu [94].

3.3 Euroopa Liidu tehisintellekti määruse ettepanek

Euroopa Liidus (EL) on esitatud mitmeid tehisintellektiga seotud õiguslikke ettepanekuid, tagamaks, et ELis kasutatavad tehisintellektisüsteemid oleksid turvalised, läbipaistvad, eetilised, erapooletud ja inimese poolt kontrollitavad [95].

2021. aasta aprillis esitas Euroopa Komisjon määruse ettepaneku tehisintellekti ühtlustatud õigusnormide kohta (tehisintellekti käsitlev õigusakt) [58]. Viidatud määruse ettepaneku seletuskirja kohaselt sätestatakse ühtlustatud nõuded, mis järgivad proportsionaalset riskipõhist lähenemisviisi tehisintellektisüsteemide arendamiseks, turule laskmiseks ja kasutamiseks ELis [58]. 8. detsembril 2023 saavutati poliitiline kokkulepe viidatud õigusakti lõpliku teksti osas [96, 97], ja toimusid tehnilised arutelud teksti viimistlemiseks. Eriti pöörati tähelepanu suure mõjuga üldotstarbeliste tehisintellekti (GPAI) mudelite künnise küsimusele, mis otsustati kindlaks määrata treenimiseks kasutatud arvutusvõimsuse kumulatiivse hulga põhjal (10^25). Tulevikus on plaanis töötada välja ühtsed standardid GPAI mudelite reguleerimiseks [98].

26. jaanuaril 2024 jagas nõukogu eesistujariik Belgia ametlikult liikmesriikide esindajatega tehisintellekti määruse ettepaneku lõplikku kompromissteksti [99]. 2. veebruaril võttis alaliste esindajate komitee (COREPER) tehisintellekti määruse vastu. Kompromiss põhines mitmetasandilisel lähenemisviisil, mis hõlmas horisontaalseid läbipaistvuseeskirju kõikidele mudelitele ja täiendavaid nõudeid sellistele tehisintellektisüsteemidele, mis võivad põhjustada süsteemset riski [98].

Tehisintellekti käsitleva õigusakti ettepanekul [58] on neli peamist eesmärki.

1. Esiteks soovitakse tagada, et EL turule lastavad ja seal kasutatavad tehisintellektisüsteemid on ohutud ning vastavad õigusaktide nõuetele ja EL väärtustele.
2. Teiseks soovitakse tagada õiguskindlus, mis soodustaks tehisintellekti tehtavaid investeeringuid ja innovatsiooni.
3. Kolmandaks ajendiks on tagada põhiõiguste kaitse tehisintellektisüsteemide kasutamisel ja asjaomaste ohutusnõuete järgimine.
4. Neljandaks on soov töötada välja ühtne turg seaduslike, ohutute ja usaldusväärsete tehisintellektirakenduste jaoks.

Viidatud määruse ettepaneku kohaselt jaotatakse tehisintellektisüsteemid nelja riskikategooriasse, mis võimaldab kehtestada nõudeid vastavalt kaasnevatele riskidele (vt tabel 3). Läbirääkimiste käigus lisati tehisintellekti määruse teksti ka mittesüsteemsed ja süsteemsed riskid, mis puudutavad üldotstarbelisi tehisintellektisüsteeme [99].

Tehisintellekti määruse ettepaneku lõplikus kompromisstekstis [99] defineeritakse tehisintellekti süsteem kui masinapõhine süsteem, mis on loodud töötama erineva autonoomiatasemega ja mis võib pärast kasutuselevõttu olla kohanemisvõimeline ning mis otseste või kaudsete eesmärkide puhul järeldab saadud sisendi põhjal, kuidas luua väljundeid, nagu ennustused, sisu, soovitused või otsused, mis võivad mõjutada füüsilist või virtuaalset keskkonda (vt artikkel 3(1)).[1]

[1] Siin ja edaspidi on käsitletud AI-süsteemidele kohalduvaid nõudeid tehisintellekti määruse ettepaneku lõplikus kompromisstekstis esitatud kujul, kuivõrd käesoleva aruande koostamise ajal ei olnud tehisintellekti määruse ametlik vastuvõetud tekst veel Euroopa Liidu Teatajas (ELT) avalikustatud. Arvestada tuleb, et siin ja alljärgnevalt viidatud kompromissteksti konkreetsed artiklid, lõiked või punktid võivad erineda ELT-s avalikustatavas tehisintellekti määruse tekstist, kuna kompromissteksti numeratsioon on korrigeerimata. – Internetis: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf Külastatud: 26. veebruaril 2024

Viidatud kompromisstekstis [99] on kirjas, et tehisintellekti määruse kehtestamise eesmärk on soodustada inimkeskse ja usaldusväärse tehisintellekti kasutuselevõttu, soodustades innovatsiooni ning tagades tervise, ohutuse, põhiõiguste, demokraatia, õigusriigi ja keskkonna kaitse tehisintellektisüsteemide kahjulike mõjude eest. Määruses sätestatakse ühtlustatud nõuded tehisintellektisüsteemide (siin aruandes ka Äl-süsteemid") turule laskmise, kasutusele võtmise ja kasutamise kohta ELis. Selles keelustatakse teatud tehisintellekti kasutusviisid, sätestatakse erinõuded kõrge riskiga AI-süsteemidele ja selliste süsteemide operaatorite kohustused. Lisaks kehtestatakse teatud AI-süsteemidele ühtlustatud läbipaistvusnormid ning üldotstarbeliste AI mudelite turule laskmise nõuded. Määrusega kehtestatakse ka turuseire ja -järelevalve normid ning meetmed innovatsiooni toetamiseks, keskendudes eelkõige väikese ja keskmise suurusega ettevõtetele, sh idufirmadele.

3.3.1 Tehisintellekti määruse kohaldamisalasse kuuluvad järgmised isikud

Tehisintellekti määruse kohaldamisalasse kuuluvad järgmised isikud [99]:

1. teenustaja (ingl k *provider*), kes ELis AI-süsteemi turule laseb või seda oma teenuses kasutab või kes laseb turule üldotstarbelise AI mudeli, olenemata sellest, kas teenustaja on asutatud või asub ELis või kolmandas riigis;
2. AI-süsteemide juurutaja (ingl k *deployer*), kelle tegevuskoht on või kes asub ELis;
3. AI-süsteemide teenustaja ja juurutaja, kelle tegevuskoht on või kes asub kolmandas riigis, kuid kelle AI-süsteemi väljundit kasutatakse ELis;
4. AI-süsteemi importija (ingl k *importer*) või turustaja (ingl k *distributor*);
5. tootja (ingl k *product manufacturer*), kes laseb turule või võtab kasutusele AI-süsteemi koos oma tootega oma nime või kaubamärgi all;
6. väljaspool ELi asutatud teenustaja volitatud esindaja;
7. mõjutatud või puudutatud isik (ingl k *affected person*), kes asub ELis.

Tehisintellekti määruse ettepaneku artiklis 3 esitatakse suurel hulgal uusi termineid, sh defineeritakse süvavõltsing ja AI kirjaoskus, aga ka näiteks treening-, valideerimis-, testimis- ja sisendandmed. Kusjuures AI kirjaoskusele on pühendatud täiesti eraldi artikkel (artikkel 4b), mis panenb AI-süsteemide teenustajatele ja juurutajatele kohustuse võtta meetmeid näiteks selleks, et tagada oma töötajate, kes puutuvad kokku AI-süsteemide haldamise ja kasutamisega, piisavad teadmised AI-süsteemidest.

Järgnevalt on tehisintellekti määruse ettepaneku lõpliku kompromissteksti [99] alusel välja toodud mõned olulisemad nõuded tehisintellektisüsteemidega seotud osapooltele.

3.3.2 Tehisintellekti määruse kohaldamisala välistused

Tehisintellekti määruse kohaldamisalast on välistatud füüsilisest isikust teenustaja, kes kasutab AI-süsteeme üksnes isiklikus mittekutselises tegevuses. Kohaldamisalast on välja jäetud näiteks ka sellised AI-süsteemid, mida kasutatakse eranditult sõjalistel, kaitse- või riikliku julgeoleku eesmärkidel. Samuti ei kohaldata viidatud määrust AI-steemidele ja -mudelitele, sh nende väljundile, mis on spetsiaalselt välja töötatud ja kasutusele võetud üksnes teadusliku uurimis- ja arendustegevuse eesmärgil. Määrust ei kohaldata AI-süsteemide või -mudelitega seotud uurimis-, testimis- ja arendustegevuse suhtes enne nende turule laskmist või kasutuselevõttu, v.a. AI-süsteemide katsetamine reaalsetes tingimustes. Lisaks on kohaldamisalast vä-

listatud ka tasuta ja avatud lähtekoodiga litsentside alusel välja antud AI-süsteemid, v.a. juhul, kui need lastakse turule või võetakse kasutusele nt kõrge riskiga AI-süsteemidena.

3.3.3 Keelatud tehisintellekti praktikad ja kasutusviisid

Määrusega keelustatakse mitmeid AI-süsteemide praktikaid (vt täpsemalt artiklist 5). Näiteks on keelatud selliste AI-süsteemide kasutusviisid, mis sihipäraselt manipuleerivaid isikuga eesmärgiga moonutada oluliselt tema käitumist ja kahjustada märgatavalt tema võimet teha teadlikku otsust. Samuti keelustatakse sellised AI-süsteemid, mis kasutavad ära inimese haavatavust tingituna tema vanusest, puudest või konkreetsest sotsiaalsest või majanduslikust olukorrast. Keelatud on ka biomeetriliste kategoriseerimissüsteemide kasutamine, mis liigitavad füüsilised isikud nende biomeetriliste andmete alusel, et teha järeldusi nende rassi, poliitiliste seisukohtade, ametiühingusse kuulumise, usuliste või filosoofiliste veendumuste, suguelu või seksuaalse sättumuse kohta. Lubatud ei ole ka sellised AI-süsteemid, mida kasutatakse füüsiliste isikute hindamiseks nende sotsiaalse käitumise või isiksuseomaduste põhjal koos sotsiaalse skooriga, mille tulemuseks on isikule kahjulik või ebasoodne kohtlemine.

3.3.4 Nõuded kõrge riskiga tehisintellektisüsteemidele

Kõrge riskiga AI-süsteemide klassifitseerimise nõuded on sätestatud määruse ettepaneku artiklis 6. Näiteks loetakse AI-süsteemi alati kõrge riskiga süsteemiks, kui see teostab füüsiliste isikute profileerimist. Teenustaja, kes leiab, et tehisintellekti määruse III lisas osutatud AI-süsteem ei ole kõrge riskiga, peab oma hinnangu dokumenteerima enne kõnealuse süsteemi turule laskmist või kasutuselevõttu. Sellise teenustaja suhtes kohaldatakse määruse kohast registreerimiskohustust (vt artikkel 51(1a)) ja riiklike pädevate asutuste nõudmisel peab teenustaja esitama asjaomase hindamise kohta tehtud dokumentatsiooni. Euroopa Komisjon peab hiljemalt 18 kuud peale tehisintellekti määruse jõustumist esitama suunised, mis täpsustavad määruse artikli 6 praktilist rakendamist, koos põhjaliku loeteluga praktilistest näidetest kõrge riskiga AI-süsteemide kasutamise juhtumite kohta.

Artiklis 9 on kehtestatud nõuded kõrge riskiga AI-süsteemi riskihaldussüsteemile. Selle lõike 2 kohaselt tuleb riskihaldussüsteemi mõista kui pidevat iteratiivset protsessi, mida kavandatakse ja viiakse läbi kõrge riskiga AI-süsteemi kogu elutsükli jooksul ning mis nõuab korrapärast süstemaatilist läbivaatamist ja ajakohastamist, hõlmates järgmisi samme:

a) teadaolevate ja mõistlikult prognoositavate riskide kindlakstegemine ja analüüs, mida kõrge riskiga AI-süsteem võib kujutada inimeste tervisele, ohutusele või põhiõigustele, kui kõrge riskiga tehisintellekti süsteemi kasutatakse ettenähtud otstarbel;
b) riskide prognoosimine ja hindamine, mis võivad tekkida, kui kõrge riskiga AI-süsteemi kasutatakse ettenähtud otstarbel ja mõistlikult prognoositava väärkasutuse tingimustes;
c) muude võimalike tekkivate riskide hindamine, mis põhineb turustamisjärgsest seiresüsteemist (vt artikkel 61) kogutud andmete analüüsil;
d) sobivate riske leevendavate meetmete rakendamine.

Kõrge riskiga AI-süsteemid peavad vastama tehisintellekti määruses kehtestatud nõuetele (vt pt 2), võttes arvesse nimetatud süsteemide otstarvet ning tehisintellekti ja sellega seotud tehnoloogiate taset. Täpsustatud on, et riskihaldusmeetmed peavad olema sellised, et iga ohuga seotud jääkriski ja üldist jääkriski on võimalik aktsepteerida (artikkel 9(4)).

Kõrge riskiga AI-süsteeme tuleb ka testida, et teha kindlaks kõige sobivaimad riskihaldussmeet-

med (artikkel 9(5)). Testimise protseduurid võivad hõlmata ka testimist reaalsetes tingimustes (artikkel 9(6); vt ka artikkel 54a). Lisaks tuleb hinnata ka võimalikke mõjusid alla 18-aastastele isikutele ja teistele haavatavamatele sihtrühmadele (artikkel 9(8)).

Kõrge riskiga AI-süsteemid, mille mudeleid treenitakse andmete peal, peavad kasutama mudeli treenimiseks, valideerimiseks ja testimiseks selliseid andmestikke, mis vastavad tehisintellekti määruses toodud kvaliteedikriteeriumidele (artikkel 10(1)). Kohaldada tuleb ka AI-süsteemi kavandatud otstarbele vastavaid ja asjakohased andmehalduspraktikaid, näiteks võimalike kõrvalekallete tuvastamiseks, ennetamiseks ja leevendamiseks (artikkel 10(2)(fa)).

Treening-, valideerimis- ja testimisandmed peavad olema asjakohased, piisavalt esinduslikud ja parimal võimalikul määral vigadeta ning kavandatud eesmärki silmas pidades täielikud ning sobivate statistiliste omadustega (artikkel 10(3)).

Eriliigiliste isikuandmete töötlemine algoritmiliste kallutatuste tuvastamiseks ja korrigeerimiseks kõrge riskiga AI-süsteemide puhul on rangelt reguleeritud. See peab esiteks vastama kõikidele EL andmekaitsealastele nõuetele ja sellise töötlemise toimumiseks peavad olema täidetud tehisintellekti määruse artikli 10(5) punktides (a)-(f) sätestatud tingimused. Esiteks on vaja põhjendada, miks ei ole algoritmiliste kallutatuste tuvastamine võimalik muude andmetega, sh sünteetiliste või anonüümsete andmetega.

Eriliigiliste andmete kasutamisel tuleb kasutada kaasaegseid turvalisuse ja privaatsust säilitavaid meetmeid, sh pseudonüümimist või privaatsuskaitse tehnoloogiaid. Tagada tuleb viidatud andmete turvalisus, sh tugevad kontrollid ja dokumentatsioon juurdepääsude osas, et välistada andmete väärkasutus ja tagada, et üksnes volitatud isikud, kellel on vastavad konfidentsiaalsuskohustused, omavad juurdepääsu eriliigilistele isikuandmetele. Selliseid andmeid ei tohi edastada, üle anda ega teha kättesaadavaks muudele isikutele. Viidatud andmed tuleb kustutada kohe pärast seda, kui algoritmiline kallutatus on parandatud või kui isikuandmete säilitamisperiood on lõppenud, olenevalt sellest, kumb sündmus saabub varem.

Kõrge riskiga AI-süsteemi tehniline dokumentatsioon koostatakse enne selle süsteemi turule laskmist või kasutuselevõttu, seda hoitakse ajakohasena ja see sisaldab vähemalt tehisintellekti määruse IV lisas sätestatud elemente (artikkel 11(1)). Kõrge riskiga AI-süsteemid peavad tehniliselt võimaldama sündmuste automaatset salvestamist (logimist) kogu süsteemi eluea jooksul (artikkel 12(1)).

Kõrge riskiga AI-süsteeme tuleb kavandada ja arendada nii, et oleks tagatud nende toimimine piisavalt läbipaistvalt, et võimaldada juurutajal süsteemi väljundit tõlgendada ja seda asjakohaselt kasutada (artikkel 13(1)). Kõrge riskiga AI-süsteemide kohta peavad olema olemas asjakohases digitaalvormingus kasutusjuhised, mis sisaldavad lühikest, täielikku, õiget ja selget teavet, mis on kasutajatele asjakohane, juurdepääsetav ja arusaadav (artikkel 13(2)). Viidatud kasutusjuhised peavad vastama määruse artiklis 13(3) kehtestatud minimaalsetele nõuetele.

Kõrge riskiga AI-süsteemide puhul tuleb ette näha võimalused, et inimesel oleks võimalik viidatud süsteemi selle kasutusaja jooksul tõhusalt monitoorida (vt artiklis 14 sätestatud inim-järelevalve nõudeid ja printsiipe). Näiteks peab inimesel olema võimalik sekkuda kõrge riskiga AI-süsteemi töösse või katkestada süsteemi töö ”stopp–nupu” või sarnase protseduuri abil (artikkel 14(4)(e)).

Kõrge riskiga AI-süsteeme tuleb kavandada ja arendada nii, et need saavutaksid asjakohase täpsuse, töökindluse ja küberturvalisuse taseme ning toimiksid selles osas järjepidevalt kogu oma elutsükli jooksul (artikkel 15(1)). Sellised süsteemid peavad olema vastupidavad volitamata kolmandate isikute katsetele muuta süsteemide kasutust, väljundeid või toimivust (artikkel 15(4)).

Määruse artiklis 21 sätestatakse, et kõrge riskiga tehisintellektisüsteemide teenustajad, kellel on põhjust arvata, et AI-süsteem, mille nad on turule lasknud või kasutusele võtnud, ei vasta tehisintellekti määruse nõuetele, peavad võtma viivitamata vajalikud parandusmeetmed, näiteks AI-süsteemi nõuetega vastavusse viimiseks või selle tegevuse peatamiseks. Teenustaja peab teavitama olukorrast ka turustajaid ja asjakohasel juhul juurutajaid, volitatud esindajaid ja importijaid.

3.3.5 Nõuded AI väärtusahelas osalejatele

Tehisintellekti määruses on kehtestatud erinevaid nõudeid ka teistele AI-süsteemiga seotud osapooltele, näiteks juurutajatele, EL väliste teenustajate volitatud esindajatele, importijatele ja turustajatele. Seetõttu on oluline hinnata, milline on konkreetse isiku roll AI väärtusahelas vastavalt tehisintellekti määrusele, et oleks selge, milliseid nõudeid ta järgima peab.

Tehisintellekti määrus on värske EL õigusperekonna liige, mistõttu vajavad uued normid ja nende rakendajad mõningast aega, et kohaneda uue olukorraga. Loodetavasti on sellel teekonnal abiks ka Euroopa tehisintellekti büroo – AI ekspertiisikeskus ELis. Viidatud büroo omab võtmerolli tehisintellekti määruse rakendamisel, mis soodustab usaldusväärse tehisintellekti arendamist ja kasutamist ning rahvusvahelist koostööd [100].

3.4 Tehisintellektiga seotud vastutuse direktiivi ettepanek

Tehisintellektiga seotud riskide leevendamiseks esitati lisaks tehisintellekti käsitlevale õigusakti ettepanekule 2022. aasta septembris ka tehisintellektiga seotud vastutuse direktiivi ettepanek [101], millega soovitakse tagada, et tehisintellekti põhjustatud kahju kandnud isikud saaksid mõistlikult oma õiguseid kaitsta. Selleks ühtlustatakse riigisiseseid lepinguvälise süülise vastutuse norme. Samuti soovitakse tagada suurem õiguskindlus ettevõtjate jaoks, kes arendavad või kasutavad tehisintellekti.

Ühe meetmena soovitakse lihtsustada kannatanute jaoks kohtumenetlust tehisintellektisüsteemide tekitatud kahju korral. Kannatanutel võimaldatakse saada hüvitist nii individuaalselt kui ka asjakohasel juhul kollektiivselt. Kui on toiminud nõuete rikkumine ja põhjuslik seos tehisintellektisüsteemiga on tõenäoline, siis kehtib ümberlükatav põhjusliku seose eeldus. Täpsustatakse, et põhjusliku seose eeldusele on võimalik tugineda vaid juhul, kui on tõenäoline, et konkreetne süü on mõjutanud asjaomase tehisintellektisüsteemi väljundit või selle puudumist, ja seda on võimalik hinnata kaasuse üldiste asjaolude põhjal. Vaatamata eelnevale on hagejal siiski tõendamiskohustus, et tehisintellektisüsteem, st selle väljund või selle suutmatus väljundit luua, on kahju põhjustanud [101].

Samuti antakse kavandatava direktiiviga suuremad võimalused õiguskaitse tagamiseks. Näiteks võib kohtu korralduse kaudu saada kannatanu teavet, et teha kindlaks kahju põhjus, ja seeläbi tuvastada, millise isiku poole kahju hüvitamiseks pöörduda.

3.5 Tooteohutus

Euroopa Parlamendi ja nõukogu määrusega (EL) 2023/988 üldise tooteohutuse kohta [102] kehtestatakse põhilised ohutusnormid tarbijale mõeldud toodete jaoks, mis on turule lastud või turul kättesaadavaks tehtud (määrus (EL) 2023/988 artikkel 1(2)). Nimetatud määruse põhjenduspunktis 5 selgitatakse, et *”[o]htlikel toodetel võivad olla tarbijatele ja kodanikele väga negatiivsed tagajärjed. Kõigil tarbijatel, sealhulgas kõige haavatavamatel, nagu lastel, eakatel või puude-*

ga inimestel, on õigus ohututele toodetele. Tarbijate käsutuses peaks olema piisavalt vahendeid kõnealuse õiguse maksma panemiseks ning liikmesriikide käsutuses peaks olema piisavalt vahendeid ja meetmeid käesoleva määruse täitmise tagamiseks.”.

28. septembril 2022 avalikustas Euroopa Komisjon ettepaneku kehtestada direktiiv, mis käsitleb vastutust puudustega toodete eest [103]. Sellega soovitakse kehtestada normid, mis reguleerivad ettevõtjate vastutust puudusega toodete põhjustatud kahju eest, samuti tingimused, mille alusel füüsilistel isikutel on õigus saada hüvitist. Direktiiviga nähakse ette ka solidaarne vastutus. Puudustega toote eest vastutavad ettevõtjad direktiivi kohaselt kümne aasta jooksul pärast toote turule laskmist.

Direktiivi ettepaneku seletuskirjas täpsustatakse, et selle eesmärk on tagada ka vastutus tehisintellektisüsteemidel esinevate puuduste korral, mis põhjustavad füüsilist või varalist kahju või andmekadu. Sellisel juhul tekib kasutajal õigus taotleda hüvitist tehisintellektisüsteemi pakkujalt või mis tahes tootjalt, kes integreerib tehisintellektisüsteemi teise tootesse. Ettepanek hõlmab ka tarkvara pakkujaid, ettevõtjaid, kes teevad toodetes olulisi muudatusi, volitatud esindajaid ja tellimuste täitmise teenuse osutajaid, mis annab kannatanule tõhusamad võimalused saada kahju eest hüvitist[2].

3.6 Intellektuaalomand

Intellektuaalomandi õiguse eesmärk on kaitsta intellektuaalse loomingu tulemusi. Generatiivne tehisaru on muutnud ühiskonna arusaama loovusest ja omandiõigusest, tõstatades küsimusi inimsisendi ja intellektuaalse omandi kohta [104]. Intellektuaalomandi õiguste koostoime tehisintellektiga on aruande kirjutamise perioodil üks intellektuaalomandiõiguse peamisi arenguvaldkondi, eeskätt tänu tehisintellektiga seotud arengutele, asjakohasele esialgsele kohtupraktikale ning rahvusvaheliste organisatsioonide ja seadusandjate poliitilistele algatustele [105].

Õigusteadlaste huviorbiiti on viimastel aastatel kerkinud tehisintellekti ja intellektuaalomandi õigusega seotud õigusküsimused. Need võib jaotada suures plaanis kaheks.

1. Automaatse loomingu õiguskaitse – näiteks, kas teatud juhtudel on võimalik omistada tehisintellekti loodud teostele autoriõigusi või patenteerida nende poolt loodud leiutisi?
2. Intellektuaalomandi õiguste rikkumised – näiteks, kuidas kaitsta efektiivselt intellektuaalomandi õiguste omajaid tehisintellektisüsteemide arendajate eest, kes kasutavad intellektuaalomandi õigustega kaitstud loomingut tehisintellektisüsteemide treenimiseks ilma õiguste omajate teadmise ja/või nõusolekuta?

Generatiivne tehisintellekt, mis suudab kirjutada seostatud teksti, luua kunsti või arhitektuurilisi kavandeid, on toonud kaasa mitmesuguseid küsimusi intellektuaalomandi olemuse kohta ja andnud ainest õigusvaidlustele. Näiteid on nii olukordadest, kus autorid on kohtusse kaevanud tehisintellekti arendajad, kes on tehisintellekti süsteeme arendanud keelatud andmete või teostega (nt loata kasutanud intellektuaalomandi õigustega kaitstud teksti, pilte vms) [106, 107], aga ka juhtumitest, kus taotletakse tehisintellekti poolt loodud teostele intellektuaalomandi õigusi [108].

Tänapäevane intellektuaalomandi õigus ei arvesta üldiselt selliste loojatega nagu seda on tehisintellekt. Praegune intellektuaalomandi süsteem loodi selleks, et ergutada inimeste loomingut ja innovatsiooni. Kuna tehisintellekt töötab autonoomsemalt, tekitab see intellektuaalomandi süsteemi jaoks põhimõttelisi küsimusi kõigi intellektuaalomandi õiguste osas [109]. Samas on tehisintellektisüsteemide ja intellektuaalomandi õiguse vahel selge mõju ja korrelatsioon [110].

[2] Vt direktiivi ettepaneku seletuskiri, alapeatükk 1.2. ja peatükk 2.

Üldiselt arvestatakse järgnevate printsiipidega – teose originaalsus, idee ja väljendusviisi dihhotoomia ning eelneva fikseerimine inimesele tajutavas vormis [111].

Näiteks loetakse autoriõiguse seaduse §4 lõike 2 kohaselt teoseks *"mis tahes originaalset tulemust kirjanduse, kunsti või teaduse valdkonnas, mis on väljendatud mingisuguses objektiivses vormis ja on selle vormi kaudu tajutav ning reprodutseeritav kas vahetult või mingi tehnilise vahendi abil. Teos on originaalne, kui see on autori enda intellektuaalse loomingu tulemus"*. Ühe lahendusena on välja pakutud näiteks hübriidne omandimudel (AiLE) [111]. Samas on leitud, et täiendavate kihtide lisamine olemasolevale intellektuaalomandi õiguste süsteemile ei ole hea lahendus tehnoloogilise progressi ühiskondliku mõju tasakaalustamiseks [112] ja ka seda, et tehisintellekti poolt loodu ei ole kaitstav [113]. Euroopa Parlamendi arvates on oluline teha selget vahet tehisintellekti abil loodud inimeste loomingul ja tehisintellekti poolt iseseisvalt genereeritud loomingul [114].

Milline saab olema intellektuaalomandi õiguste tulevik seoses tehisintellektisüsteemide arenguga, näitab vaid aeg. Selge on see, et intellektuaalomandi õiguste osas valitseb arvamuste paljusus ja lihtsat valikut laual ei ole. Võimalik on, et praegu ei olegi otsuste tegemiseks õige aeg, tehisintellektisüsteemidega seotud arengud vajavad hoolikat läbimõtlemist ja teatud küpsustaset ühiskonnas, enne, kui hakatakse muutma toimivaid õigussüsteeme.

3.7 Õiguslikud nõuded küberturbele

Nagu ka iga muu teema korral, algab tehisintellektisüsteemide turvalisus konfidentsiaalsuse, käideldavuse ja terviklikkuse tagamisest. Tehisintellektiga seotud osapooled peaksid lähtuvalt oma rollist, kontekstist ja tegutsemisvõimest rakendama tehisintellektisüsteemi elutsükli igas etapis süstemaatilist riskijuhtimist, et käsitleda riske privaatsusele, digitaalsele turvalisusele ja ohutusele ning vältida algoritmilisi kallutatusi [25].

OECD soovituste kohaselt peaksid tehisintellektisüsteemid terve oma elutsükli vältel olema turvalised, töökindlad ja ohutud. Seda nii tavapärase, planeeritud kasutuse, kui ka väärkasutuse ja ebasoodsate tingimuste korral. Eelneva saavutamiseks on oluline tagada tehisintellektisüsteemi jälgitavus. See on oluline nii kasutatavate andmete või andmekogumite, erinevate protsesside ja tehtud otsuste osas, ning võimaldab teostada kontekstile vastavat analüüsi tehisintellektisüsteemi toimimise kohta, näiteks selle väljundite või päringutele reageerimise puhul [25].

ENISA on järjestanud informatsiooni ja kommunikatsiooni tehnoloogiate infrastruktuuride ohud järgmiselt [85]:

- ründed – need kujutavad endast pahatahtlikke kavatsusi (nt teenustõkestusründed, volitamata juurdepääs, identiteedi maskeerimine);
- juhuslikud ohud – need on põhjustatud kogemata, nt inimlik viga, või legaalsete süsteemikomponentide kaudu. Tavaliselt tekivad need seadmete või infosüsteemide seadistamise või protsesside teostamise käigus;
- keskkonnaohud – need hõlmavad looduskatastroofe, nt üleujutused või maavärinad, inimtegevusest põhjustatud hävinguid, nt põlengud, plahvatused, ja tugiinfrastruktuuride rikkeid, nt voolu- või sidekatkestus;
- haavatavused – tehisintellektisüsteemi nõrkus, mida ründaja võib ära kasutada.

Sellistele ohtudele reageerimiseks on Euroopa Liidus vastu võetud mitmeid õigusakte. Küberturvalisuse teist direktiivi (NIS2) [21] ja küberturvalisuse määrust [115] peetakse Euroopa kaheks kõige olulisemaks küberjulgeolekualaseks õigusaktiks. Samuti on võtmetähtsusega ka isi-

kuandmete kaitse üldmäärus (edaspidi üldmäärus või IKÜM) [60]. Need õigusaktid rõhutavad tarneahela turvalisust, privaatsust ja isikuandmete kaitset, mis on olulised ka tehisintellektisüsteemide elutsüklis [85].

Küberturvalisuse teine direktiiv NIS2 jõustus 16.01.2023 ja selles on käsitletud ka tehisintellektisüsteeme. Nimelt soovitakse viidatud direktiiviga ergutada tehisintellekti kasutamist näiteks küberrünnete avastamisel, ennetamisel ja eelnevaga seotud ressursside planeerimisel[3]. Elutähtsatele ja olulistele üksustele soovitatakse rakendada küberhügieeni põhitavasid ja asjakohasel juhul, turvalisuse suurendamiseks, võtta kasutusele tehisintellekti või masinõppesüsteemi tehnoloogiaid[4]. NIS2 kohaselt peab tehisintellekti kasutamine olema kooskõlas ka EL andmekaitseõigusega, sh põhimõtetega nagu andmete täpsus, võimalikult väheste andmete kogumine, õiglus ja läbipaistvus ning infoturve (nt tipptasemel krüpteerimine). Samuti tuleb kinni pidada IKÜMis sätestatud lõimitud ja vaikimisi andmekaitse nõuetest [21]. NIS2 kohta on ülevaatliku teabe oma veebilehel avaldanud Belgia küberturvalisuse keskus [116].

Euroopa Parlamendi ja nõukogu määruse ettepanekuga, mis käsitleb digitaalseid koostisosi sisaldavate toodete küberturvalisuse horisontaalseid nõudeid (CRA), kehtestatakse toodetele ja teenustele küberturvalisuse sertifitseerimise raamistik [117]. Määruse kehtestamise vajadust on selgitatud toodete ja teenuste üldise madala küberturvalisuse tasemega ning kasutajate ebapiisava arusaamise ja juurdepääsuga teabele toodete ja teenuste turvalisuse kohta. CRA artiklis 8 kehtestatakse nõuded suure riskiga tehisintellektisüsteemidele.

Küberturvalisus on ka tehisintellekti määruse ettepanekus [118] kesksel kohal. Näiteks on sellel oluline roll, et tagada tehisintellektisüsteemide vastupidavus katsetele muuta nende kasutamist, käitumist, jõudlust või ohustada nende turvaomadusi pahatahtlike kolmandate osapoolte poolt, kes võivad süsteemi haavatavusi ära kasutada. Ründajad võivad võtta sihikule näiteks treeningandmed (andmemürgitus), treenitud mudelid (pahatahtlik rünne või kuuluvuse tuvastamise rünne) või kasutada ära tehisintellektisüsteemi digitaalsete varade või selle aluseks oleva IKT infrastruktuuri haavatavusi. Riskidele vastava küberturvalisuse tagamiseks tuleb rakendada sobivaid ja tõhusaid meetmeid, võttes arvesse ka praegust tehnoloogia taset.

3.8 Andmekaitse ja privaatsus

Kui tehisintellektisüsteemi mistahes elutsüklis (nt mudeli treenimisel või rakendamisel) töödeldakse isikuandmeid, siis tuleb arvestada andmekaitse ja privaatsuse tagamise nõuetega. Peamine õigusakt, mis reguleerib isikuandmete töötlemist ELis, on isikuandmete kaitse üldmäärus [60]. 4. juulil 2023 avalikustas Euroopa Komisjon määruse ettepaneku, millega soovitakse kehtestada IKÜMi täitmise tagamisega seotud täiendavad menetlusnormid [119]. Lisaks on kehtestatud erinõuded ka õiguskaitseasutustele [120] ja EL institutsioonidele [121].

Lisaks eelnevale tuleb arvestada ka riigisiseste andmekaitse ja privaatsusnormidega, teatud juhtudel võivad kohalduda ka sektorispetsiifilised nõuded. Seega tuleb iga konkreetse valdkonna ja tegevusala puhul hinnata, millised on lisaks IKÜMis kehtestatud nõuetele ka vastava valdkonna erinormid. Arvestada tuleb ka osapoolte vahel sõlmitud tingimusi (nt lepingud, andmekaitsekokkulepped, teenuse osutamise tingimused).

Tehisintellekti juurutamine toob endaga kaasa vajaduse keeruliste õiguslike ülesannete lahendamiseks. Ühed kõige akuutsemad teemad on privaatsus ja andmekaitse, eriti IKÜMi nõuete valguses. Üldmäärus kehtestab kõrged andmekaitse standardid, millel omakorda on suur mõ-

[3] Vt NIS2 põhjenduspunkt 51.
[4] Vt NIS2 põhjenduspunkt 89.

ju tehisintelelktisüsteemidele, mis sõltuvad suurtest andmemahtudest [122]. Tehisintellektisüsteemi andmekaitsealastele nõuetele tagamiseks peab see võtma arvesse IKÜMi artikli 5 lõikes 1 kehtestatud isikuandmete töötlemise põhimõtteid, mille täitmise eest vastutab ja peab olema võimeline nõuete täitmist tõendama vastutav töötleja (IKÜM artikkel 5(2)). Isikuandmete töötlemisel tagatakse, et:

(a) *töötlemine on seaduslik, õiglane ja andmesubjektile läbipaistev („seaduslikkus, õiglus ja läbipaistvus“)*;
(b) *isikuandmeid kogutakse täpselt ja selgelt kindlaksmääratud ning õiguspärastel eesmärkidel ning neid ei töödelda hiljem viisil, mis on nende eesmärkidega vastuolus; isikuandmete edasist töötlemist avalikes huvides toimuva arhiveerimise, teadus- või ajaloouuringute või statistilisel eesmärgil ei loeta artikli 89 lõike 1 kohaselt algsete eesmärkidega vastuolus olevaks („eesmärgi piirang“)*;
(c) *isikuandmed on asjakohased, olulised ja piiratud sellega, mis on vajalik nende töötlemise eesmärgi seisukohalt („võimalikult väheste andmete kogumine“)*;
(d) *isikuandmed on õiged ja vajaduse korral ajakohastatud ning et võetakse kõik mõistlikud meetmed, et töötlemise eesmärgi seisukohast ebaõiged isikuandmed kustutaks või parandataks viivitamata („õigsus“)*;
(e) *isikuandmeid säilitatakse kujul, mis võimaldab andmesubjekte tuvastada ainult seni, kuni see on vajalik selle eesmärgi täitmiseks, milleks isikuandmeid töödeldakse; isikuandmeid võib kauem säilitada juhul, kui isikuandmeid töödeldakse üksnes avalikes huvides toimuva arhiveerimise, teadus- või ajaloouuringute või statistilisel eesmärgil vastavalt artikli 89 lõikele 1, eeldusel et andmesubjektide õiguste ja vabaduste kaitseks rakendatakse käesoleva määrusega ettenähtud asjakohaseid tehnilisi ja korralduslikke meetmeid („säilitamise piirang“)*;
(f) *isikuandmeid töödeldakse viisil, mis tagab isikuandmete asjakohase turvalisuse, sealhulgas kaitseb loata või ebaseadusliku töötlemise eest ning juhusliku kaotamise, hävitamise või kahjustumise eest, kasutades asjakohaseid tehnilisi või korralduslikke meetmeid („usaldusväärsus ja konfidentsiaalsus“)*.

Arvestades tehisintellektisüsteemi arendamisel ja testimisel kasutatavaid suuri andmemahte, võib tehisintellektisüsteemide teatud andmekaitsealastele nõuetele (näiteks võimalikult väheste andmete kogumine, eesmärgi ja säilitamise piirangud) vastavuse tagamine olla keerukas. Generatiivse tehisintellekti ja suurte keelemudelite kiire arenguga seoses on kerkinud küsimus, kuidas sobitada olemasolevaid andmekaitsealaseid nõudeid uude konteksti.

Mitmed andmekaitseasutused on koostanud suuniseid, kuidas rakendada andmekaitsealaseid põhimõtteid ja nõudeid tehisintellektisüsteemide arendamisel, juurutamisel ja kasutamisel. Seda on teinud näiteks Prantsusmaa andmekaitseasutus (CNIL) [123] ja Ühendkuningriigi teabevoliniku büroo (ICO) [124]. ICO algatas ka 2024. aasta alguses konsultatsioonisarja generatiivse tehisintellekti teemal, et uurida, kuidas andmekaitsealaseid nõudeid tuleks viidatud tehnoloogia arendamisel ja kasutamisel kohaldada [125]. Konsultatsioonide käigus uuritakse andmekaitsealaseid aspekte, mis puudutavad näiteks veebikoorimise teel saadud andmete kasutamist mudeli treenimiseks, generatiivsete tehisintellekti väljundite täpsust, eesmärgi piiramise põhimõtte rakendamist, andmesubjekti õiguste tagamist [126]. Konsultatsioonide tulemusena koostatakse asjaomased soovitused.

Privaatsus ja andmekaitse tuleb tagada kogu tehisintellekti süsteemi elutsükli vältel [45]. Eriti oluline on privaatsuse ja andmekaitse tagamine just seetõttu, et tehisintellekti süsteemid võivad inimese käitumisega seotud andmetest järeldada mitte ainult eelistusi, vaid ka muud isiklikku, ja küllaltki privaatset teavet, näiteks informatsiooni inimese seksuaalse sättumuse kohta, tema

vanust, sugu, usulisi veendumusi või poliitilisi vaateid. Privaatsuse ja andmekaitse tagamine on seega tehisintellektisüsteemide puhul äärmiselt oluline, sh nii süsteemi kasutaja algselt antud teabele, kui ka teabele, mis tekib süsteemi kasutamise käigus (väljundid, soovitustele reageerimine jne). Vältida tuleb andmete pinnalt mistahes ebaseaduslikku ja ebaõiglast diskimineerimist [45]. On olnud juhtumeid, kus tehisintellektisüsteemid on lekkinud sensitiivset informatsiooni, näiteks infot vestlusajalugudest [88].

EL AI ekspertrühma hinnangul on kahju tegemisest hoidumise põhimõttega põimitud tihedalt privaatsusküsimused. Privaatsuse tagamiseks tuleb rakendada asjakohast andmehaldust, mis hõlmab mh kasutatavate andmete kvaliteeti, terviklust ja pääsuprotokolle [45].

Tehisintellekti määruse ettepanekus on hinnatud vajadust teatud juhtudel hinnata tehisintellekti süsteemi mõjusid põhiõigustele ja koostada andmekaitsealane mõjuhinnang [118]. Leitud on, et viidatud mõjuanalüüside koostamine tuleks kavandada protsesside osasks selliselt, et vältida dubleerimist ja liigset halduskoormust. Loodava tehisintellekti ameti ülesandeks jääb tulevikus töötada välja küsimustik, mida tehisintellektisüsteemide juurutajad saavad asjakohaste nõuete täitmiseks kasutada [118]. Igal juhul tuleb tehisintellekti süsteeme arendada ja kasutada kooskõlas kehtivate privaatsus- ja andmekaitsereeglitega.

Kuna tehisintellektisüsteemid baseeruvad andmetel, siis on andmete kvaliteet ülioluline. See mängib olulist rolli ka tehisintellektisüsteemide struktuuri loomisel ja toimivuse tagamisel. Treening-, valideerimis- ja testandmed peaksid olema asjakohased, piisavalt esinduslikud, võimalikult veatud ja täielikud, arvestades tehisintellektisüsteemi loomise eesmärki. Nõue, et andmestikud peavad olema võimalikult täielikud ja veatud, ei tohiks mõjutada privaatsust säilitavate tehnoloogiate kasutamist tehisintellektisüsteemide arendamise ja testimise kontekstis [118].

Arvestada tuleb ka seda, et andmestike koostamine peab põhinema andmete seaduslikul kasutamisel andmekaitsealaseid nõudeid järgides [127]. Isikuandmete töötlemine on seaduslik ainult juhul, kui on täidetud vähemalt üks IKÜM artikli 6 lõikes 1 toodud tingimustest (punktid a-f). On olnud juhtumeid, kus pädevad asutused on nõudnud ebaseaduslikult kogutud andmetel põhinevate mudelite kustutamist [128]. Selleks, et vältida mistahes diskrimineerimist, peaksid andmestikud olema ka asjakohaste statistiliste omadustega ning võtma arvesse tunnuseid, mis on omased konkreetsele olukorrale või isikute rühmale.

Üldmääruse nõuete järgimiseks tuleb tehisintellekti süsteem välja töötada, treenida ja selgelt määratletud eesmärgiga kasutusele võtta. Prantsusmaa andmekaitseasutus (CNIL) on soovitanud tehisintellekti eesmärgi kindlaks määrata juba projekti kavandamise käigus. Eesmärk peab olema seaduslik, selge ja arusaadav, ning sellest lähtuvalt on võimalik määrata, milliseid andmeid konkreetse eesmärgi täitmiseks on vaja töödelda, samuti, kui kaua neid andmeid on vaja säilitada, et eesmärgid saavutada [127].

Kuigi eesmärgi piiramise põhimõte nõuab isikuandmete kasutamist ainult eelnevalt määratletud konkreetse eesmärgi saavutamiseks, võib see tehisintellektisüsteemi puhul osutuda keerukaks. CNIL on leidnud, et algoritmi treenimise etapis ei ole alati võimalik määratleda kõiki tehisintellekti tulevasi kasutusvõimalusi, samas peaksid süsteemi tüüp ja peamised võimalikud funktsioonid olema siiski võimalikult täpselt määratletud [129].

Üldmääruse ekstraterritoriaalse jõustamise ümber toimuvad arutelud annavad põhjust arvata, et viidatud õigusaktis juurutatud jurisdiktsioonimudel, mis on ühtlasi leidnud tee ka ELi tehisintellekti määrusesse, ei pruugi praktikas toimida [130, 131, 132, 133]. IKÜMi artikli 3 lõike 2 punktide a ja b kohaselt kohaldatakse IKÜMit ka ELis asuvate andmesubjektide isikuandmete töötlemise suhtes mujal kui liidus asuva vastutava töötleja või volitatud töötleja poolt, kui andmete töötlemine on seotud liidus asuvatele andmesubjektidele kaupade ja teenuste pakkumisega või nende

tegevuse jälgimisega, kui see tegevus toimub liidus.

Üldmääruse rakendamise ajal on olnud mitmeid vaidlusi, mis puudutavad just isikuandmete töötlemist kolmandate riikide vastutavate või volitatud töötlejate poolt, kes kuuluvad IKÜMi artikli 3 lõike 2 reguleerimisalasse, kuid kes ei soovi teha koostööd Euroopa andmekaitseasutustega või ei tunnusta ELi jurisdiktsiooni (nt Clearview AI kaasus) [134, 132]. Ka tehisintellekti määruse ettepanekus kasutatakse üldmäärusele sarnast lähenemisviisi, kus kohaldamisalasse on kirjutatud kolmandate riikide ettevõtjad (vt artikkel 2(1)(c)) [99]. Praktikas võivad selle normi rakendamisel pädevaid asutusi ees oodata sarnased probleemid, mis on tekkinud üldmääruse ekstraterritoriaalse jõustamisega.

Isikuandmete edastamine kolmandatele riikidele ja rahvusvahelistele organisatsioonidele on reguleeritud IKÜMi peatükis V. Andmete edastamine on üldiselt lubatud siis, kui selleks on olemas sobiv õiguslik alus (IKÜM artiklid 6, 9) ning rakendatakse asjakohaseid ja tõhusaid kaitsemeetmeid [135]. Euroopa Komisjonil on õigus IKÜM artikli 45 alusel kindlaks teha, kas väljaspool ELi asuv riik või rahvusvaheline organisatsioon pakub piisavat andmekaitse taset [136], [137]. Näiteks võttis 2023. aasta juulis komisjon vastu adekvaatsusotsuse ELi-USA andmekaitseraamistiku piisavuse kohta [138].[5] Kui vastav komisjoni otsus on olemas, ei ole andmete edastamiseks eriluba vaja (IKÜM artikkel 45(1)). Euroopa Majanduspiirkonna (EMP) riike (Norra, Island, Liechtenstein) loetakse piisava andmekaitsetasemega riikideks.

Mittepiisava andmekaitsetasemega riiki andmete edastamisel tuleb kindlasti rakendada lisakaitsemeetmeid (vt nt [139]) või tegemist peab olema IKÜMi nõuetele vastava erandolukorraga (IKÜM artiklid 46-49) [140]. Euroopa Andmekaitsenõukogu (EDPB) on leidnud, et teatud juhtudel võivad kõnealuseks andmeedastuseks kvalifitseeruda ka kaugjuurdepääs kolmandast riigist (nt tugiteenused, tõrkeotsing), aga ka andmete salvestamine väljaspool EMPd asuvas pilvteenuses [141]. Seetõttu on tehisintellektiga seotud taristu läbimõtlemine enne teenuseosutajatega lepingute sõlmimist rangelt soovitatav, et vältida hilisemad õigusvaidlusi või sanktsioone.

3.9 Õigusraamistiku olulisus

Tehisintellektisüsteemi elutsüklis rolli mängiv isik peab olema kursis õiguslike ja regulatiivsete nõuetega, mis kujundavad õigusraamistiku, kus tegutsetakse. See määrab ära, millistele nõuetele peab AI-süsteem, aga ka seda opereeriv isik, vastama. Eelnevaga sõltuvuses on ka aspektid, kuidas korraldatakse ja juhitakse AI-süsteemiga seotud protsesse, näiteks süsteemi arendamist, testimist ja kontrollimist.

Üha olulisemaks muutuvad organisatsioonis infotehnoloogia, turvalisuse ja õiguslike teemade kombineeritud käsitlus. See tähendab ka vastavaid rolle täitvate isikute tihedat koostööd alates AI-süsteemi kavandamisest kuni selle elutsükli lõpuni. See omakorda võimaldab avardada juristide teadmisi tehnoloogiast ja tehnoloogia spetsialistide teadmisi õiguslikest nõuetest, panustades organisatsiooni teadmuse kasvu.

Mida rohkem ollakse teadlikud õigusraamistikust tulenevatest nõuetest, seda juba AI-süsteemi kavandamisel, ja viidatud nõuetele vastavuse tagamist ka päriselt tehakse, seda väiksem on tõenäosus ebasoovitavate stsenaariumite realiseerumiseks. Arvestada tuleb, et tehisintellektisüsteemidega sotud õigus on alles arenemisjärgus ja õiguskeskkonna muudatusi on oodata ka edaspidi.

[5] Varasemad sellised lepped ja otsused Euroopa Liidu ja Ameerika Ühendriikide vahel on korduvalt õigustühiseks tunnistatud. Soovitame uuringu lugejal enne Euroopa Liidu kodanike andmete Ameerika Ühendriikidesse edastamist jälgida kehtivat õiguslikku seisu.

4 Tehisintellekti rakenduste levitusmudelid

4.1 Sissejuhatus

AI rakenduste arendajatel on palju erinevaid levitusvõimalusi. On AI mudeleid, mis on vabavaraliselt kättesaadavaid ning on AI mudeleid, millele saab ligi tasuliste rakendusliideste kaudu. Järgnevalt on rõhuasetus pilvteenustel, kuna andmete liikumisega eri andmetöötlejate vahel kaasnevad privaatsusele täiendavad riskid. Samuti on pilvandmetöötlus (või teise osapoole andmekeskuste kasutamine üldiselt) ka väga levinud.

Tehniliselt lihtsaim on õhuke äriloogikat teostav rakendus, mis kasutab mõnda olemasolevat rakendusliidest. Näitena võime kujutleda OpenAI GPT mudeli rakendusliidest kasutavat juturobotit, kus põhiline väärtuspakkumine seisneb kasutajamugavuses ja pakutavates viipades. Need õhukesed lahendused võivad olla piiratud rakendusliidese taga olevate mudelite kontekstõppe võimekusega.

Keerukamad ja kallimad lahendusedkasutavad mõnda mudelit API väljakutsete kaudu, kuid haldavad seejuures kasutaja olekut ja teenindavad tema andmeid, mis võivad ka domeenispetsiifilised olla. Need eeldavad andmebaaside integreerimist, kasutajahaldust või sisend-väljundi valideerimist. Näiteks võib rakenduse levitaja kasutada mõnda *Retrieval-Augmented Generation* ehk RAG lahendust, kus mudeli üldteadmisi täiendatakse andmebaasis sisalduva infoga. Selliseid lahendusi käsitletakse alamjaotises 4.4.2.

On ka lahendusi, kus teenuseandja levitab oma mudelit ise. See eeldab, et teenuseandja kas treenib oma mudeli ise, peenhäälestab olemasolevat või võtab üle mõne välise mudeli, kuid igal juhul käivitab selle peal ise inferentsi (arvutab tehisintellekti väljundid enda taristul). See nõuab mudeli ja kasutajabaasi suurusega kasvavaid investeeringuid infotaristusse, kuid võib leevendada riske andmete konfidentsiaalsusele ja privaatsusele, kuna väheneb andmetöötlejate arv. Kui eesmärk ei ole teenindada suurt kasutajabaasi, siis on kvantimise jm. optimeerimismeetoditega võimalik käivitada inferentsi paljude vabalt kättesaadavate mudelite peal ka võimsamast personaalarvutist. Selliseid lahendusi käsitletakse alamjaotistes 4.4.3 ja 4.4.4.

Kõikidel siintoodud levitusmudelitel on ühiseid omadusi. Näiteks võib teenuseandja kasutada oma äriloogika, mudeli ja andmete haldamiseks IaaS (taristu teenusena), CaaS (arvutusjõudlus teenusena) ja PaaS (platvorm teenusena) teenuseid. Nende teenuseandjad on kasutajate andmete töötlemisel IKÜM kontekstis volitatud töötlejad. Kui kasutaja andmeid kasutatakse lisaks teenuste pakkumisele ka mudeli kvaliteedi parandamiseks või muudeks kõrvalisteks ülesanneteks, peab kasutaja andma nendeks otstarveteks teadliku nõusoleku. See puudutab võimalust liidestada teenus muude teenuste ja andmetega.

4.2 Metoodika

Levitusmudelite väljatöötamisel lähtusime võimaliku teenuseandja kaalutlustest ja vajadustest ning nende tavapraktikatest. Pöörasime erilist tähelepanu seadustes olevatele nõuetele ning kasutaja andmete liikumisele erinevate töötlejate vahel. Järgnevalt toodud levitusmudelite käsitlus ei ole ammendav, sest viise erinevate teenuste, rakendusliideste ja andmeallikate kokkuühendamiseks on väga palju. See võiks siiski olla piisav andmaks ülevaate sagedasemate lähenemiste valupunktidest, mis on seotud kasutaja ja teenusandja(te) rollide ja vastutusalaga levitusmude-

li ülesehituse ning andmevoo kontekstis. Lisaks aitavad lihtsamad mudelid anda kiiremini nõu riskianalüüsi läbiviimiseks.

Noolte abil kujutame andmevoogu, mis näitab andmete liikumist levitusmudeli erinevate komponentide vahel. Selle kujutamine on oluline, sest andmete liikumisega üle vastutusalade piiride kaasnevad riskid (nt privaatsusele), millega tuleb arvestada. Privaatsuse ja vastutusala käsitlemisel lähtume üldmäärusest. Et paremini mõista vastutusalade piire, ning muid konkreetse tehisintellekti tarneahela ülesehitusest tulenevaid levitusmudeli omadusi, kujutame levitusmudeli komponentidena nii teenuseid kui tähtsamaid tehisintellektisüsteemi andmeelemente (treenimisandmed, mudel, sisend, väljund). Oleme seadnud rõhuasetuse tehisintellekti kasutavatele pilvteenustele, sest jõudlusnõuete tõttu kasutavad tehisintellektisüsteemid pilvteenustes arvutusi kiirendavat eririistvara. Siiski tuleb arvestada, et on AI-süsteemides, mida ei levitata pilve kaudu, on riskide ulatus mõnevõrra piiratum ning neid käsitleme eraldi. Me ei ole levitusmudelite joonistel eraldi kujutanud IaaS, CaaS ja PaaS komponente, sest neid võib vabalt kasutada paljude levitusmudeli elementidega, kuid käsitleme nende kasutuselevõtu tagajärgi.

Tehisintellekti rakenduste levitusmudelite täpsemaks analüüsiks kasutame talitluse analüüsi. Modelleerime mudelid Business Process Modelling Notation (BPMN) abil. See aitab meil täpsustada mudelis töödeldavaid andmeobjekte ning osapooli, kes neid töötlevad.

4.3 Tehisintellektisüsteemi osapoolte õiguslikud rollid

Nii üldmääruse kui ka tehisintellekti määruse seisukohast on esmatähtis hinnata määruste kohaldamisala. Üldmääruse nõudeid tuleb hinnata juhul, kui AI-süsteem töötleb mistahes selle elutsüklis isikuandmeid. Tehisintellekti määruse nõudeid tuleb hinnata näiteks juhul, kui tegeletakse tehisintellektisüsteemi arendamisega või kasutatakse oma teenuses kellegi teise arendatud AI-süsteemi või API-t. Tehisintellektisüsteemiks loetakse tehisintellekti määruse kohaselt masinapõhist süsteemi, mis on loodud töötama erineva autonoomia tasemega ja mis võib pärast kasutuselevõttu olla kohanemisvõimeline ning mis otseste või kaudsete eesmärkide puhul järeldab saadud sisendi põhjal, kuidas genereerida väljundeid, nagu ennustused, sisu, soovitused või otsused, mis võivad mõjutada füüsilist või virtuaalset keskkonda [99].

Kui ollakse veendunud, et tehisintellektisüsteem või seda opereeriv isik kuulub määrus(t)e kohaldamisalasse, tuleb hinnata, millised konkreetsed nõuded määrus(te)st tulenevad. Üldmääruse kohaselt on oluline hinnata, kas kvalifitseerutakse näiteks isikuandmete vastutavaks või volitatud töötlejaks, tehisintellekti määruse puhul aga näiteks AI-süsteemi teenustajaks (*”provider”*) või juurutajaks (*”deployer”*). Määruste kohaseid rolle on veelgi ja ka need on mõistlik üle vaadata. Eelnevalt viidatud rollid on siiski põhilised, eriti vastutav töötleja üldmääruse ja teenustaja tehisintellekti määruse kohaselt, kuivõrd nendele on kehtestatud ranged vastavusnõuded. Samuti võib juhtuda, et ühel isikul on samaaegselt mitu rolli vastavalt erinevatele protsessidele, pooltevahelistele suhetele või kokkulepetele. Kuna rollidest tulenevalt sõltub vastutus, on nende tuvastamine äärmiselt oluline.

Isikuandmete kaitse üldmääruse kohaselt loetakse isikustatavate andmete vastutavaks töötlejaks isikut, kes üksi või koos teistega määrab kindlaks isikuandmete töötlemise eesmärgid ja vahendid (IKÜM artikkel 4(7)). Volitatud töötlejaks loetakse sellist isikut, kes töötleb isikuandmeid vastutava töötleja nimel ja juhistele vastavalt (IKÜM artikkel 4(8)).

Teenustajaks[1] loetakse isikut, kes töötab välja tehisintellektisüsteemi või üldotstarbelise tehisin-

[1] Tehisintellekti määruse esialgse ettepaneku eestikeelses versioonis oli kasutusel termin ”pakkuja” (*”provider”*), kuid see ei ole õnnestunud tõlge, mistõttu kasutatakse siin ja mujal läbivalt terminit teenustaja või teenuse andja.

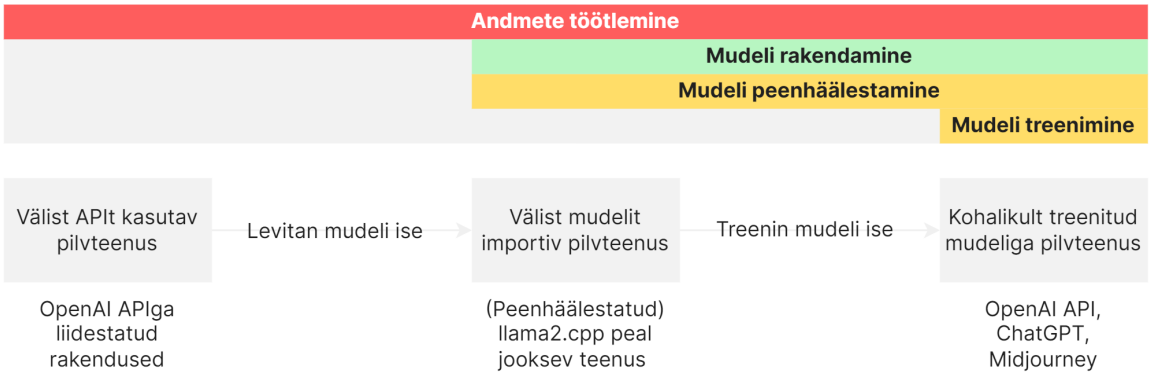
tellektimudeli või laseb eelnimetatud AI-süsteemi või AI mudeli välja töötada, eesmärgiga see turule lasta või kasutusele võtta oma nime või kaubamärgi all kas tasu eest või tasuta [99]. Juurutajaks loetakse isikut, kes kasutab AI-süsteemi oma volituste alusel, välja arvatud juhul, kui viidatud süsteemi kasutatakse isikliku, mitte kutselise tegevuse jaoks [99].

Konkreetsete nõuete tuvastamiseks on vaja teada ka seda, mis on andmetöötluse ja tehisintellekti kasutamise eesmärk, millised andmetöötlusprotsessid süsteemis toimuvad, millised andmed ja kelle vahel liiguvad ning millist AI-süsteemi või komponenti (sh selle riskitase) süsteemis kasutatakse.

4.4 Levitusmudelid

4.4.1 Mudelite ülevaade

Oleme välja töötanud kolm tehisintellekti rakeduste levitusmudelit, mis erinevad osapoolteva-helise andmete liikumise, levitava osapoole ja tehisintellekti mudeli päritolu poolest. Nende mudelite omavahelist seost ning näidisrakendusi esitab joonis 9.



Joonis 9. Levitusmudelite ülevaade levitaja ülesannete perspektiivist seoses AI mudeliga

Oleme mudelid järjestanud lähtuvalt sellest, kui suures osas saab AI rakenduse teenuseandja toetuda juba olemasolevatele AI teenustele ja toodetele. Mida spetsiifilisemaks ja keerulisemaks läheb äriline otstarve ning mida rangemad on nõudmised andmete töötlemise osas, seda suurema osa vajalikest teenustest tuleb harilikult arendada ise. See heuristika on siiski ligikaudne – andmevoo ülesehituselt hõlmab viimane levitusmudel nii levituselt lihtsaid kui keerulisi lahendusi.

Joonise ülemises osas on näidatud levitaja eri ülesannete ulatust üle erinevate levitusmudelite. Kõigil juhtudel töötleb levitaja mingeid andmeid. Alates välist mudelit importivast pilvteenusest levitab lisaks oma äriloogikale levitaja ka mudelit ennast, vajadusel seda peenhäälestades. Kohalikult treenitud mudeli puhul ei vastuta ükski kolmas osapool enam mudeli loomise ja treenimise eest, vaid see (ning treeningandmete haldus) on täielikult levitaja kätes.

4.4.2 LM1: AI-d rakendusliidese kaudu kasutav teenus

AI-põhiste teenuste loomisel on levinud arhitektuurivalik oma äriloogikas kolmanda osapoole antavat AI rakendusliidese kasutamine. Vajadusel saab teenuseandja täiendavalt töödelda või hoida kasutaja andmeid, milles võib levitaja samuti tugineda pilvteenustele. Kolmanda osapoole mudeli treenimisel kasutatud algandmed võivad seejuures olla pärit omakorda välistest allikatest.

Samuti võib see kolmanda osapoole AI pilvteenus või rakendusliides kasutada oma mudelite treenimiseks teenuseandjalt saadud kasutaja andmeid. Kõik eelmainitud pilvteenused võivad kasutada mõnda IaaS ehk taristu teenusena lahendust.

Selline mudel on olnud kasutusel näiteks masinnägemise rakendustes. Mudel muutus populaarsemaks pärast OpenAI rakendusliidese avaldamist, sest võimaldas mugavalt liidestada oma teenust võimsate keele- ja piltmudelitega. AI mudel on selle puhul väline (st ei ole teenuseandja kontrolli all). Samuti on välise päritoluga mudeli treeningandmed. Kasutaja andmed liiguvad teenusesse, teenusest AI rakendusliidese andjale, sealt tagasi teenusesse ning teenusest tagasi kasutajale. Kui teenus on liidestatud kolmandate osapoolte teenuste ja andmetega, siis võivad andmed sattuda ka sinna. Kasutaja andmeid võidakse seejuures talletada nii teenuseandja kui ka AI rakendusliidese andja poolt (nt sisendi-väljundi hoidmine vahemälus, aga ka treeningandmete andmebaasis). Käesoleva levitusmudeli erijuhuks on olukord, kus AI rakendusliidese andja pakub ka mudeli peenhäälestamise võimalust teenuseandja andmete põhjal, kuid peenhäälestatud mudelit levitab ikkagi ise. See langeb osaliselt määral kokku järgmise levitusmudeliga (vt peatükk 4.4.3).

LM1: AI rakendusliidest kasutav teenus

Lühidalt: Teenus liidestub välise rakendusliidesega, et töödelda kasutaja andmeid AI rakendusliidese andja mudeliga. Seejuures võib nii teenus kui rakendusliidese andja jagade andmaeid kolmandate osapooltega täiendavaks töötlemiseks. Mudeli treenimisel kasutatud algandmed võivad pärineda kolmandatest allikatest.

Näited: copy.ai, Streamlit ja Gradio AI demorakendused, OpenAI rakendusliidest kasutavad teenused

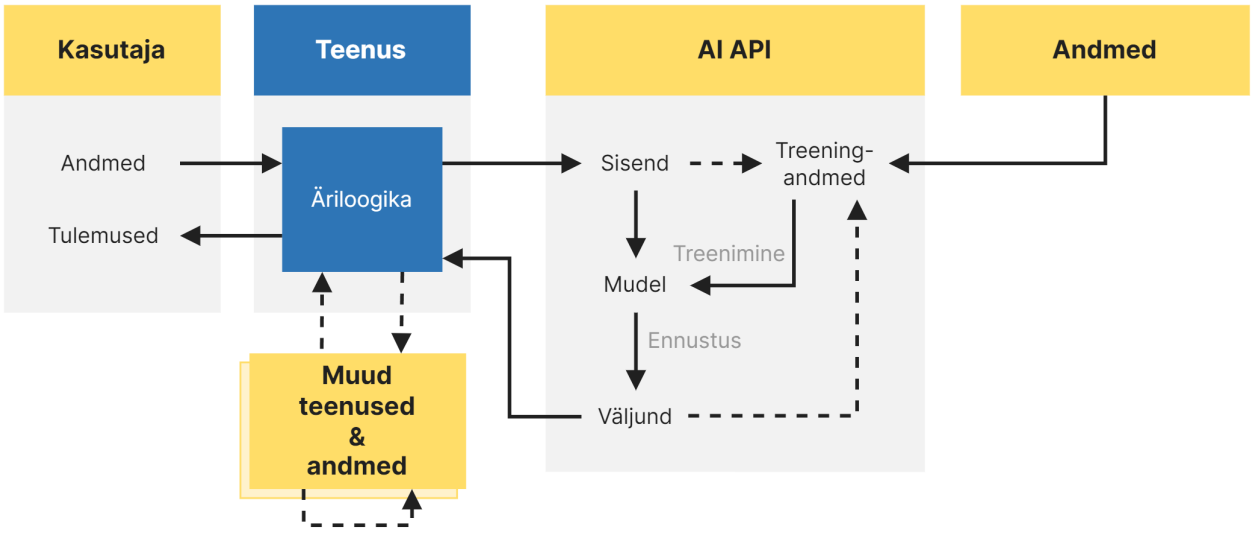
Mudeli päritolu: väline

Treeningandmete päritolu: välised

Kasutaja andmeid hoiustatakse: võimalik

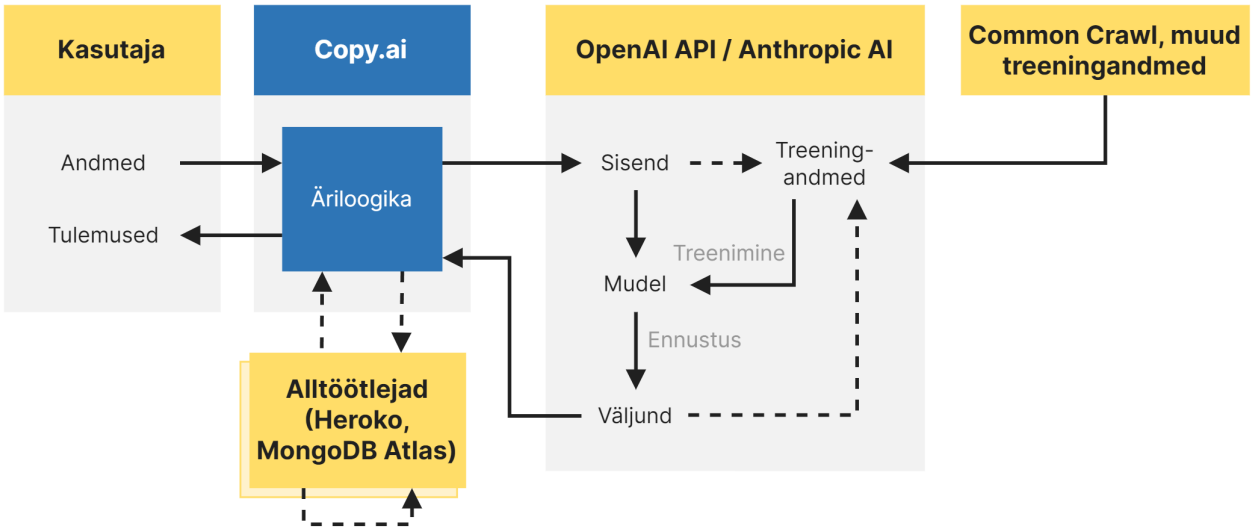
Kasutaja andmete liikumine: pilvteenusesse, sealt vajadusel muudele teenustele ning rakendusliidesesse, tagasi teenusesse, tagasi kasutajale, võimalik, et läbi erinevate taristute.

Joonis:



Ohud ja kaalutlused:

1. Isikuandmete töötlemine teenuseandja või rakendusliidese pakkuja pool.
2. Rakendusliidese andja kasutatava mudeli seletatavuse kohta võib olla teave puudulik.
3. Kasutaja päringut ja mudeli väljundit kontrollitakse teenuseandja juures.
4. Teenuseandjast sõltumatud tõrked AI API töös on oht käideldavusele.
5. Kõige madalamate kapitaliinvesteeringutega ja tehniliselt lihtsam levitusmudel.



Joonis 10. Copy.ai kui näide AI-d rakendusliidese kaudu kasutavast teenusest

Näide LM1 tüüpi teenusest on copy.ai[2]. Copy.ai kasutab OpenAI rakendusliidest ja aitab kasutajal kirjutada turundus- ja reklaamtekste. Kasutaja annab teenusele vajaliku teksti kirjelduse ning omadused (näiteks tekstistiili), teenus töötleb kirjeldusi ning esitab need päringuna tehisintellekti rakendusliidesele. Seejuures kasutaja saab ise valida, millist APIt ta soovib kasutada (Anthropic või OpenAI). Olles saanud päringu vastuse, töötleb teenus seda täiendavalt ning tagastab tulemuse kasutajale. Copy.ai levitusmudel on kujutatud joonisel 10.

Joonis 11 kirjeldab esimest tüüpi levitusmudeli (LM1) andmevooge. Esimest tüüpi levitusmudelil kasutab kasutaja teenust, mis omakorda kasutab väljundi loomiseks AI rakendusliidest (API). API pakkuja jaguneb vastavalt ülesannetele kaheks - mudeli arendamine ja teenuse levitamine. Mudeli arendamise eesmärgiks on välja töötada mudeli arhitektuur, mudelit treenida ja testida ning vajadusel luua peenhäälestusandmestikud ja peenhäälestada mudelit. Lisaks vastutab mudeli arendamine ka mudeli monitooringu eest.

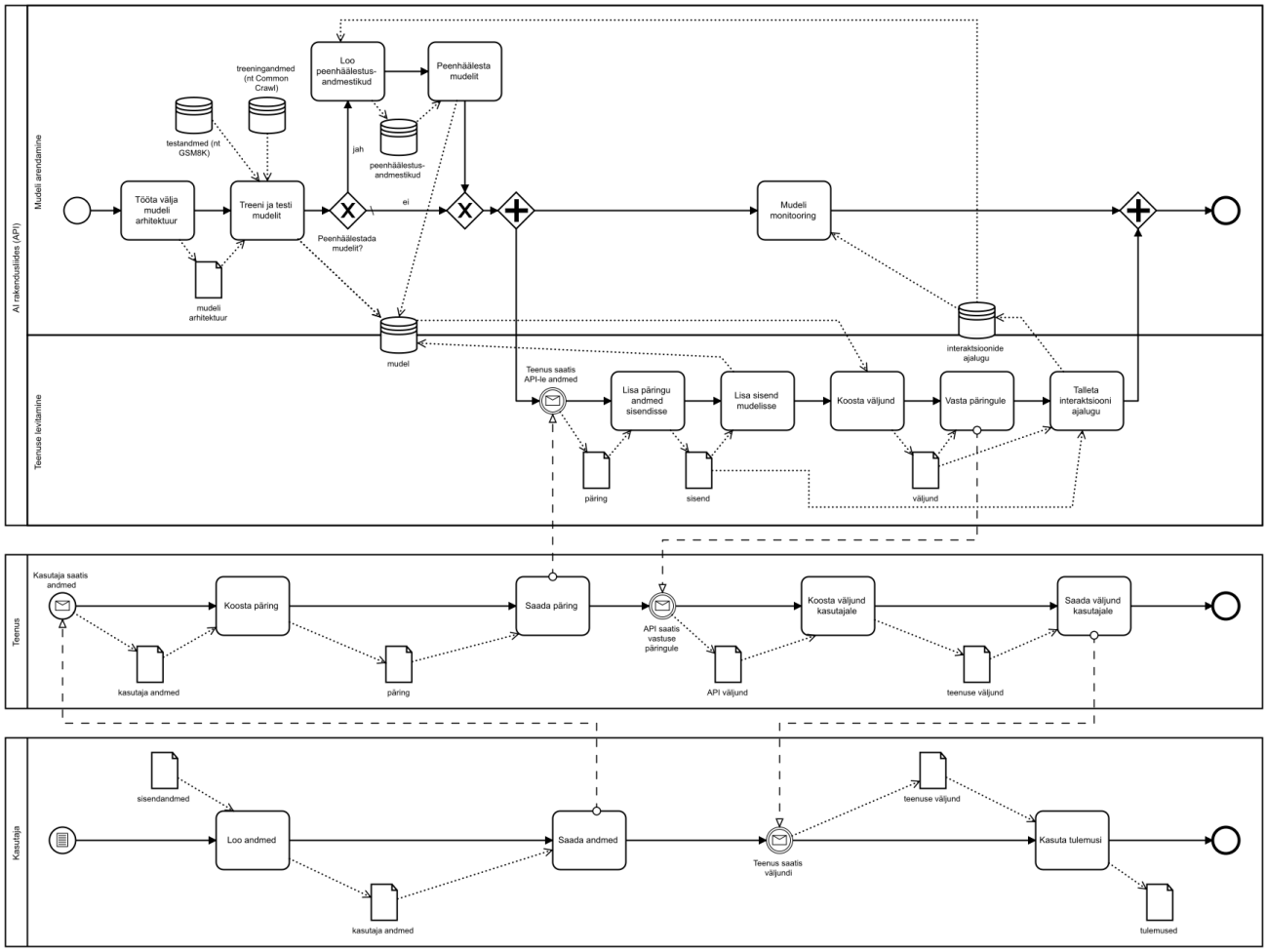
Teenuse levitamise protsess saab alguse kasutajast, kes loob enda sisendandmete põhjal vastavad andmed, mille ta edastab teenusele. Selle põhjal koostab teenus päringu ning saadab selle API-le. Kui teenus on saatnud API-le päringu, lisab API päringu andmed sisendisse ning sisendi mudelisse. Mudeli abiga koostatakse väljund ning vastatakse päringule. Interaktsioonide ajalugu võidakse talletatada ning kasutatada nii mudeli monitooringul kui peenhäälestusandmestike loomisel, kuid see oleneb teenustingimustest.

Kui API on edastanud teenusele vastuse päringule ehk väljundi, koostab teenus omakorda kasutajale väljundi ning edastab selle kasutajale. Kasutaja saab AI teenuselt saadud väljundit kasutada enda soovitud tulemuste saavutamiseks.

4.4.3 LM2: Välist AI mudelit rakendav teenus

Liidestumine välise AI rakendusliidese või veebiteenusega teeb levitaja sõltuvaks tolle teenuse käideldavusest. Samuti võib levitajal tekkida vajadus mudeli peenhäälestamiseks, mida kõik AI API pakkujad ei võimalda. Nende probleemide lahendamiseks võib levitaja võtta üle mõnelt mudeli pakkujalt (või vabavaraliselt leviva) juba eeltreenitud mudeli ning integreerida selle oma rakendusega otse. Juhul, kui levitaja peenhäälestab mudelit - seda nimetatakse ülekandeõppeks - tekib tal täiendav vajadus treeningandmete haldamiseks ja mudeli turvalisuse- ning kvaliteedinäitajate monitoorimiseks. See mudel rakendub ka erijuhul, kui mudeli pakkuja on tsentraliseeritud komponentidega liitõppe teenuseandja.

[2] https://www.copy.ai/



Joonis 11. Levitusmudel LM1 andmevood

LM2: Välist AI mudelit rakendav teenus

Lühidalt: Teenuseandja kasutatav (ja vajadusel peenhäälestatav) väljaspoolt imporditud mudel. Algne mudel on väline; mudeli looja treenib mudeli ning edastab teenuseandjale. Teenusandja levitab mudelit, kasutades seda enda ja oma klientide andmete peal, seejuures võib ta mudelit oma andmetega peenhäälestada. Pilvteenus võib liidestuda muude andmete ja teenustega, näiteks RAG-lahenduse puhul vektorandmebaasiga.

Näited: Huggingface'i hoidlast mudelit importiv teenus, Android Gboard (liitõppe näitena)

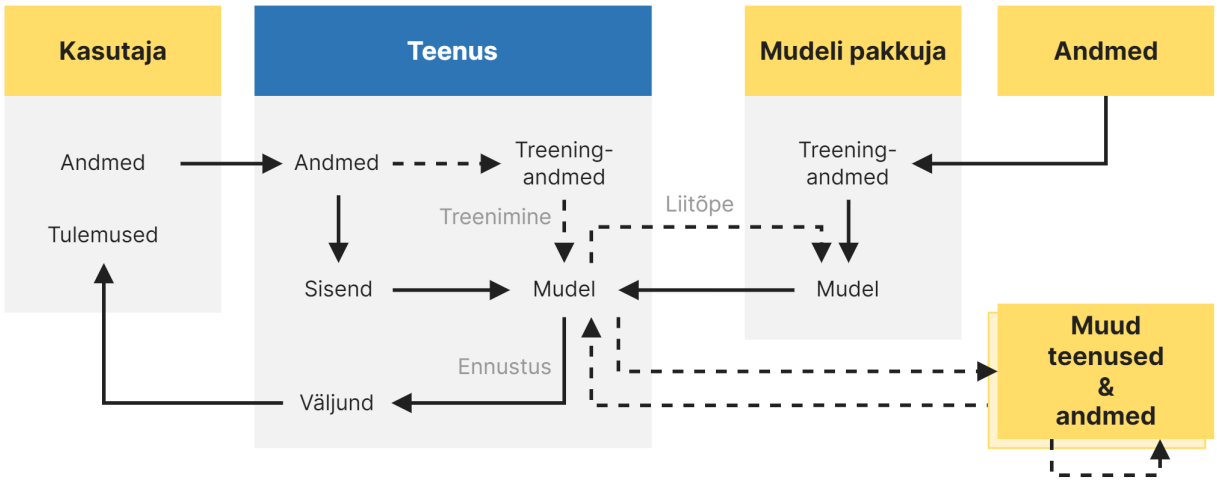
Mudeli päritolu: väline

Treeningandmete päritolu: välised, teenusandja ja kasutajate omad

Kasutaja andmeid hoiustatakse: võimalik

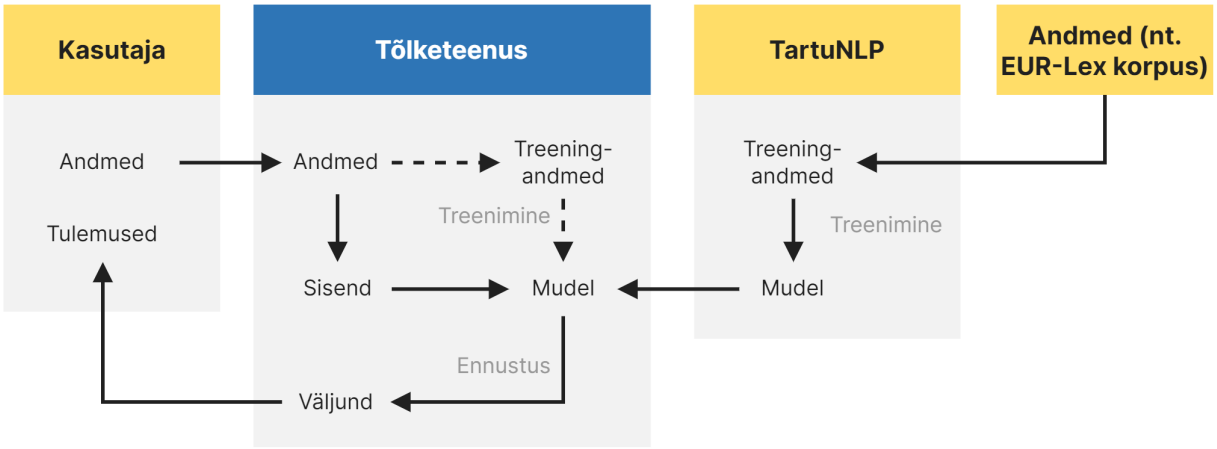
Kasutaja andmete liikumine: pilvteenusesse, vajadusel muudesse teenustesse ja kasutajale tagasi, võimalik, et läbi mõne eraldi osapoole taristu. Liitõppe korral liiguvad kaalude uuendused ka mudelitreenijale.

Joonis:



Ohud ja kaalutlused:

1. Peenhäälestamisel jälgida ohutuse ja kvaliteedimõõdikuid ning oma andmete kvaliteeti ning muudatusi nende jaotuses.
2. Teise osapoole treenitud mudeli ohutuse ja selgitatavuse kohta tuleb koguda teavet.
3. Liitõppe korral saab teatud juhtudel pidada kaalude uuendusi isikuandmeteks.



Joonis 12. Tõlketeenuse levitusmudel kui näide välist AI mudelit rakendavast teenusest

Joonisel 12 on kujutatud näide tõlketeenusest, kus AI teenusandjalt levitatav tõlketeenus kasutab mudelipakkuja (kelleks on TartuNLP) poolset eeltreenitud mudelit. Kasutaja saadab teenusele päringu (nt. sisestades selle rakenduse veebiliideses), päring liigub teenusesse, kus selle andmeid töödeldakse (tõlgitakse). Tõlgitud vastus tagastatakse kasutajale. Andmed ei liigu teenusandjalt mudelipakkujale. Teenusandja võib mudelit kasutaja andmete põhjal täiendavalt peenhäälestada.

Joonis 13 kirjeldab teist tüüpi levitusmudeli andmevooge. Selle mudeli protsessis on kolm osapoolt - kasutaja, AI teenus ning mudelipakkuja. Mudelipakkuja töötab välja mudeli arhitektuuri, treenib ja/või peenhäälestab ning testib mudelit ja seejärel pakub mudelit AI teenusele.

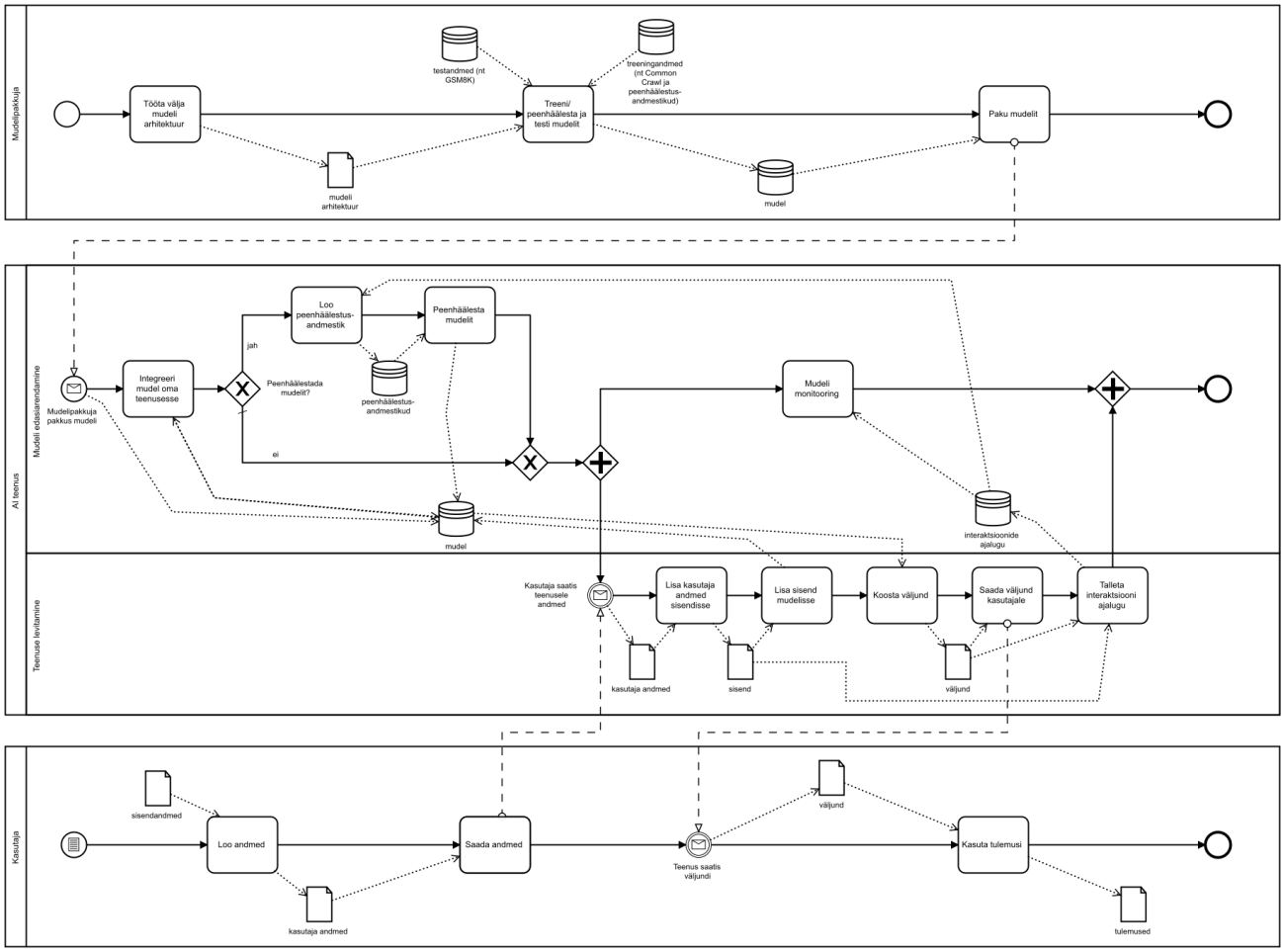
AI teenus jaguneb kaheks - mudeli edasiarendamine ja teenuse levitamine. Mudeli edasiarendamisel integreerib AI teenus mudelipakkuja poolt pakutud mudeli oma teenusesse, vajadusel loob peenhäälestusandmestiku ja peenhäälestab mudeli. Seejärel jätkab mudeli monitooringuga. AI teenusest andmed mudelipakkujale tagasi ei liigu.

Kui kasutaja on loonud andmed ning saatnud need AI teenusele, lisab AI teenuse levitamise pool need andmed sisendisse, seejärel mudelisse ja siis koostab väljundi, mille saadab kasutajale. Seejärel talletab AI teenus interaktsioonide ajaloo ning seda kasutatakse mudeli monitooringul ja seda võidakse kasutada ka peenhäälestusandmestike loomisel. Kasutaja saab AI teenuselt saadud väljundit kasutada enda soovitud tulemuste saavutamiseks.

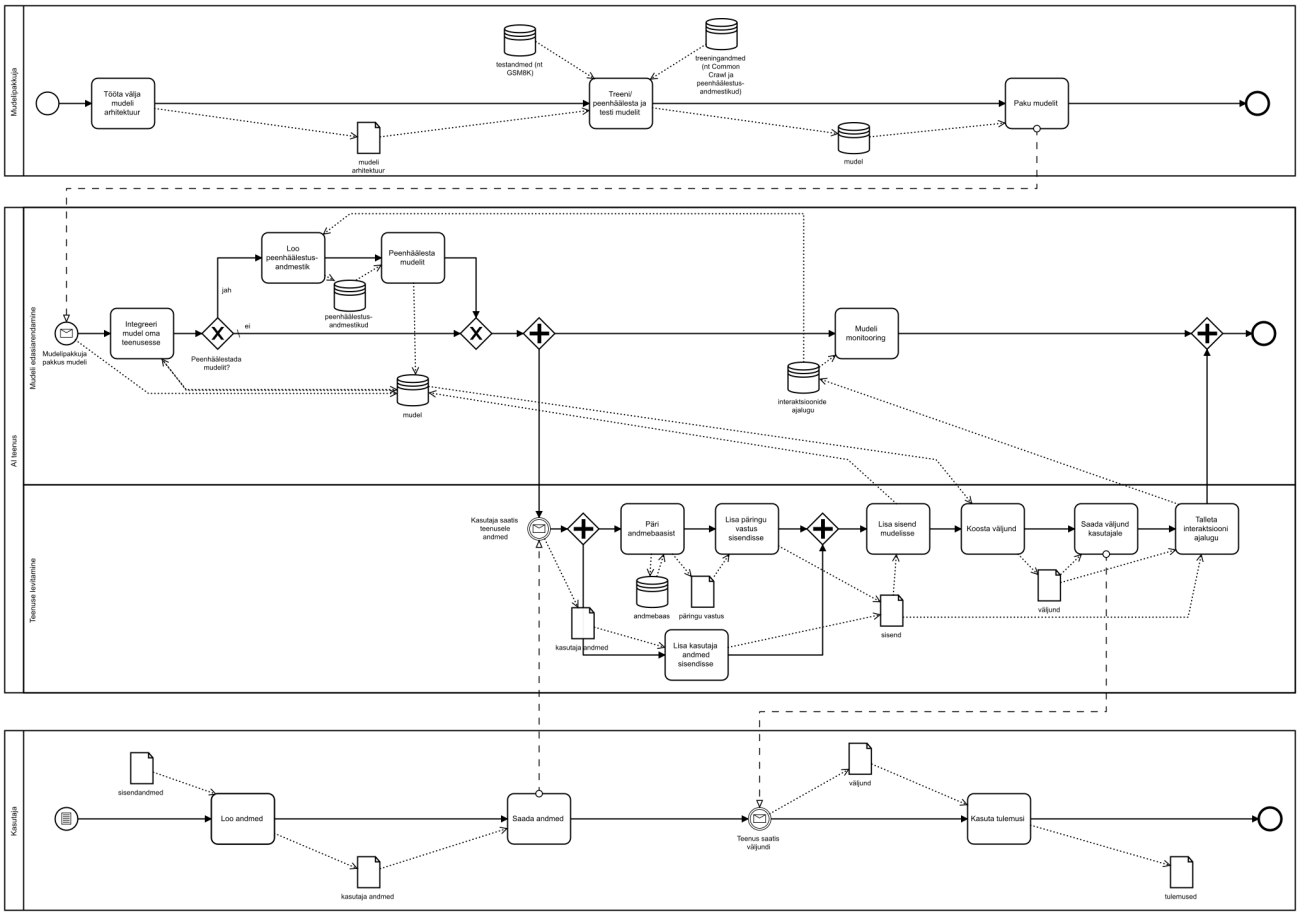
Joonis 14 on teise levitusmudeli erijuht. Kahe joonise erinevuseks on AI teenuse juures oleva teenuse levitamise osas toimuv päring andmebaasile, mille vastus lisatakse kasutaja saadetud sisendile. Seda nimetatakse RAGiks (*Retrieval Augmented Generation*) ning see saab esineda ka levitusmudelites LM1 ja LM3.

4.4.4 LM3: Ise treenitud mudeliga AI teenus

Kolmas levitusmudel hõlmab lahendusi, kus tehisintellekti mudel treenitakse ja levitatakse teenuseandja poolt. See hõlmab nii lihtsaid lahendusi, näiteks otsustuspuid ja regressioone kasutavaid lahendusi, mille lihtsuse tõttu ei pruugi mudeli väljast importimine olla otstarbekas, kui ka suurte AI-tootjate väljatöötatud lahendusi. Suurte AI mudelite treenijad pakuvad enamasti ainult enda väljatöötatud mudelitele tuginevaid teenuseid ja neil on piisavalt ressursse nende iseseisvaks levitamiseks.



Joonis 13. Levitusmudel LM2 andmevood

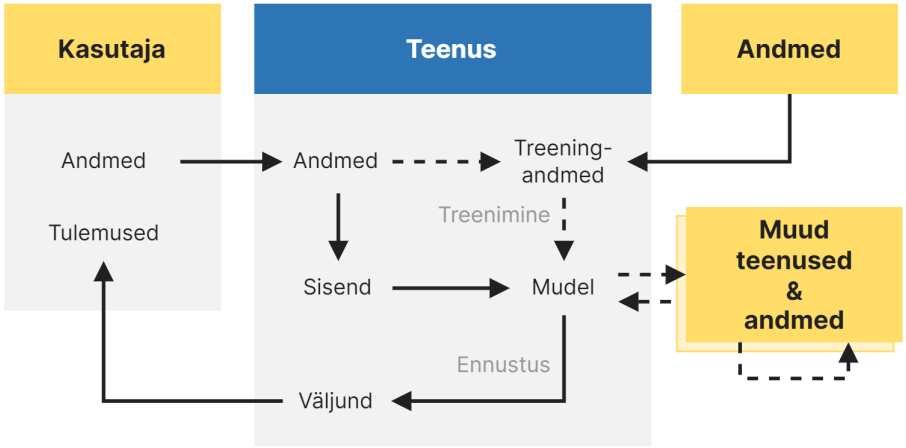


Joonis 14. Levitusmudel LM2 andmevood RAG rakendamise puhul

LM3: Ise treenitud mudeliga AI teenus

Lühidalt: Mudeli treenija kogub andmeid, treenib, levitab (ning soovi korral rakendab) mudelit. Enda treenitud mudeli ise kasutamine on oluline kasutusjuhtum. Võimalik on saavutada olukord, kus treenimisandmed, mudel ega kasutajate andmed ei liigu kolmandatele osapooltele.	**Näited:** Neurotõlge, ChatGPT ja OpenAI API, Grok, Dall-E, Midjourney
	Mudeli päritolu: sisemine
	Treeningandmete päritolu: kasutajate, teenuseandja ning kolmandate osapoolte omad
	Kasutaja andmeid hoiustatakse: võimalik
	Kasutaja andmete liikumine: pilvteenusesse, vajadusel muudesse teenustesse ja kasutajale tagasi, võimalik, et läbi taristuteenuse andja.

Joonis:

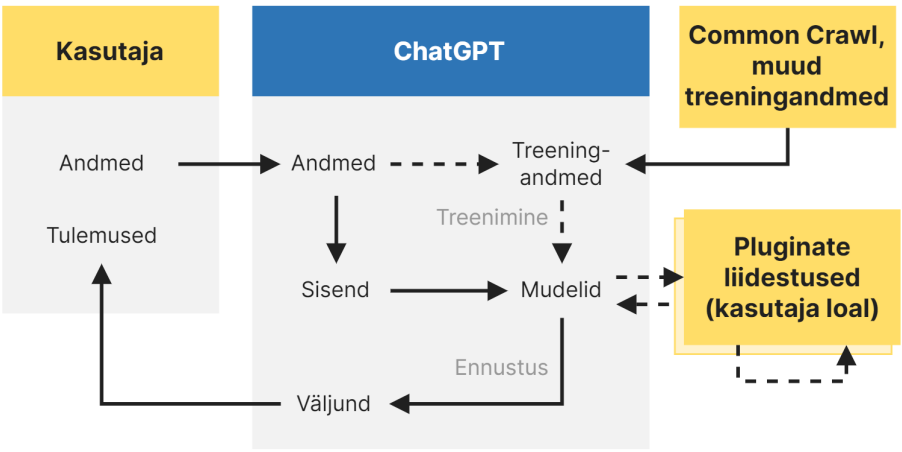


Ohud ja kaalutlused:

1. Mudeli treenijalt oodatakse teavet mudeli seletatavuse ja kvaliteedi kohta.
2. Mudeli treenijal peab olema õiguslik alus treeningandmete töötlemiseks.
3. Suurte mudelite ja treeningandmete mahtude korral on sellise lahenduse ehitamine kõige kulukam.

Selle levitusmudeli on kõik organisatsioonid, mis kasutavad tehisintellekti sisemiste teenuste loomiseks, aga näiteks ka OpenAI. Kasutaja teeb päringuid teenusele, teenus vastab päringule valitud mudeli väljundiga. OpenAI levitab treenitud mudeleid sõltuvalt sihtrühmast, üle rakendusliidese, aga ka kaaludena (nt Microsoftile). OpenAI kogub ja ostab treenimisandmed ise. Samas ei tea me kõiki detaile nende andmete päritolust. OpenAI rakendusliidese mudeleid ei treenita (2023. novembri seisuga) rakendusliidese kaudu tulnud päringute põhjal, küll aga ChatGPT kaudu tehtud päringute põhjal, kui just ei kasutata ChatGPT Enterprise versiooni[3].

ChatGPT sobib samuti käesoleva levitusmudeli alla, sest see kasutab sisemisi (OpenAI väljaarendatud) mudeleid. ChatGPT puhul on tähelepanuväärne, et kui kasutaja kasutab pistikprogramme, siis need võivad teha päringuid kolmandatele osapooltele andmete täiendavaks töötlemiseks või hankimiseks. On oluline mõista, et mudel ei suhtle pistikprogrammide ega nendega liidestatud teenustega otse, vaid see andmevahetus toimub teenuse äriloogika ühe osana. Reeglina tähendab see seda, et kui mudel teeb otsuse pistikprogrammi kasutamiseks, siis koostab ta eelviibas oleva info ja kasutaja soovi põhjal pistikprogrammi abil liidestatud teenuse suunas päringu. Mudeli koostatud päringu põhjal saadakse vastus, mis saadetakse tagasi mudelisse, kus see vormistatakse kasutajale sobilikuks vastuseks. ChatGPT levitusmudel on kujutatud joonisel 15.



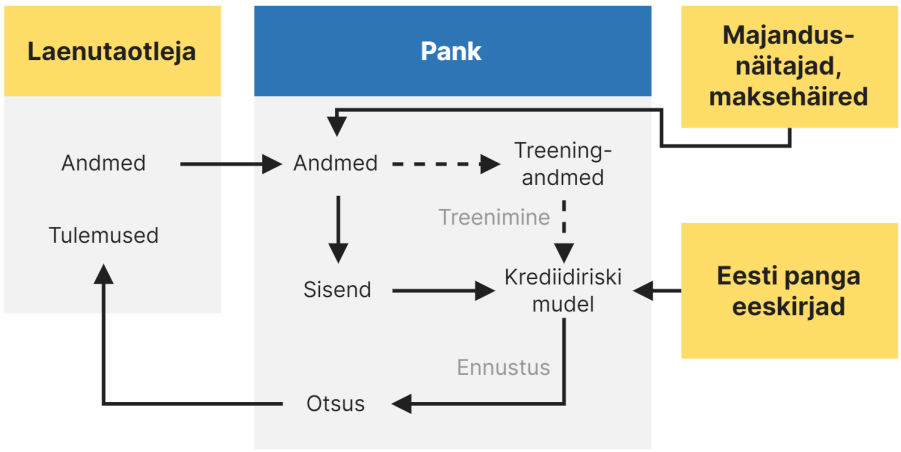
Joonis 15. ChatGPT levitusmudel

Teine variant rakendusest, mis sobib käesoleva levitusmudeliga, on arvutuslikult mittenõudlik ja oma andmete peal kergesti treenitav reeglipõhine või muu lihtne masinõppealgoritm (näiteks lineaarregressioon, otsustuspuu või naiiv-Bayesi klassifikaator). Panga krediidiriskimudel sobib käesoleva levitusmudeliga: pank treenib mudeli oma (klientide) andmete põhjal ning rakendab seda ise. Seejuures kasutab pank täiendavaid andmeid: maksehäirete andmeid, majandusnäitajaid ning panga sisemisi andmeid. Sellise teenuse levitusmudelit kujutab joonis 16.

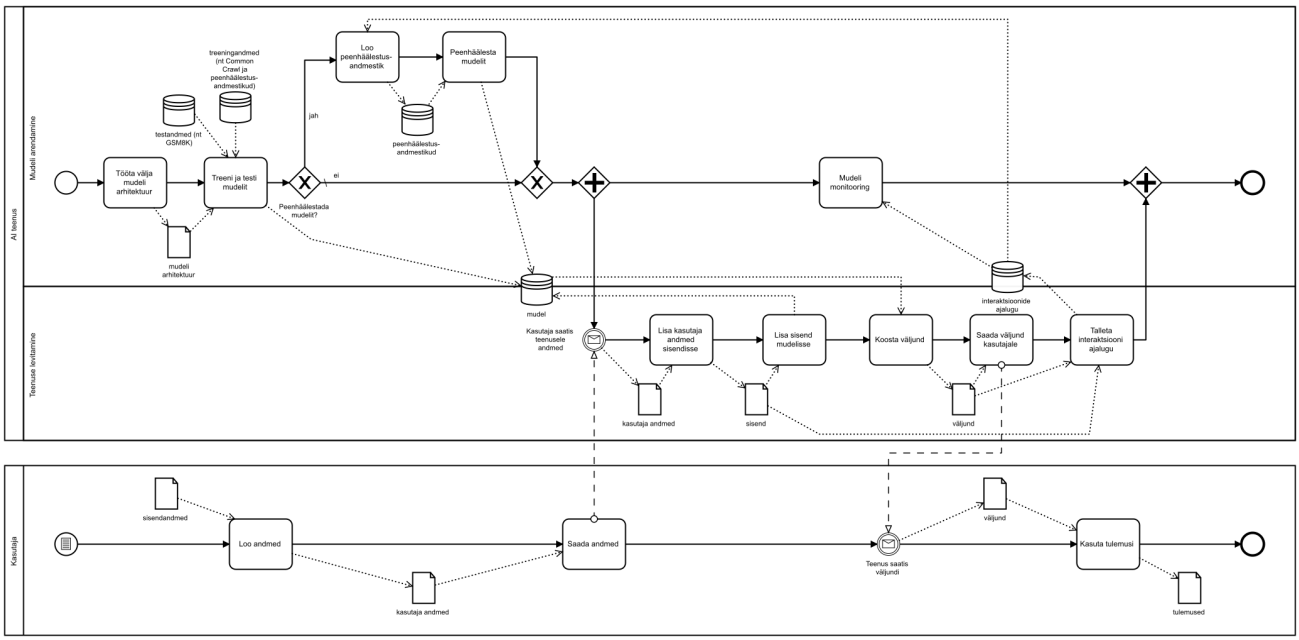
Joonis 17 kirjeldab kolmandat tüüpi levitusmudelit. Selle mudeli protsessis on kaks osapoolt - AI teenus ja kasutaja. Selles levitusmudelis on AI teenus, mudeli levitaja ja mudelipakkuja kõik üks ja sama osapool.

AI teenus jaguneb kaheks - mudeli arendamine ja teenuse levitamine. Mudeli arendamise käigus leiavad aset samad tegevused, mis eelnevates levitusmudelites - mudeli arhitektuuri väljatöötamine, mudeli treenimine ja testimine ning vajadusel selle peenhäälestamine ja mudeli monitooring. Kui kasutaja saadab AI teenusele andmed, siis lisab AI teenuse levitamise osapool need andmed sisendisse ja mudelisse, koostab väljundi ning saadab selle kasutajale. Interaktsiooni ajalugu talletatakse ning seda saab kasutada mudeli monitooringul ja peenhäälestusandmestike loomisel.

[3]Enterprise privacy at OpenAI. https://openai.com/enterprise-privacy Külastatud: 01.12.2023



Joonis 16. Krediidiasutuse eraisiku krediidiriski hindamise mudeli levitusmudel



Joonis 17. Levitusmudel LM3 andmevood

5 Tehisintellekti rakenduste riskid

5.1 Riskihalduse metoodika

Põhilised standardid, mis riskikontrolli kirjeldavad, on ISO 31000 riskihalduse standard [142] ja NIST SP 800-37 riskihalduse raamistik (RMF) [143]. Infoturvariskide eripärasid kirjeldab ISO/IEC 27005 [144] ja küberturbe eripärasid NIST küberturbe raamistik (CSF) [145]. Tehisintellektispetsiifilisi riskihalduse juhiseid kirjeldab ISO/IEC 23984 [146], mis kirjeldab, kuidas laiendada standardiga ISO 31000 vastavsuses olevat riskihalduse protsessi organisatsioonile, mis kasutab, arendab või juurutab tehisintellektisüsteeme. Kui organisatsioonil on olemas ISO/IEC 27001 sertifikaat ja töötav infoturbe halduse süsteem (ISMS), soovitame AI-süsteemid lisada olemasolevasse riskihalduse protsessi.

Selles aruandes kirjeldatud lihtsustatud metoodika sobib kokku standardite ISO 31000 ja ISO/IEC 27005 töövoogudega, aga on vajadusel ka sobiv NIST RMF ja CSF raamistikega. Seega, kui organisatsioonil soovib tulevikus kasutada keerukamat riskihalduse strateegiat, on lihtne olemasolevat AI süsteemide riskihaldust üldisemasse raamistikku kaasata. Eesti infoturbestandard (E-ITS) [147] joondub ISO/IEC 27001 sarjaga ning E-ITS-i rakendajatel on samuti võimalik kasutada riskipõhist lähenemist. Seega sobib kirjeldatud metoodika ka E-ITSi rakendavatele organisatsioonidele.

Riskihaldusprotsess koosneb kolmest sammust: konteksti loomine, riskikontroll ja riskikäsitlus. Selle aruande riskihaldusmetoodika käsitlusalaks on IT-süsteemid, mis sisaldavad tehisintellekti komponenti.

5.1.1 Tehisintellekti kaalutlused konteksti loomisel

Konteksti loomise käigus selgitatakse välja ja dokumenteeritakse huvipooled ja protsessiga seotud varad. Organisatsioon määrab oma riskivalmiduse, riskinormi ja riskiomanikud, selgitab välja huvipooltele kehtivad sisemised, riiklikud ja seadustest tulenevad nõuded. Organisatsioon määrab kindlaks riskide aktsepteerimise tingimused ning valib asjakohase riskihaldusmetoodika.

Tehisintellektisüsteemide jaoks konteksti loomise käigus tuleb tähelepanu pöörata kõigi huvipoolte tuvastamisele ja dokumenteerimisele. Tähelepanu tuleb pöörata ka sellistele osapooltele, kes ei tundu otseselt teenuseandmisega seotud olevat (näiteks treeningandmetes esinenud isikud, teoste omanikud või ka kolmanda osapoole taristu- ja teenuseandjad). Nii juhul, kui organisatsioon loob ise AI-süsteemi ning kasutab seda organisatsioonisiseselt, kui juhul, kui AI-süsteemi kasutatakse teenusena, peab arvesse võtma:

- andmesubjekte või andmeomanikke, kelle andmeid on kasutatud masinõppemudeli treenimisel,
- mudeli treenijat,
- teenuseandjat,
- teenuse kasutajat.

Organisatsioon peab huvipooled kindlaks tegema ning arvestama nende õiguste ja huvidega riskikontrolli ja -käsitluse käigus. Nende uute huvipooltega võib kaasneda vajadus arvestada uute alusdokumentide või määrustega. Olluline on kindlaks teha, kas uued huvipooled kuuluvad organisatsiooni sisse või jäävad sellest välja. Organisatsioon peab kaardistama, kellel on millised seaduslikud õigused ja kohustused ning kes millist süsteemi osa käitab.

Organisatsioon peab kindlaks tegema, kust pärinevad eri liiki andmed (mudelid, treening-, sisend- ja väljundandmed) ja tarkvarakomponendid ning milline on andmevoog eri komponentide vahel. Huvipoolte ja komponentide kaardistamine on vajalik AI-süsteemi konteksti mõistmiseks. Mõned riskid võivad kaasneda teatud tüüpi andmete või süsteemide kasutamisega. Kaardistuse visualiseerimiseks võib kasutada tööriistu, mida tavaliselt kasutatakse süsteemide modelleerimisel (UML, BPMN). On olemas erilised modelleerimistööriistad (nt PE-BPMN [148]), mille eesmärk on kirjeldada andmeobjektide liikumist ja nähtavust erinevate huvipoolte vaatepunktist.

Huvipoolte juurdepääsu andmetele on võimalik dokumenteerida nähtavustabelite abil. Tabelis 1 on näide

nähtavustabelist, mis kirjeldab, millistel huvipooltel on AI-süsteemis andmetele juurdepääs. Selles näites on kolm huvipoolt: lõppkasutaja, teenuseandja (tehisintellekti klientrakendus) ja AI rakendusliidese (API) andja (treenib ja jagab mudelit). Kõik huvipooled näevad lõppkasutajate sisendandmeid ja mudeli väljundit. Teenuse- ja AI rakendusliidese andjal on juurdepääs teenuseandja äriandmetele. Mudelit näeb ainult AI rakendusliidese andja.

Tabel 1. Nähtavustabeli lihtsustatud näide

	Kasutaja sisend	Teenuseandja äriandmed	Mudel	Väljund
Lõppkasutaja	X			X
AId kasutava teenuse andja	X	X		X
AI API andja	X	X	X	X

5.1.2 Tehisintellektisüsteemide riskikontroll

Tihti väljendatakse riski ohusündmuse realiseerumise võimalikkuse ja potentsiaalse kahju kaudu. Riskikontroll koosneb riskide identfitseerimisest, analüüsist ja hindamisest. Riskide identifitseerimise käigus otsitakse riske, selgitatakse välja, millised riskid on relevantsed ning kirjeldatakse need. Igale väljaselgitatud riskile määratakse riskiomanik. Riskianalüüsi käigus tehakse kindlaks riskide põhjused, allikad ning hinnatakse potentsiaalset kahju ja riski realiseerumise võimalikkust. Riskide hindamisel võrreldakse analüüsi tulemusena saadud riskitaset konteksti loomisel määratud vastuvõetavate riski kriteeriumitega, et teha kindlaks, kas riski tase on talutav ja aktsepteeritav.

Tehisintellekti riskikontroll põhineb loodud kontekstil. Iga AI-süsteemi komponendi puhul hindame riski huvipoolte kontekstis. Selle seose leidmine on lihtne konteksti loomise käigus koostatud nähtavustabeli järgi. Iga tuvastatud huvipool-komponent paari puhul analüüsime ja hindame kolme tüüpi riske – küberturvalisuse, regulatsioonide ja tehisintellektiga seotud riske. Küberturvalisuse riskid käsitlevad tavaliselt AI-süsteemi protsesside piisavust või AI-süsteemi komponentide (tarkvara, andmed ja teenused) konfidentsiaalsust, terviklikkust ja kättesaadavust. Määrustest tulenevad riskid käsitlevad õiguslikke kohustusi, mida kohaldatakse tehisintellekti süsteeme käitavate huvipoolte (AI-spetsiifilised määrused) või nende komponentide (nt isikuandmete, autoriõigusega kaitstud andmete, kriitilise infrastruktuuri) suhtes. Tehisintellekti riskid on seotud AI algoritmide eripäradega, AI-süsteemide mõjuga ühiskonnale ja eetiliste aspektidega. AI-süsteemide riskikontrollist on üksikasjalikumalt räägitud alamjaotises 5.2.

Tabelis 2 on näide, kuidas riski määratleda turvanõrkuste ja ohtude kaudu. Organisatsioon peab iga ohu puhul hindama ohu realiseerumise tõenäosust ja potentsiaalset kahju. Sama sündmuse tõenäosus ja kahju võivad olla erinevate organisatsioonide puhul erinevad. Vahel on kasulik võrrelda erinevate lahenduste riske, et valida organisatsiooni sobivaim lahendus. Näiteks, kuigi pilveteenuse andja pakub paremaid turvameetmeid kui väike organisatsioon ise rakendada saaks, võib pilvteenusest sõltumine kaasa tuua käideldavusriski, kui peaks kaduma side pilvteenustajaga.

5.1.3 Tehisintellektisüsteemi riskikäsitlus

Riskikäsitluseks on erinevad võimalused: riskide vältimine, riskide leevendamine, riskide üleandmine või riskide säilitamine. Sobiv viis valitakse riskianalüüsi tulemustele tuginedes.

Tõenäoliselt ei õnnestu organisatsioonil kõiki riske leevendada. Riskid järjestatakse olulisuse järgi ning selle järjekorra alusel valitakse sobivad infoturvameetmed, AI-spetsiifilised meetmed või õiguslikud meetmed, mille abil on võimalik riskid viia vastavusse organisatsiooni riskinormiga.

Riski saab vältida riskiallika kõrvaldamisega, funktsionaalsusest loobumisega, või äriprotsessi ümberkorraldamisega. Riske leevendatakse turvameetmete kasutuselevõtu abil. Alati ei ole lisaturvameetmete ka-

Tabel 2. Turvanõrkuste ja ohude näited

Andmed	Riski tüüp	Turvanõrkus	Oht
Väljund	AI risk	Kallutatud või vigane mudel	Lõppkasutaja saab väljundi, mis suunab neid ennast või teisi kahjustama
Treeningandmed	Regulatsiooniga seotud risk	puudub õiguslik alus isikuandmete töötlemiseks	Trahv andmekaitsemääruse rikkumise eest
Mudel	Infoturvarisk	Vigane identiteedihaldus	AI API andja kaotab juurdepääsu oma taristule ega saa inferentsiteenust anda

sutamine riski puhul võimalik või otstarbekas. Riskide leevendamise meetmeid on kirjeldatud peatükis 6. Riski üleandmine tähendab riski jagamist mõne teise organisatsiooniga või riskist põhjustatud kahjude kompenseerimist, näiteks kindlustust kasutades.

Kui riskikäsitluse lõpuks on riskid vastavuses organisatsiooni riskinormiga, siis ülejäänud riskid aktsepteeritakse. See tähendab, et selle riskiga rohkem ei tegeleta ning risk säilitatakse. Selleks, et riskihaldus oleks ajakohane, on vajalik riskide perioodiline seire ja läbivaatus. Protsessi oluliseks osaks on ka riskidest teavitamine, mille eesmärgiks on töötajate kursishoidmine infoturvariski halduse protsessi ja tulemustega.

5.2 Riskikontroll

5.2.1 Infoturvariskid

Digitaalsed riskid on kõige tõenäolisemad ja kõige suurema mõjuga [149]. Ohtuks on eeskätt küberkuritegevus [149, 150]. Samas võimaldavad generatiivse tehisintellekti tehnoloogiad digitaalsete riskidega ka edukamalt toime tulla [149], kui neid selleks tarbeks arendada ja rakendada. Küberturvalisuse tagamiseks kasutatavate automatiseeritud või poolautomaatsete vahendite loomisega seotud teadus- ja arendustegevust on soovitatud ka küberturvalisuse teises direktiivis (NIS2) [21].

Infoturvariske tuvastatakse ja analüüsitakse ohtude, ohusündmuse tõenäosuse ning potentsiaalse kahju põhjal. Eesti infoturbestandard E-ITS kirjeldab etalonturbe protsessi, mille osaks on etalonturbe kataloog. See kataloog sisaldab protsessimooduleid ja süsteemimooduleid. Need omakorda sisaldavad ohtude nimekirja ja meetmete kirjeldust. Etalonturbe protsessi kasutamine lihtsustab riskituvastust ning on (kõrgtasemel kasutades) kooskõlas ka ISO/IEC 27000 sarja standarditega.

E-ITS etalonturbe moodulitest [151] on tehisintellektisüsteemide juurutamise ja kasutamise puhul asjakohased järgmised protsessimoodulid: ORP (organisatsioon ja personal), CON (kontseptsioonid ja metoodikad), OPS (käidutööd) ja DER (avastamine ja reageerimine) ning järgmised süsteemimoodulid: SYS (IT-süsteemid) ja APP (rakendused). See nimekiri sisaldab ainult mooduleid, mille puhul on vaja eraldi samme tehisintellekti süsteemide juurutamisel või kasutamisel. Nimekiri ei sisalda ettevõtte ülejäänud taristu või turbehalduse ülesseadmiseks vajalikke mooduleid. Kui ettevõte ümbritsevates süsteemides riske ei kontrolli ega käsitle, siis ei ole kasu ka tehisintellektisüsteemi tugevast kaitsmisest.

Tehisintellektisüsteemide lisandumisel organisatsiooni töövoogu tekivad tõenäoliselt järgmised protsessidega seotud ohud. Samas ei ole ohtude loetelu kuidagi piiratud standardis loetletutega.

- ORP 1. AI-süsteemide kasutamiseks puuduvad selged eeskirjad; AI-süsteem ei ühildu teiste töövahenditega;
- ORP 2. töötajad ei tunne AI-süsteeme piisavalt hästi; nad on hooletud andmete kasutamisel AI-süsteemides; nende kvalifikatsioon ei ole piisav;
- ORP 3. töötajaid ei ole piisavalt AI-süsteemide ohtude ja rünnete teemal koolitatud;

- ORP 5. AI-süsteemi kasutamisega rikutakse seadust või lepingusätteid; teavet avaldatakse AI-süsteemis lubamatult; välisele AI-süsteemile avaldatakse kogemata siseteavet;
- CON 2. AI-süsteemidele sisendi andmisel eiratakse andmekaitsealaseid nõudeid; andmetöötlusprotseduurid on puudulikud ega arvesta AI-süsteemide tööpõhimõtetega; puuduvad ressursid, et tegeleda isikuandmete kaitsega AI-süsteemides; andmesubjektide privaatsus ei ole tagatud AI-süsteemides andmete töötlemisel; andmete konfidentsiaalsus ei ole AI-süsteemis tagatud seetõttu, et andmed satuvad volitamata isikute kätte või on treenitud mudelist kättesaadavad; andmetöötleja mainekahju;
- CON 3. probleemid AI-süsteemi andmete (nii sisendite, mudeli kui mõnel juhul ka väljundite) varundamisega;
- CON 6. AI-süsteemi andmete puudulik kustutus ja hävitamine;
- CON 8. ebasobiv tarkvaraarendusmetoodika AI-süsteemi arendamisel; ebapiisav kvaliteedihaldus; puudulik dokumentatsioon; ebapiisav arenduskeskkonna turve; AI-süsteemi kavandamise vead; puudulikud AI-süsteemi testimis- ja vastuvõtuprotseduurid; tootmiskeskkonna andmete kasutamine AI-süsteemi testimisel;
- CON 10. AI-süsteemi kasutamisel veebirakendusena: AI-süsteemis sisalduva tundliku taustainfo avaldamine veebirakenduses; automaatsete ründevahendite kasutamine AI-süsteemi veebirakenduse ründamiseks;
- OPS 2.2. Kehtivad kõik pilvteenuste kasutamisega seotud ohud: AI pilvteenuse kasutamise strateegia puudulikkus; sõltuvus AI pilvteenuseandjast; puudulik nõuete haldus AI pilvteenuste kasutamisel; õigusaktide nõuete rikkkumine; AI pilvteenuse andjaga sõlmitud lepingu puudulikkus; AI pilvteenuste puudulik integreerimine omaenda IT-süsteemidega; AI pilvteenuse kasutamise lõpetamise puudulik reguleerimine; avariivalmenduse kontseptsiooni puudulikkus; AI pilvsüsteemi andja süsteemi tõrge;
- OPS 2.3. Kehtivad kõik väljasttellimisega seotud ohud: AI-süsteemi väljasttellimise strateegia puudulikkus; ärikriitiliste protsesside kontrolli puudulikkus; sõltuvus AI teenuseandjast; AI teenuseandja ebapiisav infoturbe tase; ebapiisav kontroll hangitava AI teenuse üle; AI teenust reguleerivate lepete puudulikkus; pääsuõiguste halduse puudulikkus; AI teenuse andja allhangete kontrollimatus; sooritusindikaatorite (KPI) puudumine; puudulikud sätted AI-süsteemi väljasttellimise lõpetamiseks; väljasttellitava AI teenuse puudulik avariihaldus;
- OPS 3.2. Kehtivad kõik teenuseandja infoturbega seotud ohud: AI teenuse andja puudulik infoturbe haldus; AI teenuse andja puudulik avariihaldus; puudulikud teenuselepingud AI teenuse saajatega; AI teenuse andja IT-süsteemidega liidestuse nõrkused; AI teenuse saaja sõltuvus teenuseandjast; pääsuõiguste puudulik haldamine; simultaanteeninduse võime puudumine AI teenuseandjal; AI teenuse andja sõltuvus allhankijatest; AI teenuse lepingu lõpetamise puudulik kord; AI teenuse andja IT-süsteemi tõrge; suhtlusrünne;
- DER 2.1. AI-süsteemidega seotud turvaintsidentide puudulik käsitlus; tõendusjälgede hävitamine turvaintsidentide käsitlusel;
- DER 3.1. AI-süsteemides turvameetmete puudulik või plaanimata rakendamine; kontrollija puudulik kvalifikatsioon; puudulik auditi plaanimine ja kooskõlastamine; isikuandmete kasutamise kooskõlastamatus; sihilik turvaprobleemide varjamine.

Süsteemimoodul SYS kirjeldab ohud IT-süsteemidele, sealhulgas serveritele (SYS 1.1, 1.2, 1.3, 1.9), virtualiseerimissüsteemidele (SYS 1.5), konteineritele (SYS 1.6), salvestislahendustele (SYS 1.8), klientarvutitele (SYS 2.1, 2.2, 2.3, 2.4), sülearvutitele (SYS 3.1), nutitelefonidele ja tahvelarvutitele (SYS 3.2), printeritele (SYS 4.1), sardsüsteemidele (SYS 4.3), esemevõrgu seadmetele (SYS 4.4) ja irdandmekandjatele (SYS 4.5). Lisaks on SYS moodulis kirjeldatud ka ohud, mis on seotud Eesti X-tee turvaserveri (SYS.EE 1) ja eID komponendide kasutamisega (SYS.EE 2). Olenevalt loodava või kasutatava AI-süsteemi või teenuse iseloomust saab vastavad ohud leida vastavatest moodulitest.

Süsteemimoodul APP kirjeldab ohud rakendustele: mobiilirakendustele (äppidele) (APP 1.4), veebirakendustele (APP 3.1), andmebaasisüsteemidele (APP 4.3), Kubernetesele (APP 4.4), tarkvarale üldiselt (APP 6) ja tellimustarkvara arendusele (APP 7). Lisaks kirjeldab APP EE 1 ohud Eesti X-tee andmeteenusele.

Tehisintellektisüsteemi arendaja või juurutaja saab konteksti tuvastamisel teha kindlaks, millised neist ohtudest neid puudutavad. Väljaselgitatud ohtude põhjal on võimalik riskid kirjeldada, analüüsida ning hinnata.

5.2.2 Õiguslikud riskid

Tehisintellektisüsteemidega seotud õiguslike riskidena võib välja tuua õigusaktide nõuetele mittevastavuse, mis võib kaasa tuua:

1. kahjunõuete esitamise;
2. kohtuvaidlused;
3. pädevate järelevalveasutuste sanktsioonid, sh ettekirjutus vastavuse tagamiseks, sunniraha määramine, tegevuse peatamine või lõpetamine.

Eelnevad riskid võivad endaga kaasa tuua näiteks töötajate täiendava ajakulu kahjunõudega või kohtuvaidlusega tegelemisel, õigusteenuse sisseostmisega seotud kulud, rahalise kahju kahjunõude või kohtuotsuse täitmisel, kohtukulude tasumisel, saamata jäänud tulu tegevuse peatamisel või mainekahju. Viimane võib realiseeruda klientide arvu languses ja vähenenud tulus, halvimal juhul usalduse kaotuses ja tegevuse lõpetamises.

Isikuandmete töötlemisega seotud trahvid võivad ulatuda kuni 20 000 000 euroni või ettevõtja puhul kuni 4%-ni tema eelneva majandusaasta ülemaailmsest aastasest kogukäibest, olenevalt sellest, kumb summa on suurem. Tehisintellekti määruse ettepaneku kohaselt võib teatud rikkumiste puhul määrata trahvi kuni 35 000 000 eurot või kui tegemist on ettevõtjaga, kuni 7%-i tema eelneva majandusaasta ülemaailmsest aastasest kogukäibest, olenevalt sellest, kumb summa on suurem. Ebatäpse, mittetäieliku või eksitava teabe esitamise eest võib oodata rahatrahv vastavalt kuni 7 500 000 eurot või ettevõtja puhul kuni 1% tema eelneva majandusaasta ülemaailmsest aastasest kogukäibest, olenevalt sellest, kumb summa on suurem.

Euroopa Komisjon võib generatiivse tehisintellektisüsteemi teenustajale tehisintellekti määruse ettepaneku kohaselt nõuete rikkumise eest määrata trahvi kuni 3% tema eelneva majandusaasta ülemaailmsest aastasest kogukäibest või 15 miljonit eurot olenevalt sellest, kumb summa on suurem. Tehisintellekti määruse ettepanekuga on pädevale asutusele ette nähtud ka õigus tehisintellektisüsteemi turult eemaldamiseks.

Samuti on oluline, et tehisintellektisüsteemi osapooltel oleksid paigas kirjalikud lepingud, kus oleksid välja toodud poolte õigused, kohustused ja vastutus. Isikuandmete töötlemisel on olulised ka pooltevahelised andmetöötluskokkulepped. Lepingu nõuete rikkumisel võivad samuti kaasneda nii leppetrahvide nõudmine, kahjunõuete esitamine, aga ka kohtuvaidlused.

Viimastel aastatel on olnud palju kohtukaasusi, mille keskmes on olnud vaidlus tehisintellektisüsteemi treenimiseks kasutatud sisendi (tekst, fotod jne) üle (vt nt [152, 153, 154]). Eeskätt on need puudutanud autoriõiguste rikkumisi. Samas on olnud vaidlusi ka tehisintellektisüsteemidega seotud vastutuse osas. Näiteks leidis kohus kaasuses Moffatt vs Air Canada [155], et ettevõte vastutab kogu oma veebisaidil oleva teabe eest, kusjuures ei ole vahet, kas teave pärineb staatiliselt lehelt või vestlusrobotist. Kohtukaasused testivad tehisintellekti õiguslikke piire ja loodetavasti loovad lähiaastatel selgust, aidates luua normide tõlgendamise ühtlasemaid praktikaid.

5.2.3 Tehisintellekti riskid

Arengud tehisintellektis, eriti suurtes keele- ja pildisünteesimudelites on tõstnud päevakorda arutelu nende tehnoloogiate riskidest. Seejuures võivad riskid olla seotud nii (üld)võimekate mudelite ohtlike või tahtmatute väljunditega kui selliste mudelite leviku ja laiema kasutuselevõtuga ning selle ühiskondlike tagajärgedega.

Kõige võimsamaid pilt- ja keelemudeleid on kulukas treenida, kuid vabavaraliselt ulatuslikult levivad väiksemad mudelid ei jää neist võimekustelt kaugele maha ning lähitulevikus on oodata, et need saavad veelgi võimsamaks. Tehisintellekti mudelite kasutuselevõtt automatiseeritud otsuste tegemiseks kriitilistes valdkondades, näiteks meditsiinis või sõjanduses on tekitanud täiendavaid riske ja arvukalt eetilisi dilemmasid.

Eraldi tuleb käsitleda riske, mis on seotud tehisliku superintellektiga, mis on võimeline iseseisvaks tegutsemiseks ning inimese võimega selle käitumist kontrollida ja suunata. Tehisintellekti edasine areng võib

tekitada uusi, senitundmatuid riske ja võimendada olemasolevaid, seetõttu peab nende leevendamine olema järjepidev, iteratiivne protsess.

5.2.3.1 Riskide klassifikatsioon tehisintellekti määruse ettepaneku järgi

Tehisintellekti määrus on üles ehitatud riskipõhisele lähenemisele, eristades nelja riskitaset: vastuvõetamatu, kõrge, piiratud ja minimaalne (vt tabel 3). Tasemele vastavalt tulenevad nõuded AI-süsteemile. Lisaks eristab tehisintellekti määrus üldotstarbelise AI-süsteemi puhul mittesüsteemset ja süsteemset riski.

Tabel 3. Tehisintellektisüsteemi määratletud riskitasemed

Nr	Risk	Kirjeldus	Näited tehisintellekti süsteemidest
1	vastuvõetamatu risk	keelatud tehisintellekti süsteemid	AI-süsteemid, mis põhjustavad märkimisväärseid riske inimeste tervisele ja ohutusele või põhiõigustele (manipuleerivad, ekspluateerivad AI-süsteemid), nt sotsiaalsed hindamissüsteemid
2	kõrge risk	reguleeritud kõrge riskiga tehisintellekti süsteemid	nt biomeetrilised kaugtuvastussüsteemid, emotsioonide tuvastamise süsteemid, kriitilise infrastruktuuri turvakomponendid, värbamissüsteemid, polügraafid, õiguse tõlgendamise süsteemid kohtus
3	piiratud risk	vastavusnõuded	AI-süsteem ei mõjuta oluliselt otsuste sisu ega tulemust. AI-süsteem on ette nähtud kitsa protseduurilise ülesande täitmiseks, nt struktureeritud andmete loomiseks, sissetulevate dokumentide grupeerimiseks teema järgi või duplikaatide tuvastamiseks suure hulga taotluste hulgast
4	minimaalne risk	kohustused puuduvad	AI-süsteemid, mida võib kasutada piiranguteta, nt rämpsposti filtrid, AI-põhised video- või heliparandussüsteemid

5.2.3.2 Algoritmilised riskid

Järgnevalt käsitleme riske, mis on seotud konkreetsete tehisintellekti süsteemide ja nende kasutamise vahetute tagajärgedega. Kui nende riskide realiseerumine tarvitseb ründeid tehisintellektisüsteemide vastu, siis käsitleme neid ründeid alapeatükis 5.3.

Piiratud üldistusvõime. Automatiseeritud tehisintellektisüsteemide rakendamisel väga kriitilistes valdkondades (nt meditsiinis, sõjanduses või isesõitvates autodes) esineb oht, et mudel ei anna mõistlikku väljundit sisendile, mis on liiga erinev treeningandmetes esinenust. Suurtes keelemudelites on näiteks täheldatud "hallutsinatsioonide" esinemist, kus mudel tagastab usutavalt kõlava, kuid faktiliselt põhjendamatu väljundi [156]. Ohtu võimendab tehisintellektisüsteemide läbipaistmatus ja nii võib tekkida kahjuliku või eksliku väljundi "pimesi usaldamise" risk.

Ülemäärane sõltumine AIst ja inimjärelvalve kadumine. Tehisintellekti üha ulatuslikum kasutuselevõtt, sealhulgas kriitilistes süsteemides ähvardab jätta inimese tagaistmele. Mida keerukamaks lähevad tehisintellekti mudelid ja süsteemid, seda raskem on neid inimesele hoomata, mis võib vähendada järelvalvet nende üle. Inimjärelvalve vähenemisega kaasneb kahanev võime sekkuda AI-süsteemide töösse ning vältida tahtmatuid tulemeid. Seejuures võib neist saadav kasu olla nii suur, et järelvalve ja kontrollitavuse kaotuse hinda hakatakse pidama vastuvõetavaks. Et keerulised süsteemid kipuvad olema hapramad, kui lihtsad süsteemid kujutab neist ülemäärane sõltumine kuid mittemõistmine endast suurt riski.

Kallutatud ja ohtlikud vastused. Ka stiimulõppe abil turvalisemaks häälestatud mudeleid on viibasüstimistehnikaid kasutades võimalik panna genereerima diskriminerivat, solvavat või muidu potentsiaalselt kahjulikku sisu [157]. Lisaks viibasüstimisele võib mudeli kaitsemehhanisme peenhäälestamisega üsna odavalt või isegi tahtmatult maha võtta [158], ning mõned (eeskätt vabavaralised) mudelid ei sisaldagi arvestataval määral selliseid mehhanisme. Kuna mudelid on treenitud suuresti inimtekkeliste andmestike peal, milles esineb inimesele iseloomulikku kallutatust, on ka nende andmestike peal treenitud mudelid olemuslikult kallutatud. Seejuures tuleb algoritmilise diskrimineerimise korrigeerimisel olla ettevaatlik sihtnäitajate valimisel, sest need võivad olla samuti kallutatud. Korrigeerimismeetmete liigagar rakendamine võib mudeli või rakenduse võimekusi liigselt halvendada, nagu juhtus Google'i Gemini AI pildisünteesitööriistaga[1].

5.2.3.3 Ühiskondlikud riskid.

Tehisintellekti areng on kiire. Selle laialdane kasutuselevõtt tõotab tuua palju majanduslikku ja ühiskondlikku kasu, kuid samas ähvardab muuta maailma vähemalt sama põhjalikult kui seda tegi interneti lai kasutuselevõtt. Tehisintellekti (inim)ühiskondlikud riskid on seotud nii tehisintellekti vahendatud inimagentsuse laienemise ning sellega kaasnevate ühiskondlike muutuste ettearvamatusega kui ka tehisliku superintellekti (ASI) tekke võimalusega.

Autonoomne tehislik superintellekt. Teadusulmest tuttav tehisliku superintellekti teema on viimasel ajal kerkinud tehisintellekti eksistentsiaalse riski arutelu kontekstis. Trendid tehisintellekti mudelite suuruse ning arvutusvõimsuse kasvus ning emergentsete omaduste kerkimises tekitavad ootusi veelgi võimsamate, multimodaalsete ning parema üldistusvõimega mudelite või rakenduste suhtes. Kui sellisel mudelil on piisavalt autonoomiat, ligipääs kriitilistele (nt finantsilistele) süsteemidele ning võime jääda märkamatuks või vältida võimalikke vastumeetmeid, siis on risk sellevõrra suurem [5].

Piisavalt võimas ja autonoomne AI agent võib (inimese kaasabiga või ilma selleta) muutuda ohtlikuks kõigest omandades ligipääsu internetile ning võimet teha GET päringuid, kasutades ära turvaauke nagu Log4Shell [159]. Täiendavateks riskifaktoriteks on agendi võime enesetäiustamiseks ning situatsiooniline teadlikkus. Teadlaste seas puudub konsensus selliste võimekuste tekke ajalise perspektiivi üle, mis aga ei välista nendega seotud riskide võimalikkust ega võimalikku eksistentsiaalset riski inimkonnale.

AI mudelite kontrollimatu levik. AI alusmudelite vaba levik ja ulatuslik kasutuselevõtt, mis tundub nüüdseks paratamatu, mõjub AIga seotud riskide võimendajana. Mida rohkematele kasutajatele ja arendajatele on mudel kättesaadav, seda rohkem on võimalikke kuritarvitajaid ning seda suurem on vajaliku regulatsiooni ulatus [160]. Risk on veel suurem eeltreenitud mudelite leviku puhul, sest nad ei ole peenhäälestatud andma ohutuid vastuseid.

Bioloogilised ja keemilised relvad. Võimsate alusmudelite levikuga on osutatud riskile, et terrorirühmitused saavad endale vahendi, mis aitab neid senisest kergemini omandada keemilisi või bioloogilisi relvi [161]. Erilist tähelepanu tuleb pöörata autonoomsetele AI agentidele, kes võivad viia läbi vajalikku uurimistööd iseseisvalt [162]. Mõned teadlased on välja toonud, et AI-spetsiifilise bioloogiliste ja keemiliste relvade leviku riski hindamisel tuleb seda võrrelda ligipääsuga internetile, mis on kuritahtlikel osapooltel nagunii olemas ning küsida, millist[2] pudelikaela[3] nende tootmisprotsessis tehisintellekt lahendab [163].

Pudelikaelaks pole üldiselt info kättesaadavus, seejuures saab suurt keelemudelit nagunii pidada (kadudega) kokkupakitud versiooniks internetis juba niigi avalikust infost, vaid infohulgas orienteerumises ja tootmisprotsessis. Kaasaegse tehisintellekti võimed vastata oma laiadest üldteadmistest lähtuvalt küsimustele ning orienteeruda tekstilistes andmetes ja võtta neid kokku võivad protsessi kiirendada. Ehkki skeptikud on osutanud, et spetsiifilised võimekused keemiliste või bioloogiliste relvade tootmiseks on praegu pigem tagasihoidlikud, tõotab tehisintellekti mudelite ja lahenduste võimekuste kasvu oodatav

[1] Google's 'Woke' Image Generator Shows the Limitations of AI https://www.wired.com/story/google-gemini-woke-ai-image-generation/ Külastatud 23.02.2024

[2] Anthropic: Frontier Threats Red Teaming for AI Safety: https://www.anthropic.com/index/frontier-threats-red-teaming-for-ai-safety Külastatud: 09.11.2023

[3] Propaganda or Science: Open Source AI and Bioterrorism Risk: PropagandaorScience: OpenSourceAIandBioterrorismRisk Külastatud: 09.11.2023

jätkumine neid riske võimendada.

Tehisintellekt infosõjas. Kvaliteetsed teksti-, pildi-, kõne- ja videosünteesimudelid võimaldavad viia läbi ulatuslikke automatiseeritud desinformatsionikampaaniaid, mis omakorda ähvardab tekitada usaldamatust kogu veebisisu suhtes. See on mureks riikidele ja ametiasutustele [164], mis peavad nüüd otsima viise tõendamaks oma sõnumite autentsust. Tehisintellekt annab infosõjas kõigile osapooltele võimsad relvad, kuid kaitsemeetmed ei arene nendega samas tempos.

Tehisintellekt ja pettus. Generatiivse tehisintellekti levik on andnud uued vahendid ka petturitele [165]. Pildi- ja tekstisünteesimudelid võimaldavad luua usutavaid võltsidentiteete, sealhulgas passe ja muid isikut tõendavaid dokumente. Kõnesüntees võimaldab jäljendada teise inimese häält ning hõlbustab seeläbi identiteedivargust. Keelemudelid suudavad automatiseeritult luua üha usutavamaid isikupärastatud kalastuskirju. Süvavõltsingu abil loodud videod võivad tekitada neis kujutatud isikutele tõsist kahju.

5.2.3.4 Eetilised dilemmad.

Tehisintellekti kasutuselevõtuga kaasnevad paljud eetilised küsimused. Kas tehisintellekt saab olla lõplik otsustaja elu ja surma küsimustes? Kui tehisintellekti põhjustatud majanduslik üleminek on liiga järsk, kas seda peaks aeglustama? Kas tehisintellekti süsteemi või mudelit saab pidada mingi töö autoriks? Kes vastutab tehisintellektisüsteemi vigade või selle poolt tekitatud kahju eest?

Töökohtade kadumine. Suured keele- ja pildisünteesimudelid ähvardavad asendada mitmes valdkonnas inimest. Kaasaegsete keelemudelite üldvõimekus ei jää inimesele alla ülesannetes, mis nõuavad kliendiga loomulikus keeles suhtlemist etteantud protsessieeskirjade järgi, või turundus- ja muude erialaste tekstide koostamises ja kokkuvõtmises olemasoleva infovaramu põhjal. Kõnesüntees ähvardab kõnekeskuseid, pildisüntees idee- ja muid kunstnikke, tekstisüntees turundustekstide kirjutajaid ning kasutajatoe spetsialiste. Mida võimsamaks lähevad AI lahendused, seda suurem on mõju tööturule, ulatusliku töökaotusega kaasnevad majanduslikud ja ühiskondlikud riskid. Protsessi saab pidada osa laiemast automatiseerimise trendist, mis oli seni seotud pigem robootika arenguga, kus eetiline dilemma keskendub lõivsuhtele tootlikkuse ja töösuhte kindluse vahel.

Eetilised dilemmad autonoomsetes süsteemides. Tänapäeval kasutatakse tehisintellekti süsteemides, mis langetavad automatiseeritud otsuseid, mis võivad märkimisväärselt mõjutada inimeste autonoomiat. Selliste süsteemide kasutuselevõtmise korral tuleb käsitleda tehisintellekti poolt tehtavate otsuste eetilisi ja moraalseid aspekte. Kui kiiresti liikuvale isesõitvale autole satuvad ette imik ja vanaema, siis peab AI-süsteem langetama moraalse valiku, kas seada riski alla imikut, vanaema või hoopis autojuhti. Inimeluide üle otsustamise probleemaatika esineb kõigis süsteemides, kus inimesel puudub võimalus kontrollida ja õigeaegselt sekkuda otsustusprotsessi. Eriti suurt eetilist riski kujutavad täielikult autonoomsed relvasüsteemid, näiteks turellid ja droonparved, mis peavad langetama liitlane-või-vastane otsuse sekundi murdosa jooksul [166].

Sõltuvusttekitavad vestlusrobotid. Kaasaegne pildi- ja tekstisüntees võimaldab luua väga kaasahaaravaid vestlusroboteid -ja partnereid. Sõltuvalt ärimudelist võib selliste teenuste pakkujatel esineda äriline tagamõte tegemaks teenust sõltuvusttekitavaks, sobitades AI vestluspartnerit nii, et kasutaja veedaks temaga rohkem aega. Täiendavaks ohuks on keelemudelite kalduvus takkakiitmiseks, mis tekib RLHF (inimeste tagasisidele toetuva stiimulõppe) käigus [167]. Sõltuvusttekitavate vestlusrobotie poolt pidevalt antav positiivne tagasiside tekitab kajakambriefekti ning on eriti ohtlik vaimselt ja sotsiaalselt haavatavatele isikutele.

Tehisintellekt kohtusüsteemis. Tehisintellekti tehnoloogiad puutuvad otseselt või kaudselt üha enam kokku õigusemõistmisega. AI rakendused saavad lihtsustada kohtunike ja juristide tööd ja töödelda suuri andmemassiive. Nende tehnoloogiate kasutuselevõtul tuleb kaaluda tehtavate otsuste ja soovituste läbipaistvust ning erapooletust, samuti riske inimese privaatsusele (näiteks automatiseeritud jälgimisel või infokogumisel).

Tehisintellekt ja intellektuaalomand. Generatiivne tehisintellekt on tänaseks võimeline sünteesima teksti, muusikat, pilte, videosid ja muud sisu. Need võimekused esitavad väljakutse kunstnikele, ning seda mitte ainult ähvardades neid asendada, vaid ka intellektuaalomandi seisukohast. Kui pildisünteesimudel on võimeline sünteesima pilti mõne kindla kunstniku stiilis, kas paneb ta seda tehes toime autoriõigus-

te rikkumise? Kui ei, kui sarnane peab sünteesitud pilt olema kunstniku omale, et rikkumine leiaks aset? Ning viimaks, kas generatiivse tehisintellekti mudelit saab üldse pidada millegi autoriks? Need on lahtised küsimused kunstniku seisukohast, kuid täiendav probleem esineb nt. pildipankadel, kes puutuvad kokku veebisorimisega mudeli treeningandmete kogumise tarvis. Kuidas tõendada, et mudel oli treenitud autoriõiguste või muu litsentsiga kaitstud andmete peal?

Tehisintellekt ja privaatsus. Tehisintellekti areng võimendab privaatsusriske mitmetpidi. Selle võime luua seoseid internetis juba olemasoleva info põhjal anonüümseks jääda soovivaid kasutajaid. Nii paljastati Forbese ajakirjaniku poolt X (endise Twitteri) kasutaja Beff Jezose isik, võrreldes tehisintellekti abil Beff Jezose ja endise kvantarvutusinsener Guillaume Verdoni kõnesalvestusi [168], tuvastades seeläbi, et tegemist on väga tõenäoliselt ühe ja sama isikuga. Võimalusi on teisigi – sotsiaalmeediavõrgustiku kasutamisaegu, kokkupuutuvaid kontosid ning keelekasutust analüüsides on võimalik teha hinnanguid konto taga oleva isiku kohta.

Veel üks risk privaatsusele on seotud treeningandmete lekkimisega. Keelemudelitel on niigi kalduvus oma treeningandmestiku sõnalt-sõnale reprodutseerimiseks, kuid teatud viibatehnikate abil on võimalik seda kalduvust võimendada [169]. Treeningandmestikud võivad sisaldada tundlikku või autoriõigustega kaitstud infot.

Ülereguleerimisest tingitud AI hüvedest ilmajäämine. Diskussioon tehisintellekti ohtude üle ning sellega seonduvate reguleerimisettepanekute ulatus võib tähendada, et neist mõnede rakendamise puhul võivad tehisintellekti kasutuselevõtu pakutavad hüved jääda realiseerimata. Seetõttu peab nendes diskussioonides käsitlema mitte üksnes ohtude võimalikkust, vaid tuginema põhjalikule riskianalüüsile.

5.3 Tehisintellektisüsteemide vastased ründed

Tehisintellektisüsteemid teevad otsuseid andmete põhjal. Need otsused toimuvad enamasti inimese järelvalveta, võivad olla kriitilise tähtsusega (meditsiin või isesõitvad autod), samuti võivad kasutatavad andmed olla tundlikud. Ründajad võivad kasutada tehisintellektisüsteemide omadusi nende käitumise mõjutamiseks või tundliku info eraldamiseks. See tähendab, et lisaks tavalistele IT süsteemide turvameetmetele tuleb käsitleda ka tehisintellektisüsteemidele spetsiifilisi turvameetmeid. Selleks käsitleme järgnevalt tehisintellektisüsteemidele iseloomulikke ründeid. Rünnete käsitlemisel oleme lähtunud Saksamaa Föderaalse Infoturbeameti "AI security concerns in a nutshell". [4] ning OWASP-i sihtasutuse "OWASP Top 10 for LLM Applications"[5] raportidest. Seejuures ei käsitle me siin neid tehisintellektisüsteemide vastaseid ründeid, mis said käsitletud tehisintellekti algoritmiliste ning eetiliste riskide alapeatükkides.

5.3.1 Põikeründed

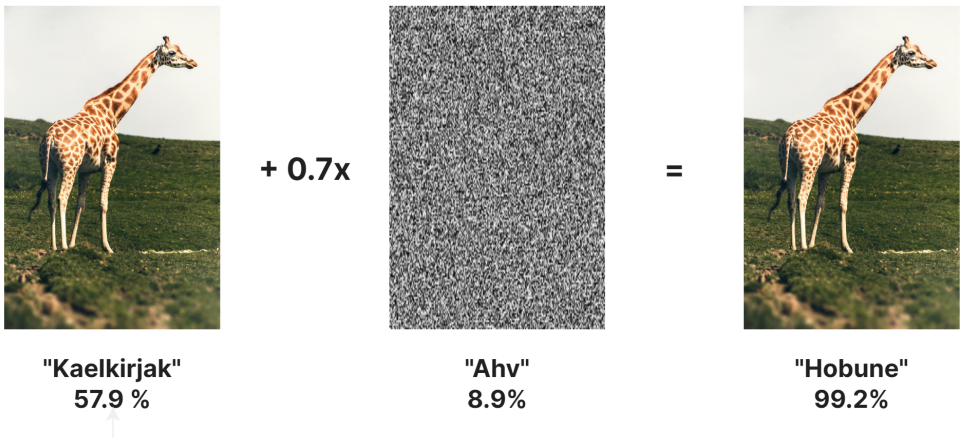
Põikeründed on ründed, kus ründaja üritab saada AI mudelilt süsteemi levitaja poolt mitte ettenähtud väljundit, seda sageli süütuna paistva ent rünnet peitva sisendiga. Seejuures võib ta võtta endale eesmärgiks mingi kindla väljundi saamise, või lihtsalt väljundi kvaliteedi kahandamise (enda valitud sisendi korral).

Pöörnäited (*adversarial examples*) on sisendid, mis peidavad põikerünnet. Näiteks juhul, kui ründajal on ligipääs kogu pildisünteesimudelile, võib ta võtta aluseks mõne hariliku sisendi, ning nihutada sisendeid mööda gradienti temale soovitava väljundklassi suunas, nagu on kujutatud joonisel 18. Selline minimaalne nihutus mõjutab mudeli väljundit, kuid võib seejuures jääda silmale märkamatuks [170, 171].

Viibasüst on suurte keelemudelite ja nende peale ehitatud AI rakenduste vastane rünne, mis kasutab ära viiba ja kontekstakna omadusi sellise väljundi saamiseks, mida mudeli levitaja ei ole ette näinud [172]. Kuna keelemudel ei suuda eristada kontekstiaknas paiknevat levitaja eelseatud eelviipa kasutaja viibast, võib kasutaja panna mudelit viibasüsti abil ignoreerima eelviibaga kaasa antud juhiseid, või neid juhiseid kasutajale väljastama. Viibasüstis sisalduvad juhised võivad käivitada koodi või pärida veebilehti ebaturvaliselt liidestatud pistikprogrammide kaudu [173].

[4] https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.html Külastatud: 08.12.2023

[5] OWASP Top 10 for Large Language Model Applications. https://owasp.org/www-project-top-10-for-large-language-model-applications/) Külastatud: 26.02.2024



Joonis 18. Pildi häirimine spetsiaalselt valitud müraga paneb mudeli ennustama valet väljundklassi [170].

Seejuures ei pruugi viibasüst pärineda pahatahtlikult kasutajalt, vaid kelleltki, kes sisestab selle veebiressursile, mida arvutivõrguga liidestatud LLM-rakendus võib pärida [174]. Sel juhul on tegemist kaudse viibasüstiga. Nii jõuavad uued juhised mudelini; ründe skeem on kujutatud joonisel 19. Käesolev rünne sarnaneb veebirakendustes levinud koodisüstile, kus ebaturvalise sisendi käitlemise korral võib rakendus käivitada sisendis olevat koodi.

Ebaturvalise väljundi käitlemise korral ei kontrollita mudeli enda koostatud päringuid ja käske. Selle tagajärjel võib ründaja saada viibasüsti läbi ligipääsu AI rakenduse backend-süsteemidele, kui mudel on nendega liidestatud. Näiteks võib kasutaja viip sisaldada juhiseid koodi täitmiseks kutsega `exec` või `eval` käsu poole. Alternatiivselt võib mudeliga liidestuv pistikprogramm või kolmanda osapoole teenus tagastada mudelile ebaturvalist väljundit, mis omakorda tagastab selle kasutajale. See võib näiteks sisaldada mõnes programmeerimiskeeles (nt JavaScript) kirjutatud koodi, mida hakatakse seejärel kasutaja veebilehitsejas täitma.

Alusmudeli nõrkuste ülekandumine on risk, mis esineb siirdeõppel [171], sest mudeli peenhäälestamisel kasutatakse tavaliselt treenimisest palju vähem andmeid. Nii võib ründaja kasutada vabavaralise mudeli teadaolevaid nõrkusi pahatahtlike sisendite suhtes selle peal peenhäälestatud mudeli vastu, omamata seejuures peenhäälestatud mudelile juurdepääsu.

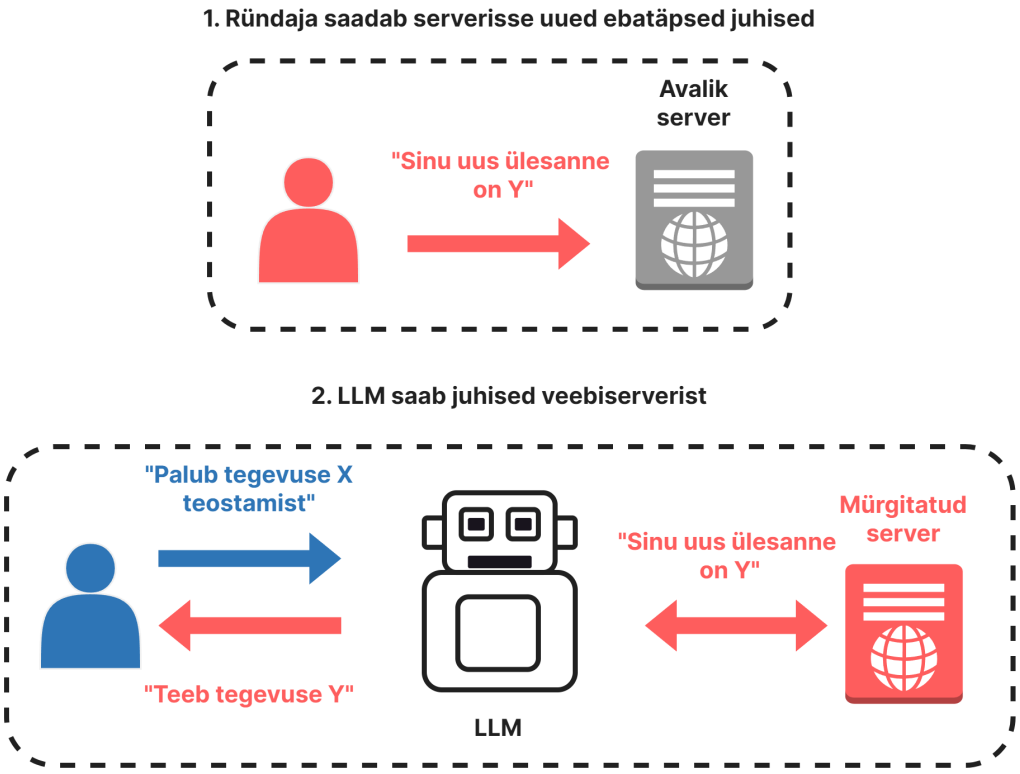
5.3.2 Infoeraldusründed

Infoeraldusründed hõlmavad ründeid, kus ründaja üritab eraldada mudelist ja selle väljunditest infot, millele tal ei tohiks olla ligipääsu. Nii võib ründaja teha järeldusi isiku kuulumise kohta treeningandmestikku, saada tema kohta teada tundlikke detaile, varastada mudelit või rekonstrueerida treeningandmestikku.

Mudeli vargus on rünne, kus üritatakse rekonstrueerida mudeli põhjal n-ö ”varimudel”, mida treenitakse ründaja sisendite ning rünnatava mudeli väljundite abil [175]. Võimas ja täpne varimudel võib avada uksed muudele rünnetele, näiteks põikerünnetele.

Kuuluvusluure (*membership inference attack*) on rünne, kus ründaja üritab tuvastada, kas mõni kirje esines treeningandmestikus või mitte [176, 177]. Et info ainuüksi kirje kuulumise kohta treeningandmestikku võib olla tundlik (näiteks meditsiiniandmetel treenitud mudeli korral), on sedasorti ründel suur risk privaatsusele. Selline rünne vajab ligipääsu mudeli väljundile ning võib lisaks kasutada teadmisi treeningandmestikus olevatest statistilistest seostest, et tuvastada, kui tõenäoline on mõni mudeli väljund kirje esinemisel treeningandmestikus või selle puudumisel sellest.

Atribuudiluure (*attribute inference attack*) on rünne, kus ründaja üritab teada saada mudeli treeningandmestikus sisalduvaid täiendavaid talle mitteteadaolevaid tundlikke atribuute kirje kohta, mille puhul ta teab, et see esineb treeningandmestikus. See töötab sarnaselt liikmelisuse järeldamisele - ründaja tu-



Joonis 19. Kaudne viibasüst (kohandataud algallikast [174]).

gineb teadmisele tundlike ja talle teadaolevate atribuutide statistilistest seostest treeningandmestikus, ning hindab nende koosesinevuse tõenäolisust kasutades mudeli väljundit.

Pöördmodelleerimine (*model inversion*) ehk treeningandmestiku rekonstrueerimine on rünne, kus ründaja eesmärk on teada saada mudeli väljundklasse iseloomustavaid omadusi (sisendeid või nende elemente) [178]. Ründajal on juurdepääs mudelile ning ta kasutab seda (näiteks treenides selle vastu generatiivse mudeli [179]), et rekonstrueerida sihtklassidele vastavaid treeningandmestiku kirjeid, mis võivad paljastada tundlikku infot.

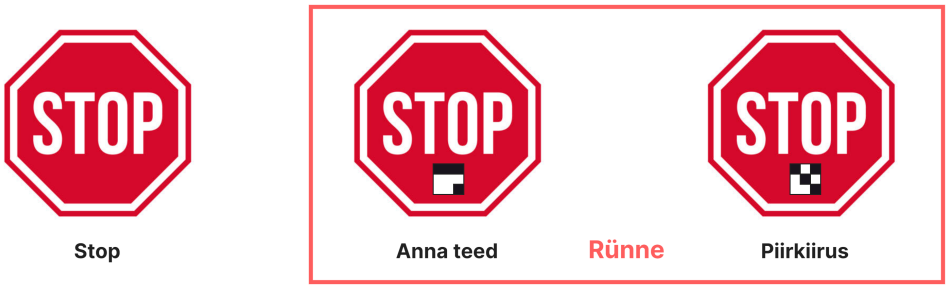
5.3.3 Mürgitus- ja tagaukseründed

Andmemürgitus (*data poisoning*) näeb ette treeningandmestiku mõjutamist eesmärgiga mõjutada mudeli jõudlust mõnes kindlas suunas või lihtsalt langetada seda. Andmemürgitus näeb ette väljundklassi muutmist treeningandmete kirjetes, püüdes tekitada maksimaalselt kahju [180, 181, 171, 182, 183]. Mürgitatud andmete peal treenitud mudel on kas üldiselt halvema jõudlusega või ei tule toime mõnede spetsiifiliste sisendikategooriatega.

Päästikrünne (*backdoor attack*) on andmemürgituse erijuht, kus treeningandmeid mürgitatakse näidete hulgaga, kus klassimärgend on vale ainult kindla päästiku (*trigger*) sisaldumisel näites [184, 185, 183]. Selle tulemusel kahaneb mudeli jõudlus või mudel ennustab valet klassi ainult siis, kui päästik sisaldub mudelile etteantavas näites. Sellisel moel mürgitatud mudel toimib muus olukorras õigesti ja seega on ründe tuvastamine harilikust andmemürgitusest raskem. Tagauksega mudel on haavatav põikerünnetele. Tagaukse rünne on kujutatud joonisel 20.

5.3.4 Teenustõkestus

Teenustõkestus ehk ummistusrünne on rünne, kus arvutisüsteemi töö halvatakse arvukate või arvutusnõudlikku protseduuri algatavate päringute abil. Suured keelemudelid on autoregressiivsed, see tähendab,



Joonis 20. Tagaukse rünne, kus mürgitatud andmete peal treenitud mudel klassifitseerib stopp-märki valesti kindla mustri esinemisel sisendis [185].

et igat järgmist väljundisõne konstrueerides võetakse arvesse kogu mudeli poolt seni koostatud väljundit. See tähendab, et vastuseks kulunud aeg (ja seeläbi arvutuslik ressurss) on seotud väljundi pikkusega [186]. Ründaja võib seda omadust ära kasutada, pärides mudelit sisenditega, mis panevad seda tagastama pikki väljundi jadu [187]. Samuti on võimalik halvata mudeli tööd saates sisendeid, mis mahuvad napilt kontekstaknasse, suurendades mudeli mälukasutust.

6 Leevendusmeetmed

6.1 Infoturvariskide leevendamise meetmed

Sarnaselt infoturvariskide aluseks olevate ohtude kirjeldusele, tugineme ka meetmete puhul E-ITS etalonturbe kataloogile [151]. Meetmed on kataloogis süstemaatiliselt kirjeldatud, lihtsasti juurdepääsetavad ning XLS- ja PDF-vormingus allalaaditavad. Selles aruandes toome välja vaid meetmete loetelu.

6.1.1 Meetmed protsessiriskide leevendamiseks

Infoturbe korralduse (ORP 1) meetmed on järgmised.

- Määratakse AI-süsteemidega seotud ülesanded ning kohustused, need tehakse teatavaks kõigile töötajatele ning neid vaadatakse regulaarselt üle (ORP.1.M1).
- Töövahendite ja seadmete nimistusse lisatakse AI-süsteem või selle olulised komponendid, nende hankimisel arvestatakse nende ühilduvust ja turvalisust (ORP.1.M8).
- AI-süsteemide kohta kehtestatakse turvalise kasutamise eeskiri, seda hoitakse ajakohasena ja tutvustatakse töötajatele (ORP.1.M16).

Personali (ORP 2) meetmed on järgmised.

- Töötajad saavad regulaarselt tegevusalale vastavat koolitust ja täiendõpet, töötajaid motiveeritakse ennast pidevalt arendama, uute töötajate otsimisel on nõutav haridus, kvalifikatsioon ja oskused selgelt kirjeldatud, ametikohale vajaliku kvalifikatsiooni kirjelduse õigsust kontrollitakse perioodiliselt (ORP.2.M15).
- Personalivalikul osalejad kontrollivad kandidaadi usaldusväärsust (ORP.2.M7).

Infoturbe teadlikkuse tõstmise ja koolituse (ORP 3) meetmed on järgmised.

- Juhtkonda teavitatakse regulaarselt AI-süsteemidega seotud riskidest, nendega seotud võimalikest kahjudest ja mõjust äriprotsessidele, juhtkond on teadlik õigusaktidega kehtestatud nõuetest AI-süsteemidele, juhtivtöötajad näitavad tehisintellektisüsteemide vastutustundlikult kasutamises eeskuju (ORP.3.M1).
- Töötajaid juhendatakse AI-süsteeme turvaliselt kasutama (ORP.3.M3).
- Luuakse AI-süsteemide riske ja õiguslikke aspekte käsitlev teadvustus- ja koolitusplaan (ORP.3.M4).
- Kavandatakse ja rakendatakse AI-süsteemide riske ja õiguslikke aspekte käsitlev teadvustus- ja koolitusplaan, kõik töötajad läbivad oma tööülesannetele ja vastutusalale vastava koolituse (ORP.3.M6).
- Õpitulemusi mõõdetakse ja hinnatakse (ORP.3.M8).
- Ohustatud isikutele ning organisatsioonidele viiakse läbi erikoolitus, mis käsitleb nii konfidentsiaalsust, terviklust kui käideldavust (ORP.3.M9).

Vastavusehalduse (ORP 5) meetmed on järgmised.

- Piiritletakse õiguslikud raamtingimused, töötatakse välja protsess kõigi turbehaldusele mõju avaldavate õigusaktide, lepingute ja muude nõuete väljaselgitamiseks, õiguslikke raamtingimusi võetakse arvesse AI-süsteemide äriprotsesside, rakenduste ja arhitektuuri kavandamisel ning AI-süsteemide või nende osade hankimisel. Õigusaktidest tulenevaid erinõudeid tehisintellektisüsteemidele arvestatakse eriti järgmistes valdkondades: isikuandmete kaitse, ärisaladuse kaitse, intellektuaalomandi kaitse (ORP.5.M1).
- Õiguslikke raamtingimusi järgitakse juba kavandamise ja disaini etappides (ORP.5.M2).
- Kavandatakse ja viiakse läbi vastavusehaldus (ORP.5.M4).
- Vastavusehaldusele tehakse regulaarselt läbivaatusi (ORP.5.M8).

Isikuandmete kaitse (CON 2) meetmed on järgmised.

- Organisatsioon on analüüsinud AI-süsteemis töödeldavate isikuandmete asukohti, liike ja kaitsetarvet (CON.2.M1).
- Isikuandmete töötlemine on AI-süsteemis kaardistatud kogu andmete elutsükli ulatuses (CON.2.M3)
- AI-süsteemid on kavandatud või protsessi lisatud nii, et isikuandmeid töödeldakse õigus- ja eesmärgipäraselt ning tagatud on andmete minimaalsuse printsiip (CON.2.M6).
- Tagatakse andmesubjekti õigused (CON.2.M8).
- Organisatsioon minimeerib AI-süsteemis isikuandmete töötlemisel isikuga otseselt või kaudselt seostatavate andmete kasutamist, võimalusel andmed pseudonüümitakse või anonüümitakse (CON.2.M9).
- Viiakse läbi AI-süsteemiga seotud andmekaitsealased mõjuhinnangud (CON.2.M13).
- AI-süsteemide loomisel ja protsessi lisamisel rakendatakse lõimitud andmekaitse ja vaikimisi andmekaitse põhimõtteid, kasutades näiteks privaatsuskaitse tehnoloogiaid (CON.2.M22).
- AI veebirakendustes kasutatavad küpsised ja jälgimisvahendid on kooskõlas isikuandmete kaitse üldmääruse ja teiste asjakohaste õigusaktidega (CON.2.M24).

Andmevarunduse kontseptsiooni (CON 3) meetmed on järgmised.

- Andmevarunduseeskiri sisaldab AI-süsteemi andmeid (CON.3.M2).
- Andmevarundusplaanid sisaldavad AI-süsteemide eripärasid (kas varundatakse treeningandmeid, mudelit, sisendeid, väljundeid) (CON.3.M4).
- Koostatakse andmevarunduse kontseptsioon AI-süsteemide jaoks (CON.3.M6)

Andmete kustutuse ja hävitamise (CON 6) meetmed on järgmised.

- Andmete kustutuse ja hävitamise kord sisaldab AI-süsteemi eripärasid (CON.6.M1).
- Protseduurid andmete turvaliseks kustutuseks sisaldavad AI-süsteemi eripärasid (CON.6.M12).

Tarkvaraarenduse (CON 8) meetmed on järgmised.

- AI-süsteemi arendamiseks on valitud sobiv tarkvaraarendusmetoodika ja metoodikale vastav protsessimudel ning seda järgitakse. Tarkvaraarenduse protsessimudel sisaldab infoturbe nõudeid. Arendusprotsessi käigus on infoturbe nõuetega arvestatud. (CON.8.M2).
- AI-süsteemi puhul on arvestatud turvalise süsteemikavandamise reegleid, need on dokumenteeritud ning nende järgimist kontrollitakse (CON.8.M5).
- AI-süsteemi arendamisel kasutatakse usaldusväärsetest allikatest saadud tarkvarateeke (CON.8.M6).
- AI-süsteeme testitakse tarkvaraarenduse käigus ning teostatakse koodi läbivaatus. Testimine viiakse läbi arendus- ja testkeskkondades, mis on käidukeskkonnast eraldatud. (CON.8.M7).
- Turvakriitilised paigad ja uuendid töötatakse välja ning paigaldatakse ilma viivituseta (CON.8.M8).
- AI-süsteemi lähtekoodi turvalisuse tagamiseks ja koodimuudatuste haldamiseks rakendatakse sobivaid versioonihalduse tööriistu (CON.8.M10).
- Välised tarkvarakomponendid ja teegid, mille turvalisuses ei saa olla täielikult kindel, läbivad enne kasutuselevõttu turvatestimise (CON.8.M20).
- AI-süsteemi kohta on olemas üksikasjalik ja põhjalik dokumentatsioon (CON.8.M12).
- AI-süsteemi arendamise esimeses etapis viiakse läbi riskikontroll (CON.8.M21).
- AI-süsteemi arhitektuuri valimisel arvestatakse nõuetega ja riskikontrolli tulemustega (CON.8.M22).

Veebirakenduste arenduse (CON 10) meetmed on järgmised.

- Tagatud on turvaline autentimine AI veebirakendues (CON.10.M1).
- Kasutajate pääsuõigused on vajadusekohaselt piiratud (CON.10.M2).
- AI veebirakendus väljastab kasutajatele vaid ettenähtud ja lubatavaid andmeid ja sisu (CON.10.M4).
- AI veebirakendus on kaitstud volitamata automatiseeritud juurdepääsu eest (CON.10.M6).
- Tagatud on konfidentsiaalsete andmete kaitse (CON.10.M7).
- AI veebirakendusse edastatud sisendandmeid käsitletakse potentsiaalselt ohtlike andmetena, neid filtreeritakse ja valideeritakse enne edasist töötlust (CON.10.M8).

- Piiratakse väljundites ja veateadetes tundliku taustainfo avaldamist (CON.10.M10).
- AI veebirakendus on arendatud turvalise tarkvaraarhitektuuri järgi, kõik komponendid ja sõltuvused on dokumenteeritud (CON.10.M11).
- AI veebirakenduse töö ajal tekkinud tõrgete lahendamisel säilitatakse veebirakenduse terviklus, kõik veateated logitakse (CON.10.M13).
- Käideldavuse tagamiseks takistatakse ressursside blokeerimist (CON.10.M17).
- Konfidentsiaalsuse ja tervikluse tagamiseks kaitstakse tundlikke andmeid krüptomehhanismidega (CON.10.M18).

Pilvteenuste kasutamise (OPS 2.2) meetmed on järgmised.

- Kehtestatakse pilvteenuste strateegia, mis sisaldab pilvteenuste eesmärke, eeliseid ja riske, nendega seotud õiguslikke, korralduslikke, majanduslikke ja tehnilisi raamtingimusi. Viiakse läbi teostatavuse, tasuvuse ja turvalisuse analüüs. Kasutuselevõtuks koostatakse etapiviisiline teenuse kasutuselevõtu plaan (OPS.2.2.M1).
- Strateegiale tuginedes koostatakse pilvteenuste turvapoliitika. Rahvusvaheliselt tegutsevate teenuseandjate puhul on arvestatud riigipõhist spetsiifikat ja õigusaktidest tulenevaid nõudeid (OPS.2.2.M2).
- Pilvteenust kasutav AI-süsteem lisatakse pilvteenuste loendisse (OPS.2.2.M3).
- Määratletakse ja dokumenteeritakse pilvteenuse kasutamisega seotud vastutusalad ja teenusepoolte tegevused (OPS.2.2.M4).
- Pilvteenuste turvapoliitika põhjal koostatakse pilvteenuste turbe programm, mis käsitleb pilvespetsiifilisi riske (nt sõltuvus pilvteenuse andjast, simultaanteenindus, fikseeritud andmevormingud, andmetele juurdepääs). Pilvteenuste turbe programm on kooskõlas pilvteenuse andja ja võrgutarnijaga sõlmitud lepingutega ning teenuse kasutustingimustega (OPS.2.2.M7).
- Pilvteenuse andja valitakse nõuete spetsifikatsiooni alusel (OPS.2.2.M8).
- Sõlmitakse kliendi vajadusele vastav pilvteenuse leping (OPS.2.2.M9).
- Pilvteenusele migreeritakse turvaliselt (OPS.2.2.M10).
- Pilvteenuste jaoks on välja töötatud avariivalmenduse programm (OPS.2.2.M11).
- Pilvteenuse vastavust teenuslepingus kokku lepitud tingimustele ja turvanõuetele ning pilvteenuse turbe programmi järgimist kontrollitakse regulaarselt (OPS.2.2.M12).
- Pilvteenuse andja tõendab infoturbe vastavust õigusaktidest tulenevatele nõuetele ja/või rahvusvaheliselt tunnustatud kriteeriumidele (OPS.2.2.M13).
- Pilvteenusleping lõpetatakse korrakohaselt (OPS.2.2.M14)
- Pilvteenuse andja vahetamiseks või siseteenusele üleminekuks on kehtestatud kriteeriumid, mis sisaldavad porditavusnõudeid ja teenuse ülekantavuse testimise kohustuse (OPS.2.2.M15).
- Pilvteenuse andjale on esitatud andmevarunduse detailsed nõuded (OPS.2.2.M16).
- Lepitakse kokku andmete krüpteerimise vajalikkus ning krüpteerimismehhanismid (OPS.2.2.M17).

Väljasttellimise (OPS 2.3) meetmed on järgmised.

- Kõikidele väljasttellitavatele teenustele on kehtestatud turvanõuded, mille määramisel on arvestatud, mis andmeid töödeldakse ning milline peab olema andmevahetusprotseduuride ja -liideste turve. Arvestatud on ka äriprotsesside vahelist sõltuvust ning protsesside sisendeid ja väljundeid (OPS.2.3.M1).
- Teenuse väljasttellimise võimalikkus otsustatakse riskipõhiselt. Teenuse jätkuvat vastavust lubatud riskiprofiilile kontrollitakse regulaarselt (OPS.2.3.M2).
- Teenuseandja valimiseks on koostatud turvanõudeid sisaldav nõuete profiil (OPS.2.3.M3).
- Sõlmitakse kliendi nõuetele vastav teenuseleping (OPS.2.3.M4).
- Teenuseandja peab tagama erinevatele klientidele sarnaste teenuste pakkumisel kliendi andmete turvalise eraldatuse (OPS.2.3.M5).
- Dokumenteeritud on väljasttellitava teenuse turbe põhimõtted ning neid järgitakse (OPS.2.3.M6).
- Välisteenuse leping lõpetatakse korrakohaselt (OPS.2.3.M7)
- Väljasttellimise strateegia sisaldab AI-süsteemide ja teenuste tingimusi (OPS.2.3.M8).

- Väljasttellimise strateegiast lähtuvalt täiendatakse hankepoliitikat AI-süsteemide ja teenuste informatsiooniga (OPS.2.3.M9).
- Väljasttellitud teenuste registrisse lisatakse AI-süsteemid ja teenused (OPS.2.3.M11).
- Teenuselepingus määratakse, millistele objektidele ja võrguteenustele tohib teenuseandja kliendi võrgus juurde pääseda. Teenuse peamised sooritusindikaatorid (KPI) dokumenteeritakse teenuselepingu osana. Teenuseleping sisaldab erinevaid võimalusi väljastellitava teenuse lõpetamiseks ning seonduvaid protseduure kliendi andmete ja varade tagastamiseks. Teenuseleping sisaldab osapoolte kohustusi ja käitumisjuhiseid avariiolukorras tegutsemiseks (OPS.2.3.M14).
- Kaardistatakse alternatiivsed teenuseandjad, kellel on sobiv ettevõtte profiil ja piisav infoturbe tase. Koostatakse tegevuskava teenuse migreerimiseks (OPS.2.3.M19).
- Väljasttellitava teenuse jaoks töötatakse välja avariivalmenduse plaan (OPS.2.3.M20).
- AI-süsteemis edastatakse teenuseandja ja kliendi vahelises andmevahetuses tundlikud andmed krüpteeritult (OPS.2.3.M23).

Teenuseandja infoturbe (OPS 3.2) meetmed on järgmised.

- AI teenuseandja on teenuste kavandamisel arvestanud teenuse saajate infoturbe vajadusi. Teenus on vastavuse seadusandlusest tulenevate (sh andmekaitse) nõuetega (OPS.3.2.M1).
- AI teenuseandja on välja töötanud teenuselepingu tüüptingimused (OPS.3.2.M2).
- Allhankijate kasutamisel järgib AI teenuseandja turvanõudeid (OPS.3.2.M3).
- AI teenuseandja on eraldab oma süsteemides ´erinevate klientide andmed ja käitluskeskkonnad üksteisest piisavalt turvaliselt (OPS.3.2.M4).
- AI teenuseandja on koostanud kõiki klientidele pakutavaid teenuseid hõlmava turvakontseptsiooni (OPS.3.2.M5).
- Teenuselepingus on kokku lepitud lepingu korralise ja erakorralise lõpetamise tingimused (OPS.3.2.M6).
- Allhankijate teenuseid kasutav AI teenuseandja koostab alternatiivsete allhankijate loendi (OPS.3.2.M7).
- AI teenuseandja on dokumenteerinud põhimõtted teenuste loomiseks, testimiseks ja kasutusele võtuks (OPS.3.2.M8).
- Teenuselepingutes sätestatud turvameetmete täitmist ja turvameetmete jätkuvat asjakohasust kontrollitakse perioodiliselt ja/või sündmusepõhiselt (OPS.3.2.M9).
- Koostatakse teenuste avariivalmenduse plaan (OPS.3.2.M11).
- AI teenuseandja protsessidele ja IT-süsteemidele on tehtud riskkontroll (OPS.3.2.M12).
- AI teenuseandja tagab tarneahela läbipaistvuse (OPS.3.2.M16).
- AI teenuseandja ja kliendi töötajate sissepääsu ruumidesse, süsteemidesse ja võrkudesse ning juurdepääsu tehisintellektisüsteemi andmetele ja tarkvarale reguleeritakse sobivate korralduslike ja tehniliste vahenditega (OPS.3.2.M17).
- Allhankija töötajaid on juhendatud nende tööülesannete täitmise osas ning teavitatud kehtivatest infoturbe nõuetest ja infoturvet reguleerivatest dokumentidest (OPS.3.2.M18).
- Andmete turvaliseks edastamiseks ning nende hoidmiseks AI teenuseandja juures on kokku lepitud turvalised krüpteerimismehhanismid (OPS.3.2.M20).

Turvaintsidentide käsitluse (DER 2.1) meetmed on järgmised.

- Võimalike turvaintsidentide määratlus sisaldab AI-süsteemidega seotud turvaintsidentide määratlust (DER.2.1.M1).
- Turvaintsidentide käsitluse juhend sisaldab tehisintellektisüsteemidega seotud turvaintsidentide käsitlust (DER.2.1.M2).
- Turvaintsidentide käsitluse metoodika sisaldab AI-süsteemidega seotud turvaintsidentide käsitlust (DER.2.1.M7).
- Turvaintsidentidest teavitamise juhend sisaldab AI-süsteemidega seotud turvaintsidentidest teavitamist (DER.2.1.M9).
- Hinnatakse AI-süsteemidega seotud turvaintsidentide mõju (DER.2.1.M10).

- IT-talituse töötajad on valmis AI-süsteemidega seotud turvaintsidentide käsitluseks (DER.2.1.M15).
- Äriprotsesside erinevast kaalukusest tulenevalt määratakse AI-süsteemidega seotud intsidentide käsitluse prioriteedid (DER.2.1.M19).

Auditite ja läbivaatuste (DER 3.1) meetmed on järgmised.

- Auditi käsitlusalasse lisatakse AI-süsteemid (DER.3.1.M2).
- Läbivaatuse käigus kontrollitakse, kas vaatlusalused infoturbemeetmed on AI-süsteemides rakendatud terviklikult, sobivalt ja ajakohaselt (DER.3.1.M4).
- Läbivaatuste objektide loend sisaldab AI-süsteemi komponente (DER.3.1.M8).
- AI-süsteeme auditeerib sobiv auditi- või läbivaatusrühm (DER.3.1.M9).

6.1.2 Meetmed süsteemiriskide leevendamiseks

Süsteemiriskide leevendusmeetmed on tavaliste IT-süsteemide ja AI-süsteemide ning tavaliste rakenduste ja AI rakenduste puhul samad. Neid meetmeid kirjeldavad E-ITS etalonturbe moodulid SYS ja APP.

6.2 Tehisintellektispetsiifiliste riskide leevendamise meetmed

6.2.1 Tehisintellektisüsteemi kvaliteedi ja ohutuse tõstmine

Tehisintellektisüsteemide väljundi kvaliteediga seotud riskide leevendamiseks on mitmeid lähenemisi. Väljastpoolt sissetoodava mudeli korral on esimene leevendusmeede lihtsalt parema AI mudeli hankimine (eeldusel, et see on võimalik). See eeldab uurimistööd mudeli pakkuja ning mudeli tarneahela suhtes (nt. andmestike kvaliteedinäitajad). Teiseks tuleb süsteemi väljundite kvaliteeti pidevalt jälgida tuvastamaks, kas AI mudeli kvaliteet on ajas stabiilne ja kas ta tuleb toime seninägemata sisenditega. Seda saab teha nii tehniliste mõõdikute või kasutajate tagasiside vahendusel. Kui nihe mudeli kvaliteedis (näiteks mingi kindla sisendklassi korral) või muu intsident on tuvastatud, siis selle tõlgendamiseks tulevad abiks mudeli seletatavuse meetodid.

Hallutsineerimise vältimiseks keelemudelites on mitmeid lahendusi. Rakenduse arhitektuuri poole pealt aitab RAG (Retrieval-Augmented Generation) lahendus, kus väljundi koostamisel päritakse infot olemasolest (tekst-)andmestikule. Liidestades AI mudel olemasoleva teadmiste baasiga saab vähendada ebatõeste või tõendamatute vastuste esinemissagedust. Et AI-süsteem jääks kontrollitavaks ja seletatavaks, kasutatakse RAG-lahendusi, kus keelemudeli väljund sisaldab viiteid otsingumootori poolt leitule. Keelemudeli väljundit saab täiendavalt suunata viibatehnikaga, juhendades seda kasutama ainult otsingumootori poolt leitud infot. Hallutsionatsioonide vältimiseks võib samuti tulla abiks mudeli täiendav peenhäälestamine ja treeningandmestiku kvaliteedi haldamine.

Vältimaks AIst sõltumist ja selle üle kontrolli kaotamist tuleb eelistada inimese kaasatusega (*human-in-the-loop*) tehnoloogiaid. Seda eriti, kui tegemist on kriitiliselt tähtsate või kõrge riskiga kasutusjuhtudega. Tuleb piirata agendipõhise tehisintellekti tegutsemisvabadust konkreetse ülesande domeeniga, muuhulgas piirates AI agendi volitusi. Tehisintellekti kasutuselevõtu osas erinevates töövoogudes tuleb olla läbipaistev, samuti tuleb pidada kinni vastavatest regulatsioonidest.

Kallutatud ja ohtlike vastustega seotud riskide leevendamine on protsess, mis ulatub kogu AI tarneahela peale. On vaja tagada treeningandmete kvaliteet ja mitmekesisus, peenhäälestada mudel vastavalt kvaliteedi- ja turvanäitajatele ning mõõta ja jälgida neid näitajaid mudelit kasutava AI rakenduse levitamisel, vajadusel piirates lubamatuid sisendeid või väljundeid.

6.2.2 Tehisintellektisüsteemide tehniliste rünnete leevendusmeetmed

Järgmises tekstis kasutame levitusmudelitele viitamiseks peatükis 4 kasutusele võetud lühendeid, sest kõik kaitsemeetmed ei ole kõikide mudelite puhul asjakohased. Iga kaitsemeetme järele lisame, milliste mudelite puhul vastav kaitsemeede on asjakohane.

Keelemudeli eelviip ei tohi sisaldada infot, millele kasutajal ei tohiks olla ligipääsu. Levitaja peab lähtuma eeldusest, et eelviiba sisu on igal juhul võimalik kätte saada. **LM1, LM2, LM3**

Kui keelemudel konstrueerib kasutaja sisendi põhjal päringuid mõne liidestatud teenuse (nt. RAG süsteemi komponentide) suhtes, ei tohi päringul olla rohkem õigusi, kui kasutajal. See tähendab, et kui mõni teenus või rakendus (nt. andmebaas) on liidestatud keelemudeliga, peab eeldama, et kasutaja on võimeline koostama päringu liidestatud teenuse suhtes käsitsi. Nii on võimalik leevendada volitamata ligipääsu ja tundlike andmete lekke riske. **LM1, LM2, LM3**

Kui kasutaja sisend sisaldab koodi, mida käitatakse, peab käitamiskeskkond olema isoleeritud. Kui aga koodi käitamine ei ole ettenähtud funktsionaalsus, siis tuleb kasutaja sisendi töötlemisel arvestama võimalusega, et see sisaldab `eval`, `exec` jm. käskude või funktsioonide väljakutseid, mis seda ikkagi taotlevad. Koodi kaugkäituse vältimiseks tuleb selliseid sisendeid filtreerida. Kaudset viibasüsti saab leevendada valideerides API väljakutsete ja muude liidestatud rakenduste suhtes teostatud päringute vastuseid. **LM1, LM2, LM3**

Tehisintellekti rakendustes kasutatakse proksi ja tulemüüri arhitektuuri, kus kasutaja päring jõuab kõigepealt proksisse, mis logib ja filtreerib pahatahtlikke päringuid, vajadusel puhastab neid, sõnastab nad ümber ning valib vastavad mudelid. Seejärel edastatakse päringud tulemüürile, mis kaitseb mudeleid ja nende taristut. Tulemüürist edastatakse päring mudelini. Mudeli vastus läbib proksit ja tulemüüri vastupidises järjekorras ning vastust kontrollitakse samuti mõlemal sammul kasutajale jõudmise hetkeni. **LM1, LM2, LM3**

Vältimaks mudeli väljundi tõlgendamist kasutaja brauseri poolt JavaScripti või Markdowni koodina (skriptisüst), tuleb mudeli väljundit kodeerida. **LM1, LM2, LM3**

Andmemürgitused ja tagaukseründed eeldavad ligipääsu treening- või peenhäälestusandmestikele. Kaitsemeetmed selle vastu ulatuvad kogu mudeli elutsükli ja tarneahela peale. Esimene kaitsemeede selliste rünnete vastu on andmestike kureerimine. Treeningandmestiku kogumisel andmesorimise läbi (andmete automatiseeritud kogumisel internetist) tuleb rakendada kvaliteedimeetrikaid, kontrollida andmeallikaid ja filtreerida neid vastavalt usaldusväärsusele, seejuures erilist tähelepanu tuleb pöörata selliste andmeklasside kvaliteedile, mis on seotud mudeli spetsiifikaga (nt. õiguslikud või meditsiinilised allikad). **LM3**

Tagaukserünnete vältimiseks võib piltmudeli treenimisel kasutada erinevaid töökindluse tõstmise tehnikaid, näiteks piltide muundamist: neile müra lisamist ja pildiosade maskeerimist - nii saab vähendada tagaust avavate sisendite mõju. **LM3**

Kui eeltreenitud mudel võetakse üle väljaspoolt tuleb veenduda, et mudeli pakkuja on usaldusväärne ning läbipaistev oma andmete tarneahela suhtes ning pakub piisavat infot mudeli võimekustest ja nõrkustest (mudelikaardid). **LM1, LM2**

Mudeli kasutamisel rakenduses tuleb jälgida pidevalt selle jõudlust, ka erinevate sisendi kategooriate või klasside lõikes, et oleks võimalik tuvastada, kui mudeli jõudlus langeb alla mingi lävendi andmete mõne kategooria või klassi puhul - see võib viidata andmemürgitusele. **LM1, LM2, LM3**

Siirdeõppel nõrkuste edasikandumise riski (mis on suurim vabavaraliste eeltreenitud mudelite ülevõtmise puhul) leevendamiseks on soovitatav mudelit täiendavalt peenhäälestada, ehkki seegi ei pruugi olla piisav. Mudeli peenhäälestamise järel ei saa enam tugineda esialgse mudeli kvaliteedi- ja turvalisusnäitajatele [158] - neid tuleb rakendada uuesti. **LM2**

Keelemudeleid on võimalik panna oma treeningandmestiku sisu tsiteerima[169]. Treeningandmetes sisalduva tundliku isikustatava teabe lekitamise vastaseid meetmeid on mitmeid. Esiteks võib üritada neid treeningandmestikust välistada, seda kas kirje või andmestiku kaupa. Alternatiivina saab kasutada sünteetilisi andmeid, mis säilitavad algandmetes olevaid seoseid, kuid ei sisalda tundlikku või isikustavat teavet. Samuti võib andmed pseudonümiseerida, näiteks asendades isikustatavad andmepunktid vastavate märgenditega. Pseudonüümimist on võimalik rakendada ka mudeli väljundi poole peal ehk osana AI rakenduse loogikast, kuid väljundid võivad sel juhul osutuda ikkagi tuvastatavateks [188] ning mudeli lekkimise puhul on selline mudel infoeraldusrünnetele avatum.

Pelgalt isiklikuna paistva või isegi isikliku infoga kattuva info tagastamist mudeli poolt ei saa veel pidada privaatsuse riivamiseks, sest see võib olla juhuslik kokkusattumus mudelis esinevate seoste tõttu. Näiteks

võib keelemudel anda mingile päringule väljundiks sageliesineva nime ja sümptomitega patsiendi haigusloo. Veendumaks, et tegemist oli tõepolest juhuse ning mitte isiklike andmete lekkega, võib rakendada diferentsiaalprivaatsuse meetodit, kus võrreldakse sellise väljundi saamise tõenäosusi olukordades, kus vastavad kirjed sisaldusid või ei sisaldunud treeningandmestikus. Samuti võib kasutada diferentsiaalprivaatseid (või muid privaatsuskaitse tehnoloogiaid[189] kasutatavaid) treenimis- ja peenhäälestusmeetodeid [190]. **LM3**

Teenustõkestuse leevendamiseks rakenduse tasemel tuleb lähtuda rakenduste infoturbepraktikatest. Selleks, et takistada AI mudeli omadusi ära kasutavat teenustõkestusrünnet, tuleb piirata sisendi pikkust arvestades mudeli omadustega (nt transformerpõhise keelemudeli korral kontekstakna pikkusega), ressursside kasutamist ühele päringule vastamiseks ning alamsammude või -päringute hulka. **LM1, LM2, LM3**

Ühe kasutaja poolt tehtavate päringute arvu piiramine võib aidata võidelda pöördmodelleerimise ja mudelivarguse vastu, raskendades ründajatel piisava treeningandmestiku kogumist või *logitite* tuletamist. **LM1, LM2, LM3**

6.3 Ühiskondlike riskide leevendusmeetmed

6.3.1 Ühiskonna tasemel rakenduvad leevendusmeetmed

Tehisintellekti süsteemid on teinud viimastel aastatel läbi suure arenguhüppe. Kuigi viidatud süsteemidel on potentsiaal suurendada efektiivsust ja luua uusi võimalusi, võivad sellega kaasneda ühiskonnale ka mitmed riskid, mida käsitleme alljärgnevalt:

- **Andmekaitse ja privaatsus**. Suured andmemahud tehisintellekti süsteemides kätkevad andmete kuritavrituste ohtu, sh privaatsusriivet. Üheks võimalikuks leevendusmeetmeks on tõsta ühiskonna teadlikkust tehisintellekti süsteemidest ning seotud andmekaitse ja privaatsusteemadest, näiteks koostada suuniseid andmete kogumise, töötlemise ja säilitamise osas. Väga efektiivne viis riskide leevendamiseks on ka isikustatavate andmete töötlemise vähendamine. Seda saab teha kas äriloogikat muutes või siis privaatsuskaitse tehnoloogiate abil tehisintellekti rakendavas süsteemis.
- **Tööjõuturu muutused**. Tehisintellekti areng toob endaga kaasa muudatused ka tööjõuturul. Viidatud tehnoloogia võimaldab teatud tööprotsesse lihtsustada ja muuta efektiivsemaks, misläbi toimub tööjõu ümberstruktureerimine. Samas võivad kaduda ka teatud töökohad traditsioonilistel tööstusharudel. Tööjõuturu muudatustega toime tulemiseks on võimalik rakendada uuenduslikke haridus- või ümberõppeprogramme, mis aitavad inimestel kohaneda uue tehnoloogiaga ja õppida kasutama tehisintellekti poolt pakutavaid võimalusi.
- **Sotsiaalne lõhestumine**. Kui teatud ühiskonnagrupid ei oma juurdepääsu tehisintellekti tehnoloogiale või oskusi seda tehnoloogiat efektiivselt kasutada, siis võib see süvendada digitaalset lõhet. Seetõttu tuleks mõelda, kuidas tehisintellekti tehnoloogia oleks kättesaadav erinevatele ühiskonnagruppidele alates lastest kuni vanuriteni, näiteks rakendades laiapõhjalisi haridusprogramme.
- **Diskrimineerimine**. Tehisintellekti süsteemide arendamine eelarvamustevabalt on küllaltki keerukas protsess. Eelarvamustega ja diskrimineeriva mustriga tehisintellektisüsteem võib suurendada aga sotsiaalset ebavõrdsust ja rikkuda inimeste põhiõigusi. Seetõttu tuleb süstemaatiliselt hinnata tehisintellekti süsteemi algoritme ja vajadusel neid parendada (või halvemal juhul nende töö lõpetada), et tagada mitmekesisuse ja õigluse põhimõtete järgimine.
- **Tehnoloogiline sõltuvus ja haavatavus**. Ühiskonna sõltuvus tehisintellekti süsteemidest on tõusutrendis. See võib aga omakorda suurendada ühiskonna haavatavust. Leevendusmeetmena tuleks mitmekesistada tehnoloogilist infrastruktuuri ning investeerida tehisintellekti turvalisuse ja vastupanuvõime arendamisse.
- **Ökoloogiline jalajälg**. Tehisintellekti süsteemid põhinevad suurtel andmemahtudel ja arvutuslike ressursside intensiivsel kasutamisel. Seetõttu suurendab nende loomine ja tööshoidmine energiatarbimist ja seeläbi ka ökoloogilist jalajälge. Üheks leevendusmeetmeks võiks olla keskkonnasäästliku ja energiatõhusama tehisintellektisüsteemi alaste teadusuuringute läbiviimine. Tehisintellekti poolt tekitatud keskkonnamõju hindamiseks on vaja kokku leppida ka konkreetsed mõõdikud.

Tehisintellekti süsteemide mõjud ühiskonnale on mitmetahulised ja võimalikud kaasnevad ohud nõuavad sobivate leevendusmeetmete rakendamisel lähenemisviise, mis võtavad arvesse tehisintellekti mõju erinevatele aspektidele. Teadus- ja arendustegevus ning poliitikakujundus peaksid püüdlema selle poole, et tehisintellekti süsteemide kasutamine toetaks ühiskonna üldist heaolu, kaasamist ja säästvat arengut.

6.3.2 AI-süsteemi taseme leevendusmeetmed

Lähtuvalt tehisintellekti reguleerivate õigusaktide ja järelevalveasutuste madalast küpsusest on kõige käepärasem viis rakenduse ohutuse tagamiseks enesehindamine. Tehisintellektiteenuse või -äpi peab süsteemi loomise ajal hindama, milline on tema süsteemi mõju indiviididele ja nende kaudu ühiskonnale. Selle hinnangu efektiivsus sõltub loomulikult arendaja eetilistest tõekspidamistest ning tehnoloogilisest taustast.

Riigid ja ettevõtted üle maailma on välja töötanud mitmeid soovitusi ja juhiseid, kuidas ülesandele läheneda. Kasutatakse termineid nagu vastutustundlik (*responsible*), usaldusväärne (*trustworthy*) ja ohutu (*safe*) tehisintellekt. Toome siin välja Euroopa Liidu kõrgetasemelise ekspertrühma välja töötatud usaldusväärse tehisintellekti enesehinnangu mudeli [90], mis seab usaldusväärsusele seitse tingimust.

1. inimlik kontroll ja järelevalve,
2. tehniline töökindlus ja ohutus,
3. privaatsus ja andmevalitsemine,
4. avatus,
5. erisuste austamine, diskrimineerimise vältimine ja õiglus,
6. keskkonna ja ühiskonna heaolu tagamine ning
7. vastutavus.

Järgnevalt toome välja suunised, mida tehisintellekti süsteemide arendamisel, juurutamisel ja kasutamisel soovitame arvesse võtta.

- **Inimkesksed väärtused**. See tähendab, et tehisintellekti süsteemi arendades tuleb lähtuda inimkesksest disaini põhimõtetest, austades ja kaitstes inimese kehalist ja vaimset terviklikkust ning tema identiteeditunnet [45].
- **Kahju tegemisest hoidumine**. Selle kohaselt peavad tehisintellekti süsteemid olema ohutud ja turvalised, tehniliselt töökindlad ja välistada tuleks nende pahatahtlik kasutamine [45].
- **Õiglus**. Tuleb tagada, et tehisintellekti süsteem edendaks võrdseid võimalusi ja ei oleks ebaõiglaselt kallutatud ega diskrimineeriks inimesi või ühiskonnagruppe [45].
- **Vastutus**. Vastutus tähendab, et tehisintellekti arendamisega seotud osapooled vastutavad selle nõuetekohase toimimise eest, lähtudes oma rollist ja võttes arvesse nii süsteemi kasutamise konteksti kui ka kooskõla tehnika tasemega [191].
- **Selgitatavus**. Tehisintellekti süsteemi otstarve ja võimekus peab olema teada ning protsesse peab saama võimalikult suures ulatuses selgitada isikutele, keda need mõjutavad [191].
- **Kaasav majanduskasv, säästev areng ja heaolu.** Usaldusväärse tehisintellekti kasutamine peaks looma väärtust nii inimesele, ühiskonnale kui tervele planeedile, suurendama loovust, vähendama ebavõrdsust ja kaitsma looduskeskkonda [192].

7 Poliitikasoovitused

Järgnevate poliitikasoovituste rakendamine toetab Eestis tehisintellekti ökosüsteemi ja majandussuuna kasvu. Eetilise ja vastutustundliku tehisintellekti arendamiseks on vaja toimivat ökosüsteemi, mis julgustaks, inspireeriks ja toetaks. Oluline on laiaulatuslik koostöö erinevate avaliku- ja erasektori osapoolte vahel. Kestlikuks kasutuseks tuleb tööd teha tehisintellektisüsteemidega seotud riskide teadvustamise ja õigeaegse leevendusmeetmete rakendamise nimel.

- **Investeeringud tehisintellekti teadus- ja arendustegevusse**. Selleks, et Eestisse tekiks konkurentsivõimelisi tehisintellekti valdkonna ettevõtteid, tuleks panustada tehisintellektiga seotud uurimis- ja arendustegevustesse. Tuleks näha ette riiklikke investeeringuid ja julgustada ka erainvesteeringuid. Ka küberturvalisuse teine direktiiv (NIS2) julgustab tegelema tehisintellektiga seotud teadus- ja arendustegevusega, mis parandaks küberrünnete avastamist ja ennetamist ning selle tarbeks ressursside planeerimist.
- **Talentide järelkasv**. Luua tuleks stipendiumiprogramme ja koostööprojekte ülikoolidega, et suurendada kohalike spetsialistide arvu. See loob omakorda aluse kodumaisele tehisintellekti ekspertide kogukonna tekkimisele. Talentide koolitamine võimaldab suurendada inimvõimekust, mis on oluline ka tööturu muutustega kohanemiseks.
- **Tehisintellekti süsteemide liivakastide, arendus- või inkubatsioonikeskuste loomine**. Tehisintellekti arendajatele võib luua kontrollitud keskkonnad, mis annavad ettevõtjatele juurdepääsu vajalikele ressurssidele (nt rahastamine, taristu, mentorlus, tehniline tugi) ja kus on võimalik uusi tehisintellekti lahendusi testida ja katsetada. Sellised kontrollitud keskkonnad võimaldaksid tehisintellekti süsteemide turvalisemat üleminekut uurimis- ja arendusfaasist kasutuselevõtu- ja käitamisfaasi. Regulaatoritele annab see võimaluse saada teadmisi uutest tehisintellekti tehnoloogiatest ja neid teadmisi vajadusel poliitikakujundamise otsustel arvesse võtta. Tehisintellekti määruse kohaselt peab liikmesriik looma vähemalt ühe tehisintellekti regulatiivse liivakasti.
- **Avaliku andmeplatvormi või andmefondi loomine**. Tehisintellekti süsteemid on suures sõltuvuses andmetest. Avalikud andmeplatvormid võimaldaksid ettevõtetel ja teadlastel juurdepääsu suuremahulistele andmekogumitele, mida saab kasutada tehisintellekti algoritmide treenimiseks ja testimiseks erinevates valdkondades. Treeningandmete loomine võib olla alustavatele tehisintellekti arendajatele ajakulukas ja keeruline (nt andmekaitseõiguse ja intellektuaalomandi õiguse kontekstis). Euroopas, sh Eestis, avalikustatakse küll avaandmeid, kuid nende kasutamine tehisintellekti mudeli treenimiseks ei ole praktiline. Probleemkohaks on asjaolu, et need ei kajasta hästi tegelikku elu – avaandmete ”puhastatuse”tase on kõrge ja see omakorda ei võimalda tagada mitmekesisust ega üldjuhul sisalda ka äärejuhte. Seetõttu võiks riik aidata luua avalikke sünteetilisi andmestikke, mis oleksid esinduslikud, eelarvamustevabad, austaksid eraelu puutumatust ning arvestaksid nii isikuandmete kaitse nõuete kui intellekutuaalomandi õigusega.
- **Standardid tehisintellekti mudelite kirjeldamiseks**. Tehisintellekti mudelite standardid oleksid vajalikud selleks, et oleks võimalik tuvastada, mis sorti andmestikel need treenitud on ning kuidas on need andmed saadud. Lisaks on kasulik kohandada standardeid sünteesitud piltide, tekstide ja muu teabe märgendamiseks.
- **Tehnoloogiline tööriistakast tehisintellekti süsteemide turvalisuse tagamiseks**. Tehisintellekti süsteemide turvalisuse tagamisel tuleb arvestada tehnoloogia arengutega. Seetõttu on soovitav rakendada süsteemide kaitsmisel tõhusaid tehnoloogilisi tööriistasid, näiteks otspunktprivaatsust (ingl k *end-to-end privacy*), mis takistab kõrvalistele isikutele tehisintellekti süsteemis andmetele juurdepääsu (nt volitamata andmete lugemist või nende salajast muutmist).
- **Kujundada tehisintellekti jaoks soodne poliitiline keskkond**. Selge õigusraamistik julgustab ettevõtteid investeerima AI-süsteemidesse. Selleks tuleks koostada suuniseid ja jagada parimaid tehisintellektiga seotud praktikaid, nt jagades riigi kogemusi ja õppetunde Krattide arendamisest. Läbi poliitikakujundamise tuleks julgustada uuendusi ja konkurentsi usaldusväärse tehisintellekti arendamisel. Soovituslik on korraldada ka innovatsioonikonkursse, et innustada uuenduslike tehisintellekti rakenduste loomist erinevates valdkondades.

- **Edendada koostööd teiste riikidega**. Rahvusvahelised partnerlussuhted on olulised, et jagada teadmisi, kogemusi ja ressursse (nt koostööprojektid). See omakorda võimaldab ka kiiremat tehnoloogilist arengut ja suurendab ekspordivõimalusi.
- **Hoida rahvuskeelt ja edendada selle arengut digiajastul**. Nii tehisintellekti treenimiseks kasutatavad andmestikud kui internetisisu laiemalt on valdavalt ingliskeelsed. Sellegipoolest avab tehisintellekt suuremad võimalused keele arengusse panustamiseks, seda kvaliteetsete automaattõlgete, arhiivmaterjalide automatiseeritud digiteerimise ja nendest struktureeritud info eraldamise, innovatiivsete õppematerjalide ja muude digihumanitaaria meetodite võimendamise läbi. Eesti kultuuri püsimajäämise jaoks on suur väärtus Eesti keele teksti- ja kõnekorpuste jätkuv arendus.
- **Tõsta ühiskonna teadlikkust tehisintellektisüsteemidest**. Hoogustada tuleks tehisintellektiga seotud avalikku arutelu ja teha teavituskampaaniaid. See on vajalik, et selgitada tehisintellekti eeliseid, aga ka väljakutseid. Samuti on oluline koguda tagasisidet kodanikelt, et kujundada poliitikat, mis vastaks ühiskonna vajadustele.

8 Rakendaja kiirjuhis

8.1 Kirjelda oma tehisintellektisüsteem

Kasuta joonisel 21 näidatud töölehte ja täida selle neli osa järgmiste juhiste põhjal.

Loetle tehisintellektisüsteemi lõppkasutajad (vormi jaotised A1-An).

1. Kes tehisintellektisüsteemi otseselt kasutavad? Nii teenuseandja kui kasutaja poolel. Hinda peamised rollid, kelle andmeid AI-süsteem töötleb või kes selle tulemusi tarbivad. Oluline – lisa ka võimalikud automaatotsuseid tegevad infosüsteemid lõppkasutajate sekka, sest hiljem on seda infot mõjuanalüüsis vaja.
2. Pane kirja, milleks kasutaja süsteemi vajab. Hiljem on see info abiks mõjude hindamisel.
3. Pane kirja, milliseid andmeid kasutaja tehisintellektisüstemile edastab ja mida vastu saab. Nende andmete põhjal on võimalik hiljem teha riskide ja mõjude analüüs. Võimalusel märgi ka ära, kas andmed on struktureeritud, tabeli kujul, tekstilised, pildid, heli, video või nende kombinatsioon.

Kirjelda tehisintellekti tehnoloogiat rakendavat teenust (vormil jaotised B1 ja B2).

1. Millise ülesande täitmiseks on tehisintellektisüsteem (äpp või teenus) loodud, millist väärtust see loob?
2. Pane kirja enda parima teadmise kohaselt, milliseid mudeleid ja tehnoloogiaid kasutab teenuseandja, kelle mudeli peale äpp või teenus on ehitatud.
3. Kirjelda, millise taristu peal (oma andmekeskus, pilvteenus) teenus töötab ning mis riigis taristu asub.
4. Toetudes AI-süsteemi kasutajate kohta kirja pandule, tee kokkuvõte sellest, milliseid andmeid teenus annab AI-komponendile edasi ja mida vastu saab.

Selgita, kas tehisintellektimudeli käivitamist ostetakse sisse teenusena või tehakse seda enda taristul.

1. Kui tehisintellektimudeli käivitamist ostetakse sisse teenusena (nt rakendusliidese kaudu), siis täida alamosa C1.
   a. Kes on teenuseandja ja millisest riigist ta on?
   b. Milliste andmete peal on mudelit treenitud? Eesmärk on hinnata, et mudeli treenimine on olnud seaduslik (nt ei ole kasutatud ilma loata autoriõigustega kaitstud teavet).
   c. Millisest riigist on teenuseandja pärit ja millises riigis asub tema taristu?
   d. Pane ka kirja viide antava teenuse tingimustele või teie vahel sõlmitud lepingutingimustele.
2. Kui loodav tehisintellektisüsteem käivitab mudelit ise (vahet pole, kas see on ise treenitud, litsentseeritud või ostetud), siis täida alamosa C2.
   a. Kes on mudeli treeninud ja millisest riigist on see pärit?
   b. Milliste andmete peal on mudelit treenitud? Eesmärk on hinnata, et mudeli treenimine on olnud seaduslik (nt ei ole kasutatud ilma loata autoriõigustega kaitstud teavet).
   c. Millist tehnoloogiat mudel kasutab (nii palju kui on teada)?
   d. Millise riigi taristul mudelit käivitatakse (on see oma andmekeskus või pilvtaristu?)

Lõpuks, pane kirja kõik, mida tead mudeli treenimise kohta, olgu ta teiste või enda treenitud.

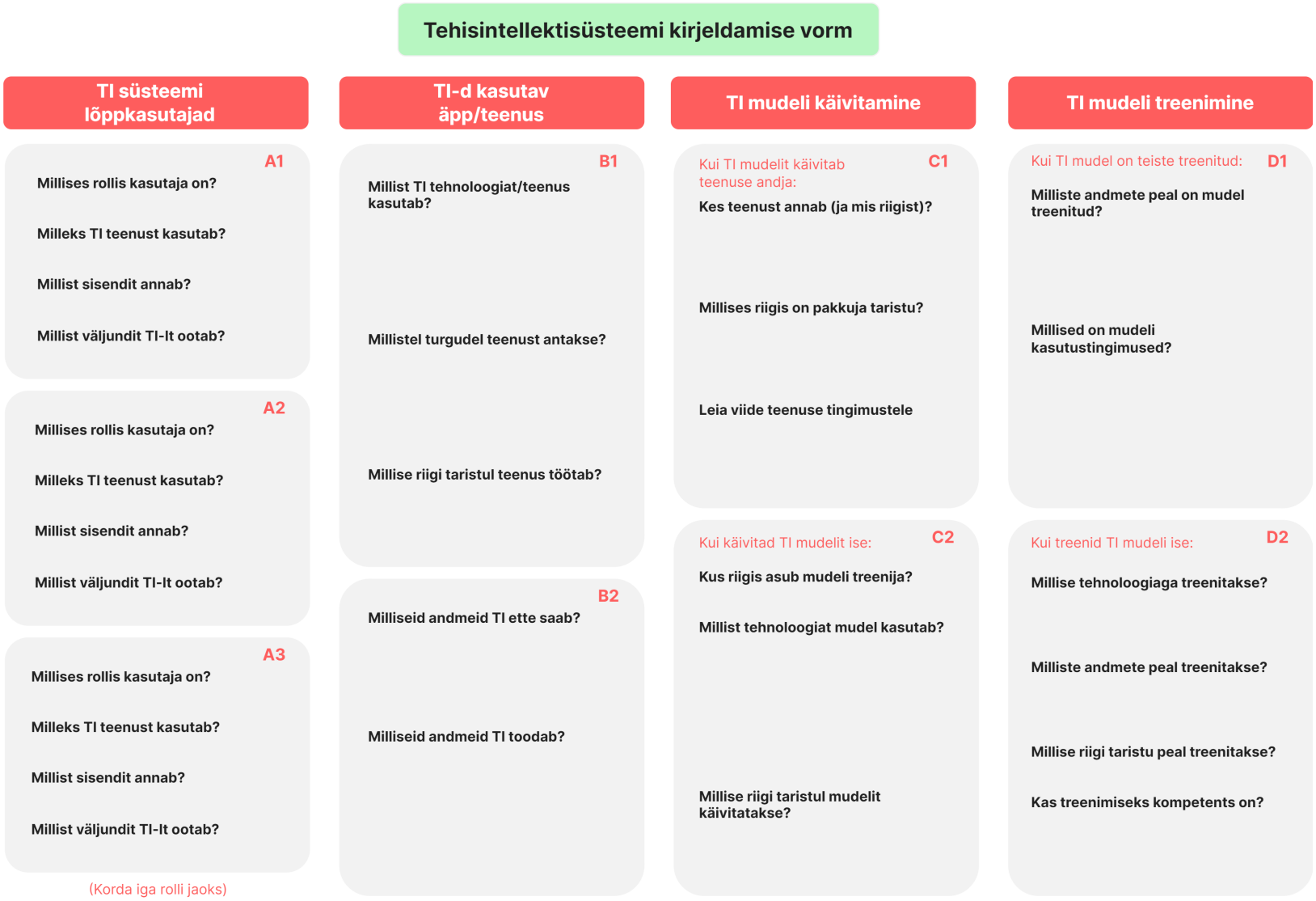
1. Kui tehisintellektimudel on ostetud, litsentseeritud või rakendusliidese kaudu kasutusel, siis täida alamosa D1.
   a. Nii palju, kui teada on: milliste andmete peal on mudel treenitud?
   b. Millised on mudeli kasutustingimused? Näiteks, milliseid vastutusi võtab ja garantiisid annab mudeli treenija.
2. Kui tehisintellektisüsteemi andja treenib mudeli ise, siis tuleb täita alamosa D2.
   a. Milliste andmete peal mudelit treenitakse? Kust on need saadud ja millistel tingimustel?

b. Millise tehnoloogiaga mudelit treenitakse? Võimalusel loetleda algoritmid ja tööriistad.
c. Millise riigi taristul mudelit treenitakse?
d. Kirjelda, millised kompetentsid teenuse andjal mudeli treenimiseks on.

8.1.1 Kuidas minna põhjalikumaks?

Joonisel 21 kirjeldatud vorm aitab teha esimese mõtete struktureerimise ja küsida asjakohaseid küsimusi. Kui see töö on tehtud, on otstarbekas kirjutada vastused paremini lahti. Seda võib teha eraldi dokumendis. See on sobiv koht protsessi sidumiseks organisatsioonis kehtiva kvaliteedi-, juhtimis- või küberturbesüsteemiga. Kui see näeb ette teatud protsesside läbimist, siis aitab siin aruandes antud vorm sellele protsessile vajalikku teavet koguda.

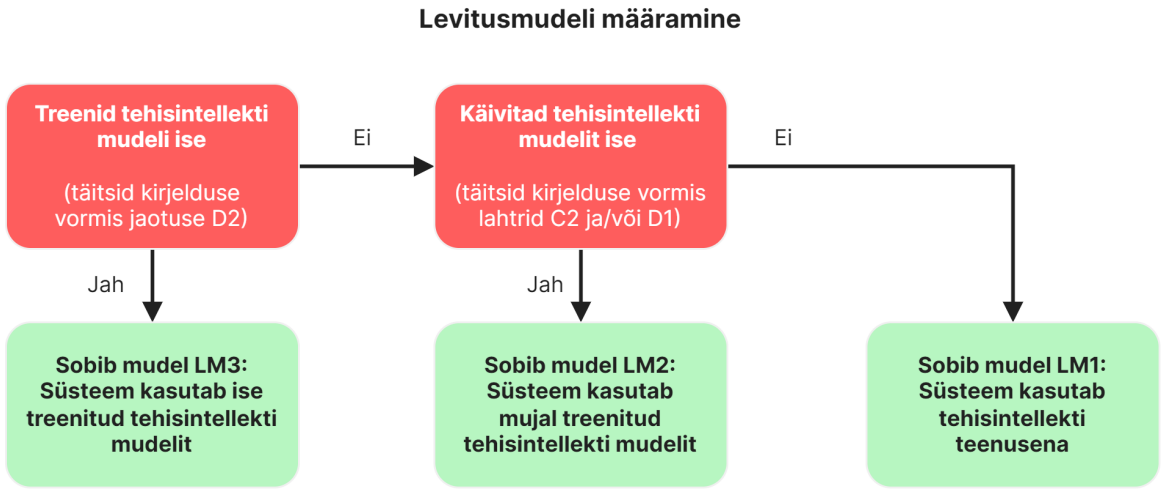
Veel üks samm edasi on juurutada tehisintellekti juhtimissüsteem, näiteks ISO/IEC 42001. Seda saab vajadusel ka integreerida ISO 9001 ja ISO/IEC 27001 standarditel põhinevate juhtimissüsteemidega.



Joonis 21. Tehisintellektisüsteemi kirjeldamise vorm

8.2 Leia oma süsteemiga sobiv levitusmudel

Kui tehisintellektisüsteem on vormi järgi kirjeldatud, on järgmine samm leida, millise levitusmudeli järgi saab riskikontrollli läbi viia. Kui vorm on täidetud, saab valiku teha lihtsalt, kahele küsimusele vastates. Sobiv otsustusdiagramm on joonisel 22.



Joonis 22. Tehisintellektisüsteemi levitusmudeli määramine

Esimese küsimuse eesmärk on välja selgitada, kas mudeli loomine on tehisintellektisüsteemi omaniku kontrolli all. Kui see on nii, peab hilisemas riskianalüüsis mudeli loomist käsitlema teistest mudelitest erinevalt.

Teise küsimuse eesmärk on leida, kas mudeli rakendamine on tehisintellektisüsteemi omaniku kontrolli all. Sellega saame riskikontrollis suunata tähelepanu mudeli valiku ja käitlemisega seotud riskidele.

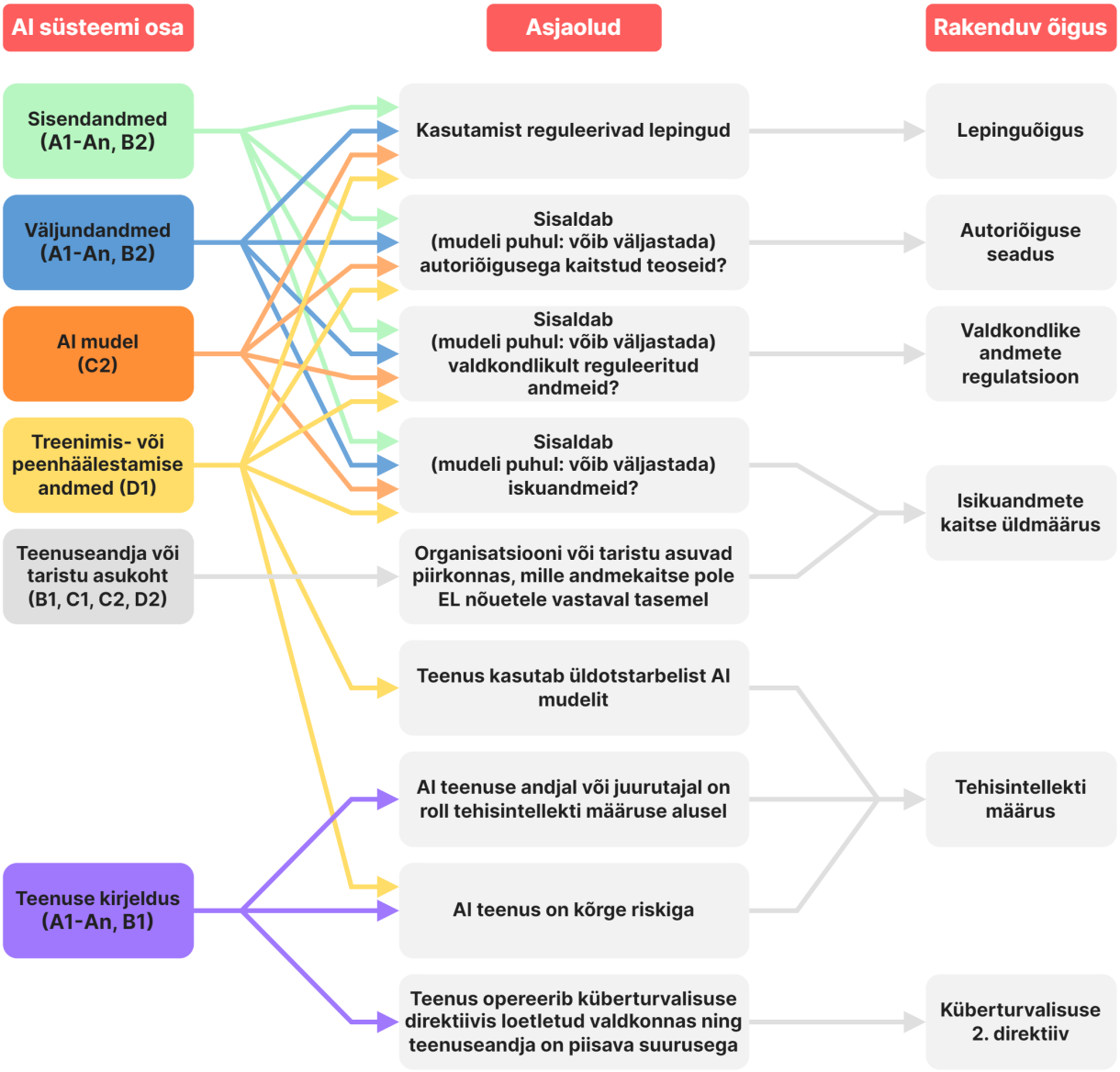
Kui tehisintellektisüstemi omanik ei treeni ega käivita ise mudelit, on kõrge tõenäosusega kasutusel levitusmudel, kus tehisintellektikomponent ostetakse teenusena sisse.

Märgime ära, et kõikidel neil juhtudel võib taristuna kasutada kas enda andmekeskust, era- või avalikku pilvandmetöötlust. See ei mõjuta levitusmudeli valikut, riskikontrolli käigus käsitletakse taristu asukohta eraldi.

8.3 Tuvasta rakenduvad õigusnormid

Oluline on mõista, et käesolevas aruandes toodud juhised ei ole kvalifitseeritavad õigusnõuandena ja neid ei saa käsitleda õigusnõu andmisena ega õigusteenuse osutamisena. Juhiste peamiseks eesmärgiks on aidata leida, milliste õigusaktidega peaks kindlasti arvestama. Iga teenuseandja kohustus on tagada oma teenuse vastavus õiguslikele, lepingulistele ja muudele nõuetele.

Joonisel 23 on lihtsustatud juhis, kuidas tuvastada, millised õigusnormid Euroopa Liidus võivad tehisintellektisüsteemile kohalduda. Keskendume olukorrale, kus juhist kasutab tehisintellektipõhise teenuse andja.



Joonis 23. Lihtsustatud juhendtabel EL rakenduva regulatsiooni tuvastamiseks

8.3.1 LM1: AI-d rakendusliidese kaudu kasutav teenus

Kas AI-d kasutav äpp/teenus töötleb isikustatud andmeid (vt vormi jaotiseid A1-An ja B2)?
Kui jah, siis on loodav teenus isikuandmete kaitse üldmääruse ning muu rakenduva andmekaitseõiguse käsitlusalas.

Kas AI-d kasutav äpp/teenus töötleb autoriõigusega kaitstud teoseid (vt vormi jaotiseid A1-An ja B2)?
Kui jah, siis on loodav teenus autoriõiguse seaduse ning muu rakenduva autoriõiguste kaitse õiguse käsitlusalas.

Kas AI-d kasutav äpp/teenus töötleb mõne valdkonna kaitstud andmeid (nt maksusaladust, pangasaladust, salastatud teavet)? (vt vormi jaotiseid A1-An ja B2)?
Kui jah, siis tuleb teenuse arendamisel jälgida vastavaid valdkondi reguleerivate õigusaktide nõudeid.

Kas AI-d kasutav äpp/teenus töötleb mingeid andmeid lepingute alusel? (vt vormi jaotiseid A1-An ja B2)?
Kui jah, siis tuleb teenuse arendamisel jälgida vastavate lepingute nõudeid.

Kas AI-d kasutav äpp/teenus või AI mudeli käivitamist pakkuv teenus töötab taristul, mis asub ebapiisava andmekaitsetasemega territooriumil (vt aruande jaotist 3.8 ning vormi jaotiseid B1 ja C1)?
Kui jah, siis tuleb eraldi käsitleda ja hinnata andmekaitse nõudeid, mis puudutavad isikustatavate andmete töötlemist sellisel taristul.

Milline on ettevõtte või organisatsiooni roll tehisintellekti määruse järgi?
Hinda tehisintellekti määruse kohaldamisala, ja tuvasta, kas kvalifitseerud näiteks teenustajaks, juurutajaks või muuks kohaldamisalas olevaks isikuks, kellel on roll AI-süsteemi elutsüklis. Sellest lähtuvalt järgi vastavale rollile kehtestatud nõudeid.

Millise riskitasemega on on AI-d kasutav äpp/teenus tehisintellekti määruse kohaselt (vt tabel 3 ning vormi jaotised A1-An ja B2)?
Viidatud tabelist saad esmase võimaliku hinnangu AI-süsteemi riskitaseme osas, mida valideeri tehisintellekti määruse konkreetsetel nõuetel. Sellest tulenevalt hinda tehisintellekti määruse järgi, milliseid nõudeid kohaldatakse vastava riskitasemega AI-süsteemile.

Kas kasutatav tehisintellekti tehnoloogia kasutab üldotstarbelist AI mudelit (vt vormi jaotist B1)?
Üldotstarbelisi AI mudeleid kasutavatele süsteemidele kohaldatakse tehisintellekti määruse järgi täiendavaid nõudeid.

8.3.2 LM2: süsteem kasutab mujal treenitud tehisintellekti mudelit

Vasta kõikidele peatükis 8.3.1 esitatud küsimustele ja seejärel järgmistele lisaküsimustele.

Kas AI mudeli treenimisel on kasutatud isikustatud andmeid, autoriõigusega teoseid või muid andmeid, mille töötlemiseks on vaja eraldi luba (vaata vormi jaotist D1)?
Kui jah, siis tuleb selgitada välja, kas mudeli kasutamisel võib see väljastada andmeid, mille töötlemiseks vajab teenuse või äpi looja vastavat õiguslikku alust.

Kas äpi või teenuse looja plaanib AI mudelit täiendada või edasi treenida?
Kui jah, siis peab teenuse arendaja veenduma, et tal on õigus neid andmeid AI mudeli treenimiseks kasutada.

8.3.3 LM3: süsteem kasutab ise treenitud tehisintellekti mudelit

Vasta kõikidele peatükkides 8.3.1 ja 8.3.2 esitatud küsimustele ning seejärel järgmistele lisaküsimustele.

Kas AI mudeli treenimisel kasutatakse isikustatud andmeid, autoriõigusega teoseid või muid andmeid, mille töötlemiseks on vaja eraldi luba (vt jaotise D2 vastuseid)?

Kui jah, siis tuleb selgitada välja, kas mudeli kasutamisel võib see väljastada andmeid, mille töötlemiseks vajab teenuse või äpi looja vastavat õiguslikku alust.

Kas AI mudelit kasutatakse Euroopa Liidus AI-põhise äpi või teenuse osana? (vt vormi jaotised B1 ja B2, aga kaaluda tuleb ka juhtu, kui mudeli peal teenust osutab keegi teine)?
Siin punktis vaatleme erijuhtu, kui treenitud mudelit rakendab tegelikult keegi teine. Kuigi see erijuht ei esine selle aruande levitusmudelites, soovitame sellele tähelepanu pöörata. Selline rakendus kuulub tehisintellekti määruse käsitlusalasse.

8.3.4 Kuidas minna põhjalikumaks?

Andmekaitseõiguse nõuete täitmiseks on esmane samm paika panna süsteemi osapooled isikuandmete kaitse üldmääruse mõistes ning seejärel kaardistada andmevood nende vahel. Töö tulemuseks võiks olla tabel, kus ridades on kõik AI-süsteemi tööga seotud osapooled ning veergudes nende poolt töödeldavad andmeelemendid.

Igasse tabeli lahtrisse tuleks teha märk, kui vastav osapool töötleb vastavat andmeelementi andmekaitseõiguse mõttes (sh kogumine, talletamine ja kustutamine). Kui süsteemis kasutatakse privaatsuskaitse tehnoloogiaid, saab lahtrisse lisada ka, mis tasemel on andmeelement vastava osapoole jaoks raskemini isikustatavaks muudetud.

Tehisintellekti õigus on käesoleva aruande koostamise ajal kiires arengus ning siin ei ole võimalik kiireid ja konkreetseid soovitusi anda. Oluline on jälgida, kuidas arenevad tehisintellekti regulatsioonid arendatava teenuse sihtturgudel.

8.4 Hinda ohte kasutajatele, ühiskonnale ja keskkonnale

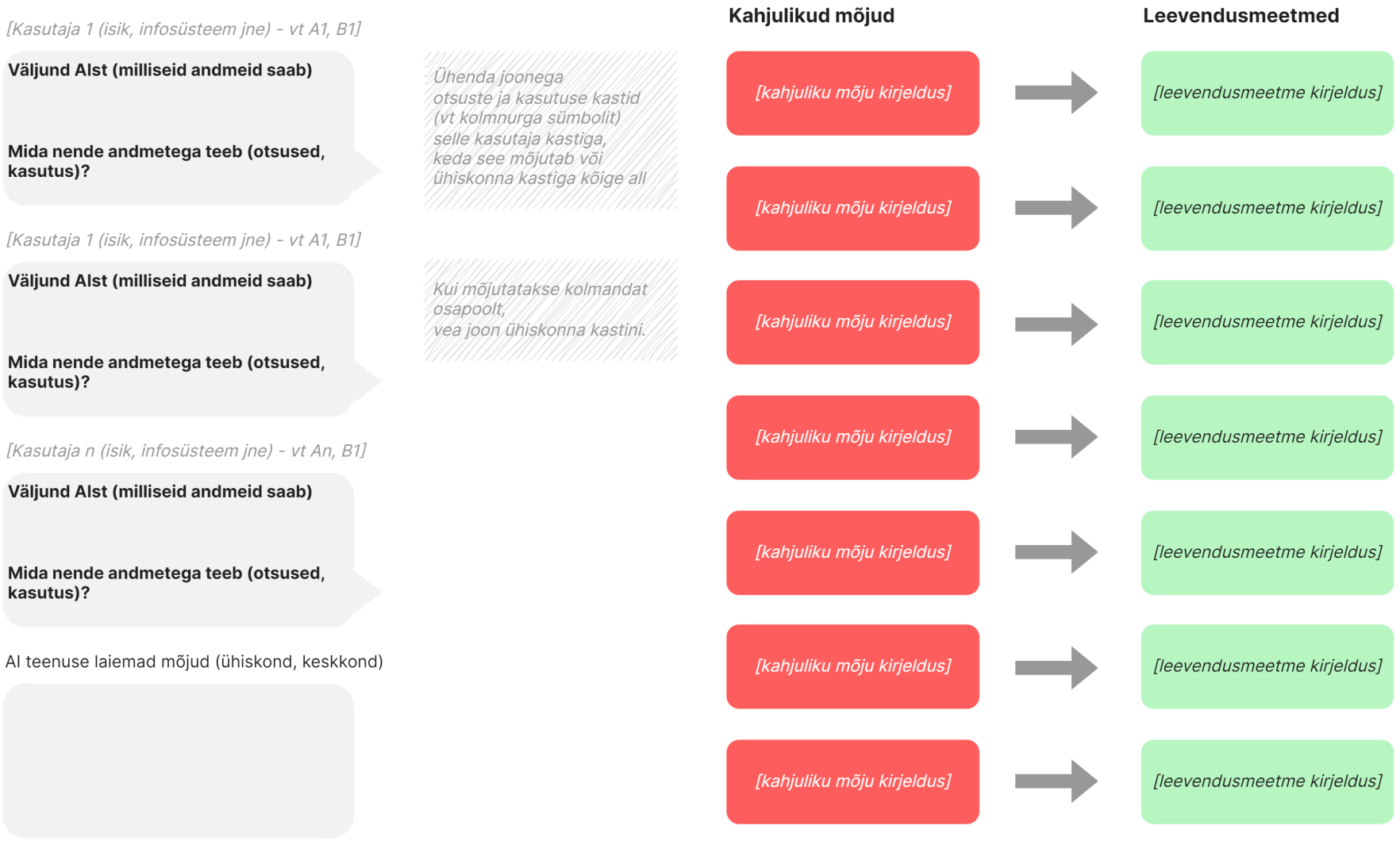
8.4.1 LM1: süsteem kasutab tehisintellekti teenusena

Mõjuanalüüsi samm 1.1: Iga lõppkasutaja kohta, vaata tema kohta vastatud küsimusi (vt vormi jaotiseid A1-An) ning üldist süsteemi kirjeldust (vt vormi jaotiseid B1 ja B2) ning pane kirja, milliseid otsuseid võib see kasutaja AI teenuselt saadud vastuste baasil teha ning kas mõni neist otsustest mõjutab kuidagi mõnda teist kasutajat või kolmandat osapoolt ja võib teda suunata tegema mingeid otsuseid või samme.

Oluline on käsitleda nii süsteemi kliendi- kui teenuseandjapoolseid kasutajaid. Süsteemi klient võib saada tehisintellekti väljundist teavet, mille põhjal ta teeb mõne enda või teiste elu mõjutava otsuse. Selliste mõttekäikude läbi analüüsimine aitab teadvustada AI-süsteemi mõju inimeste käitumisele ja selle kaudu ühiskonnale.

Eraldi olulise sammuna tuleb siin vaadelda lõppkasutajatena ka tehisintellekti põhjal automaatotsuseid tegevaid infosüsteeme ning nende mõju. Näiteks, kui tehisintellekti automaatotsuse järgi väljastatab teenus või äpp toetuseid, laene või tehakse üürile andmise otsuseid, on tehisintellektisüsteemil otsene mõju kolmandate isikute eludele ning sellest peab teenuse looja olema teadlik.

Kogu kokku kõik mõttekäigu jooksul leitud tegevused, milleni AI teenuse väljund võib isikut suunata. Joonisel 24 on näidis töölehest, mille abil analüüsi teostada.



Joonis 24. Tehisintellekti ühiskondliku mõjuanalüüsi tööleht

Mõjuanalüüsi samm 1.2: Iga sellise tegevuse kohta hinda, kas sellel võib olla negatiivne mõju indiviidile või ühiskonnale. Näiteks võivad olla kahjulikud mõjud järgmised:

1. AI teenuse kasutamine mõjutab mõne isiku või isikute grupi põhiõiguseid.
2. AI teenuse väljundi põhjal tehtud otsus diskrimineerib mõnda ühiskonnagruppi mõne tema tunnuse järgi.
3. AI teenuse väljundi põhjal tehtud otsus viib mõne ühiskonna liikme iseennast kahjustama (nt väär tervisenõuanne, väär haridusnõuanne, väär investeerimisnõuanne).
4. AI teenuse väljundi põhjal tehtud otsus viib mõne ühiskonna liikme kahjustama teist isikut (nt väär diagnoos, väär ravinõuanne, väär kahtlustus süüteos, väär hinnang oskustele või võimetele).

Mõjuanalüüsi samm 1.3: Kogu kokku võimaliku kahjuliku mõjuni viivate otsustega stsenaariumid. Analüüsi, millises ulatuses tehisintellektimudeli käivitamise teenuseandja nende eest vastutust võtab ja vastumeetmeid pakub. Leia stsenaariumid, mille puhul AI mudeli käivitamise teenuseandja vastumeetmed ja vastutus ei ole riski leevendamiseks piisav. Hinda, kas saad teenuste äriloogikat muuta või vähendada või lisa süsteemile sobivad vastumeetmed (nt avatus, inimliku järelevalvemehhanismi loomine, tugevam andmehaldus, täiendavad kontrollid äriloogikas, teavituskampaaniad, koolitused).

Mõjuanalüüsi samm 1.4: Hinda loodava süsteemi mõju loodus- ja elukeskkonnale laiemalt (ilma konkreetsetele indiviidide grupile keskendumata). Hinda, kas süsteemi loomisel on mõju meie keskkonnale – kas see mõjutab energia või loodusressursside kasutust, näiteks toetades raiskavat või saastavat käitumist. Kui mõju on kahjulik, muuda või vähenda süsteemi funktsionaalsust või rakenda kahju vältimiseks vajalikke meetmeid.

8.4.2 LM2: süsteem kasutab mujal treenitud tehisintellekti mudelit

Täida kõik peatükis 8.4.1 esitatud juhised ning seejärel täida järgmised juhised.

Mõjuanalüüsi samm 2.1: Tutvu tehisintellekti mudeli pakkuja teenuse tingimuste ning mudeli kirjelduse ning ohutusteabega (vt vormi jaotiseid C1 ja C2). Hinda, millised kahjulikud mõjud võivad mudeli kasutamisest tekkida.

Kui näed, et riske saab mudeli täiendava treenimisega tehniliselt adekvaatsel, õiguslikult korrektsel ja eetilisel moel kahandada, siis lisa mõjude haldamise meetmete sekka tehisintellektimudeli täiendav treenimine.

8.4.3 LM3: süsteem kasutab ise treenitud tehisintellekti mudelit

Täida kõik peatükkides 8.4.1 ja 8.4.2 esitatud juhised ning seejärel täida järgmised juhised.

Mõjuanalüüsi samm 3.1: Hinda tehisintellekti mudeli treenimise andmestiku tasakaalustatust. Kas see on piisavalt esinduslik, et vältida diskrimineerimist mudeli rakendamisel? Kui ei ole, leia õiguspärased ja eetilised viisid täiendavate andmestike lisamiseks treenimisse.

Mõjuanalüüsi samm 3.2: Hinda tehisintellekti mudeli treenimise kompetentsi ja tehnoloogilisi lahendusi. Kas kvaliteetse mudeli ise treenimine on jõukohane? Kui tekib kahtlus, et ei ole, tuleb kaaluda kellegi teise treenitud mudeli kasutamist ise treenimise asemel.

8.4.4 Kuidas minna põhjalikumaks?

Mõjude analüüsiks saab jälgida selleks välja töötatud juhendmaterjale. Soovitame Euroopa Liidu kõrgetasemelise ekspertgrupi välja töötatud eneseanalüüsi metoodikat [90] ning LLM rakenduste korral OWASP

sihtasutuse LLM AI Cybersecurity & Governance kontrollnimekirja [193].

On oodata, et Euroopa Liidu tehisintellekti regulatsioon liigitab mõned tehisintellektisüsteemid kõrge riskiga süsteemideks ning seab vastavatele teenuseandjatele täiendavad kohustused.

8.5 Teosta riskikäsitlus ja vali leevendusmeetmed

8.5.1 Tehisintellektisüsteemide võtmeriskid

Anname siin peatükis juhised, millele riskikäsitluse läbiviimisel pöörata ennekõike tähelepanu. Tegemist ei ole ammendavate turvasoovitustega. Iga organisatsioon on erinev ning võib vajada sügavamat käsitlust. Kui AI teenust andvas organisatsioonis on olemas toimiv riskikäsitluse praktika, siis tuleb juhinduda sellest ning siinseid juhised kasutada stardipositsioonina, millest tööd alustada.

Tabelites 4, 5 ja 6 on antud vastavalt teenuseandmise, AI mudeli käitamise ja AI mudeli treenimise võtmeriskid. Hindame nende mõju tähtsaks ning neid peab teenuseandja kindlasti käsitlema. Riskikontrolli käigus võib loomulikult selguda täiendavaid riske.

Kõigis kolmes tabelis on antud tehisintellekti rakendamise võtmeriskid etapist (sisendi ettevalmistamine teenuses, mudeli käitamine, mudeli treenimine) ja levitusmudelist lähtuvalt.

8.5.2 Soovitused küberturbe leevendusmeetmete kohta E-ITS standardist

Joonisel 25 esitame valiku tehisintellektisüsteemide kaitsmiseks sobilikest Eesti infoturbestandardi E-ITS moodulitest meetmetest. Esitame need ka süsteemi kontekstist lähtuvalt.

Põhiosa meetmetest rakenduvad teenuseandja organisatsioonile, tarkvaraarendusele ning pilvteenuste ja väljasttellimise praktikatele. Mõnede meetmete juures rõhutame ka nende olulisust kasutajatele loodud masin- või kasutajaliidestel. Samuti rõhutame mõnede praktikate tähtsust tehisintellekti rakendusliidese või mudeli andjatega suheldes.

Pilvteenuste ja väljasttellimise moodulid on antud valikulistena – kui teenuseandja ei kasuta pilvandmetöötlust ega telli midagi väljast, pole nende juurutamine loodava tehisintellekti teenuse jaoks oluline.

8.5.3 Soovitused tehisintellekti leevendusmeetmete kohta

Tehisintellektisüsteemide ohutuse tõstmiseks soovitame ka rakendada leevendusmeetmeid peatükist 6.2. Need aitavad tõsta tehisintellektisüsteemi kvaliteeti ning vältida konkreetsetest AI-tehnoloogiatest tulenevaid riske.

8.5.4 Kuidas minna põhjalikumaks?

Soovitame juurutada terviklikult mõne standarditud infoturbe-, küberturbe juhtimissüsteemi või riskikontrollimetoodika. Täpsemaid viited annab peatükk 5.1. E-ITS või ISO/IEC 27001 standardi asjakohases mahus juurutamine on tehisintellektisüsteemide turvalisuse arendamisel suureks toeks. Siinse kiirjuhise rakendamisel tehtud töö ei lähe sellisel juhul raisku ning toetab valitud standardite juurutamist.

Tabel 4. Tehisintellekti rakendava teenuse käitamise võtmeriskid levitusmudelist lähtuvalt

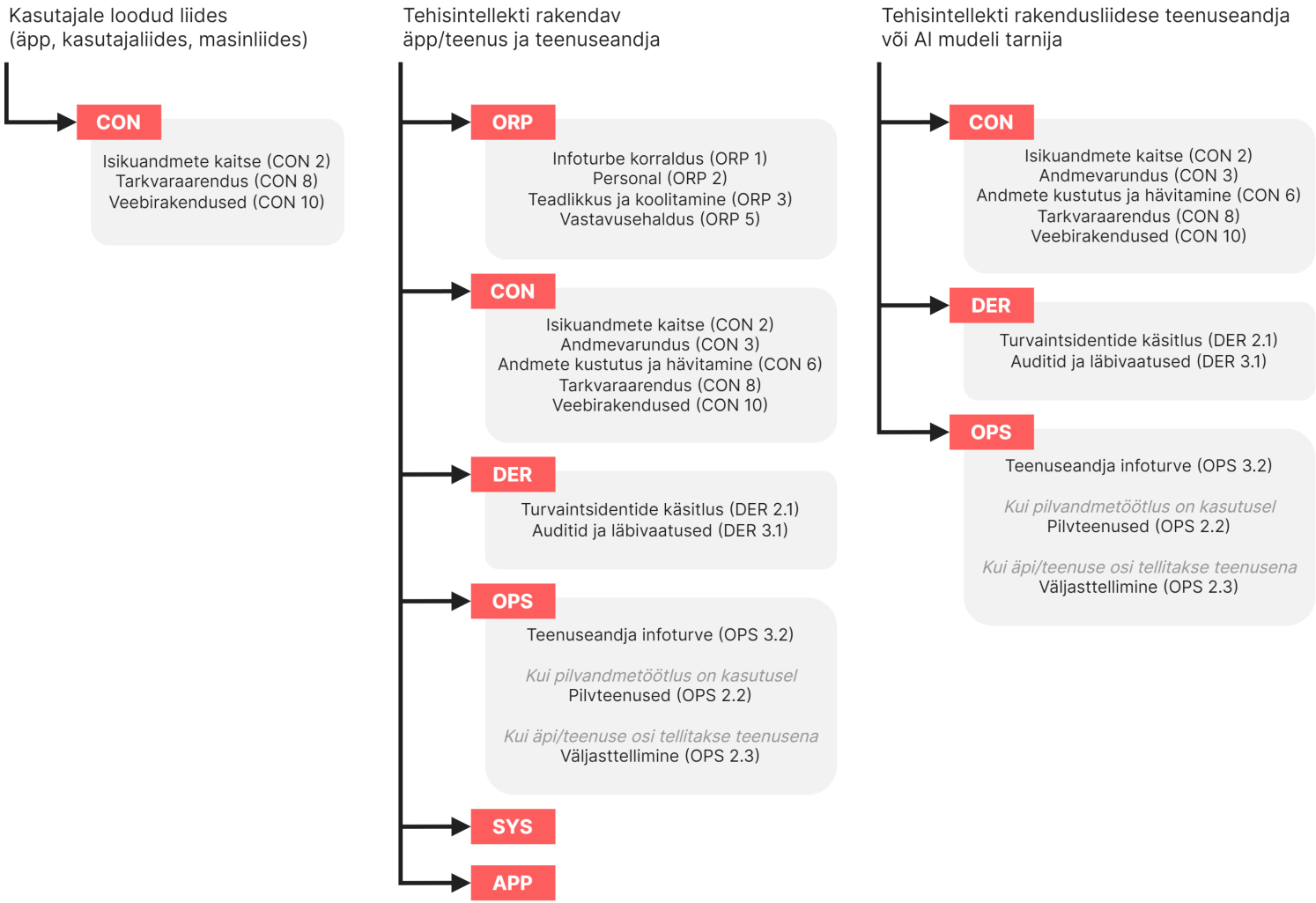
Kategooria	LM1: AI-d rakendusliidese kaudu kasutav teenus	LM2: Välist AI mudelit rakendav teenus	LM3: Ise treenitud mudeliga AI teenus
Küberturve	AI rakendusliidese käideldavus ei vasta teenuse nõuetele	Tavapärased riskid	Tavapärased riskid
Õigus	Teenuseandjal pole õigust alust AI sisend- või väljundandmete töötlemiseks või APIle edastamiseks	Teenuseandjal pole õigust alust AI sisend- või väljundandmete töötlemiseks või APIle edastamiseks	Teenuseandjal pole õigust alust AI sisend- või väljundandmete töötlemiseks või APIle edastamiseks
AI ohutus	AI API väljunditel on kahjulikud mõjud	Vt mudeli käitamise riske tabelis 5	Vt mudeli käitamise riske tabelis 5

Tabel 5. Tehisintellekti mudeli käitamise võtmeriskid levitusmudelist lähtuvalt

Kategooria	LM1: AI-d rakendusliidese kaudu kasutav teenus	LM2: Välist AI mudelit rakendav teenus	LM3: Ise treenitud mudeliga AI teenus
Küberturve	Teenuseandja ei käita mudelit ise	AI mudeli käitamise taristul pole piisavalt jõudlust (käideldavusrisk) AI mudeli tarnija ei paku mudelile täiendusi ja parandusi	AI mudeli käitamise taristul pole piisavalt jõudlust (käideldavusrisk)
Õigus	Teenuseandja ei käita mudelit ise	AI mudel või selle väljundid sisaldavad andmeid, mille töötlemiseks teenuseandjal pole õigust Teenuseandjal pole mudeli täiendamiseks kasutatavate andmete töötlemiseks õiguseid	Vt mudeli treenimise riske tabelis 6
AI ohutus	Teenuseandja ei käita mudelit ise	AI mudeli väljunditel on kahjulikud mõjud Andmed ja tööriistad, millega mudelit täiendatakse, kahandavad mudeli kvaliteeti	Vt mudeli treenimise riske tabelis 6

Tabel 6. Tehisintellekti mudeli treenimise võtmeriskid levitusmudelist lähtuvalt

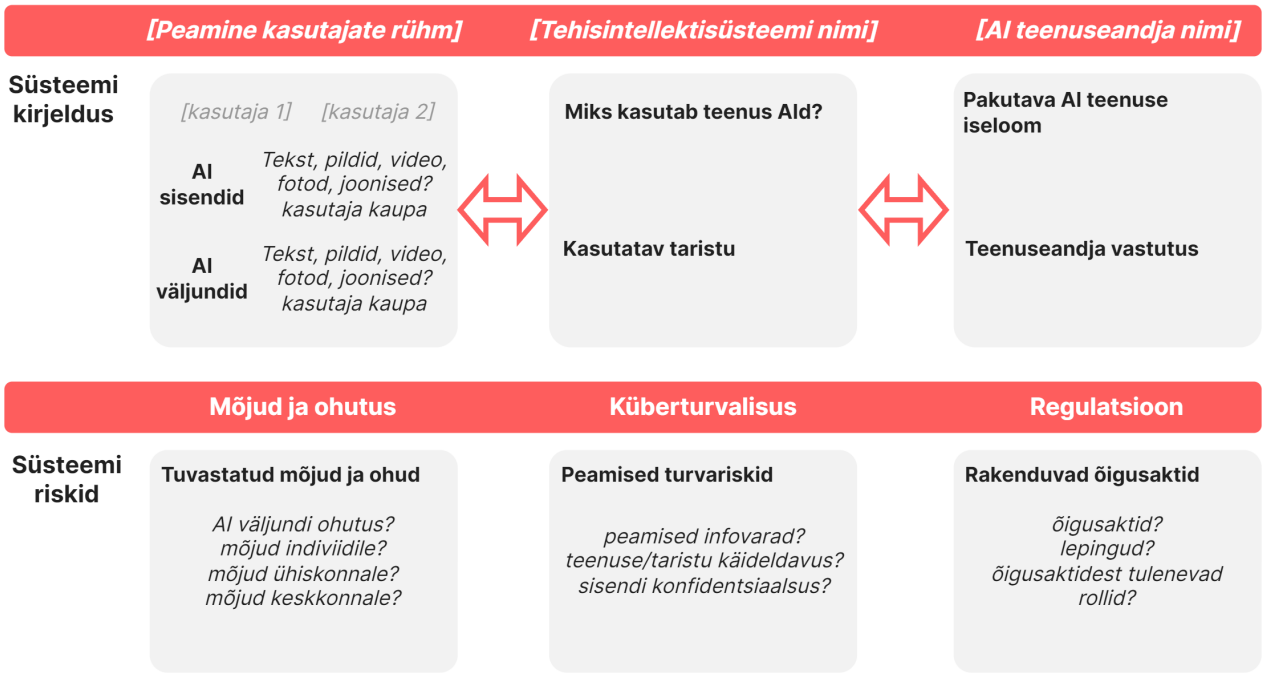
Kategooria	LM1: AI-d rakendusliidese kaudu kasutav teenus	LM2: Välist AI mudelit rakendav teenus	LM3: Ise treenitud mudeliga AI teenus
Küberturve	Teenuseandja ei treeni mudelit ise	Teenuseandja ei treeni mudelit ise	AI mudeli treenimise taristul pole piisavalt jõudlust (käideldavusrisk)
Õigus	Teenuseandja ei treeni mudelit ise	Teenuseandja ei treeni mudelit ise	Teenuseandjal pole mudeli treenimiseks kasutatavate andmete töötlemiseks õiguseid
AI ohutus	Teenuseandja ei treeni mudelit ise	Teenuseandja ei treeni mudelit ise	AI mudeli väljunditel on kahjulikud mõjud Andmed ja tööriistad, millega mudelit treenitakse, kahandavad mudeli kvaliteeti



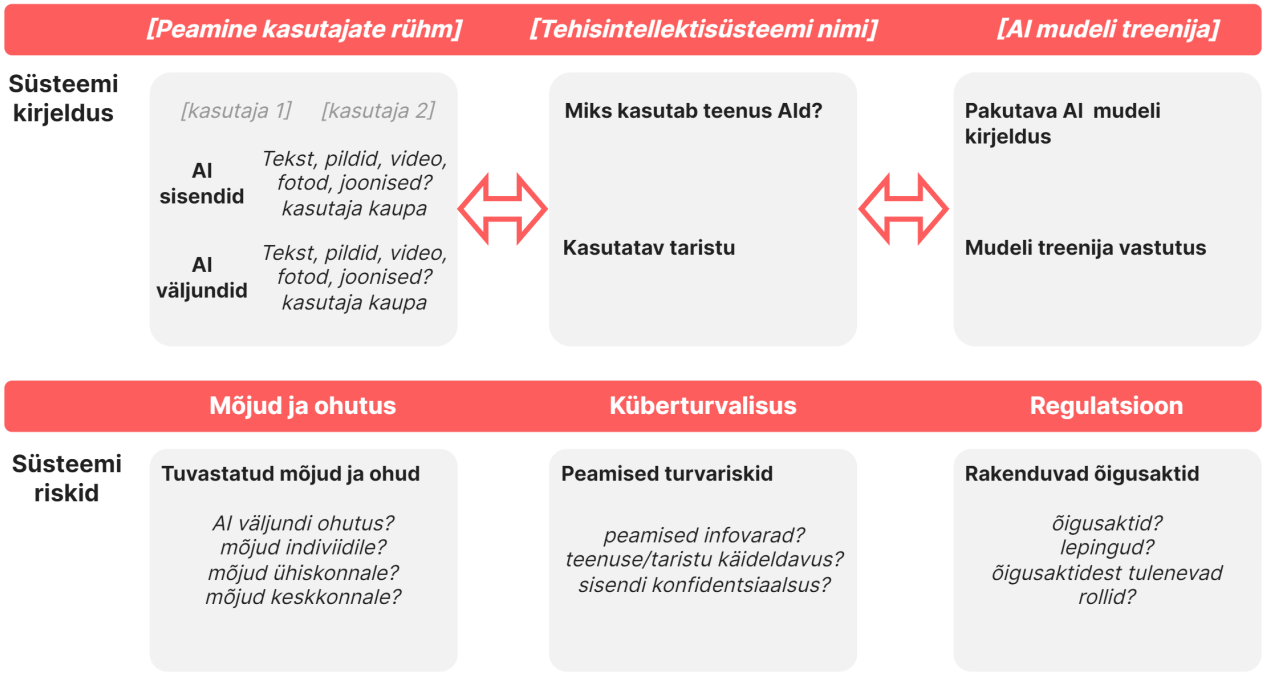
Joonis 25. Tehisintellektisüsteemide jaoks soovituslikud E-ITS moodulid ja nende rakendamise kontekst

8.6 Tehisintellektisüsteem ühe slaidiga

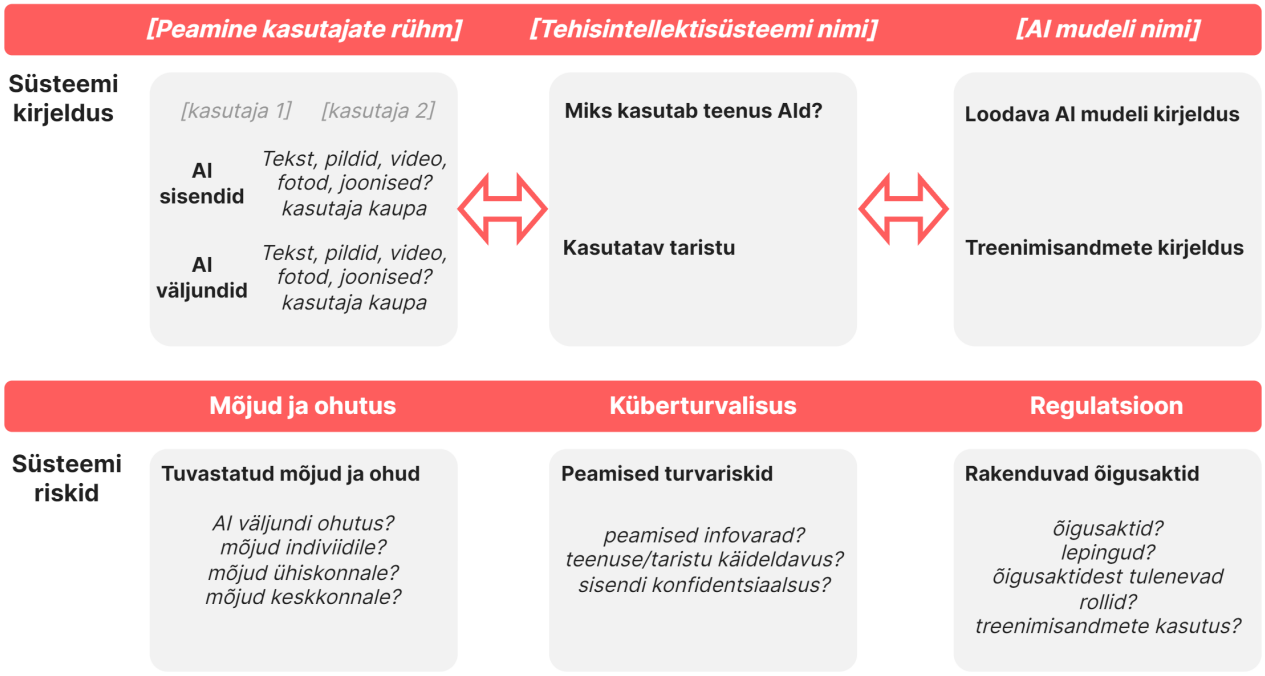
Tehisintellekti rakendamisel võib esineda olukordi, kus on vaja ühele ekraanipildile (nt ettekande leheküljele) panna loodava süsteemi ülevaade. Järgmistel joonistel pakume välja ühe võimaliku struktuuri. Igal joonisel on ühe levitusmudeli jaoks sobilik struktuur (LM1 jaoks joonis 26, LM2 jaoks joonis 27 ja LM3 jaoks joonis 28).



Joonis 26. Levitusmudeli LM1 slaidipõhi



Joonis 27. Levitusmudeli LM2 slaidipõhi



Joonis 28. Levitusmudeli LM3 slaidipõhi

bibliograafia

[1] Kai Wang *et al. Neural Network Diffusion*. 2024. arXiv: 2402.13144 [cs.LG].

[2] Yutao Sun *et al. Retentive Network: A Successor to Transformer for Large Language Models*. 2023. arXiv: 2307.08621 [cs.CL].

[3] Bo Peng *et al. RWKV: Reinventing RNNs for the Transformer Era*. 2023. arXiv: 2305.13048 [cs.CL].

[4] Albert Gu ja Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. 2023. arXiv: 2312.00752 [cs.LG].

[5] Meredith Ringel Morris *et al. Levels of AGI: Operationalizing Progress on the Path to AGI*. 2023. arXiv: 2311.02462 [cs.AI].

[6] Blaise Agüera y Arcas ja Peter Norvig. „Artificial General Intelligence Is Already Here". *Noema Magazine* (oktoober 2023). URL: https://www.noemamag.com/artificial-general-intelligence-is-already-here/.

[7] Rafael Rafailov *et al. Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. 2023. arXiv: 2305.18290 [cs.LG].

[8] Mohammad Gheshlaghi Azar *et al. A General Theoretical Paradigm to Understand Learning from Human Preferences*. 2023. arXiv: 2310.12036 [cs.AI].

[9] Tom B. Brown *et al. Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL].

[10] Tero Karras, Samuli Laine ja Timo Aila. *A Style-Based Generator Architecture for Generative Adversarial Networks*. 2019. arXiv: 1812.04948 [cs.NE].

[11] Robin Rombach *et al. High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV].

[12] Aditya Ramesh *et al. Hierarchical Text-Conditional Image Generation with CLIP Latents*. 2022. arXiv: 2204.06125 [cs.CV].

[13] Wenhui Wang *et al. Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks*. 2022. arXiv: 2208.10442 [cs.CV].

[14] Wenhai Wang *et al. InternImage: Exploring Large-Scale Vision Foundation Models with Deformable Convolutions*. 2023. arXiv: 2211.05778 [cs.CV].

[15] Chengyi Wang *et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers*. arXiv. Jaanuar 2023. URL: https://www.microsoft.com/en-us/research/publication/neural-codec-language-models-are-zero-shot-text-to-speech-synthesizers/.

[16] Matthew Le *et al. Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale*. 2023. arXiv: 2306.15687 [eess.AS].

[17] Alexei Baevski *et al. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*. 2020. arXiv: 2006.11477 [cs.CL].

[18] Hugo Touvron *et al. LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[19] Aakanksha Chowdhery *et al. PaLM: Scaling Language Modeling with Pathways*. 2022. arXiv: 2204.02311 [cs.CL].

[20] OpenAI. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[21] Euroopa Liit. „Euroopa Parlamendi ja nõukogu direktiiv (EL) 2022/2555, 14. detsember 2022, mis käsitleb meetmeid, millega tagada küberturvalisuse ühtlaselt kõrge tase kogu liidus, ja millega muudetakse määrust (EL) nr 910/2014 ja direktiivi (EL) 2018/1972 ning tunnistatakse kehtetuks direktiiv (EL) 2016/1148 (küberturvalisuse 2. direktiiv)“. *Euroopa Liidu Teataja L 333* 65 (27. detsember 2022), lk. 80–152.

[22] Waddah Saeed ja Christian Omlin. „Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities“. *Knowledge-Based Systems* 263 (2023), lk. 110273. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2023.110273. URL: https://www.sciencedirect.com/science/article/pii/S0950705123000230.

[23] Luca Nannini, Agathe Balayn ja Adam Leon Smith. „Explainability in AI Policies: A Critical Review of Communications, Reports, Regulations, and Standards in the EU, US, and UK“. Teoses: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '23. New York, NY, USA: Association for Computing Machinery, 2023, lk. 1198–1212. DOI: 10.1145/3593013.3594074. URL: https://doi.org/10.1145/3593013.3594074.

[24] Nagadivya Balasubramaniam *et al.* „Transparency and explainability of AI systems: From ethical guidelines to requirements“. *Information and Software Technology* 159 (2023), lk. 107197. ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2023.107197. URL: https://www.sciencedirect.com/science/article/pii/S0950584923000514.

[25] OECD. *OECD Legal Instruments. Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. Adopted on: 22.05.2019. Amended on: 08.11.2023*. November 2023. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

[26] Feiyu Xu *et al.* „Explainable AI: A Brief Survey on History, Research Areas, Approaches and Challenges“. Teoses: september 2019, lk. 563–574. ISBN: 978-3-030-32235-9. DOI: 10.1007/978-3-030-32236-6_51.

[27] Christoph Molnar. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Paperback. Veebruar 2022.

[28] Jason Wei *et al. Emergent Abilities of Large Language Models*. 2022. arXiv: 2206.07682 [cs.CL].

[29] Sheng Lu *et al. Are Emergent Abilities in Large Language Models just In-Context Learning?* 2023. arXiv: 2309.01809 [cs.CL].

[30] Rajat Raina, Anand Madhavan ja Andrew Ng. „Large-scale deep unsupervised learning using graphics processors“. Teoses: köide 382. Juuni 2009, lk. 110. DOI: 10.1145/1553374.1553486.

[31] Amir Gholami *et al. A Survey of Quantization Methods for Efficient Neural Network Inference*. 2021. arXiv: 2103.13630 [cs.CV].

[32] Albert Tseng *et al. Quip#: Quip with Lattice Codebooks*. Detsember 2023.

[33] Uriel Singer *et al. Make-A-Video: Text-to-Video Generation without Text-Video Data*. 2022. arXiv: 2209.14792 [cs.CV].

[34] Levon Khachatryan *et al. Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators*. 2023. arXiv: 2303.13439 [cs.CV].

[35] Andreas Blattmann *et al. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets*. 2023. arXiv: 2311.15127 [cs.CV].

[36] Joon Sung Park *et al. Generative Agents: Interactive Simulacra of Human Behavior*. 2023. arXiv: 2304.03442 [cs.HC].

[37] Edward J. Hu *et al. LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL].

[38] Yuanzhi Li *et al.* „Textbooks Are All You Need II: phi-1.5 technical report“. September 2023. URL: https://www.microsoft.com/en-us/research/publication/textbooks-are-all-you-need-ii-phi-1-5-technical-report/.

[39] Google DeepMind Gemma Team. *Gemma: Open Models Based on Gemini Research and Technology*. 2024. URL: https://storage.googleapis.com/deepmind-media/gemma/gemma-report.pdf.

[40] Albert Q. Jiang *et al. Mistral 7B*. 2023. arXiv: 2310.06825 [cs.CL].

[41] Jinze Bai *et al. Qwen Technical Report*. 2023. arXiv: 2309.16609 [cs.CL].

[42] Ben Sorscher *et al. Beyond neural scaling laws: beating power law scaling via data pruning*. 2023. arXiv: 2206.14486 [cs.LG].

[43] Europol. *Facing reality? Law enforcement and the challenge of deepfakes. An Observatory Report from the Europol Innovation Lab*. 2022. DOI: 10.2813/158794|QL-02-24-129-EN-N. URL: https://www.europol.europa.eu/cms/sites/default/files/documents/Europol_Innovation_Lab_Facing_Reality_Law_Enforcement_And_The_Challenge_Of_Deepfakes.pdf.

[44] Partha Pratim Ray. „ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope“. *Internet of Things and Cyber-Physical Systems* 3 (2023), lk. 121–154. ISSN: 2667-3452. DOI: https://doi.org/10.1016/j.iotcps.2023.04.003. URL: https://www.sciencedirect.com/science/article/pii/S266734522300024X.

[45] mille Euroopa Komisjon moodustas 2018. a juunis Sõltumatu kõrgetasemeline tehisintellekti eksperdirühm. *Eetikasuunised usaldusväärse tehisintellekti arendamiseks*. 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[46] Forbes. *24 Top AI Statistics And Trends In 2023*. Aprill 2023. URL: https://www.forbes.com/advisor/business/ai-statistics/%5C#sources_section.

[47] Statistikaamet. *Tehisintellekti tehnoloogiate kasutamine ettevõtetes on tõusutrendis*. September 2023. URL: https://www.stat.ee/et/uudised/tehisintellekti-tehnoloogiate-kasutamine-ettevotetes-tousutrendis.

[48] Ada Lovelace Institute and The Alan Turing Institute. *How do people feel about AI? A nationally representative survey of public attitudes to artificial intelligence in Britain*. Juuni 2023. URL: https://www.turing.ac.uk/sites/default/files/2023-06/how%5C_do%5C_people%5C_feel%5C_about%5C_ai%5C_-%5C_ada%5C_turing.pdf.

[49] Euractiv. *EU top court's ruling spells trouble for scoring algorithms*. Detsember 2023. URL: https://www.euractiv.com/section/data-privacy/news/eu-top-courts-ruling-spill-trouble-for-scoring-algorithms/.

[50] Politico. *Dutch scandal serves as a warning for Europe over risks of using algorithms*. Märts 2022. URL: https://www.politico.eu/article/dutch-scandal-serves-as-a-warning-for-europe-over-risks-of-using-algorithms/.

[51] Maxi Scherer. „'Artificial Intelligence and Legal Decision-Making: The Wide Open?'" *Journal of International Arbitration* 36 (5 2019), lk. 539–573. URL: https://kluwerlawonline.com/journalarticle/Journal+of+International+Arbitration/36.5/JOIA2019028.

[52] Maja Brkan. „Opinions. Artificial Intelligence and Judicial Decision-Making". *European Data Protection Law Review* 9.3 (2023). DOI: 10.21552/edpl/2023/3/5. URL: https://doi.org/10.21552/edpl/2023/3/5.

[53] Matthew Dahl *et al. Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models*. 2024. eprint: 2401.01301.

[54] Matthew Dahl *et al. Hallucinating Law: Legal Mistakes with Large Language Models are Pervasive*. Jaanuar 2024. URL: https://hai.stanford.edu/news/hallucinating-law-legal-mistakes-large-language-models-are-pervasive.

[55] Shaun Lim. „Judicial decision-making and explainable artificial intelligence. A Reckoning from First Principles". *SAcLj* 280 (2021). URL: https://law.nus.edu.sg/trail/wp-content/uploads/sites/9/2022/03/9777_09.-Shaun-Lim-Judicial-Decision-Making-and-Explainable-AI.pdf.

[56] Euroopa Komisjon. *High-level expert group on artificial intelligence*. URL: https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai.

[57] Euroopa Komisjon. *Ethics guidelines for trustworthy AI*. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[58] „Ettepanek: Euroopa Parlamendi ja nõukogu määrus, millega nähakse ette tehisintellekti käsitlevad ühtlustatud õigusnormid (tehisintellekti käsitlev õigusakt) ja muudetakse teatavaid liidu õigusakte" (). URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=CELEX:52021PC0206.

[59] Euroopa Parlament. *EU AI Act: first regulation on artificial intelligence*. Juuni 2023. URL: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence.

[60] European Union. „Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)". *Official Journal L119* 59 (4. mai 2016), lk. 1–88.

[61] The White House. *Fact Sheet: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. Oktoober 2023. URL: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/.

[62] UK Parliament. *Parliamentary Bills*. 2023. URL: https://bills.parliament.uk/bills/3464.

[63] UK Parliament. *Artificial intelligence and employment law*. URL: https://commonslibrary.parliament.uk/research-briefings/cbp-9817/.

[64] Official Website of the International Trade Administration. *UK AI regulations 2023*. 2023. URL: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach.

[65] Australian Government. *Supporting responsible AI: discussion paper*. Juuni 2023. URL: https://consult.industry.gov.au/supporting-responsible-ai.

[66] Australian Government. *Positioning Australia as a leader in digital economy regulation. Automated Decision Making and AI Regulation. Issues Paper*. Märts 2023. URL: https://storage.googleapis.com/converlens-au-industry/industry/p/prj211c4e81fb27d14public%5C_assets/automated-decision-making-ai-regulation-issues-paper.pdf.

[67] Dentons. *Australian Government requests public feedback on regulating 'Safe and Responsible AI'*. Juuni 2023. URL: https://www.dentons.com/en/insights/alerts/2023/june/6/australian-government-requests-public-feedback-on-regulating-safe.

[68] eSafety Commissioner. *Tech Trends Position Statement. Generative AI*. URL: https://www.esafety.gov.au/sites/default/files/2023-08/Generative%5C%20AI%5C%20-%5C%20Position%5C%20Statement%5C%20-%5C%20August%5C%202023%5C%20.pdf.

[69] *Australia to require AI-made child abuse material be removed from search results*. September 2023. URL: https://www.reuters.com/technology/australia-require-ai-made-child-abuse-material-be-removed-search-results-2023-09-08/.

[70] The Guardian. *Search engines required to stamp out AI-generated images of child abuse under Australia's new code*. Juuni 2023. URL: https://www.theguardian.com/technology/2023/sep/08/search-engines-required-to-stamp-out-ai-generated-images-of-child-abuse-under-australias-new-code.

[71] Government of Canada. *Artificial Intelligence and Data Act*. September 2023. URL: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act.

[72] *Government of Canada launches consultation on the implications of generative artificial intelligence for copyright*. Oktoober 2023. URL: https://www.canada.ca/en/innovation-science-economic-development/news/2023/10/government-of-canada-launches-consultation-on-the-implications-of-generative-artificial-intelligence-for-copyright.html.

[73] Runway Strategies. *Global AI Regulation Tracker*. Detsember 2023. URL: https://www.theguardian.com/technology/2023/sep/08/search-engines-required-to-stamp-out-ai-generated-images-of-child-abuse-under-australias-new-code.

[74] ISO. *ISO/IEC 22989:2022*. 2022. URL: https://www.iso.org/standard/74296.html.

[75] International Electrotechnical Commission. *Two new foundational standards for artificial intelligence*. Juuli 2022. URL: https://www.iec.ch/blog/two-new-foundational-standards-artificial-intelligence.

[76] ISO. *ISO/IEC 23053:2022*. Juuli 2022. URL: https://www.iso.org/standard/74438.html.

[77] ISO. *ISO/IEC FDIS 5259-1. Artificial intelligence. Data quality for analytics and machine learning (ML). Part 1: Overview, terminology, and examples*. URL: https://www.iso.org/standard/81088.html.

[78] ISO. *ISO/IEC DIS 5259-2. Artificial intelligence. Data quality for analytics and machine learning (ML). Part 2: Data quality measures*. URL: https://www.iso.org/standard/81860.html.

[79] ISO. *ISO/IEC TS 4213:2022. Information technology. Artificial intelligence. Assessment of machine learning classification performance*. URL: https://www.iso.org/standard/79799.html.

[80] ISO. *BS 30440:2023. Validation framework for the use of artificial intelligence (AI) within healthcare. Specification. Current. Published: 31 Jul 2023*. Juuli 2023. URL: https://knowledge.bsigroup.com/products/validation-framework-for-the-use-of-artificial-intelligence-ai-within-healthcare-specification.

[81] IEEE. *IEEE Standards Association. Ethically aligned design, Version 1, Translations and reports*. URL: https://standards.ieee.org/industry-connections/ec/ead-v1/.

[82] Google. *Google AI. Responsibility: Our principles*. URL: https://ai.google/responsibility/principles/.

[83] Google. *Google AI. Responsibility: Responsible AI practicess*. URL: https://ai.google/responsibility/responsible-ai-practices/.

[84] Microsoft. *Microsoft Responsible AI Standard, v2. General Requirements. For external release. June 2022*. Juuli 2022. URL: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf.

[85] OECD. *Multilayer Framework for Good Cybersecurity Practices for AI*. Juuni 2023. URL: https://www.enisa.europa.eu/publications/multilayer-framework-for-good-cybersecurity-practices-for-ai.

[86] Euroopa Komisjon. *Ethics guidelines for trustworthy AI*. Aprill 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[87] mille Euroopa Komisjon moodustas 2018. a juunis Sõltumatu kõrgetasemeline tehisintellekti eksperdirühm. *Eetikasuunised usaldusväärse tehisintellekti arendamiseks*. Aprill 2019. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[88] A. Vassilev *et al. AAdversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. (National Institute of Standards and Technology, Gaithersburg, MD) NIST Artifcial Intelligence (AI) Report, NIST Trustworthy and Responsible AI NIST AI 100-2e2023*. 2024. DOI: https://doi.org/10.6028/NIST.AI.100-2e2023.

[89] High-level expert group on artificial intelligence. *Policy and investment recommendations for trustworthy Artificial Intelligence*. Juuni 2019. URL: https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence.

[90] High-level expert group on artificial intelligence. *The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment*. Juuli 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[91] European AI Alliance. *Welcome to the ALTAI portal!* URL: https://futurium.ec.europa.eu/en/european-ai-alliance/pages/welcome-altai-portal.

[92] High-level expert group on artificial intelligence. *AI HLEG - Sectoral Considerations on Policy and Investment Recommendations for Trustworthy AI*. Juuli 2020. URL: https://futurium.ec.europa.eu/en/european-ai-alliance/document/ai-hleg-sectoral-considerations-policy-and-investment-recommendations-trustworthy-ai.

[93] Euroopa Komisjon. *Komisjoni aruanne Euroopa Parlamendile, nõukogule ning Euroopa Majandus- ja Sotsiaalkomiteele. Aruanne selle kohta, milline on tehisintellekti, asjade interneti ja robootika mõju ohutusele ja vastutusele. Brüssel, 19.2.2020, COM(2020) 64 final*. Veebruar 2020. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=CELEX%5C%3A52020DC0064&qid=1664351307510.

[94] Euroopa Komisjon. *Valge raamat. Tehisintellekt: Euroopa käsitus tipptasemel ja usaldusväärsest tehnoloogiast. Brüssel, 19.2.2020, COM(2020) 65 final*. Veebruar 2020. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/?uri=CELEX%5C%3A52020DC0065&qid=1664351767552.

[95] Euroopa Parlament. „ELi digipöörde strateegia“ (). URL: https://www.europarl.europa.eu/news/et/headlines/society/20210414STO02010/eli-digipoorde-strateegia.

[96] Euroopa Komisjon. „Commission welcomes political agreement on Artificial Intelligence Act“ (detsember 2023). URL: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.

[97] Euroopa Parlament. „Legislative Train Schedule. Artificial intelligence act“ (). URL: https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence.

[98] Luca Bertuzzi. *EU countries give crucial nod to first-of-a-kind Artificial Intelligence law*. Veebruar 2024. URL: https://www.euractiv.com/section/artificial-intelligence/news/eu-countries-give-crucial-nod-to-first-of-a-kind-artificial-intelligence-law/.

[99] *Analysis of the final compromise text with a view to agreement. Interinstitutional File: 2021/0106(COD). No. Cion doc.: 8115/21*. Jaanuar 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[100] European Commission. *European AI Office*. 2024. URL: https://digital-strategy.ec.europa.eu/en/policies/ai-office.

[101] Euroopa Komisjon. *Ettepanek: Euroopa Parlamendi ja nõukogu direktiiv lepinguvälise tsiviilvastutuse normide tehisintellektile kohandamise kohta (tehisintellektiga seotud vastutuse direktiiv). Brüssel, 28.9.2022. COM(2022) 496 final. 2022/0303(COD)*. September 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:52022PC0496&from=EN.

[102] Euroopa Liit. „Euroopa Parlamendi ja nõukogu määrus (EL) 2023/988, 10. mai 2023, milles käsitletakse üldist tooteohutust ja millega muudetakse Euroopa Parlamendi ja nõukogu määrust (EL) nr 1025/2012 ja Euroopa Parlamendi ja nõukogu direktiivi (EL) 2020/1828 ning tunnistatakse kehtetuks Euroopa Parlamendi ja nõukogu direktiiv 2001/95/EÜ ja nõukogu direktiiv 87/357/EMÜ“. *Euroopa Liidu Teataja L 135* 66 (23. mai 2023), lk. 1–51.

[103] Euroopa Komisjon. *Ettepanek: Euroopa Parlamendi ja nõukogu direktiiv, mis käsitleb vastutust puudusega toodete eest. Brüssel, 28.9.2022. COM(2022) 495 final. 2022/0302(COD)*. September 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:52022PC0495.

[104] Jude Ume. „What AI Means for Intellectual Property". *ITNOW* 65.2 (mai 2023), lk. 44–45. ISSN: 1746-5702. DOI: 10.1093/combul/bwad059. eprint: https://academic.oup.com/itnow/article-pdf/65/2/44/50367250/bwad059.pdf. URL: https://doi.org/10.1093/combul/bwad059.

[105] Peter Georg Picht ja Florent Thouvenin. „AI and IP: Theory to Policy and Back Again – Policy and Research Recommendations at the Intersection of Artificial Intelligence and Intellectual Property". *IIC - International Review of Intellectual Property and Competition Law* 54.6 (1. juuli 2023), lk. 916–940. DOI: 10.1007/s40319-023-01344-5. URL: https://doi.org/10.1007/s40319-023-01344-5.

[106] Winston Cho. *AI Companies Take Hit as Judge Says Artists Have "Public Interest" In Pursuing Lawsuits A federal judge rejected arguments from Stability AI, Midjourney and DeviantArt that the suit is intended to suppress its free speech*. Veebruar 2024. URL: https://www.hollywoodreporter.com/business/business-news/artist-lawsuit-ai-midjourney-art-1235821096/.

[107] Bobby Allyn. *'New York Times' considers legal action against OpenAI as copyright tensions swirl*. August 2023. URL: https://www.npr.org/2023/08/16/1194202562/new-york-times-considers-legal-action-against-openai-as-copyright-tensions-swirl.

[108] Will Bedingfield. *The Inventor Behind a Rush of AI Copyright Suits Is Trying to Show His Bot Is Sentient. Stephen Thaler's series of high-profile copyright cases has made headlines worldwide. He's done it to demonstrate his AI is capable of independent thought*. September 2023. URL: https://www.wired.com/story/the-inventor-behind-a-rush-of-ai-copyright-suits-is-trying-to-show-his-bot-is-sentient/.

[109] European Parliament. *World Intellectual Property Organization. WIPO CONVERSATION ON INTELLECTUAL PROPERTY (IP) AND ARTIFICIAL INTELLIGENCE (AI). WIPO/IP/AI/3/GE/20/INF/5*. Jaanuar 2021. URL: https://www.wipo.int/edocs/mdocs/mdocs/en/wipo_ip_ai_3_ge_20/wipo_ip_ai_3_ge_20_inf_5.pdf.

[110] S. Suganya ja E. Prema. „Human centric intellectual property rights and legal personality of artificial intelligence". *International Journal of Intellectual Property Management* 13.3-4 (2023), lk. 252–267. DOI: 10.1504/IJIPM.2023.134051. eprint: https://www.inderscienceonline.com/doi/pdf/10.1504/IJIPM.2023.134051. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJIPM.2023.134051.

[111] Atif Aziz. „Artificial Intelligence Produced Original Work: A New Approach to Copyright Protection and Ownership". *European Journal of Artificial Intelligence and Machine Learning* 2.2 (märts 2023), lk. 9–16. DOI: 10.24018/ejai.2023.2.2.15. URL: https://www.ej-ai.org/index.php/ejai/article/view/15.

[112] Mauritz Kop. „TAI & Intellectual Property: Towards an Articulated Public Domain". *University of Texas School of Law, Texas Intellectual Property Law Journal (TIPLJ)* 28.1 (juuni 2019), lk. 44–45. ISSN: 1746-5702. DOI: http://dx.doi.org/10.2139/ssrn.3409715. eprint: https://academic.oup.com/itnow/article-pdf/65/2/44/50367250/bwad059.pdf.

[113] Winston Cho. *AI-Created Art Isn't Copyrightable, Judge Says in Ruling That Could Give Hollywood Studios Pause. A federal judge on Friday upheld a finding from the U.S. Copyright Office that a piece of art created by AI is not open to protection*. August 2023. URL: https://www.hollywoodreporter.com/business/business-news/ai-works-not-copyrightable-studios-1235570316/.

[114] European Parliament. *European Parliament resolution of 20 October 2020 on intellectual property rights for the development of artificial intelligence technologies (2020/2015(INI))*. Oktoober 2020. URL: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_EN.html.

[115] Euroopa Liit. „Euroopa Parlamendi ja nõukogu määrus (EL) 2019/881, 17. aprill 2019, mis käsitleb ENISAt (Euroopa Liidu Küberturvalisuse Amet) ning info- ja kommunikatsioonitehnoloogia küberturvalisuse sertifitseerimist ja millega tunnistatakse kehtetuks määrus (EL) nr 526/2013 (küberturvalisuse määrus)". *Euroopa Liidu Teataja L 151* 62 (7. juuni 2019), lk. 15–69.

[116] Centre for Cyber security Belgium. *The NIS 2 Directive: What does it mean for my organization?* 2024. URL: https://ccb.belgium.be/en/nis-2-directive-what-does-it-mean-my-organization.

[117] Euroopa Komisjon. *Ettepanek: Euroopa Parlamendi ja nõukogu määrus, mis käsitleb digielemente sisaldavate toodete küberturvalisuse horisontaalseid nõudeid ja millega muudetakse määrust (EL) 2019/1020. Brüssel, 15.9.2022. COM(2022) 454 final. 2022/0272(COD)*. September 2022. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:52022PC0454.

[118] EU Presidency. *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. Analysis of the final compromise text with a view to agreement. Brussels, 26 January 2024. Interinstitutional File: 2021/0106(COD). No. Cion doc.: 8115/21*. Jaanuar 2024. URL: https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf.

[119] Euroopa Komisjon. *Ettepanek: Euroopa Parlamendi ja nõukogu määrus, millega kehtestatakse määruse (EL) 2016/679 täitmise tagamisega seotud täiendavad menetlusnormid. Brüssel, 4.7.2023. COM(2023) 348 final. 2023/0202(COD)*. Juuli 2023. URL: https://eur-lex.europa.eu/legal-content/ET/TXT/HTML/?uri=CELEX:52023PC0348.

[120] Euroopa Liit. „Euroopa Parlamendi ja nõukogu direktiiv (EL) 2016/680, 27. aprill 2016, mis käsitleb füüsiliste isikute kaitset seoses pädevates asutustes isikuandmete töötlemisega süütegude tõkestamise, uurimise, avastamise ja nende eest vastutusele võtmise või kriminaalkaristuste täitmisele pööramise eesmärgil ning selliste andmete vaba liikumist ning millega tunnistatakse kehtetuks nõukogu raamotsus 2008/977/JSK". *Euroopa Liidu Teataja L 119* 59 (4. mai 2016), lk. 89–131.

[121] Euroopa Liit. „Euroopa Parlamendi ja nõukogu määrus (EL) 2018/1725, 23. oktoober 2018, mis käsitleb füüsiliste isikute kaitset isikuandmete töötlemisel liidu institutsioonides, organites ja asutustes ning isikuandmete vaba liikumist, ning millega tunnistatakse kehtetuks määrus (EÜ) nr 45/2001 ja otsus nr 1247/2002/EÜ". *Euroopa Liidu Teataja L 295* 61 (21. november 2018), lk. 39–98.

[122] N. AllahRakha. *AI and the Law: Unraveling the Complexities of Regulatory Frameworks in Europe*. August 2023. URL: https://irshadjournals.com/index.php/ibys/article/view/115/102.

[123] *AI: ensuring GDPR compliance*. September 2022. URL: https://www.cnil.fr/en/ai-ensuring-gdpr-compliance.

[124] Information Commissioner's Office. *Guidance on AI and data protection*. URL: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

[125] Information Commissioner's Office. *Information Commissioner's Office launches consultation series on generative AI*. Jaanuar 2024. URL: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/01/information-commissioner-s-office-launches-consultation-series-on-generative-ai/.

[126] Information Commissioner's Office. *ICO consultation series on generative AI and data protection*. Jaanuar 2024. URL: https://ico.org.uk/about-the-ico/ico-and-stakeholder-consultations/ico-consultation-series-on-generative-ai-and-data-protection/.

[127] CNIL. *AI: ensuring GDPR compliance*. September 2021. URL: https://www.cnil.fr/en/ai-ensuring-gdpr-compliance.

[128] Federal Trade Commission. *FTC Takes Action Against Company Formerly Known as Weight Watchers for Illegally Collecting Kids' Sensitive Health Data*. Märts 2022. URL: https://www.ftc.gov/news-events/news/press-releases/2022/03/ftc-takes-action-against-company-formerly-known-weight-watchers-illegally-collecting-kids-sensitive.

[129] CNIL. *Artificial intelligence: CNIL unveils its first answers for innovative and privacy-friendly AI*. Oktoober 2023. URL: https://www.cnil.fr/en/artificial-intelligence-cnil-unveils-its-first-answers-innovative-and-privacy-friendly-ai.

[130] Dan Svantesson. *The European Union Artificial Intelligence Act: Potential implications for Australia*. 2022. DOI: https://doi.org/10.1177/1037969X211052339.

[131] Adèle Azzi. „The Challenges Faced by the Extraterritorial Scope of the General Data Protection Regulation". *JIPITEC* 9.2 (2018), lk. 126–137. ISSN: 2190-3387. URL: http://nbn-resolving.de/urn:nbn:de:0009-29-47231.

[132] Michal Czerniawski ja Dan Svantesson. „Challenges to the extraterritorial enforcement of data privacy law - EU case study". Teoses: jaanuar 2024, lk. 127–153.

[133] Federico Fabbrini ja Edoardo Celeste. „The Right to Be Forgotten in the Digital Age: The Challenges of Data Protection Beyond Borders". *German Law Journal* 21.S1 (2020), lk. 55–65. DOI: 10.1017/glj.2020.14.

[134] Chris Burt. „Clearview denies jurisdiction of French regulator in response to €20M fine". *Biometric Update* (oktoober 2022). URL: https://www.biometricupdate.com/202210/clearview-denies-jurisdiction-of-french-regulator-in-response-to-e20m-fine.

[135] IAPP. *Toward a risk-based approach? Challenging the 'zero risk' paradigm of EU DPAs in international data transfers and foreign governments' data access schedule*. Veebruar 2024. URL: https://iapp.org/news/a/towards-a-risk-based-approach-challenging-the-zero-risk-paradigm-of-eu-dpas-in-international-data-transfers-and-foreign-governments-data-access/.

[136] European Commission. *Adequacy decisions. How the EU determines if a non-EU country has an adequate level of data protection*. URL: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en.

[137] European Commission. *Commission finds that EU personal data flows can continue with 11 third countries and territories*. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_161.

[138] European Commission. *Adequacy decision for the EU-US Data Privacy Framework*. URL: https://commission.europa.eu/document/fa09cbad-dd7d-4684-ae60-be03fcb0fddf_en.

[139] EDPB. *Recommendations 1/2022 on the Application for Approval and on the elements and principles to be found in Controller Binding Corporate Rules (Art. 47 GDPR). Adopted on 20 June 2023*. 2023. URL: https://edpb.europa.eu/system/files/2023-06/edpb_recommendations_20221_bcr-c_v2_en.pdf.

[140] Andmekaitse Inspektsioon. *Isikuandmete edastamine välisriiki*. URL: https://www.aki.ee/isikuandmed/andmetootlejale/isikuandmete-edastamine-valisriikikuidas-aru-saada-mi#edastamineameerika.

[141] Euroopa Andmekaitsenõukogu. *Suunised 05/2021 isikuandmete kaitse üldmääruse artikli 3 ja V peatüki rahvusvahelist edastamist käsitlevate sätete koostoime kohta. Versioon 2.0. Vastu võetud 14. veebruaril 2023*. Veebruar 2023. URL: https://edpb.europa.eu/system/files/2023-09/edpb_guidelines_05-2021_interplay_between_the_application_et.pdf.

[142] *Risk management — Guidelines*. en. Standard ISO 31000:2018. International Organization for Standardization, 2018. URL: https://www.iso.org/standard/65694.html.

[143] *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*. en. Standard NIST SP 800-37 Rev. 2. US National Institute of Standards ja Technology, 2018. URL: https://csrc.nist.gov/pubs/sp/800/37/r2/final.

[144] *Information technology — Information security, cybersecurity and privacy protection — Guidance on managing information security risks*. en. Standard ISO/IEC 27005:2022. International Organization for Standardization, 2022. URL: https://www.iso.org/standard/80585.html.

[145] *NIST Cybersecurity Framework 1.1*. en. Standard NIST CSF v. 1.1. US National Institute of Standards ja Technology, 2018. URL: https://www.nist.gov/cyberframework/framework.

[146] *Information technology — Artificial intelligence — Guidance on risk management*. en. Standard ISO/IEC 23984:2023. International Organization for Standardization, 2023. URL: https://www.iso.org/standard/77304.html.

[147] Riigi Infosüsteemi Amet. *Eesti infoturbestandard (E-ITS)*. 2023. URL: https://eits.ria.ee/.

[148] Pille Pullonen, Raimundas Matulevičius ja Dan Bogdanov. „PE-BPMN: Privacy-Enhanced Business Process Model and Notation“. Teoses: *Business Process Management*. Springer International Publishing, 2017, lk. 40–56. DOI: 10.1007/978-3-319-65000-5_3.

[149] HM Government. *Safety and Security Risks of Generative Artificial Intelligence to 2025*. URL: https://assets.publishing.service.gov.uk/media/653932db80884d0013f71b15/generative-ai-safety-security-risks-2025-annex-b.pdf.

[150] Richard Fang *et al. LLM Agents can Autonomously Hack Websites*. 2024. arXiv: 2402.06664 [cs.CR].

[151] Riigi Infosüsteemi Amet. *Eesti infoturbestandardi etalonturbe kataloog*. 2023. URL: https://eits.ria.ee/et/versioon/2023/eits-poohidokumendid/etalonturbe-kataloog.

[152] The New York Times. *The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said*. Detsember 2023. URL: https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

[153] TechTarget. *AI lawsuits explained: Who's getting sued? Authors, artists and others are filing lawsuits against generative AI companies for using their data in bulk to train AI systems without permission*. Jaanuar 2024. URL: https://www.techtarget.com/whatis/feature/AI-lawsuits-explained-Whos-getting-sued.

[154] The Fashion Law. *From ChatGPT to Getty v. Stability AI: A Running List of Key AI-Lawsuits*. URL: https://www.thefashionlaw.com/from-chatgpt-to-deepfake-creating-apps-a-running-list-of-key-ai-lawsuits/.

[155] Civil Resolution Tribunal of British Columbia. *Moffatt v. Air Canada, 2024 BCCRT 149 (CanLII)*. Veebruar 2024. URL: https://www.canlii.org/en/bc/bccrt/doc/2024/2024bccrt149/2024bccrt149.html.

[156] Lei Huang *et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. 2023. arXiv: 2311.05232 [cs.CL].

[157] Rusheb Shah *et al. Scalable and Transferable Black-Box Jailbreaks for Language Models via Persona Modulation*. 2023. arXiv: 2311.03348 [cs.CL].

[158] Xiangyu Qi *et al. Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!* 2023. arXiv: 2310.03693 [cs.CL].

[159] José Luis Ricón. *Set Sail For Fail? On AI risk*. Nintil. Available at https://nintil.com/ai-safety/. August 2022.

[160] Government Office of Science, UK. *Future Risks of Frontier AI*. Tehniline raport. Technology & Science Insights ja Foresight, oktoober 2023.

[161] Jonas B. Sandbrink. *Artificial intelligence and biological misuse: Differentiating risks of language models and biological design tools*. 2023. arXiv: 2306.13952 [cs.CY].

[162] Daniil A. Boiko, Robert MacKnight ja Gabe Gomes. *Emergent autonomous scientific research capabilities of large language models*. 2023. arXiv: 2304.05332 [physics.chem-ph].

[163] OpenAI Research Team. *Building an Early Warning System for LLM-Aided Biological Threat Creation*. OpenAI. URL: https://openai.com/research/building-an-early-warning-system-for-llm-aided-biological-threat-creation (vaadatud 08.02.2024).

[164] Joseph R. Biden Jr. *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. Executive Order. 14110. 2023.

[165] T. C. King, N. Aggarwal, M. Taddeo *et al.* „Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions". *Sci Eng Ethics* 26 (2020), lk. 89–120. DOI: 10.1007/s11948-018-00081-0.

[166] Brigitta Dresp-Langley. „The weaponization of artificial intelligence: What the public needs to be aware of". *Frontiers in Artificial Intelligence* 6 (2023), lk. 1154184. DOI: 10.3389/frai.2023.1154184.

[167] Mrinank Sharma *et al. Towards Understanding Sycophancy in Language Models*. 2023. arXiv: 2310.13548 [cs.CL].

[168] Emily Baker-White ja Forbes Staff. „Who Is @BasedBeffJezos, The Leader Of The Tech Elite's 'E/Acc' Movement?" *Forbes* (detsember 2023). Külastatud 04.12.2023. URL: https://www.forbes.com/sites/emilybaker-white/2023/12/01/who-is-basedbeffjezos-the-leader-of-effective-accelerationism-eacc/.

[169] Milad Nasr *et al. Scalable Extraction of Training Data from (Production) Language Models*. 2023. arXiv: 2311.17035 [cs.LG].

[170] Ian J. Goodfellow, Jonathon Shlens ja Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2015. arXiv: 1412.6572 [stat.ML].

[171] Jing Lin *et al. ML Attack Models: Adversarial Attacks and Data Poisoning Attacks*. 2021. arXiv: 2112.02797 [cs.LG].

[172] Jonas Geiping *et al. Coercing LLMs to do and reveal (almost) anything*. 2024. arXiv: 2402.14020 [cs.LG].

[173] Joseph Lucas. „Mitigating Stored Prompt Injection Attacks Against LLM Applications". *NVIDIA Technical Blog* (august 2023). URL: https://developer.nvidia.com/blog/mitigating-stored-prompt-injection-attacks-against-llm-applications/.

[174] Kai Greshake *et al. Not what you've signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection*. 2023. arXiv: 2302.12173 [cs.CR].

[175] Daryna Oliynyk, Rudolf Mayer ja Andreas Rauber. „I Know What You Trained Last Summer: A Survey on Stealing Machine Learning Models and Defences". *ACM Computing Surveys* 55.14s (juuli 2023), lk. 1–41. ISSN: 1557-7341. DOI: 10.1145/3595292. URL: http://dx.doi.org/10.1145/3595292.

[176] Reza Shokri *et al. Membership Inference Attacks against Machine Learning Models*. 2017. arXiv: 1610.05820 [cs.CR].

[177] Boris van Breugel *et al. Membership Inference Attacks against Synthetic Data through Overfitting Detection*. 2023. arXiv: 2302.12580 [cs.LG].

[178] Ngoc-Bao Nguyen *et al. Re-thinking Model Inversion Attacks Against Deep Neural Networks*. 2023. arXiv: 2304.01669 [cs.LG].

[179] Kuan-Chieh Wang *et al. Variational Model Inversion Attacks*. 2022. arXiv: 2201.10787 [cs.LG].

[180] Liam Fowl *et al. Adversarial Examples Make Strong Poisons*. 2021. arXiv: 2106.10807 [cs.LG].

[181] Battista Biggio, Blaine Nelson ja Pavel Laskov. *Poisoning Attacks against Support Vector Machines*. 2013. arXiv: 1206.6389 [cs.LG].

[182] Shawn Shan *et al. Prompt-Specific Poisoning Attacks on Text-to-Image Generative Models*. 2023. arXiv: 2310.13828 [cs.CR].

[183] Micah Goldblum *et al. Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses*. 2021. arXiv: 2012.10544 [cs.LG].

[184] Xinyun Chen *et al. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning*. 2017. arXiv: 1712.05526 [cs.CR].

[185] Ruixiang Tang *et al. An Embarrassingly Simple Approach for Trojan Attack in Deep Neural Networks*. 2020. arXiv: 2006.08131 [cs.CR].

[186] Megha Agarwal *et al. LLM Inference Performance Engineering: Best Practices*. https://www.databricks.com/blog/llm-inference-performance-engineering-best-practices. Accessed on 11 Dec 2023. 2023.

[187] Jiaqi Ruan *et al. Applying Large Language Models to Power Systems: Potential Security Threats*. 2023. arXiv: 2311.13361 [cs.AI].

[188] Robin Staab *et al. Beyond Memorization: Violating Privacy Via Inference with Large Language Models*. 2023. arXiv: 2310.07298 [cs.AI].

[189] Cybernetica AS. *Privaatsuskaitse tehnoloogiate kontseptsioon*. Tehniline raport. 2023.

[190] Rouzbeh Behnia *et al.* „EW-Tune: A Framework for Privately Fine-Tuning Large Language Models with Differential Privacy“. Teoses: *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, november 2022. DOI: 10.1109/icdmw58026.2022.00078. URL: http://dx.doi.org/10.1109/ICDMW58026.2022.00078.

[191] OECD. *Accountability (Principle 1.5)*. URL: https://oecd.ai/en/dashboards/ai-principles/P9.

[192] OECD. *Inclusive growth, sustainable development and well-being (Principle 1.1)*. URL: https://oecd.ai/en/dashboards/ai-principles/P5.

[193] OWASP Foundation. *LLM AI Cybersecurity & Governance Checklist*. 2024. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.pdf.